The Mechanical Psychologist:

How Computational Techniques Can Aid Social Researchers in the Analysis of High-Stakes Conversation

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy by

Darren Cook

October 2022

Abstract

Qualitative coding is an essential observational tool for describing behaviour in the social sciences. However, it traditionally relies on manual, time-consuming, and error-prone methods performed by humans. To overcome these issues, cross-disciplinary researchers are increasingly exploring computational methods such as Natural Language Processing (NLP) and Machine Learning (ML) to annotate behaviour automatically. Automated methods offer scalability, error reduction, and the discovery of increasingly subtle patterns in data compared to human effort alone (N. C. Chen et al., 2018). Despite promising advancements, concerns regarding generalisability, mistrust of automation, and value alignment between humans and machines persist (Friedberg et al., 2012; Grimmer et al., 2021; Jiang et al., 2021; R. Levitan & Hirschberg, 2011; Mills, 2019; Nenkova et al., 2008; Rahimi et al., 2017; Yarkoni et al., 2021).

This thesis investigates the potential of computational techniques, such as social signal processing, text mining, and machine learning, to streamline qualitative coding in the social sciences, focusing on two high-stakes conversational case studies. The first case study analyses political interviewing using a corpus of 691 interview transcripts from US news networks. Psychological behaviours associated with effective interviewing are measured and used to predict conversational quality through supervised machine learning. Feature engineering employs a Social Signal Processing (SSP) approach to extract latent behaviours from low-level social signals (Vinciarelli, Salamin, et al., 2009). Conversational quality, calculated from desired characteristics of interviewee speech, is validated by a human-rater study. The findings support the potential of computational approaches in qualitative coding while acknowledging challenges in interpreting low-level social signals.

The second case study investigates the ability of machines to learn expert-defined

behaviours from human annotation, specifically in detecting predatory behaviour in known cases of online child grooming. In this section, the author utilises 623 chat logs obtained from a US-based online watchdog, with expert annotators labelling a subset of these chat logs to train a large language model. The goal was to investigate the machine's ability to detect eleven predatory behaviours based on expert annotations. The results show that the machine could detect several behaviours with as few as fifty labelled instances, but rare behaviours were frequently over-predicted. The author next implemented a collaborative human-AI approach to investigate the trade-off between human accuracy and machine efficiency. The results suggested that a human-in-the-loop approach could improve human efficiency and machine accuracy, achieving near-human performance on several behaviours approximately fifteen times faster than human effort alone.

The conclusion emphasises the value of increased automation in social sciences while recognising the importance of social scientific expertise in cross-disciplinary research, especially when addressing real-world problems. It advocates for technology that augments and enhances human effort and expertise without replacing it entirely. This thesis acknowledges the challenges in interpreting computational signals and the importance of preserving human insight in qualitative coding. The thesis also highlights potential avenues for future research, such as refining computational methods for qualitative coding and exploring collaborative human-AI approaches to address the limitations of automated methods.

Acknowledgements

I want to begin by extending my sincerest thanks to my supervisory team of Prof. Simon Maskell and Prof. Laurence Alison for all their support over the last five years. Simon, your guidance, support, patience, and unwavering belief in my ability as a researcher have brought me to this point. For that, I will remain forever thankful. Laurence, your dedication to research that addresses significant real-world problems inspired me to pursue this PhD. I'm sure it will continue to shape my future career. I have gained numerous invaluable experiences working alongside you both, and I have grown both as a researcher and person.

Special mention must also go to Dr Miri Zilka, who has had as significant an impact on my career as any over the last few years. Miri, you already know how you have inspired me to complete this thesis. It is a privilege to know you both as a friend and colleague, and I am glad the PhD is just the beginning of our time working together.

I also thank the University of Liverpool's Institute for Risk and Uncertainty for funding this project. In particular, thank you to Dr Alejandro Diaz, Prof. Scott Ferson, and Dr Adam Mannis for helping me navigate those rocky early years. Thanks also to those in the Psychology Department, particularly Dr Michael Humann, Dr Freya O'Brien, Dr Susan Giles, and Dr Hayley Rhodes, for always remembering 'one of your own' sat away in the IEEE department.

Some of my biggest thanks go to my family, particularly Mum and Dad – I promise I won't do another degree after this one. To Lew, Lucy, Sim, Sophie, Paisley, Hannah, Felix, Mary, Simon, Rog, Stan, and Bloss, thank you for always supporting me, keeping me entertained, fed, watered, caffeinated, and just about sane.

My final thanks go to my partner Christina. Thank you for always supporting me, listening to me and caring for me, especially during the countless times when I doubted I would ever reach this stage. I hope it was worth it.

Publications

- Cook, D., Zilka, M., Maskell, S., & Alison, L. (2021). A Psychology-Driven Computational Analysis of Political Interviews. *21st Annual Conference of the International Speech Communication Association*, 1942–1946. https://doi.org/10.21437/Interspeech.2021-2249
- Cook, D., Zilka, M., de Sandre, H., Giles, S., Weller, A., & Maskell, S. (2022). The Mechanical Psychologist: Can We Automate the Analysis of Online Child Sexual Exploitation Discourse? *arXiv preprint*
- Butcher, B., Zilka, M., Cook, D., & Weller, A. (2023). Optimising Human-Machine Cooperation for Efficient High-Precision Information Extraction from Text [Submitted to]. *Sixth ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT)*
- Cook, D., Zilka, M., DeSandre, H., Giles, S., & Maskell, S. (2023). Protecting Children from Online Exploitation: Can a trained model detect harmful communication strategies? [Submitted to]. *Sixth AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society*

Manuscript in Preparation

- Cook, D., Zilka, M., Alison, L., & Maskell, S. (n.d.). Does multimodality improve unimodal performance in predicting the outcome of a political interview? [In progress]
- Cook, D., & Humann, M. (n.d.). Casualty perceptions of first response communication during a major incident [In progress]

List of Figures

List of Tables

Acronyms

ANOVA Analysis of Variance

ASR Automatic Speech Recognition

CA Conversation Analysis

CPU Central Processing Unit

CSE Child Sexual Exploitation

CAT Communication Accommodation Theory

CAQDAS Computer-Assisted Qualitative Data Analysis Software

CSS Computational Social Science

CA Conversation Analysis

GLoVe Global Vectors for word representation

GPU Graphics Processing Unit

HNR Harmonics to noise ratio

IPU Inter-pausal units

ICC Intra-Class Correlation

LF Labelling Function

LpOCV Leave-pair-Out Cross-Validation

LSTM Long short-term memory

LLA Local Linguistic Alignment

LIWC Linguistic Inquiry Word Count

LSM Language Style Matching

MCC Matthews Correlation Coefficient

NER Named Entity Recognition

NLTK Natural Language Tool-Kit

NLI Natural Language Inference

NLP Natural Language Processing

POS Part-of-Speech

QA Question and Answer

RMSE Root Mean Squared Error

SCP Subtractive Conditional Probability

SSP Social Signal Processing

SVM Support Vector Machine

TF-IDF Term Frequency-Inverse Document Frequency

TTR Type-Token-Ratio

WER Word Error Rate

Contents

Chapter 1

Introduction and Background

The rise of the digital age and the widespread availability of powerful computational resources have ignited academic interest in harnessing technology to address complex social and behavioural questions. This thesis stands out by uniquely situating itself at the intersection of social science and computational technology, aiming to explore how researchers can integrate domain knowledge with cutting-edge computational techniques to revolutionise the analysis of social interactions. Traditional approaches, such as conversation (Sacks et al., 1974; Schegloff, 1993), discourse (Potter & Wetherell, 1987), thematic (V. Braun & Clarke, 2012), and content analysis (Leech & Onwuegbuzie, 2008), have long been the dominant research methods in qualitative social science. However, the coding of behavioural observations is an activity typically performed manually by experts (Maclin & Maclin, 2005). As such, it severely restricts research scalability and hinders its application to more extensive and diverse settings (Karamshuk et al., 2017).

The novel aspect of this thesis lies in its application of technology to automatically examine social behaviour during conversations, taking a cross-disciplinary approach that synergistically combines social science methodologies with the immense power of computational techniques. Two innovative computational methods, Social Signal Processing (SSP) and Sequence-to-Sequence Transformers, are explored throughout the thesis and applied to two high-stakes case studies: effective interviewing and online child safety. The author has chosen these domains for their relevance to current social and political issues. They are also research domains that typically require large amounts of manual effort for analysis.

By creatively fusing computational approaches with domain expertise, this thesis contributes to understanding how research can harness technology to tackle real-world problems. It offers insight into the future of large-scale social scientific research, demonstrating the transformative potential of a cross-disciplinary approach in revealing insights that would be unattainable using traditional methodologies alone. The remainder of this chapter will provide an overview of the problem space within which automated conversation analysis exists and describe the underlying social scientific literature that informs the thesis. The chapter concludes with a statement of the thesis's objectives and a summary of its main findings, underscoring the novelty of work in this space and its potential to reshape the landscape of social interaction analysis.

1.1 Automating the study of social interactions

1.1.1 What is conversation analysis?

Conversation Analysis (CA) is the systematic study of behaviour during real-world social interaction. It was originated by sociologist Harvey Sacks and colleagues (1974) as a method of describing how speakers in otherwise free-flowing discourse engage in the structured organisation of speech turns. When two speakers interact, the interaction involves more than a simple back-and-forth of speaking and listening positions. Instead, when interlocutors engage with one another, they do so in a collaborative effort in order to produce meaningful and naturally-occurring dialogue (Hutchby & Wooffitt, 2008). According to CA, how speakers perform and organise turn-taking reveals hidden information about the social dynamics within the conversation (Sacks et al., 1974). Rather than focusing explicitly on language use, which can be misleading (Housley et al., 2019), a central objective of CA is to explore how speakers use language to achieve a particular social action (Rytsarev, 2020).

Traditional CA is, first and foremost, a descriptive discipline. At its core, it defers the notion of quantification in favour of a detailed examination of social behaviour (Housley et al., 2019). It is considered a predominantly bottom-up methodology (Stivers, 2015). That is, analysis is based on a highly detailed exploration of human idiosyncrasies within a particular set of social interactions. The aim of CA is to describe in precise detail what people are *doing* with language rather than establishing broad patterns across datasets. For this reason, formalised coding using a

pre-defined knowledge framework (otherwise known as a deductive approach) is typically regarded as antithetical to pure CA. Strict interpretation regarding what CA is and is not has led some to reject the notion of a quantifiable approach as overly simplistic. In an early criticism, Emanuel Schegloff, one of the central figures in CA notes: "the fact we *can* do quantitative analysis does not entail that we *should*" (Schegloff, 1993, p.116). Others have warned against operationalising arbitrary links between lexical forms and semantic intent. For instance, in Housley et al. (2019), the authors argue that simple rules such as counting the occurrence of a phrase like 'thank you' as an indicator of gratitude are baseless in a real-world setting. People say extend their thanks for various reasons depending on context (e.g., sarcasm). At the centre of these criticisms is the idea that using formalised codes reduces and flattens the complexity of the underlying behaviour within the speech turn. Whilst efforts have been made to mobilise the approach, attempts to digitise CA through formal coding procedures have been criticised for not sufficiently adhering to the ethos of "real" CA (D. Giles et al., 2015, p.45).

However, a counter-argument suggests that structured coding has a place in traditional CA. Stivers (2015) argues that a formalised coding framework must adhere to the following in order to remain CA consistent: (a) the behaviour codes should be derived from underlying theory that motivates the research, (b) attention is paid to significant contextual factors, such as where a code appears within the conversation, and the social role of the speakers themselves. Over the decades since CA's inception, frameworks that adhere to and disregard these rules have appeared under a single umbrella of 'conversation analysis'.

In conversation analysis, as with broader social science, there has been an ongoing debate about the balance between qualitative and quantitative methods (Karamshuk et al., 2017). Some researchers argue that the richness and complexity of social interaction can only be captured through in-depth qualitative analysis (Housley et al., 2019; Schegloff, 1993). On the other hand, proponents of quantitative analysis assert that using structured coding and statistical methods can reveal broader patterns and trends, offering valuable insights into various aspects of social interaction, such as socio-demographic and psychological factors like age, gender, or culture (Heritage & Greatbatch, 1986; Hopkins & King, 2010).

One notable example of a CA-consistent analysis is the work performed by Heritage and Greatbatch (1986), who used a formalised coding framework to classify an audience's applause behaviour during a political speech. The researchers based their coding dictionary on prior qualitative research exploring applause behaviour

in public events, satisfying the first criteria mentioned earlier. Equally, emphasising speakers' behaviour, specifically during turn exchanges, satisfies the second criterion. Both this and similar work demonstrate that while CA, as a largely qualitative methodology, has tended to explore the minutiae of social interaction, a broader viewpoint can also be sought. Indeed, a primary advantage of a formalised coding framework is its suitability for a statistical approach (Hopkins & King, 2010). Statistical methods make it possible to look outwards across a larger spectrum of social interaction, comparing coding frameworks in different settings and exploring the impact of socio-demographic and psychological factors. Consequently, this view of CA forms the social scientific underpinnings of the present thesis.

1.1.2 Existing coding methods and their drawbacks

Like any qualitative methodology, CA relies on manual effort to identify social phenomena in conversation. Records of social interactions, whether in the format of an audio recording, video recording, or transcript, represent unstructured data. Consequently, the researcher must first convert the data into an accessible and analysable format. Suppose the unstructured data exists in audio or video format. In that case, this must first be transcribed before any actual behaviour coding can occur. Transcription can produce a substantial amount of text amongst even a modestly-sized corpus. According to Maclin and Maclin (2005), one fifteen-minute interaction can yield up to 450 pages of annotation, depending on the type of transcription performed and the depth of analysis desired.

Once transcripts are generated, the researcher can begin systematically coding each speech turn or utterance with behaviours of interest. This process is, without a doubt, extremely time-consuming and laborious, and the amount of work required increases dramatically as the coding scheme becomes more intricate. Take, for instance, a simplistic coding scheme based on a small number of mutually-exclusive dialogue acts (i.e., greetings, questions, answers, and farewells). As each dialogue act is mutually exclusive, coding the entire corpus may take only as long as the time it would take to read each utterance – perhaps an ambitious goal that excludes the effects of fatigue and any time spent verifying the quality of the annotations. Given the requisite time and resources, such a task will require considerable effort, but it is achievable for most researchers. However, the task becomes inherently more challenging if a speech turn can exhibit multiple behaviours (Dönmez et al., 2005). For instance, perhaps the utterance "Hi, how are you?" should be labelled

as both a greeting and a question. Coding, in this instance, would require multiple passes of the corpus, as each speech turn and behaviour pairing would need to be considered independently. Under such circumstances, the limitations of a wholly manual approach become evident. It seems reasonable to estimate that the time required to complete coding would increase by a factor equal to the number of behaviours of interest.

Researchers in this field face a critical decision: who should perform the coding? In the social sciences, two predominant approaches address this question. The first approach involves utilising experts, who are often the researchers themselves. These experts boast extensive domain knowledge and may have previous coding experience. However, their specialised expertise makes their numbers relatively limited compared to non-experts. This scarcity can lead to a small pool of experts taking on substantial coding tasks, which may not be suitable for large-scale projects. The second approach relies on crowd-sourcing platforms, such as Amazon Mechanical Turk[1]. This method recruits large numbers of untrained annotators from the public to carry out coding tasks (Shaw et al., 2011). This approach aids scalability, as the number of workers can be adjusted to accommodate research requirements (Haug et al., 2021). However, inexperience, boredom, and fatigue may affect the quality of non-expert annotations if not adequately addressed (Snel et al., 2012). The subsequent section delves further into the advantages and disadvantages of each approach.

Expert annotation

We can regard an expert as someone with sufficient background in a particular subject matter or who possesses a high degree of skill or competency in a particular domain (Ericsson et al., 2018). In behaviour coding, an expert may be a practitioner or someone with substantial experience in a particular field. Examples of expert coders could include doctors or nurses in a medically-focused study. It could also include police officers, judges or lawyers in law enforcement. More typically, research team members, including postgraduate students with sufficient theoretical or methodological understanding, perform coding under the role of an 'expert' (Maclin & Maclin, 2005). For the most part, annotations generated by experts are considered to represent 'ground-truth' – a gold standard description that most closely represents an objective picture of the real world (Allen et al., 2017).

[1] https://www.mturk.com/

Utilising a pool of experts for behaviour coding offers several advantages. First, experts are presumed to be more accurate than non-experts (Snow et al., 2008), making them preferable when a high-level accuracy is required. Second, complex annotation schemes may necessitate skills and experiences only a few possess. For instance, the ORBIT framework for annotating interrogator and terrorist behaviour (see Alison et al., 2013) demands expertise in psychological research methods, interpersonal behaviour, and effective interviewing strategies. Employing non-experts in such cases could lead to inconsistencies due to unfamiliarity with the domain and insufficient methodological understanding. Additional constraints, such as data access restrictions in the defence and security sector, may further limit the pool of potential coders. Lastly, having fewer expert annotators simplifies the resolution of edge cases, as there are fewer sources of disagreement (Artstein & Poesio, 2008).

However, this approach also has its drawbacks, primarily stemming from the limited pool of available resources, which may constrain the scale of the research. Experts might be unavailable or unwilling to undertake tedious, fatigue-inducing work. Relying on a small group of annotators makes the analysis time-consuming and resource-intensive, even for modestly-sized corpora. Consequently, researchers often need to adapt the amount of data to the available resources (Guo et al., 2016). Analysing conversational data can be laborious. Some estimates suggest that one hour of spoken dialogue may necessitate up to ten hours of manual analysis (Chi, 1997). Distributing work among a limited number of experts may contribute to excessive cognitive load (Mclaren et al., 2007), increasing fatigue, especially during more extended conversations (McGhee et al., 2011). Outsourcing to professional services, such as manual transcription services, may be an option. However, it can be expensive[2] and is unsuitable for niche settings or sensitive data.

Non-expert annotation

When it is infeasible for experts to perform the annotation, researchers may outsource coding activity to non-experts. Crowd-sourcing options such as Amazon Mechanical Turk are becoming increasingly common to quickly annotate a large corpus of (usually text-based) data (Boghrati et al., 2018; Fiore, 2019; R. Levitan et al., 2012; Snow et al., 2008). Here, tasks can be advertised for non-experts to perform in return for a small financial reward. The researcher can specify the number of workers they need and the fee they are willing to pay. Depending on the

[2]This is based on transcription estimates (without additional analysis) quoting costs of up to £10,000 to transcribe a corpus of twenty-five audio recordings.

task, this can provide an efficient and cost-effective alternative to using experts, as a resource can be easily scaled up or down to meet the project's demands. Studies have used crowd-sourcing to perform a variety of tasks such as transcription (Pérez-Rosas et al., 2017), dialogue act tagging (Wen et al., 2015), and to provide quality ratings (Lin & Walker, 2017).

Questions have been asked, however, over the quality of the work generated via crowd-sourcing. Workers' ability is not certified or validated in any meaningful sense. Coupled with the small reward, there is little to prevent workers from quickly submitting meaningless answers to acquire a fee (Hu et al., 2016). However, assuming that most workers are sincere in their efforts, do they serve as a genuine alternative to experts? The answer appears to be 'it depends'. In the study by Snow et al. (2008), it was found that the performance of a single expert in an emotion classification task could only be approximated by utilising four Turkers. This result indicates a significantly higher degree of variation in the label distribution when employing non-experts for the task. In a similar crowd-sourced emotion classification task by Snel et al. (2012), researchers found that non-experts were increasingly prone to spurious labelling practices due to boredom and fatigue effects. A potential solution to this issue could be emphasising worker engagement in the research design, ensuring their contributions are valued and meaningful. Despite the limitations of using novice annotations in some settings, a crowd-sourcing approach can help quickly acquire large amounts of subjective data such as ratings or scores. For example, in an employment interview study, Naim et al. (2018) used Mechanical Turk to collect ratings of interview performance, noting that ratings generated in this manner are more akin to 'audience' perception than to 'expert' labels.

This section has shown that whilst expert and non-expert annotators can both serve a valuable purpose in detecting behaviour during social interactions, considerable limitations exist when relying on a wholly manual approach. These limitations often prohibit the scalability of coding to more extensive and diverse settings and reduce the complexity of behaviours that can be analysed. As an alternative, the following section highlights the advantages of outsourcing these tasks to a computer to perform automatically.

1.1.3 Advantages of an automated solution

Assuming that machines can reliably detect relevant behaviours, the primary advantage of an automated approach lies in its enhanced efficiency. This efficiency

enables the exploration of social phenomena on a much larger scale, offering a more comprehensive perspective than what manual methods alone can achieve. The integration of computation into the social sciences has arrived at a critical moment. Grimmer et al. (2021) observes that while social scientists have historically had to contend with a scarcity of data, the digital age has provided a wealth of real-world material in the form of social media and the internet. In today's context, a well-equipped undergraduate has the potential to work at a scale that would have previously necessitated entire research teams and hundreds of hours of effort. Indeed, while a direct comparison between manual and automated effort may be futile, the author estimates that manually processing the content within this thesis would have taken longer than the entire funding period. However, it is crucial to acknowledge that increasing data volume does not automatically ensure improved research quality. Meticulous planning, a rigorous methodology, and proper data analysis remain indispensable for guaranteeing that such large-scale projects are meaningful and valuable (Lazer et al., 2020).

A related benefit of an automated approach is the ability to perform large-scale quantitative research in real-world settings. Traditional social science has typically obtained knowledge of human behaviour through methods including experiments, interviews, surveys and questionnaires (Karlgren et al., 2020). A disadvantage of these methods is that they only partially reflect the natural world in which they are designed to measure. For instance, researchers design experiments in artificial settings to eliminate confounding variables (Dalati & Marx Gómez, 2018). Equally, as closed questionnaire responses are predefined, respondents may be forced into selecting an answer that only approximates their view (Bowling & Ebrahim, 2005). Computational approaches, by contrast, are equipped to gather structured knowledge of human behaviour direct from observation (Vinciarelli, Pantic, et al., 2009). This increases the ecological validity of research compared to traditional methods (Qiu et al., 2018). Combining a computational approach with appropriate statistical and machine learning techniques presents the opportunity to quickly extract large amounts of high-dimensional data to infer correlation or predict outcomes in real-world settings (Shu & Ye, 2023).

Another advantage enabled by a large-scale computation analysis is the enhanced ability to reproduce research findings. A long-standing issue within the social sciences, most notably fields that more commonly apply statistical techniques such as psychology, is the so-called 'replication crisis' (Maxwell et al., 2015). This is the general observation that many, usually positive, findings reported in psychological

research are not replicated in follow-up studies. For example, a meta-analysis of 100 studies across three prestigious psychology journals, the Journal of Personality and Social Psychology, the Journal of Experimental Psychology, Learning, Memory and Cognition and Psychological Science, reported that only a third of articles could be successfully replicated (Open Science Collaboration, 2015). Failure to replicate can arise for several reasons, including an underpowered sample, post-hoc theorising or *p*-hacking, or more severe cases of malpractice (Adjerid & Kelley, 2018). In the case of behaviour coding, latent or hard-to-define behaviours require a significant degree of interpretation on the part of the annotator. Given this subjectivity, a failure to replicate can arise due to inherent disagreement between those performing the annotation. This can be particularly true when the frequency of a particular item is low (Moyers et al., 2016). By contrast, the efficiency afforded via computation enables analyses to be quickly re-run with ease. A reproducible output can be guaranteed by setting a random seed within the programming script (Dutta et al., 2022). Anyone accessing the script, data, and appropriate seed values can faithfully recreate an identical output.

1.1.4 Disadvantages of automated methods

While the advantages of automated methods for qualitative coding are significant, interdisciplinary researchers should also consider some notable disadvantages. One limitation is that automated algorithms can lack the deep understanding and nuance that human researchers bring to the coding process. For example, automated systems can easily misinterpret subtleties in language, cultural context, and sarcasm, leading to poorer results than those derived from human effort (Sileo et al., 2022). Furthermore, automated methods are highly dependent on the quality of the input data. Any errors or inconsistencies in the data can propagate throughout the analysis, potentially impacting the validity of the findings (Rose & Fischer, 2011).

Another disadvantage is the potential for lack of transparency in automated algorithms, especially when using complex machine learning models. These "black box" models can make it difficult for researchers to understand and explain the rationale behind the decisions made by the algorithm (Rudin, 2019). This lack of interpretability can lead to scepticism in the research community and hinder the adoption of such methods in qualitative research.

Additionally, the reliance on automated methods can contribute to devaluing the importance of human expertise and judgment in the social sciences. Research indi-

cates that there is potential for researchers to become overly reliant on automated methods that appear to perform well (Ashktorab et al., 2021). It is reasonable to suggest that this may reduce the development of critical thinking skills and methodological rigour in studies with inflated or unwarranted confidence in automation. In sum, while automated solutions offer numerous benefits regarding efficiency, scalability, and reproducibility, it is crucial to remain cognizant of their limitations and continue to value human researchers' essential role in qualitative coding.

So far in this chapter, the author has highlighted the advantages and disadvantages of an automated approach to behaviour coding. Advantages include increased scalability, validity, and reliability of analyses, addressing several significant and long-standing limitations within the social sciences. However, as noted by the disadvantages, there is no one-size-fits-all approach to automatically coding behaviour data with computers. This is evident in the various computational approaches routinely deployed in fields such as computational linguistics, where tools from simple dictionaries (Niederhoffer & Pennebaker, 2002) to complex neural networks (Ebrahimi et al., 2016) are used. Each of the approaches utilised over the last few decades has its own merits and drawbacks. Therefore, the next section is dedicated to highlighting these.

1.1.5 Overview of automated coding methods

This section provides an overview of recent literature on social interaction from a cross-disciplinary perspective. As noted in the section above, traditional social science has relied heavily on manual observation, transcription and analysis of small samples of behaviour, partly due to the scarcity of large datasets and the high computational resource costs (Grimmer et al., 2021). This landscape, however, is rapidly shifting. The internet, particularly social media, has increased an individual's digital footprint, presenting opportunities to study social interaction at a previously unimaginable scale. According to reports, over 500 million Tweets are sent per day (Shepherd, 2023). There are just under three billion active Facebook users (Dixon, 2023) sending 14 billion messages per day (Daniel, 2023), and uploading almost 5000 messages every second (Formosa, 2023). As a consequence of this abundance of data, cross-disciplinary fields such as Social Signal Processing (SSP), Behavioural Signal Processing (BSP), and Computational Social Science (CSS) are becoming increasingly important tools in this space. Over the last two decades, the number of cross-disciplinary approaches applied to behavioural coding has steadily grown.

Increasingly, techniques are being developed to incorporate domain knowledge into a scalable computational solution. Given this shift towards increased human-AI collaboration, it is worth outlining the methods and studies that have influenced this re-direction.

Computer-aided technologies

The limitations of relying solely on manual effort in an increasingly digital world have inevitably led to a discussion about what researchers can do to scale up behaviour coding within the social sciences. As the size of available corpora grows, traditional approaches become increasingly restrictive and expensive. Consequently, there has been a growing interest in, and an increasing need for, scalable solutions using computers to automatically analyse the underlying structure of text on more extensive and complex data (Krippendorff, 2018). Increased availability of affordable computational resources has resulted in several technological solutions designed to outsource some of the more time-consuming coding features (Kosinski et al., 2015). These solutions vary in the level of assistance they provide.

At one end of this scale is a collection of software known as Computer-Assisted Qualitative Data Analysis Software (CAQDAS). Here, the researcher still performs the task of classifying behaviours in text, with the software responsible for the organisation and storage of data. Proprietary software[3] such as NVivo, ATLAS.ti, and MAXQDA provide a user interface for researchers to extract text segments from documents, such as transcripts, and organise them into thematic nodes. This software makes it easy to collaborate and share data, with claims that it speeds up coding and enhances academic rigour (Oliveira et al., 2013). It can also quantify the frequency of codes in a text and visualise common terms that characterise a code through word clouds (see Figure 1.1 for an illustration). This approach enables a degree of efficiency unmatched by earlier pen and paper methods, as the software performs 'clerical' tasks that would have been previously conducted by the researcher (Cypress, 2019; Wiedemann, 2013). However, whilst CAQDAS tools are an efficient alternative over traditional methods, the researcher still performs most of the decision-making. Therefore, it remains prone to annotator fatigue and is limited in handling complex coding schemes and generalising to new data.

At the opposite end of this scale are a series of techniques designed to outsource

[3]Equivalent open-source packages exist, most notably RDQA in `R`. However, these appear to be rarely used in comparison to NVIVO. This is presumably due to the steeper learning curve.

Figure 1.1: Example of word cloud describing a "Natural Environment" theme. The word cloud has been generated using NVivo.

the allocation of behaviour codes to a computer. Doing so reduces the heavy demands placed on the human annotator and, as a result, increases the scalability of analysis to potentially billions of input texts (Chuang et al., 2015). Computers, however, do not understand text in the same way that humans do. It is necessary first to transform the text to be understood and analysed algorithmically. Several ways of achieving this differ in the level of specialist knowledge and resources each requires to perform appropriately. Three in particular that have influenced this thesis are (1) dictionaries, (2) social signal processing, and (3) supervised machine learning. The remainder of this section will describe the relative merits of each as they relate to the study of social interactions.

Dictionaries

Many of the earliest automated solutions to coding involved applying dictionaries to text-based data. A dictionary is a list of words that represent and describe a topic or theme. For instance, the concept of a "dog" might be represented by the words `{`Dog', `Hound', `Labrador', `Poodle', `Pedigree', `Bulldog', `Greyhound', `Kennel'}`. Features of a dictionary are generally single words, but sub-words (Haslinger, 1997) and multi-word phrases (Higashinaka & Isozaki, 2008) have also been used. Curating the features that form a particular dictionary can be performed from a top-down or bottom-up perspective. A top-down approach is derived from theory and often requires domain knowledge to compile a list of eligible words that studies validate through experiments (Pennebaker et al.,

2015). Contrastingly, a bottom-up approach requires a training set to extract salient words representing a given concept. Statistical measures such as Term Frequency-Inverse Document Frequency (TF-IDF)[4] can be applied to the training data in order to find the most relevant words to a particular class or document (Fan & Ilk, 2020).

Once a dictionary has been curated, it can be used to classify an unlimited number of unseen texts by comparing the words within each text to those within each dictionary. This step is most often done via a one-hot-encoding approach – where each document or unit of analysis is given a value of 1 if it contains an eligible token from a particular dictionary, and 0 otherwise (Seger, 2018). Other approaches represent each document based on the percentage of words corresponding to a particular dictionary (Niederhoffer & Pennebaker, 2002; Pennebaker & King, 1999; B. H. Richardson et al., 2019), or build topic models based on the distribution of eligible tokens (Gaston et al., 2018). The advantages of dictionaries are that they are straightforward to implement, require limited statistical knowledge, and can be easily used to process millions of input texts with minimal hardware requirements (Wiedemann, 2013). Because of this, they have been used in many applied settings, including law enforcement (B. H. Richardson et al., 2014; Taylor & Thomas, 2008), political speech (Guo et al., 2016), legal settings (Bayram & Ta, 2018), pedagogy (Aspinwall & VerBurg, 2007), and in both spoken (Sun et al., 2020) and text-based communication (Drouin et al., 2017).

However, a common criticism of a dictionary-based approach is its inability to generalise to new data – an issue known as overfitting. Overfitting occurs when a machine learning model learns to capture the noise in the training data, resulting in an overly complex model that performs well on the training data but poorly on new, unseen data (Bishop & Nasrabadi, 2006). Dictionaries overfit because they are predicated on co-occurring words appearing in the dictionary and the input text. A bottom-up or data-driven approach, where salient words are modelled based on their distribution within a single training set, will also likely overfit the specific texts they were trained on. Equally, a top-down approach derived from theory is unlikely to capture the totality of a given concept. For instance, how can we be sure we have captured every possible descriptor of "dog" in the above example? As a result of these issues, dictionaries are prone to an increased number of false positives and false negatives (J. Lee & Hamilton, 2022). This issue is especially true with verbal data

[4]Term Frequency-Inverse Document Frequency is a numerical measure that reflects the importance of a word in a document or a collection of documents. It considers both the frequency of the word in the document and its rarity across the entire collection, helping to identify important words that provide meaningful insights into the text (Ramos et al., 2003)

(as opposed to written data such as documents), as speakers use various spelling and linguistic choices when framing utterances.

Similarly, because dictionaries are typically based on string-matching, they often fail to distinguish between homonyms - words spelt the same but with different meanings. Accordingly, homonymy can also increase the false positive rate. Unless studiously maintained and updated, many dictionaries will fall behind as language trends evolve over time. Consequently, most dictionaries tend to be single or low-use beyond their initial implementation (Scharkow, 2013).

However, a possible exception to this rule is a proprietary word-count tool known as Linguistic Inquiry Word Count (LIWC) (Pennebaker et al., 2015). LIWC is a collection of dictionaries based on high-frequency function words. Function words are semantically-neutral words that include pronouns, articles, propositions, and auxiliary verbs. Collectively they account for over half of the words spoken in conversation, despite representing less than 1% of the English vocabulary (Tausczik & Pennebaker, 2010). Prior research has shown that, despite their regularity, the use of function words is associated with differences in personality (Pennebaker & King, 1999), indicates attractiveness (Ireland et al., 2011), and predicts conflict in relationships (Niederhoffer & Pennebaker, 2002), politeness (Holtgraves & Perdew, 2016), and deception (M. T. Braun et al., 2015). Unlike dictionaries based on content words, focusing on words that appear regularly in speech means LIWC can be easily applied to new domains.

Close inspection of the LIWC dictionaries reveal a substantial overlap between those words classified as function words with stop-words (Saini & Rakholia, 2016b). Whilst there is no universally agreed definition of a stop-word, they are generally defined as words with little semantic utility and include articles (e.g., 'a', 'the', 'an'), prepositions (e.g., 'at', 'by', 'for'), and pronouns (e.g., 'him', 'she', 'they') (Saini & Rakholia, 2016a). Because they are words commonly used in everyday language, researchers often remove stop-words during the pre-processing steps of most Natural Language Processing (NLP) tasks (Jurafsky & Martin, 2019).

LIWC has been regularly maintained, going through repeated iterations since its inception almost thirty years ago (see Pennebaker, 1993). The current version contains over 70 categories, including clout, emotional tone, power, and psychological processes such as anger, anxiety, and affect. Its simplicity has meant that LIWC has remained an influential tool in the field of psycholinguistics[5], despite the increased

[5]According to Google Scholar, the 2001 version of LIWC (see Pennebaker et al., 2001) continues

availability of more powerful techniques (Bahgat et al., 2022; Biggiogera et al., 2021; Boyd et al., 2022; Brandt & Herzberg, 2020; Lumontod III, 2020; Lyu et al., 2023).

Social Signal Processing

Despite their utility, CAQDAS software and dictionaries such as LIWC require the annotator or researcher to work in a manner driven by the technology. Audio recordings, documents, or transcripts must be processed and stored correctly, and dictionaries must be studiously maintained. In both cases, the criteria regarding how a researcher performs behaviour coding are oriented towards the computer's capabilities. This is undoubtedly a weakness, as it does not reflect how humans think. In contrast, when humans interpret social interactions, they attend to a variety of different social cues based on linguistics (*what* is said), paralinguistics (*how* something is said), and a host of other signals such as gaze, posture, and physical gestures. Interpreting these social cues is not performed independently but collectively processed by our brains. Because of this, humans can interpret both behaviour and intent in social interactions with minimal prior knowledge (Vinciarelli, Salamin, et al., 2009) – something that machines, for the most part, are incapable of doing.

There is a growing interest in building models to analyse the *social signals* that individuals exchange during real-world interactions. Social Signal Processing (SSP) is a growing cross-disciplinary domain incorporating aspects of psychology, cognitive science, and signal processing (Narayanan & Georgiou, 2013; Vinciarelli, Pantic, et al., 2009). SSP is designed to develop modern computers capable of social intelligence. Historically, building complex behavioural models has proven challenging to measure beyond strict experimental control (A. Pentland, 2007). However, general improvements in automated technologies such as body sensors (Narayanan & Georgiou, 2013), conversational agents (Zheng et al., 2022), and speech recognition (Malik et al., 2018) have meant that researchers are increasingly able to automatically build sophisticated models of behaviour based on low-level activity acquired from sensors or other raw data such as transcripts.

In developing a scalable alternative to subjective human effort, a SSP approach is helpful for several reasons. First, social signals are observable low-level behaviours (Vinciarelli, 2017). Focusing on observation means behaviours can be measured ob-

to receive an average of 600+ new citations each year. LIWC has also been extended into many languages, including Spanish, French, German, Dutch, Russian, Portuguese, Romanian, Japanese and Korean.

jectively rather than requiring a subjective interpretation by a human annotator. Using computers means that the researchers can reliably reproduce the detection of behaviours by setting seed values and saving models (Loper & Bird, 2002). Therefore, such an approach contrasts typical manual effort, which is prone to variation within and across annotators due to inherently human issues such as fatigue (Maclin & Maclin, 2005). Avoiding fatigue issues also means the approach is scalable to larger datasets and adaptable to new domains without extensive re-training. Moreover, using algorithms to define behaviours of interest means greater feature complexity without relying on a human-interpretable definition (Bengio et al., 2013).

A second advantage is a capacity to analyse social interactions by examining how speakers coordinate their low-level social activities in response to one another. Social Signal Processing (SSP) is grounded in the idea that the social signals produced by speaker B will change according to the social signals generated by speaker A (Vinciarelli, 2017). This suggests that signals are not random or independent events but follow specific principles and patterns. Consequently, it may be possible to deduce social motivations by modelling how speakers coordinate their behaviour throughout an interaction. For example, from a social psychological perspective, increasing similarity along linguistic or non-verbal dimensions is often considered an affiliative action (Chartrand & Bargh, 1999; Lakin et al., 2003). In general, we tend to amplify similarities to enhance rapport with others. At the same time, we emphasise differences or disagreement by acting more dissimilar (H. Giles et al., 1991). The extent, effects, and direction of coordination have been extensively studied in SSP research (see Bonin et al., 2013; Burgoon et al., 2017; C. C. Lee et al., 2014; R. Levitan & Hirschberg, 2011; Lord et al., 2015), making it a central component of the current thesis.

A third related advantage is the potential to examine this behavioural coordination from a multi-modal standpoint. Coordination is not unidirectional; we might converge on some behaviours while diverging on others (Danescu-Niculescu-Mizil et al., 2011; Danescu-Niculescu-Mizil & Lee, 2011). The high cognitive demands placed on manual annotators inevitably require a simplified perspective of social activity, as it is unrealistic to expect humans, even experts, to account for the simultaneous coordination of lexical, semantic, acoustic, hand, eye, and facial behaviours without investing significant time, effort, and resources.

Social signals represent low-level activity that, in isolation, offers minimal explanatory or interpretive value in a real-world context. However, there is increasing cross-disciplinary interest in using SSP to model conversational outcomes. For ex-

ample, in Georgiou et al. (2011), the authors used a combination of acoustic and lexical signals to detect behaviours such as blame and acceptance in a corpus of couples therapy sessions. In another example from the counselling domain, Imel et al. (2014) used a vocal feature-set to measure ratings of therapist performance. Other research has modelled collaborative task success from a combination of lexical (Friedberg et al., 2012; Nenkova et al., 2008), syntactic (Reitter & Moore, 2007, 2014), and acoustic markers (R. Levitan & Hirschberg, 2011). Furthermore, facial expressions, topic models, and vocal and linguistic behaviour have been used to predict performance and outcome in a job interview (N. C. Chen et al., 2018; Naim et al., 2018; Nguyen & Gatica-Perez, 2015). SSP has also been used to explore behavioural differences in deceptive versus non-deceptive dialogue (Dunbar et al., 2014), and rates of confession in police interrogations (B. H. Richardson et al., 2014).

Machine Learning

A weakness of the Social Signal Processing approach is that it focuses on surface-level behaviours (Kochanowicz et al., 2016), and therefore requires the explicit engineering of features of interest at a granular level (i.e., defining rules for individual words or phrases). These features are based on predefined rules or algorithms that the researcher sets. A simple example is TF-IDF, where each word in a document or transcript is weighted by its relative frequency in the remainder of a corpus (Ramos et al., 2003). The granularity of SSP features can mean that individual features themselves, whilst providing a particular predictive capability, offer minimal interpretability and explainability from a social scientific or psychological perspective. Consequently, much of the knowledge gained using SSP can be challenging to align with the types of behaviour most often found in the social sciences. Social science researchers are invariably interested in more abstract and latent behaviours (Bass & Semetko, 2021; Netemeyer et al., 2003). Examples of these behaviours, such as empathy and rapport, can be challenging even for humans to define (Heide, 2013). Correspondingly, previous attempts to use an SSP approach to detect these behaviours have received mixed results (Gaume et al., 2019; Imel et al., 2014).

An alternative approach is to avoid the need to explicitly define a behaviour algorithmically by providing positive examples to a machine. Analogous to traditional supervised machine learning, a dataset comprising both positive and negative class instances is provided as training input, which the machine then uses to learn a representation of a given concept (Hastie et al., 2009). Whilst this requires initial

manual effort to create a labelled dataset, a sufficiently trained language model can perform coding on a much more extensive collection of unlabelled data. Furthermore, modern advances in NLP, particularly deep learning architectures, indicate that machines are increasingly capable of this task. In one study, Ando et al. (2017) used neural networks to code a corpus of online dialogues with verbal behaviours such as gratitude, agreement, greeting and replies. Another example combined verbal behaviours (i.e., words) with additional contextual and sentiment features to classify speech acts in a customer service setting (Fan & Ilk, 2020). Similarly, using k-Nearest Neighbour, Flor and Andrews-Todd (2022) coded online chat messages with social and cognitive behaviours, including negotiation, sharing information, and planning. Researchers within the computer sciences have also used this approach to classify documents, with Scharkow (2013) using a Naive Bayes classifier to perform an automatic content analysis on German news articles. Moreover, Burnap et al. (2015) used support vector machines to detect racial tensions on Twitter.

Machine learning, however, is not a faultless solution. First, manual methods performed by experts are typically more accurate than even the best-automated approaches (Learning, 1997). As a result, computers can fail in cases where humans would not (Rosé et al., 2008). Improvements in accuracy can often be attained by providing more training input. However, gathering sufficient manual data is expensive and often impractical (Grimmer et al., 2021). Models trained with insufficient data are likely to overfit, meaning they may not generalise well beyond the initial corpus. Overfitting can pose a challenge for niche or novel frameworks, as acquiring sufficient training data will be non-trivial and highly time-consuming. Additionally, many machine learning algorithms are considered black boxes, making it difficult to understand the internal workings and decision-making processes (Lipton, 2018). Consequently, it can be challenging to explain why a particular prediction has been made (Rudin, 2019). Supplementary algorithms such as LIME or SHAP exist to aid the explainability of machine learning models (Suvorova, 2022). However, such techniques often require statistical knowledge that may exceed that of the typical qualitative researcher (N. C. Chen et al., 2018).

Low model explainability can impact the level of trust a user may have in an automated solution. An absence of sufficient understanding can mean that users mistrust the predictions made (Ribeiro et al., 2016). The inverse of this can be equally problematic, however, whereby an absence of sufficient knowledge can result in users too-readily accepting a machine prediction – a concept known as 'overtrust' (Yang et al., 2020). Studies have shown that people express more trust in ML

solutions that are demonstrably accurate (Yin et al., 2019).

However, it is essential to acknowledge that machine learning models, especially those dealing with human behaviour, can be influenced by subjective reasoning and biases in the training data (Celis et al., 2020). These biases can affect the accuracy and fairness of the models, ultimately impacting user trust (Barocas et al., 2019). Consequently, any predictions a machine makes must not be automatically accepted as correct but are inspected and appropriately interpreted. Given the subjectivity of human behaviour coding and potential biases in ML models, performing error analyses on automated predictions or comparing the level of agreement between humans and machines is highly important (Ameisen, 2020).

Transformer Architecture

A significant breakthrough in natural language processing and machine learning has been the development of the Transformer architecture, which has led to a new generation of more powerful and efficient models for text analysis (Vaswani et al., 2017). The Transformer architecture addresses several limitations of traditional machine learning models and offers promising opportunities for qualitative coding tasks in the social sciences.

Unlike earlier neural network architectures such as recurrent neural networks (RNNs) and long short-term memory (LSTM) networks, Transformers do not rely on sequential processing (Tabani et al., 2021). Sequential processing means the language model processes words or sentences sequentially and updates its internal state after each step. As the model updates its understanding at each step, these approaches can encounter difficulties detecting dependencies between words or sentences over varying ranges (Hochreiter & Schmidhuber, 1997). To overcome this, Transformers employ self-attention mechanisms to capture long-range dependencies and complex patterns within the text more effectively (Vaswani et al., 2017). Self-attention can also mean that Transformers are faster to train than other neural networks, as models can process input texts in parallel (Xu et al., 2020). Capturing complex linguistic patterns is particularly important for social science methodologies such as grounded theory, as the meaningful interpretation of human behaviour often requires understanding the context in which it occurs (Pidgeon et al., 1991).

Transformer-based models, such as BERT (Devlin et al., 2018), GPT-2 (Radford et al., 2019), GPT-3 (Brown et al., 2020), and the recent GPT-4 (OpenAI,

2023) have demonstrated impressive performance in a wide range of natural language understanding tasks, including sentiment analysis, question-answering, and text classification. In addition, their ability to generate high-quality text representations makes them well-suited for qualitative coding tasks, as they can capture the underlying semantics and context required for accurate behaviour annotation.

However, despite the many advantages of Transformer architectures, some challenges must be addressed. One issue is the computational complexity of these models, which often require large amounts of computational resources for training and inference (Strubell et al., 2019). This can pose a barrier for social science researchers with limited access to high-performance computing infrastructure. Moreover, like other machine learning models, Transformers are also vulnerable to overfitting. Consequently, they may struggle to generalise when trained on limited or highly unbalanced datasets (Behera & Dash, 2022). This highlights the importance of carefully curating training data and applying regularisation techniques to ensure the robustness and generalisability of the resulting models.

Lastly, the interpretability of Transformer-based models remains a challenge, as their complex attention mechanisms and the vast number of parameters can make it challenging to understand the reasoning behind their predictions (Du et al., 2019). Developing methods to increase the transparency and explainability of Transformer models is an ongoing area of research. Social scientists must engage with these efforts to maximise the potential of these powerful tools for qualitative coding tasks.

In summary, the Transformer architecture offers a promising direction for automating qualitative coding tasks in social science research. Despite the challenges associated with computational complexity, generalisation, and interpretability, these models have demonstrated the ability to capture complex patterns and context within the text. With continued advancements in NLP and machine learning, Transformer-based models hold great potential for enhancing the efficiency, accuracy, and scalability of qualitative coding in social sciences.

This section has introduced the fundamental technologies inherent in an automated approach to behaviour coding. Additionally, the relative merits and drawbacks of each approach have been presented. However, one of the central issues in this space is the challenges associated with cross-disciplinary research. Whilst these challenges are not the focus of the current thesis, they are worth highlighting. As such, they are briefly described in the next section.

1.1.6 Barriers to cross-disciplinary research

This thesis adopts a cross-disciplinary approach, exploring aspects of the social sciences and utilising techniques more often found within the computational sciences. The synthesis of these disciplines is a growing area of research but one very much in its infancy, with many pertinent issues yet to be fully explored.

Several recent review papers have outlined some of the research gaps between traditional social science and machine learning (Adjerid & Kelley, 2018; N. C. Chen et al., 2018; Grimmer et al., 2021; Mahmoodi et al., 2017; Yarkoni et al., 2021; Yarkoni & Westfall, 2017). These studies have suggested a natural tension between the two disciplines along several dimensions. These include a lack of methodological overlap, misalignment on explanation versus prediction, and whether research should seek to optimise accuracy over efficiency.

As a growing area of interdisciplinary study, such disparities highlight the challenges faced by researchers working in this area and warn of the dangers posed by accepting a simple 'plug-in and play' mindset. Indeed, Lazer et al. (2014) warn against big data hubris, remarking that the increased scalability offered by automation does not decrease a dependency on foundational aspects of social science such as sampling and ethical considerations. Instead, given these disparities, understanding the necessary trade-offs between the two disciplines is vital for encouraging an interdisciplinary approach with sound methodological rigour.

One of the main issues preventing greater collaboration between the social and computational sciences has been the historical lack of overlap between the two fields. According to Allen et al. (2017), this disparity exists for several reasons. First, as each discipline originates from a different epistemology, they are designed to ask different questions. The design of social science research can be inductive or deductive, with its central purpose being to test hypotheses, develop new or existing theories of behaviour, and understand the intricacies of human and societal interactions. Conversely, computer science research aims to create innovative algorithms, technologies, and automated solutions that address complex technical challenges while striving for continuous performance and efficiency improvement (Cioffi-Revilla, 2014).

Second, the two fields' contrasting aims call for different data collection practices. Much of the social sciences is based on collecting an *adequate* data sample (Grimmer et al., 2021). Researchers typically obtain data using qualitative methods such as surveys, interviews, and observation (Mohajan et al., 2018). Data collected tends to

be single use, as the repeated use of data is considered unethical within the social sciences (Allen et al., 2017). By contrast, a computer science approach obtains low-level behavioural data, such as that obtained from sensors (Vinciarelli, Pantic, et al., 2009). Data samples tend to be many orders of magnitude greater than even the most extensive social science studies due to the automated nature of data collection. Unlike the social sciences, datasets tend to be made available to the broader research community and the public (i.e., Kaggle) (Allen et al., 2017).

Lastly, the types of analysis differ between the two fields. As discussed in Section 1.1.5 of this chapter, software such as CAQDAS has become an increasingly popular way for social scientists to quantify codes attributed from observed behaviour. Objectivity and reliability of the coding effort are obtained by conducting multiple data passes using independent annotators and measuring the pairwise agreement between them (Jordan & Miller, 2003). Within machine learning, it is expected that performance is tested on an out-of-sample dataset (Tsamardinos et al., 2018). In other words, data that did not feature in the initial training. Evaluation metrics such as accuracy, precision, and recall are then used to automatically measure the ability of a model to predict new data. Generalisability is obtained by performing cross-validation - repeating the process of data splitting into many training and testing regions and averaging performance (Domingos, 2012).

Another underlying tension between the two disciplines concerns how they seek to understand and model behaviour. This is based on whether the research goal is oriented towards *explanation* or *prediction*. For much of the social sciences, emphasis is placed on explanation. When the primary goal is an explanatory one, researchers seek to describe a behaviour's causal underpinnings (Yarkoni & Westfall, 2017). For instance, the factors that result in biases in job recruitment or disparities in police stop-and-search. Conversely, the core objective of machine learning is to learn a representation of data that can generalise beyond an initial training set (Domingos, 2012). In other words, to predict future behaviour. For example, predicting the outcome of elections from economic data (Sinha et al., 2020).

The challenges faced in integrating the two domains have meant that social and computer science have historically remained siloed from one-another (Lazer et al., 2020). However, sufficient computational power and large amounts of behavioural data have become more widely available for social researchers (Watts, 2007). As such, to continue working in isolation would represent a missed opportunity. There is an increasing demand for social research that utilises computational technologies (Jager et al., 2020). Recognising the value of interdisciplinary expertise, researchers

and institutions are actively promoting collaboration and skill development to bridge the gap between these disciplines (Endo, 2017; Lazer et al., 2009; Lazer et al., 2020). This collaboration can foster a new generation of experts proficient in both social and computational sciences.

However, the social sciences are lagging behind other fields concerning utilising computational resources to address theoretical problems (Yarkoni et al., 2021). This reticence is partly because many qualitatively-focused researchers lack the technical expertise to work with complex computational techniques such as machine learning and NLP (N. C. Chen et al., 2018). Machine learning typically requires knowledge of applied mathematics, linear algebra, and programming languages such as Python or R. Many, if not all, of these skills, are developed within universities' computer science and statistics departments (Yarkoni et al., 2021). By contrast, a lack of knowledge of the so-called 'big data' approaches is a significant entry barrier for students and researchers within the social sciences (Metzler et al., 2016). For researchers from a primarily qualitative background, learning complex technical skills may evoke anxiety towards statistical approaches, which further decreases the potential for cross-disciplinary collaboration (Jager et al., 2020). The inverse problem is equally pertinent, however. Whilst most computer scientists can comfortably generate machine learning models from social data, they may be less likely to possess sufficient domain awareness. This lack of knowledge can lead to subtle but impactful features being overlooked in the analysis. The previously mentioned function word/stop-word utility is a simple example of this - where a computer scientist might routinely remove something that turns out to be important from a social scientific perspective. Similarly, a feature selection approach that remains agnostic to the domain in question might lead to model overfitting (Lazer et al., 2014). The complexity of social phenomena also means that behaviours often cannot be easily approximated through algorithmic means alone (N. C. Chen et al., 2018). This is especially true when behaviours are abstract, hard to define, or require a high degree of contextual knowledge. Therefore, the role of social scientific knowledge is vital to ensuring that automated solutions are modelling behaviour in both an impactful and appropriate manner.

1.2 Thesis statement

The following thesis statement summarises the approach and claims made throughout this thesis:

> Analysis of social interactions that infuse domain knowledge with modern computational technologies can: (claim 1) increase the scale of analysis compared to manual coding, (claim 2) improve the efficiency and reliability of classifications by reducing manual effort, and (claim 3) increase the complexity of the behaviours analysed. Consequently, (claim 4) a cross-disciplinary approach can be exploited to perform large-scale behaviour coding consistent with groups of experts.

These claims align with the research questions addressed in the following chapters. A summary of the questions covered throughout this thesis is provided in the following section.

1.3 Research overview

A summary of the main research themes and questions examined throughout this thesis are presented in Table 1.1.

Research Question	Covered in...
Can computational techniques be used to overcome limitations of expert labelling?	Chapter 3 4 5
What tasks are more or less suited to a computational solution?	Chapter 5 6
Do machines perform comparably with humans?	Chapter 3 3 5
Can an automated solution extend domain knowledge and advance theories of social behaviour?	Chapter 2 3 4 5 6

Table 1.1: Summary of research questions

This work is comprised of two research components to examine the validity of the thesis statement. The first part of the thesis spans the three chapters between Chapter 2 to Chapter 4. Leveraging the SSP approach discussed above (see Lazer et al., 2014; A. Pentland, 2007; Vinciarelli, Pantic, et al., 2009), psychological behaviours are automatically extracted from transcripts and audio recordings of a corpus of political interviews. The research aims of this section of the thesis are to automatically model behaviours associated with effective interviewing techniques and examine whether they contribute to the conversational quality of the interaction.

By focusing on low-level features such as the linguistic and acoustic behaviour of the interviewer, this section aims to provide a scalable alternative to traditional social scientific methods that tend to utilise high levels of manual effort. In particular, this section of the thesis focuses on the effects of the interviewer mirroring the verbal and non-verbal behaviour of the interviewee. As a growing area of cross-disciplinary research, prior research has developed an increasing number of measures to identify behaviour mirroring. However, they often achieve conflicting results. Therefore, Chapter 2 aims to aggregate several of these measures into a single Python package named `CoPyCAT`. This package measures the degree of verbal mirroring between two speakers across several lexical, stylistic, syntactic, and semantic dimensions. The work covered in this section is heavily influenced by a social psychological theory of behaviour mirroring known as Communication Accommodation Theory (CAT) (H. Giles et al., 1991). According to CAT, speakers that share a communicative goal will increasingly tend to mirror each others' behaviour as the conversation unfolds. Based on this theory, the thesis assumes that the interviewer mirroring the previous interviewee's utterance will predict conversational quality.

Chapter 3 formally examines this hypothesis by applying `CoPyCAT` to a corpus of $N \approx 700$ political interviews. Research in other types of conversation has suggested that verbal mirroring predicts positive social outcomes (Kulesza et al., 2014; Nenkova et al., 2008; Reitter & Moore, 2007, 2014; van Baaren et al., 2003). As the domains typically explored focus on collaborative mirroring, this chapter aims to assess whether more adversarial types of conversation follow a similar pattern. To examine this, a supervised machine learning approach is applied to a feature vector of interviewer speech generated from `CoPyCAT`. An automated measure of conversational success is sought to minimise reliance on human effort. Four outcomes derived from the interviewee's speech (specificity, clarity, diversity, and relevance) are proposed as outcome measures. The suitability of each outcome is assessed and confirmed in a small validation study with eight non-expert human raters. The four success measures are then used as the target variable in the aforementioned supervised machine learning task.

Encouraging results in Chapter 3 motivate the experiments conducted in Chapter 4. Here, features derived from `CoPyCAT` are combined with non-verbal mirroring features to explore whether a multi-modal approach enhances prediction performance. The chapter begins by describing the methodological framework used to measure non-verbal mirroring. The chapter then examines the reliability of the non-verbal mirroring feature set by comparing values obtained against augmented

data. Three experiments are then performed using supervised machine learning in a manner similar to the previous chapter.

The first experiment was influenced by CAT research that indicates that rather than a ubiquitous behaviour, the extent that mirroring occurs depends on contextual factors such as gender (Bilous & Krauss, 1988; R. Levitan et al., 2012). To test this in the confines of a political interview, a multi-modal feature set comprised of verbal and non-verbal mirroring by the interviewer was used to classify the gender composition of the interview participants (same-gender or mixed-gender).

The second experiment followed a similar methodology to explore differences in political partisanship (i.e., did the interviewee belong to the same political party as the interviewer?). The motivation behind this experiment was the consensus that mirroring is associated with rapport-building (Lubold & Pon-Barry, 2014) and affiliation toward an in-group (Lakin et al., 2008). It was therefore anticipated that interviewers would mirror differently based on whether the interviewee represented the same or different political orientation.

The final experiment in this section sought to build upon the findings of Chapter 3 by attempting to predict the outcome of a political interview from the mirroring behaviour of the interviewer. The same outcome variables as before were used. A multi-modal feature set based on verbal and non-verbal features was used as predictors.

The second research component shifts from a SSP based approach requiring no human labelling to a deep learning approach that seeks to leverage a small amount of expert-annotated data. The motivation was to avoid the need to define behaviours (i.e., rapport-building) algorithmically and explicitly. To better match the domain knowledge of the author, the conversational domain also changes. Focus shifts from behaviours of effective interviewing to predatory behaviours in online Child Sexual Exploitation (CSE). In addition to this, as no explicit feature engineering is required, the number of behaviour categories explored increases from one (mirroring) to eleven. These behaviours were identified collaboratively with Investigative Psychological experts from the University of Liverpool. In Chapter 5, a small ($n \approx 5000$) corpus of predatory messages addressed to a decoy posing as a child is manually annotated before being used as training input to a pre-trained language model. The main objective of this chapter is to explore whether a language model can learn a sufficient representation of predatory behaviour to perform automatic labelling of a much larger corpus ($n \approx 1.3m$ messages). Classification performance is compared

to ground truth using a hold-out set. Given the time-intensive nature required to establish ground truth, classification accuracy is examined with various training data, from zero-shot to full-shot. Additional tests included expanding the message window to determine whether additional contextual information provided by victim messages aided model accuracy.

In the final research chapter, Chapter 6 examines the trade-off between model accuracy and efficiency through a collaborative human-AI approach. Using the same domain and data as the previous chapter, the problem is re-conceptualised as an information extraction task. This chapter aims to determine whether a human-in-the-loop approach can extract predatory behaviour with the precision of human expertise whilst maintaining the efficiency of machine automation. A weak supervision methodology is constructed, whereby a series of labelling functions extract segments of the chat log for a user to verify. The findings of this chapter highlight both the promise of computation whilst emphasising the importance of expert knowledge in these increasingly technical systems.

The thesis concludes with a short discussion of how the main research findings have contributed to the increased integration of computational solutions within the social sciences. The author provides additional commentary highlighting the limitations of the methods used within the thesis and the potential direction of future research in this area. The final section of the current chapter now provides a summary of the thesis contributions.

1.4 Summary of contributions

The primary aim of this thesis is to examine the suitability of computational techniques and automation to the task of behaviour coding within the social sciences. The thesis focuses on two conversational domains: (a) a face-to-face political interview[6], and (b) online interactions between an adult and decoy posing as a child. These domains were chosen because of their real-world significance, although the approaches used have been intentionally designed to remain adaptable to other forms of dyadic interaction. The decision to focus on these types of conversation specifi-

[6]It was initially expected that a central component of this thesis would focus on suspect interviewing during a police investigation. This ultimately proved impossible due to constraints placed on data access. Political interviewing was chosen as a proxy due to its similar structure (dyadic, conflicting goals, outcome oriented), and the relative ease of accessing a sufficient quantity of second-hand data from online repositories.

cally was influenced, in part, by conversations with domain practitioners. Experts in both interviewing and child protection services cited the detrimental impact of having to manually label transcripts and chat-logs. Considerable time requirements and excessive cognitive load were both given as substantial blockages in their respective domains. In both cases, the value of automatically labelling relevant behaviours was deemed useful. Potential applications included evaluating and training interviewers with enhanced conversation-management skills, and automatically flagging potentially harmful online behaviours for human assessment.

The findings generated from this thesis are an encouraging sign that a cross-disciplinary approach can reduce several of the more time-consuming aspects practitioners faced. However, findings also indicate that the extent an entirely automated approach can or should be exploited is unlikely. Findings from the thesis indicate that in the case of high-stakes settings such as those explored here, automation must work with, not in place of, the human. This is especially true in the case of CSE, where anything less than perfect recall would arguably place children in considerable danger. Instead, the greatest application of an automated approach appears to be one that can filter or flag potential interesting data-points to experts for further examination. With this objective in mind, the contribution to knowledge claimed throughout this thesis is summarised accordingly:

Chapter 3

- Demonstrate an automated approach can extract meaningful interviewer behaviour.
- Develop a measure of interview success from the verbal behaviour of the interviewee.
- Demonstrate evidence that verbal mirroring by the interviewer predicts the conversational quality of the interview.
- Highlight important features, and perform an error analysis indicating where the model fails.

Chapter 4

- Build on the previous analysis to include non-verbal mirroring.

- Demonstrate that interviewers mirror the non-verbal behaviour of the interviewee
- Explore differences in mirroring based on gender composition and political partisanship.
- Compare unimodal and multimodal models on predicting conversational quality.

Chapter 5

- Examine performance of a deep learning model on classifying predatory behaviour in a corpus of known child sexual exploitation.
- Explore the minimum number of expert-labelled instances required to achieve a suitable level of performance.
- Apply the model to a larger unlabelled corpus.

Chapter 6

- Explore the trade-off in efficiency versus accuracy when detecting predatory behaviour through a collaborative human-AI approach.
- Compare differences in accuracy when the model generated more versus fewer suggestions for human verification.

Contributions made by others

Where appropriate, a section has been added to the end of each chapter to highlight the contributions made by others throughout this thesis. For clarity, these contributions are also summarised here:

In Chapter 3, one other researcher provided valuable insight and support in addition to the author's supervisory team. Several outcomes of this chapter were included in a co-authored manuscript (see Cook et al., 2021, for details). The author wishes to acknowledge the co-authors' contributions, particularly their support in reviewing several drafts of the accepted manuscript.

Chapters 5 and 6 utilised an unpublished coding dictionary created by a postgraduate student at the University of Liverpool. This dictionary was created as part of the student's MSc. dissertation in Investigative and Forensic Psychology, as was the effort to manually label the chat logs later used as training data in the current thesis. The postgraduate student, alongside their supervisor, are both co-authors on a manuscript currently under review (see Cook et al., 2023, for details).

Lastly, the weak supervision approach used in Chapter 6 was developed in collaboration with a fellow PhD student at the University of Surrey. A manuscript has been produced detailing this work, of which the author of the current thesis is included as a co-author (see Butcher et al., 2023, for details).

Aside from the details mentioned here, unless otherwise stated, all work included within this thesis was performed by the author.

Chapter 2

Developing a Python Package for Measuring Verbal Mirroring in Naturalistic Settings

2.1 Introduction

The primary aim of this thesis is to explore the suitability of computation as an alternative to manual behaviour coding traditionally used within the social sciences. As a growing area of cross-disciplinary research, several potential solutions have been proposed. Throughout the thesis, two popular methods of automated coding are explored: (1) an approach based on Social Signal Processing (SSP), and (2) an approach based on the transformer architecture. The following three chapters focus on SSP.

As highlighted in the previous chapter, SSP is a growing cross-disciplinary framework at the intersection of psychology, computer vision, and speech and signal processing (Vinciarelli, Pantic, et al., 2009). At its core is the notion that human behaviour can be inferred automatically through machine-detectable 'social signals' (Vinciarelli, Salamin, et al., 2009). Prior research defines a social signal as a behaviour that directly or indirectly conveys information about a social action, emotion, or intent (Moreno et al., 2013). Social signals include overt signals, such as words and vocal mannerisms, and covert signals, such as heart rate (Georgiou et al.,

2011). A social signal that has received considerable attention from both the social scientific and SSP communities are referred to as 'mirroring'[1] – the tendency for speakers to adopt both the vocal and non-vocal mannerisms of their fellow interlocutor (Chartrand & Bargh, 1999). Studies of mirroring have linked its presence to success in conversation (Garrod & Pickering, 2004; Nenkova et al., 2008; Reitter & Moore, 2007, 2014), with suggestions it reflects underlying emotions (Niederhoffer & Pennebaker, 2002), and contributes to prosocial effects such as generosity (van Baaren et al., 2003), rapport-building (Lubold & Pon-Barry, 2014), and social attraction (Ireland et al., 2011). A wide range of studies that explore both the presence and effects of mirroring have been undertaken across multiple disciplines, including social psychology, cognitive science, linguistics, and computer science. This diversity has resulted in a scattered methodological landscape, with many competing approaches often resulting in conflicting findings. These studies often take place under conditions of strict experimental control. As such, the presence and effects of mirroring in real-world dialogue have been debated (Healey et al., 2010). Therefore, this methodological chapter aims to outline mirroring as a suitable alternative to time-intensive manual annotation. Specifically, this chapter will conduct the following:

1. Provide an overview of the theoretical background of mirroring as a social signal.
2. Describe the recent literature on automated detection of mirroring.
3. Introduce `CoPyCAT`, a python package designed to automatically measure verbal mirroring between two speakers based on multiple linguistic properties.

2.2 Related work

2.2.1 What is mirroring? The background of verbal mirroring as a social signal

A common observation regarding conversation is how often an individual will mirror the behaviour of an interlocutor. Speakers in conversation have been observed mirroring each other's linguistic choices, including word-choice (Nenkova et al., 2008),

[1] Other terms used within the literature include convergence, divergence, accommodation, alignment, entrainment, mimicry, and synchrony

linguistic style (Niederhoffer & Pennebaker, 2002), and syntax (Reitter & Moore, 2014). At a paralinguistic level, speakers mimic each other's cadence (Schultz et al., 2016), pitch (Lubold & Pon-Barry, 2014), and vocal energy (R. Levitan & Hirschberg, 2011). Speakers have also been observed mirroring each other's hand gestures (Bergmann & Kopp, 2012) and facial expressions (Chartrand & Bargh, 1999). Therefore, aligning our behaviour with that of a fellow speaker appears crucial in how we organise ourselves during conversation (Garrod & Pickering, 2004). Consequently, it is unsurprising that mirroring has been observed in many forms of social interaction. It has been observed in everyday conversation (Kulesza et al., 2014; van Baaren et al., 2003). It is also evident in professional settings such as a courtroom (Bayram & Ta, 2018), presidential debates (Romero et al., 2015), and police interrogations (B. H. Richardson et al., 2014). It appears not to be confined solely to physical (i.e., face-to-face) interactions, having also been observed in online computer-mediated-communication (Liebman & Gergle, 2016), and through dialogue with conversational agents (i.e., chat-bots) (Brennan, 1996).

The ubiquity of behavioural mirroring raises interesting questions regarding its purpose as a social tool. Pickering and Garrod (2004) argue that mirroring (also referred to as *alignment*) is an essential component of dialogue production. This is because successfully producing dialogue is considered a form of joint action that requires the coordinated efforts of both speakers. Instead of a sequence of independently-generated monologues, speech turns are produced concordantly with the rest of the conversation; they are patterned, like a dance (Niederhoffer & Pennebaker, 2002). Dialogue should pose a considerable cognitive challenge compared to monologue because humans cannot anticipate what a fellow interlocutor will say. Instead, humans can seamlessly listen-think-respond to an utterance with minimal difficulty. Garrod and Pickering (2004) argue the primary reason for this is that alignment of verbal and non-verbal behaviour results in speakers forming a shared mental representation of the conversation. In other words, it helps speakers establish a common grounding (Clark, 1996).

There is also a considerable social element to why mirroring occurs in conversation. A prominent theory in this area is CAT[2] (H. Giles, 1973; H. Giles et al., 1991), which postulates that mirroring is a social tool intended to reduce psychological distances between speakers. It is subsequently argued that humans tend to mirror in settings where socially aligning with an interlocutor is self-beneficial. For instance, in Lakin et al. (2003), the authors found that the rate of non-verbal mir-

[2]Also known as Speech Accommodation Theory in earlier work.

roring increased when an individual was excluded from a social group from which they were initially a member.

Consequently, CAT claims that mirroring reflects power dynamics in a relationship (H. Giles, 1973), with evidence indicating that increased mirroring reflects subservience to the perspective of the other speaker (Niederhoffer & Pennebaker, 2002). Similarly, considerable evidence indicates that mirroring is a crucial component in building rapport (Lubold & Pon-Barry, 2014), particularly where the conversation is cooperative (B. H. Richardson et al., 2019). Research has demonstrated that mirroring increases in conversations where affiliation is the central objective of the conversation (Lakin & Chartrand, 2003). Perhaps unsurprisingly, therefore, intentionally mirroring the behaviour of an interlocutor (also known as a reflection) is a much-advocated technique in several professional settings where building rapport is vital. Examples of this include client-centred therapy (H. Giles et al., 1991; W. R. Miller & Rollnick, 2012) and police interrogation (Alison et al., 2013).

In addition to the social advantages already discussed, a large body of experimental research within social psychology has demonstrated how behaviour mirroring often results in prosocial outcomes at an individual and group level. For instance, in a study of tipping behaviour, van Baaren et al. (2003) found that servers who mirrored a customer's order verbatim received a larger tip than those who paraphrased or acknowledged the order through other means. Presumably, repeating the order back to the customer signalled a level of attentiveness that resulted in increased generosity towards the waitress. This finding also appears to extend beyond the remit of the dyad, increasing generosity and prosociality more broadly. A similar study found that when a teller at a currency exchange centre repeated a customer's request verbatim, the customer was more likely to comply with a later request by the teller to donate to an unrelated charity (Kulesza et al., 2014). Findings such as these have subsequently been used to prove that mirroring behaviour is an effective strategy for persuading and influencing others (Jacob et al., 2011; Maddux et al., 2008).

A central criticism of this claim and the studies mentioned above is that experiments are typically based on short, discrete conversations. Under these conditions, the transactional nature of a sales pitch or restaurant interaction likely constrains the production of free-flowing, naturally occurring language. As such, verbatim repetition in these settings may be more readily associated with attentiveness, particularly in highly-specific social interactions such as ordering in a restaurant. However, how this translates into more prolonged everyday interactions needs to be explored. This

criticism is partly supported by research in more naturalistic settings, which suggested that the presence of mirroring as a ubiquitous behaviour may be overstated in a real-world context (Healey et al., 2010). Intuitively, this makes sense. Merely parroting back the last thing a fellow speaker has said is unlikely to be a marker of successful interaction and is instead likely to come across as irritating.

Rather than verbatim repetition, researchers have instead argued for a more nuanced explanation regarding both the presence and effects of mirroring. One of the most prominent perspectives on this topic, particularly from a psycholinguistic standpoint, concerns the work of James Pennebaker (Ireland et al., 2011; Niederhoffer & Pennebaker, 2002; Pennebaker, 1993; Pennebaker et al., 2015; Pennebaker & King, 1999; Pennebaker et al., 2003). Pennebaker notes that mirroring content words such as nouns can result from topicality effects. For instance, a conversation about sports will likely include both speakers making duel-references to sport-related concepts. Instead, he argues for monitoring the alignment of semantically-neutral function words such as particles (e.g., 'a', 'the', 'in', 'out', 'over', 'under') in conversation (Pennebaker et al., 2003). A similar approach was also used in Nenkova et al. (2008) and was found to predict outcomes in a group task. Due to these terms possessing little semantic content, they are less impacted by topicality effects. Therefore, mirroring these words is considered an instance of Language Style Matching (LSM) (Ireland et al., 2011). Furthermore, being very common in everyday language, function words can be explored in various settings. Over the last three decades, a substantial body of psycholinguistic research has explored the social benefits of mirroring these words. Taylor and Thomas (2008) found that increased LSM correlated with a peaceful resolution in hostage negotiation. In a study of police interrogation outcomes, B. H. Richardson et al. (2014) found evidence that a suspect who consistently mirrored the interrogator's use of function words was likelier to confess to a crime than a suspect who maintained their innocence. Moreover, a study of speed-dating interactions found that couples who matched displayed higher rates of LSM than those who did not.

However, there are limits to how converging with another speaker is considered valuable. At the centre of CAT is the notion that the strength and direction of accommodation mediate the relationship between speakers (H. Giles et al., 1991). In other words, it acts as a non-conscious social signal that communicates a speaker's intent. There are many naturally-occurring social situations where affiliation is neither the objective nor the desired outcome of a conversation (i.e., arguments). Consequently, speakers in these situations might become increasingly dissimilar in

their behaviour as a sign of disagreement (as in Culpeper et al., 2003) – a process known as *divergence*.

2.2.2 Prior computational approaches

Given the myriad ways mirroring has been measured from a computational perspective, it is worth outlining some of the desired criteria. Previous review papers have been written on this topic to standardise computational measures and enable comparisons between different approaches (see Doyle et al., 2016; Xu & Reitter, 2015). For brevity, this section only explores criteria pertinent to the research objectives covered by this thesis. Other desiderata are not explored but are highlighted where appropriate. This section aims not to identify a single measure that satisfies every requirement. Instead, this section highlights prominent and commonly-used measures' relative strengths and weaknesses.

Directionality

As a measure of linguistic influence, any verbal mirroring measure must indicate *who is mirroring who*. Mirroring is not synonymous with general similarity. For example, in a dyadic (i.e., two-speaker) interaction, speaker A can converge towards the verbal behaviour of speaker B without speaker B reciprocating. This distinction can be exemplified in the work of Bilous and Krauss (1988), who, in a study of mixed versus same-gender dyads, found that whilst males *converged* towards females on certain dialogue acts (i.e., frequency of laughter and back-channels), females *diverged* from males. In order to measure directionality, the temporal nature of the interaction must be considered. It should not be possible to base mirroring on speech turns that happen later in the conversation. For instance, mirroring of linguistic style is often captured by a measure introduced in Ireland et al. (2011). This measure (LSM) compares the weighted average of function words between two speakers irrespective of the order in which they occur in the conversation. Despite its wide application in various domains, this measure does not distinguish between a CAT inspired interpretation of mirroring and general linguistic similarity. By contrast, an amended version of this measure was introduced in Müller-Frommeyer et al. (2019) that does satisfy this criterion. The main difference between this approach and the previous one was that the latter compared pairs of speech turns between speaker A and the *previous* turn produced by speaker B. This makes it

possible to infer the cause and effect of mirroring rather than simply correlation.

Baselining

Even if the temporal nature of the conversation is factored in, the co-occurrence of words in adjacent speech turns does not necessarily indicate mirroring. Two speakers may use similar words or phrases for various reasons, particularly at a linguistic level. A degree of similarity is required for the conversation to progress (Pickering & Garrod, 2004) smoothly. If two speakers are engaged in a conversation about sports, the likelihood that they will each use sport-related concepts will be inflated. This does not indicate mirroring but reflects a topicality effect (Reitter & Moore, 2014).

Similarly, two speakers can coincidentally appear linguistically similar based on a concept known as 'homophily'. This is a general tendency for two speakers with similar knowledge and beliefs to behave similarly (Rogers & Bhowmik, 1970). Curiously, homophily is associated with many of the same social benefits as mirroring, although based on a different theoretical framework (Doyle et al., 2016).

To satisfy baselining criteria and to distinguish mirroring from homophily, the typical word usage (i.e., the baseline) of each speaker must be considered. One of the ways this can be achieved is through the measure first introduced by Danescu-Niculescu-Mizil et al. (2011). This approach measures the probability of word co-occurrence between consecutive and time-ordered speech turns (conveniently satisfying the directionality condition above). It subtracts from that value a particular speaker's average word use. The left-over value can then be interpreted as the change in behaviour (i.e., the degree of mirroring) in speaker A based on the prior behaviour of speaker B. Thus, effectively studying both the presence and impact of mirroring requires accounting for these baseline frequencies.

Feature Separability

Theories of communication accommodation have highlighted the multifaceted nature of mirroring (H. Giles et al., 1991). Speakers do not converge across all behaviour equally. Instead, they may converge on some behaviours whilst diverging on others (Bilous & Krauss, 1988; Danescu-Niculescu-Mizil & Lee, 2011). A criticism of several approaches is that they create a single mirroring score across all linguistic tokens

(i.e., overall words). For instance, the previously cited LSM is usually based on an averaged score over nine categories of function words[3]. Similarly, Local Linguistic Alignment (LLA) is a measure that creates a single aggregate score overall shared words or syntactic structures (Fusaroli et al., 2012). Whilst these approaches give a general indication of convergence (where speakers become more similar to one another), they often fail to acknowledge any divergent behaviours (where speakers become increasingly dissimilar).

Non-separability between features also overlooks the inherent properties of the words or features themselves. Linguistic structures follow a pattern known as *Zipf's law* (Reverdy & Vogel, 2017). This is where the frequency of a word or feature is inversely proportional to its rank. In other words, the most frequent word in a corpus is twice as common as the second most frequent (Saif et al., 2014). This can mean that commonplace words naturally tend to co-occur not through reasons associated with mirroring but because they are naturally high-frequency words.

Moreover, different behaviours are likely to have different rates of alignment. As noted by Doyle et al. (2016), we would not necessarily expect a word like 'you' to have exceptionally high levels of convergence as its mirror opposite in conversation is the word 'me'. Previous measures that have factored in feature separability include the approach described in Danescu-Niculescu-Mizil et al. (2011). This measure is equally applicable at both the level of individual words or similar linguistic structures and at an aggregated or group level (see Danescu-Niculescu-Mizil & Lee, 2011, for an illustration of the latter).

Other criteria

Briefly, several other factors have been cited as desired characteristics in a measure of verbal mirroring. Doyle et al. (2016) cites robustness to sparse data as one example. In their work, the authors focus on short online social media interactions that typically comprise a small number of messages exchanged between speakers. It is undoubtedly the case that many of the existing methods, particularly those based on weighted averages (see Fusaroli et al., 2012; Ireland et al., 2011; Müller-Frommeyer et al., 2019), may struggle in this setting. However, given that the domains explored in this thesis are based on more extended conversations, this aspect was not

[3]The most common application of linguistic style is based on the following grammatical categories: articles, negations, non-personal pronouns, quantifiers, auxiliary verbs, adverbs, personal pronouns

particularly relevant. Similarly, Xu and Reitter (2015) suggested that an optimal measure of mirroring should extract normally distributed features and that features at different linguistic levels should correlate (i.e., increase lexical convergence should increase at a rate similar to syntactic convergence). These criteria were derived from a cognitive science-based theory of mirroring, the exploration of which was beyond the scope of the current thesis (see Garrod & Pickering, 2004; Pickering & Garrod, 2004, for details of the Interactive Alignment Model).

2.3 *CoPyCAT*: A python package for measuring verbal mirroring in dyadic conversation

This section introduces `CoPyCAT`[4], a python package designed to measure verbal mirroring from dyadic transcripts. The name *CoPyCAT* is intended to reflect the idea of closely mirroring the behaviour of another speaker in conversation – a copy-cat being a well-known colloquialism for mimicking others. The capitalisation of 'CAT' is intended to highlight that the measures included are consistent with Communication Accommodation Theory (see H. Giles et al., 1991). The remainder of this chapter describes critical steps in the package's development and the rationale behind pre-processing and feature engineering decisions.

As per similar packages, most notably the ALIGN package introduced in Duran et al. (2019), CoPyCAT extracts text-based features of verbal mirroring from transcripts of dyadic conversations. These features can then be used for downstream tasks, such as measuring the impact of mirroring on a collective activity (Friedberg et al., 2012; Nenkova et al., 2008; Rahimi et al., 2017; Reitter & Moore, 2007, 2014).

There are, however, several aspects that distinguish CoPyCAT from related methods. First, as highlighted in the section above, many existing approaches quantify verbal mirroring as a single value by averaging behaviours over all speakers. Here, the measures included satisfy the directionality criterion – that is, it has been made possible to isolate mirroring performed by a single speaker. Second, CoPyCAT explores mirroring from multiple linguistic levels, including lexical, syntactic, stylistic, and semantic. Each of these is also further broken down into *macro features* – features that calculate a single mirroring value overall behaviours at that level, and *micro features*, which measure the amount of mirroring of a particular

[4]See https://github.com/cookie1986/CoPyCAT

linguistic object (i.e., an individual word). Micro-features satisfy the feature separability criteria, whereas macro-features are included to enable comparison with similar research.

2.3.1 Data types

CoPyCAT is designed to work with transcripts. As such, it can accept files in either plain text (.txt), comma separated values (.csv) or JavaScript Object Notation (.json) format.

Plain text files (.txt)

Plain text files should comprise the speaker label and the content of each speech turn. Speech turns are distinguished based on the presence of a newline. The author observed that transcripts designed for human use often use newline characters to split longer speech turns to aid readability. These should be removed to ensure that newlines coincide with a speaker change.

The format of individual speech turns should include the name of the speaker followed by a single colon symbol. The colon is then used to parse each line into a tuple of speaker, label and speech content. Figure 2.1 details an example of the accepted format.

```
MARGARET BRENNAN: Welcome back to Face the Nation. We now go to Senator Kirsten Gillibrand. She's a Democrat from New
York. And the newest contender in the 2020 Democratic presidential primary. She joins us this morning from Des Moines.
Senator welcome to Face the Nation.

SEN GILLIBRAND: Thank you Margaret.

MARGARET BRENNAN: You have said that you are a no on the president's proposal here to end the shutdown. Why isn't the
offer of three years of protection for DAC-DACA on TPS recipients enough of a start for you?
```

Figure 2.1: Example of transcript in plain text format.

Comma Separated Values (.csv)

CSV files should be passed with at least two columns labelled “speaker” and “content” respectively. Each row should indicate a complete speech turn. See Figure 2.2 for an illustration.

speaker	content
MARGARET BRENNAN	Welcome back to Face the Nation. We now go to Senator Kirsten Gillibrand. She's a Democrat from New York. And the newest contender in the 2020 Democratic presidential primary. She joins us this morning from Des Moines. Senator welcome to Face the Nation.
SEN GILLIBRAND	Thank you Margaret.
MARGARET BRENNAN	You have said that you are a no on the president's proposal here to end the shutdown. Why isn't the offer of three years of protection for DAC-DACA on TPS recipients enough of a start for you?

Figure 2.2: Example of transcript in CSV format.

JavaScript Object Notation (.json)

JSON files should contain "speaker" and "content" keys and a "speech_turn" key indicating the position of the turn within the chat. The resulting format should resemble the illustration in Figure 2.3.


```
{
  {
    "speech_turn":"1",
    "speaker":"MARGARET BRENNAN",
    "content":"Welcome back to Face the Nation. We now go to Senator Kirsten Gillibrand. She's a Democrat from New
              York. And the newest contender in the 2020 Democratic presidential primary. She joins us this morning
              from Des Moines. Senator welcome to Face the Nation."
  },
  {
    "speech_turn":"2",
    "speaker":"SEN GILLIBRAND",
    "content":"Thank you Margaret."
  },
  {
    "speech_turn":"3",
    "speaker":"MARGARET BRENNAN",
    "content":"You have said that you are a no on the president's proposal here to end the shutdown. Why isn't the
              offer of three years of protection for DAC-DACA on TPS recipients enough of a start for you?"
  }
}
```


Figure 2.3: Example of transcript in JSON format.

Once read into CoPyCAT using one of the three methods described above, the output is a $3xN$ data frame, where the three columns denote the speaker ID, the position of the speech turn in the transcript, and the verbal content. An illustration of the processed transcript is referenced in Figure 2.4.

	speaker	content	speech_turn
0	MARGARET BRENNAN	Welcome back to Face the Nation. We now go to...	0
1	SEN GILLIBRAND	Thank you Margaret. \n	1
2	MARGARET BRENNAN	You have said that you are a no on the presid...	2
3	SEN GILLIBRAND	Well, he's the one who got rid of DACA protec...	3
4	MARGARET BRENNAN	The president said this morning in a tweet th...	4
5	SEN GILLIBRAND	No I don't. I don't take him on his word on a...	5
6	MARGARET BRENNAN	Well Democrats would have to agree to anythin...	6
7	SEN GILLIBRAND	Well what I wish he would do is just open up ...	7
8	MARGARET BRENNAN	And that means money order a border barrier? \n	8
9	SEN GILLIBRAND	I don't think that's the most effective way t...	9

Figure 2.4: Example of the processed transcript.

2.3.2 Pre-processing

Standardising speaker labels

The next stage involves pre-processing the transcript into a standardised format. The first step involves standardising speaker labels, where all turns allocated to the first speaker in the transcript are given label 1, and all others are given label 2. Returning to the example in Figure 2.4, all speech turns generated by `MARGARET BRENNAN` are converted to 1, whilst turns attributed to `SEN GILLIBRAND` would be labelled with a 2. Given that CoPyCAT measures mirroring in a single direction, this step is essential for tracking who is mirroring who. As part of this, consecutive speech turns produced by the same speaker are concatenated. This ensures that the previous utterance in the transcript is always produced by the other speaker in the dyad.

Cleaning speech turns

Several common pre-processing steps are used to standardise the content of speech turns and remove unnecessary noise. First, any characters not appearing in the printable ASCII character set are removed[5]. Because of this, CoPyCAT is designed to work primarily with English language transcripts. In addition to this, annotations commonly generated during the transcription effort, such as *INAUDIBLE*, *CROSSTALK*, *LAUGHTER*, and *EXPLETIVE*, are identified and filtered out of the transcript using Regular Expressions. Although these behaviours have been used in prior research (Gervais et al., 2013; Mehu, 2011; Pamungkas et al., 2020), the author found that they had been inconsistently applied throughout the corpora studied in this thesis. There is also an option to include additional annotations by providing a user-specified list of exclusion terms. Hesitation markers such as 'erm', and 'um' are also removed, as are possessive nouns, which are standardised by removing the possessive *'s* marker. In addition, all text was converted to lowercase, numbers were converted to word format (i.e., 100 becomes 'one hundred'), and all punctuation was removed. Contracted words, such as "don't" and "can't", are expanded into long-form "can not" and "will not" using a dictionary provided in (Duran et al., 2019). Lastly, an automatic spell-checker based on Levenstein's Distance was considered (Norvig, 2016). It was, however, found to perform poorly on the data used in this thesis and so was omitted.

[5]The complete list of printable ASCII characters can be found at https://www.ascii-code.com/

Tokenization, Part of Speech Tagging, and Lemmatization

Measuring mirroring from transcripts requires standardising how language is represented within the corpus. All speech turns are subjected to the following steps: first, the verbal content of each speech turn is parsed into individual word units known as *tokens*. Next, each token is matched with its corresponding grammar tag in a process known as Part-of-Speech (POS) tagging. This step helps distinguish between homonyms – words spelt the same but with different meanings. Additionally, word inflexions were standardised using a step known as *lemmatisation*. This reverts a word to its canonical form – for example, the words `` `run', `ran', `running', `runner' `` are all inflexions of the lemma `` `run' ``. These steps were performed using `spaCy` (Honnibal & Montani, 2017). Alternative libraries such as `Natural Language Tool-Kit (NLTK)` (Loper & Bird, 2006) and `Stanza`[6] (Qi et al., 2020) were also considered, performing comparably. `spaCy` was ultimately chosen as it is simple to construct the described pre-processing steps in a pipeline object, which minimises the amount of Python code required. An illustration of each of the steps discussed thus far in this section is provided in Table 2.1.

Step	Processed Text
Original text	Usain Bolt is running in a 100 metre race.
Lower-casing	usain bolt is running in a 100 metre race.
Punctuation removal	usain bolt is running in a 100 metre race
Convert numerical	usain bolt is running in a one hundred metre race
Tokenization	[‘usain’, ‘bolt’, ‘is’, ‘running’, ‘in’, ‘a’, ‘one’, ‘hundred’, ‘metre’, ‘race’]
Lemmatize	[‘usain’, ‘bolt’, ‘is’, ‘run’, ‘in’, ‘a’, ‘one’, ‘hundred’, ‘metre’, ‘race’]
Part-of-Speech (POS) tagging	[(‘usain’, ‘NNP’), (‘bolt’, ‘NNP’), (‘is’, ‘VBZ’), (‘run’, ‘VB’), (‘in’, ‘IN’), (‘a’, ‘DT’), (‘one’, ‘CD’), (‘hundred’, ‘CD’), (‘metre’, ‘NN’), (‘race’, ’NN’)]

Table 2.1: Example of text after each pre-processing step using a dummy sentence.

[6]Formerly and more commonly known as `StanfordNLP`

2.3.3 Calculating mirroring values

Micro and Macro-level mirroring

This section describes how the pre-processed speech turns described above are used to identify instances of mirroring between two speakers. Mirroring is based on the degree of similarity between adjacent speech turns. CoPyCAT encompasses two broad levels, which have been termed micro-level mirroring and macro-level mirroring. A micro-level mirroring analysis involves calculating an individual mirroring score for every word or token uttered by a speaker. It accounts for the idea that speakers will not converge on all behaviours to the same degree (Thakerar et al., 1982). Thus, it enables a more granular understanding of the specific words or phrases that are more likely to be reflected by a speaker. Micro-level analyses have been used to detect verbal mirroring in both physical (Danescu-Niculescu-Mizil et al., 2011) and online settings (Doyle et al., 2016).

By contrast, a macro-level mirroring analysis generates a single indicator of mirroring by aggregating across all words or phrases. A commonly deployed approach averages the mirroring values across all behaviours of interest. For instance, if a micro-level analysis calculates mirroring for nouns, verbs, and adjectives individually, a macro-level analysis could be calculated by averaging the three values together. A macro-level analysis of mirroring has been used to predict attraction in romantic relationships (Ireland et al., 2011) and success in collaborative tasks (Nenkova et al., 2008; Reitter & Moore, 2014).

As discussed in this chapter, mirroring behaviour can be represented at various linguistic levels. With this in mind, CoPyCAT includes features based on lexical, stylistic, syntactic and semantic mirroring. Each linguistic level includes a macro and micro-level method, except for semantic mirroring, which only exists at the macro level. To be precise, these are not new measures within the alignment or accommodation research space and have been examined and validated in several existing studies over the last two decades (Carrick et al., 2016; Danescu-Niculescu-Mizil et al., 2011; Fusaroli et al., 2012; Müller-Frommeyer et al., 2019; Murray & Oertel, 2018; Niederhoffer & Pennebaker, 2002; Taylor & Thomas, 2008). However, the accumulation of these measures within a single, easy-to-use package is novel. In the section of this chapter, each available measure is presented, with an overview in Table 2.2.

Linguistic Representation	Level	Measure	Description
Lexical	Macro	LLA	Probability of word co-occurrence in adjacent turns averaged over all speech turns
	Micro	SCP (n-grams)	Individual probability of each word being repeated over all uses of the word
Stylistic	Macro	rLSM	Negated absolute difference between speakers' use of function words over all turns
	Micro	SCP (LIWC)	Representing each word as a LIWC-token; the individual probability of each token being repeated over all uses of the token
Syntactic	Macro	Branching factor difference	Absolute difference in branching factor between adjacent turns, averaged over all speech turns
	Micro	SCP (Dep. relation tags)	Representing each word by its dependency parse tag; the individual probability of each token being repeated over all uses of the token
Semantic	Macro	GloVe cosine similarity	Representing each word as a word embedding, the average cosine similarity between adjacent speech turns

Table 2.2: Overview of the different ways verbal mirroring can be expressed using CoPyCAT.

Lexical mirroring

Lexical mirroring is based on the repetition of words between adjacent turns. The words within each speech turn are grouped into n-grams ranging from unigrams (single words) to trigrams (groups of three consecutive words). This approach is commonly referred to as a *bag-of-words* model as it disregards the order of words within the speech turn.

Micro-level. Lexical mirroring at a micro-level is based on a measure known as Subtractive Conditional Probability (SCP) (Danescu-Niculescu-Mizil et al., 2011). SCP is a probabilistic framework that measures the probability of a word appearing in consecutive speech turns and subtracts the probability over all possible speech

turns (see Equation 2.1).

$$LM^{F}_{(i,j)} \triangleq P(T^{f}_{i}|T^{f}_{j}, T_i \hookrightarrow T_j) - P(T^{f}_{i}|T_i \hookrightarrow T_j) \tag{2.1}$$

Here, the left term, $LM^{f}_{(i,j)}$, is the degree of lexical mirroring (LM) of feature f expressed by speaker i in relation to speaker j. The first term on the right is the conditional probability of speaker i uttering feature f given its usage in the previous speech turn performed by speaker j. The second term on the right is the total probability of speaker i uttering f over all replies to speaker j. A score between -1 and 1 is computed for each f, with positive values indicating convergence (i.e., speaker i increases usage of f after speaker j's usage) and negative values indicate divergence (i.e., speaker i decreases usage of f after speaker j's usage). Theoretically, a value close to 0.0 indicates no evidence of lexical mirroring – also referred to by H. Giles et al. (1991) as *language maintenance*.

SCP was chosen over other methods due to its ability to separate mirroring between different words or phrases. As mentioned, in naturally-occurring conversation, mirroring is not uniformly distributed over all verbal behaviours (Church, 2000). By subtracting the total probability of a feature appearing in a random speech turn, SCP reduces the effect of homophily – the tendency for speakers talking about the same topic to use similar words (Doyle et al., 2016). It also minimises the likelihood of detecting mirroring due to the high frequency of certain words. It is widely regarded that the frequency of word use in conversation follows a *Zipfian* distribution[7] (Reverdy & Vogel, 2017). Consequently, the highest frequency words are likely to co-occur not due to mirroring effects but as a result of being the most common words. By subtracting a speaker's average word use, the resulting value is the degree to which the speaker increased or decreased usage in response to an interlocutor (i.e., a priming effect) (Danescu-Niculescu-Mizil et al., 2011).

Macro-level. At the macro-level, a measure known as LLA (Fusaroli et al., 2012) is used. LLA measures the probability of word repetition between a target and prime speech turn over all words in a speech turn. CoPyCAT uses the implementation formalised in both Doyle et al. (2016) and Y. Wang et al. (2015):

[7] In a Zipfian distribution, the frequency of an object is inversely proportionate to its rank. In linguistics, this means that the most common word in a corpus will be twice as likely to appear as the second most common word, which will be twice as likely to appear as the third most common word.

$$LLA = \frac{\sum_{w_i \in T_b} \delta(w_i \in T_a)}{Len(T_a)Len(T_b)} \tag{2.2a}$$

$$\delta(w_i) = \begin{cases} 1, & \text{if } w_i \in T_a \\ 0, & \text{otherwise.} \end{cases} \tag{2.2b}$$

Here, LLA refers to the local linguistic alignment between two speech turns T_a and T_b. For each word w_i that appears in T_b, a value of 1 is given if that word also appears in T_a. Values are then summed over all words in a speech turn and normalised by the total length (number of words) in the combined speech turns $Len(T_a)Len(T_b)$. The resulting value is the proportion of words shared between the two speech turns. A single value is obtained by taking the average LLA overall speech turns.

Unlike SCP, LLA is solely a measure of convergence (i.e., the minimum possible value obtained is 0, which indicates no similarity between speech turns). For this reason, it was deemed a more appropriate macro-level measure compared to averaging SCP values, which due to the inclusion of a divergence measure, would result in convergence and divergence effects cancelling each other out.

Stylistic mirroring

Stylistic mirroring is a subset of lexical mirroring, focusing exclusively on non-content words. Non-content words (function words) are conceptually similar to stop-words, often removed as standard in many text-mining applications (Feldman & Sanger, 2006).

Micro-level. At the micro-level, stylistic mirroring adopts the same approach as lexical mirroring, except that words are first categorised according to LIWC dictionaries (Pennebaker et al., 2015). As a reminder, LIWC focuses on how people use words rather than the content of the words themselves. As such, LIWC dictionaries are based on psycholinguistic categories. Examples include affective processes (positive and negative emotions), social processes (references to other people), and cognitive and perceptual processes (thinking, seeing, feeling). Using LIWC, each

word in a transcript is converted into its corresponding LIWC entity before Equation 2.1. As there are fewer LIWC categories than words, this provides a coarser examination of mirroring but one that removes the effect of topicality entirely.

Macro-level. A measure known as LSM is used to analyse stylistic mirroring at a macro-level. LSM reflects the extent that two speakers converge on semantically-neutral words such as articles, prepositions, and auxiliary verbs – otherwise known as *function words* (Pennebaker et al., 2003). The conventional implementation of LSM is a speaker-independent measure. That is, it is ambivalent to the turn-taking nature of a conversation, comparing the entirety of the conversation without regard for who is performing the mirroring. CoPyCAT addresses this by adopting several steps from the implementation described in Müller-Frommeyer et al. (2019) – referred to as reciprocal LSM or rLSM. This is outlined in Equation 2.3:

$$rLSM_{B\to A}(f) = 1 - \frac{|f_B^{T=i} - f_A^{T=i-1}|}{f_B^{T=i} + f_A^{T=i-1} + .0001} \tag{2.3}$$

Here, the left-term $rLSM_{B\to A}(f)$ refers to the degree of style matching between a single speech turn performed by speaker B towards speaker A on linguistic dimension f. On the right, the numerator is the absolute percentage difference in f between a speech turn T performed by speaker B at position i in the transcript and the prior speech turn performed by speaker A at $i-1$. The denominator is the sum of the two percentages, adding .0001 to prevent division by zero (if a particular speech turn does not contain f). The result of this fraction is then subtracted from 1 so that higher values indicate greater LSM. The maximum score to be obtained is 1, indicating speakers perfectly matched on language style. The minimum potential score is 0, suggesting no mirroring. Calculating a single measure of stylistic mirroring is achieved by averaging $rLSM$ scores overall turns performed by a specific speaker.

In order to aid understanding, we can consider a simple conversation between two people, Alice (A) and Bob (B). We will use the word "the" as our linguistic dimension (f).

Conversation transcript:

Alice: "I went to the park and played with the kids."
Bob: "Oh, I also went to the park, but I didn't see the kids."
Alice: "Well, the park is quite large, so it's easy to miss people."

Now, let us calculate the rLSM value for Bob's response towards Alice (Turn 2) using the formula provided in Equation 2.3. First, we need to calculate the frequency of "the" in each turn:

Alice (Turn 1): 2 occurrences of "the" / 9 total words = 0.2222
Bob (Turn 2): 2 occurrences of "the" / 12 total words = 0.1667

Now, we can plug these values into Equation 2.3:

$$rLSM_{B \to A}(\text{the}) = 1 - \frac{|(0.1667 - 0.2222)|}{(0.1667 + 0.2222 + .0001)} \tag{2.4}$$

$$= 1 - \frac{|(-0.0555)|}{(0.3889)} \tag{2.5}$$

$$= 1 - 0.1427 \tag{2.6}$$

$$= 0.8573 \tag{2.7}$$

The rLSM value for Bob's response towards Alice (Turn 2) is 0.8573. This value is close to 1, suggesting a relatively high degree of stylistic mirroring between Bob and Alice in this specific turn with respect to the use of the word "the". To get a more comprehensive measure of stylistic mirroring for the entire conversation, we need to calculate rLSM values for all turns and average them.

Syntactic mirroring

At the syntactic level, speech is represented as syntax – the grammatical structure of speech turns.

Micro-level. At a micro-level, syntactic mirroring uses the same approach described at the lexical and stylistic levels, with Equation 2.1 used to compare adjacent speech turns. Each word is represented by its corresponding dependency relation tag. Dependency parsing is a process that seeks to identify the relationships (i.e., the dependencies) between words in a sentence. Collectively, it represents the syntactic structure of a sentence. Take the following sentence ("I prefer morning flights in the winter.") as an example:

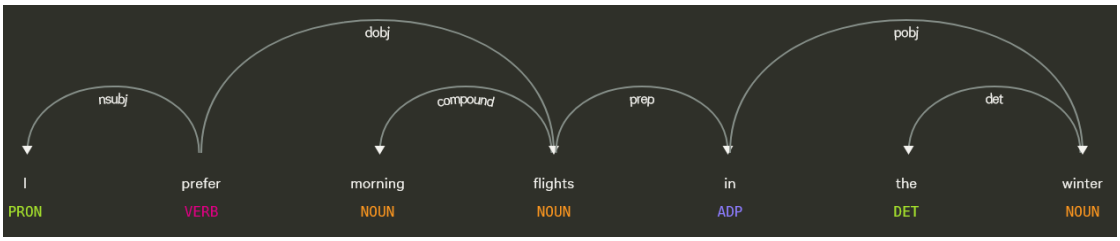

Figure 2.5: Dependency parse example.

The relationships between words in Figure 2.5 are indicated by arrows. Relationships exist between pairs of words, where one word forms the parent node and the other the child node. Here, we see that the word 'prefer' modifies the meaning of 'flights'. The relationship is categorised as `dobj`, which stands for *direct object*. A direct object is a word or phrase (in this case, the noun 'flights') that receives the action of a verb (in this case, 'prefer').

In order to utilise this information at a micro-mirroring level, each speech turn is converted into a bag-of-relations based on its constituent dependency parse tags. A similar feature set was used to predict task success in a group performance task (Murray & Oertel, 2018). Equation 2.1 is then calculated in the same manner as previously described.

To illustrate the differences in micro-mirroring features between lexical, stylistic, and syntactic levels, Figure 2.6 describes an example sentence and highlights how micro-mirroring is represented at each level.

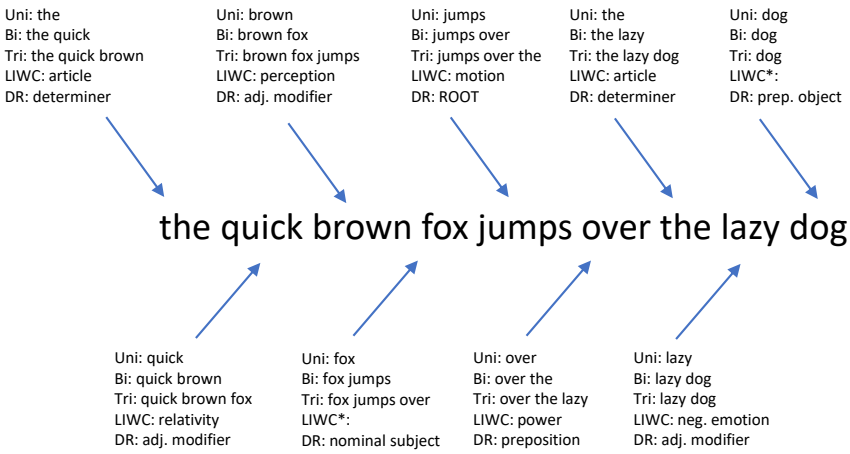

Figure 2.6: Example of text when represented as (1) n-grams (unigram to trigram), (2) LIWC categories, (3) dependency nodes. Note: LIWC (stylistic mirroring) does not include nouns (illustrated by an *).

Macro-level. Macro-level syntactic mirroring is based on differences in branching factor between adjacent speech turns. In tree-like structures, the branching factor is the number of child nodes at each parent node. Returning to the example in Figure 2.5, we observe that the node 'flights' has two children – 'morning' and 'in' – illustrated by the two arrows leading outwards. By contrast, the node for 'winter' only has one child node – 'the'. Therefore, taking the average branching factor within a speech turn serves as a proxy for syntactic depth, where longer, more complex sentences will have a higher average branching factor. The average absolute distance between adjacent speech turns is calculated to measure the degree of mirroring at this level, where larger values indicate greater dissimilarity between speakers. This number is then negated to maintain consistency with other measures in this chapter. After negation, values closer to zero can be interpreted as indicative of greater syntactic mirroring.

Semantic mirroring

Lastly, semantic mirroring is measured solely at the macro level based on the average cosine similarity between adjacent speech turns. Cosine similarity is a similarity metric used to compare two vectors by calculating the cosine of the angle between them. It ranges from -1 (entirely dissimilar) to 1 (completely similar), with a value of 0 indicating no correlation. Each word in the transcripts is converted into a fifty-dimensional Global Vectors for word representation (GLoVe) vector (see Pennington et al., 2014) and averaged over all words to create a single, turn-level embedding. GLoVe is a word embedding technique that transforms words into continuous, fixed-size vectors. It captures semantic relationships between words by leveraging co-occurrence information from large text corpora, enabling efficient representation and analysis of words in natural language processing tasks. Once each speech turn has been converted into a single GLoVe vector, adjacent turns are compared via cosine similarity, where points occupying a similar position within the vector space are more semantically similar versus those with a greater distance between them. An advantage of cosine similarity is that it is not affected by differences in turn length, as it measures the cosine of the angle between two data points. Cosine similarity generates a value between 0 and 1 for each pair of speech turns, broadly reflecting the percentage similarity between two input texts. Averaging these values over all eligible turns thus generates a single value representing semantic mirroring across the entire transcript.

2.3.4 Feature fusion and output

For each transcript, CoPyCAT outputs a single feature vector comprising mirroring at each level. A micro-level analysis, particularly at the lexical level, can result in a high-dimensional vector, as each word contained within the transcript is considered an independent feature. To address high-dimensionality, minimising dimensionality through feature selection techniques, such as recursive feature elimination or principal component analysis, is advisable prior to any downstream task (Shah & Patel, 2016). These techniques can help identify the most relevant features and discard redundant or irrelevant ones, reducing the risk of overfitting. For instance, a lexical feature vector with several thousand dimensions might be reduced to a more manageable number while preserving the text's essential information.

Extracting mirroring behaviours at all levels can easily result in a feature vector with several thousand dimensions, which will likely result in over-fitting unless feature reduction techniques are utilised (Vijayan et al., 2017). In addition to feature selection, dimensionality reduction can also be achieved through techniques like linear discriminant analysis (Kowsari et al., 2019) or t-distributed stochastic neighbour embedding (t-SNE) (Van der Maaten & Hinton, 2008), which aim to project the high-dimensional data into a lower-dimensional space while preserving the original structure and relationships among data points.

Likewise, the features are not scaled and, as a result, may require transformation prior to being applied to a downstream task. As the micro-level analyses all use the same algorithm across linguistic forms (SCP, see Danescu-Niculescu-Mizil et al., 2011), features all reside on a -1 to 1 scale. At a macro level, features exist on a 0 to 1 scale. Combining macro and micro-level features without transformation does not present a problem for algorithms such as decision trees or random forests (Breiman, 2001). However, it should be considered for kernel methods such as support vector machines.

For example, when using support vector machines or other distance-based classifiers, it is essential to ensure that all features are on the same scale to prevent features with larger ranges from dominating the classification process (Cortes & Vapnik, 1995). This can be achieved by applying feature scaling techniques such as min-max normalisation or standardisation (z-score), which transform the features into a comparable range or distribution (Hastie et al., 2009). By performing these pre-processing steps, the final feature vector will be more suitable for various machine learning algorithms, improving the classifier's performance and aiding

interpretability.

2.4 Summary

The last decade has seen various approaches to measuring verbal mirroring in dyadic conversations. This diversity of methods, however, has resulted in a scattered research landscape. Consequently, it has been suggested that mirroring, as claimed by experimental research, may have been overstated (Healey et al., 2010). It is, therefore, essential to create flexible tools that can measure various mirroring effects in naturalistic settings. This chapter has sought to provide an initial answer to this problem.

`CoPyCAT` has been developed as a flexible tool for measuring verbal mirroring in dyadic conversations. It allows the user to select from a series of different linguistic representations at both a macro and micro level. Additionally, it allows the user to infer the mirroring direction and is an extension of typical measures that do not distinguish between speaker roles. It also includes different preprocessing strategies based on standard text mining and NLP techniques.

The flexibility of `CoPyCAT` is intended to encourage usage across various dyadic interactions. Future development of the tool could sensibly extend beyond two-person conversations to multi-speaker and group settings. While existing measures have not been designed with this purpose in mind, they have been explored in several collaborative environments (Litman et al., 2016; Murray & Oertel, 2018; Rahimi et al., 2017).

The purpose of the present chapter was to design and implement a computational approach to verbal mirroring as an alternative to time-intensive manual labelling. What has been presented is a walk-through of `CoPyCAT`, a python package designed to extract aspects of verbal mirroring consistent with theories of communication accommodation (H. Giles et al., 1991). Exploring the surrounding literature on mirroring informed the decision to represent mirroring at various linguistic levels. Exploring the relevant literature on existing computational methods enabled the author to identify those most suitable to the current thesis' aims. In the next chapter, `CoPyCAT` is used as a feature engineering tool to predict the effects of verbal mirroring in the context of a political interview.

Chapter 3

Predicting the Conversational Quality of a Political Interview: A Social Signal Processing Approach

3.1 Introduction

The previous chapter introduced `CoPyCAT`, a Python tool to extract psychological behaviours from social interactions automatically. Building on this framework, the present work reports whether an automated feature set can be used to predict a social outcome. Both the current chapter, in addition to Chapter 4, focus on interviewing, and in particular, how the behaviour of the interviewer can influence the interviewee and, in turn, the outcome of the interview. Analysing interviews to deduce how an interviewer may steer a conversation in a desired direction is not new. Modern dedicated training initiatives intend to teach interviewers effective conversational strategies for this purpose (Alison et al., 2013). However, existing research has focused almost exclusively on qualitative coding performed manually by experts. This limitation hinders analysis in scale and behaviour complexity. Furthermore, the high resources required to offer continuous feedback restrict interviewer skill development in a real-world setting. By contrast, predicting the outcome of interviews using interviewer features derived by automatic means would help overcome some of the existing limitations. Consequently, the aims of the present chapter are as

follows:

1. Describe related research on effective interviewing, highlighting the potential application of an automated approach in a real-world setting.
2. Validate the use of an automatically derived outcome variable as an alternative to subjective performance ratings.
3. Present the findings of a supervised machine learning task that automatically predicts the outcome of an interview from the behaviour of the interviewer.
4. Discuss the potential limitations of the adopted approach, and suggest recommendations for future research in this area.

3.2 Related work

3.2.1 Effective interviewing

Interviewing serves as a vital form of information gathering in many domains. For example, interviews are routinely used to guide the hiring of new staff (Naim et al., 2018). An interview is also commonly used to gather information for law enforcement purposes, either from those suspected of a crime (Alison et al., 2013) or witnesses (Brönnimann et al., 2013). Equally, televised interviews with prominent political figures form the basis of public opinion, helping to shape the outcome of elections and public policy (Baum, 2005).

A substantial body of work has been generated by social psychologists interested in how an interviewer's behaviour can help shape the outcome of the interview. Skilled interviewers use various interpersonal techniques to orchestrate the flow of conversation (Gubrium & Holstein, 2003). Researchers have shown that behaviours such as expressing empathy (Papadopoulou et al., 1996), active listening (Christiansen et al., 2018; Sharpley et al., 2005), and building rapport (Collins & Carthy, 2018), are associated with positive interview outcomes. In a study of police interviews with suspected terrorists, for example, Alison et al. (2013) found that interviewers who exhibited more prosocial behaviours, such as rapport-building, acquired more case-relevant information from suspects. Equally, they also found that anti-social or maladaptive behaviour had the inverse effect, reducing compliance with

law enforcement. In another study, Wolfman et al. (2016) examined the impact of different questioning strategies (i.e., open questions, closed questions, option-posing questions) when interviewing children, finding differences in the information gathered based on the type of question asked. These findings are often used as guidelines and training material in a real-world setting.

Many existing frameworks rely on manual coding effort and subjective assessment by experts. As such, they do not typically scale well to larger datasets (Naim et al., 2018). A reliance on manual processes can also pose a significant barrier when attempting to apply research findings in a real-world setting. Many industries deploy dedicated training programs designed to teach interviewers new techniques. However, research indicates that skill usage declines after an initial training period (Lamb et al., 2002). It is thought that while regular feedback would help address this decline (Forsberg et al., 2010), the costs associated with continuous skill evaluation often prohibit ongoing evaluation. By contrast, a computational approach based on automatically extracted data could be used to monitor and provide feedback to interviewers as they acquire new skills continuously.

As discussed in the previous chapter, an advantage of an SSP approach is the ability to extract low-level behaviours from a transcript or audio recording automatically. Consequently, it would appear well-suited to a task that automatically predicts optimal interview outcomes. Increasingly, SSP systems are being developed to measure the impact of a wide range of social behaviours in naturally occurring social interactions, serving as a scalable alternative to previous manual methods. For example, both L. Chen et al. (2016) and Naim et al. (2018) found that a combination of verbal and physical features could be used to assess the suitability of candidates for a job automatically. In another study, differences in linguistic behaviour were observed in a police interview based on whether the suspect confessed to a crime (B. H. Richardson et al., 2019).

One of the advantages of job interviews and police interrogations is that they both include clearly defined indicators of a successful interaction – a job-seeker is either offered a job or not, and the suspect either confesses to the crime or does not. These outcomes can be hard-coded and binarised in a manner that perfectly encapsulates the purpose of the interaction. Therefore, it is relatively straightforward to exploit these features in a simple classification task, using an SSP feature set as predictors. The same cannot be said of other interactions, including a political interview – the domain explored here. A political interview aims to inform and entertain an overhearing third party (i.e., the watching public) (Heritage, 1985). For their part,

interviewers are tasked with holding political figures to account (Hutchby, 2011) and extracting meaningful information (Garcia, 2018) whilst managing the flow of the interaction and maintaining a neutral political stance (Vraga et al., 2012). These factors can be challenging to conceptualise algorithmically and represent a barrier to a fully automated approach.

3.2.2 Mirroring as an effective interviewing technique

The previous chapter focused on the effects of behavioural mirroring in social settings. It was demonstrated that mirroring an interlocutor in conversation can lead to a positive social response. Over the last decade, studies in fields such as computational linguistics have explored this link by measuring the effect of mirroring on producing a more successful interaction. These studies have routinely focused on collaborative dialogue, using mirroring as a proxy for interlocutors sharing a common understanding. For instance, Nenkova et al. (2008) used a list of high-frequency stop-words to predict the naturalness of dialogue in a corpus of telephone interactions. Studies have also sought to predict the success of a conversation from the amount of mirroring between speakers. In task-oriented interactions, mirroring keywords and phrases pertinent to the topic can improve common understanding between speakers (Friedberg et al., 2012). However, effects do not appear to be localised solely at the lexical level, with studies also indicating that alignment of syntactic structures can lead to improved outcomes in a collaborative setting (Reitter & Moore, 2007, 2014). Moreover, a similar pattern of behaviour has also been observed at an acoustic level, with features derived from vocal intensity and pitch predicting the outcome of a similar game (Rahimi et al., 2017).

Whilst little attention appears to have been paid to the potential benefits of interviewer mirroring in *political* interviews, there is evidence indicating that displaying similarities to the interviewee can foster a more productive dialogue. A study of counselling sessions revealed that therapists who mirrored the verbal behaviour of their patients were rated as more empathic (Lord et al., 2015). In counselling psychology, empathy is considered a conscious attempt to understand another person through active listening (W. R. Miller & Rollnick, 2012). It has also been observed in other forms of interviewing, such as talk-show interviews, where increased mirroring was interpreted as an attempt by the interviewer to develop a shared understanding with their subject (Gregory & Webster, 1996). Moreover, intentional speech mirroring has been demonstrated to increase compliance and evidence yield during a law

enforcement interrogation (Alison et al., 2013).

Given the conceptual similarities between a political interview and the domains discussed above, it is reasonable to suggest that the interviewer mirroring the interviewee might improve the quality of the interaction. A political interview can be a high-pressure and contentious conversation, and non-compliance by the interviewee is a potential outcome if the interview is not managed appropriately (Pluss, 2010). However, as professional journalists, interviewers are typically expected to remain neutral to prevent stance-taking accusations (Hutchby, 2011). Mirroring the interviewee's behaviour may be positively interpreted as an attempt to see a particular issue from the interviewee's perspective, which may improve the quality of the responses. To the best of the author's knowledge, questions such as this remain wholly unexplored via a large-scale automated analysis. It, therefore, serves as an ideal first case study in this thesis.

3.3 Corpus of political interviews

This chapter introduces a corpus of N=691 dyadic (two speakers) political interviews derived from transcripts of popular US cable news networks (see Table 3.1 for a breakdown of interviews per network). Secondary transcripts were sought from online repositories maintained by the host network. The motivation behind using secondary transcripts was to generate a corpus large enough to exploit the capabilities of automation. As secondary transcripts were both accessible and plentiful from online sources, it was regarded as a more appropriate strategy compared to generating entirely new transcripts from video footage. This decision, however, resulted in a highly imbalanced corpus, as CNN, in particular, makes their transcriptions more easily accessible compared to other networks. Indeed, CNN represented almost half (46%) of the entire corpus. The corpus is also imbalanced with respect to the political orientation of the host network, with all but Fox News leaning to the political left[1]. This imbalance, however, reflects a broader societal trend within the US that sees media networks typically occupying a liberal viewpoint (Croteau, 1998; Groseclose & Milyo, 2005).

Effort was taken to ensure a representative interviewee sample by featuring interviews with an approximately equal number of Democratic (50.01%) and Republican

[1]Knowledge of the political orientation of a network was based on a consensus agreement of two fact-checking websites: https://mediabiasfactcheck.com and www.allsides.com/media-bias

Network	Political Orientation	Num. of Interviews	% of Corpus	% Non-Partisan
CNN	Left	315	45.59	41.27
Fox	Right	141	20.41	26.24
CBS	Left-Centre	105	15.2	67.62
NBC	Left	64	9.26	46.88
MSNBC	Left	51	7.34	1.96
ABC News	Left-Centre	15	2.17	60.0

Table 3.1: Demographic information of the news networks represented in the corpus.

(49.9%) politicians. To control for potential bias effects within the corpus, the author decided to include an approximately balanced number of partisan (60%) and non-partisan (40%) interviews - that is, interviews where the speakers share or conflict on their base political beliefs[2]. Bias towards a particular political ideology may lead an interviewer to react favourably towards an interviewee (Waldman & James, 1998), which may, in turn, impact their behaviour. Interestingly, networks with the greatest proportion of non-partisan interviews within the corpus are those positioned towards the political Centre (ABC News - 60% and CBS - 70%) - suggesting that these networks encourage a diversity of viewpoints. Interviews within the corpus were originally conducted and broadcast between 2013-2020, covering US presidential campaigns for the 2016 and 2020 elections. The corpus comprised 261 participants, with 55 interviewers and 206 interviewees. Timestamps were not included in the original transcripts, so timings are based on the word length. The length of interviews varied between 549 and 9102 words ($M = 1883.3$, $SD = 1113.35$). In total, the corpus comprised just under 1.3 million words and $28,022$ speech turns ($M = 40.97$, $SD = 47.46$). To the author's knowledge, the corpus represents the largest known collection of political interviews used for academic purposes within the social sciences. By contrast, similarly focused work using the PoliModal corpus comprises 56 interviews (Trotta et al., 2019) - $< 1\%$ of the corpus generated here.

3.3.1 Transcript pre-processing

Spot-checking was performed to ensure that the transcripts faithfully reflected the interview content. This step was achieved by searching for video recordings of the interview online and comparing the footage against the transcript. Approximately 20% of the corpus was found online, with transcriptions of satisfactory quality. Seven

[2]A politician's membership with either the Republican (right) or Democratic (left) party was taken as a reasonable approximation of political orientation.

interviews were removed due to suggestions that the transcripts had been heavily edited, resulting in a final corpus of n=684 transcripts.

Transcripts were then inspected and cleaned using CoPyCAT to ensure a suitable level of inter-network consistency. All transcripts used an orthographic method of transcription. Orthographic transcription means that transcribers used standard spelling and did not include false starts or filler utterances such as 'er' or 'umm' (W. Gibson & Brown, 2009). A proportion of the corpus included symbols used to indicate hesitation or pauses. For example, an interrupted speech turn was often appended with a '–' sequence, and attempts to re-establish the conversational floor were often prepended with the same sequence. Similarly, a vocal pause was often marked with a ' - '. Whilst a possible source of mirroring in its own right (Suzuki & Katagiri, 2004), there was insufficient coverage of these non-linguistic behaviours within the transcripts to warrant their inclusion. As such, they were removed using Regular Expressions.

3.4 Designing an automatic measure of a successful political interview

The first experiment in this thesis sought to validate a computational measure of success in a political interview via comparison with a small sample of human annotators.

3.4.1 Dataset

This study utilised N=10 interviews sampled from the larger corpus described in Section 3.3. Interviews were selected semi-randomly from the following criteria: First, the corpus was filtered to include only those interviews where full video footage could be obtained online from YouTube (approximately 20% of the whole corpus). From here, interviews were grouped into categories based on the host network (no eligible videos were identified for ABC or MSNBC). As Fox and CNN had the most significant representation in the corpus, three interviews were randomly selected from each network, and two from NBC and CBS. Six of the ten interviews sampled were non-partisan. Interviews were then spot-checked to ensure a range of speaker demographics had been captured and that interviews varied in length. A description of this dataset is illustrated in Table 3.2.

Network	Interviewer	Interviewee	Duration	Partisan
Fox (r)	Chris Wallace	Stephen Miller (r)	13 : 25	P
CBS (l)	Gayle King	Ivanka Trump (r)	7 : 03	NP
NBC (l)	Chuck Todd	Beto O'Rourke (l)	7 : 53	P
CNN (l)	Jake Tapper	Rudy Giuliani (r)	15 : 21	NP
CBS (l)	John Dickerson	Marco Rubio (r)	9 : 17	NP
CNN (l)	Wolf Blitzer	Eric Swalwell (l)	9 : 08	P
Fox (r)	Tucker Carlson	Tulsi Gabbard (l)	5 : 56	NP
NBC (l)	Chuck Todd	Hakeem Jeffries (l)	8 : 11	P
Fox (r)	Chris Wallace	Val Demings (l)	10 : 41	NP
CNN (l)	Anderson Cooper	Kellyann Conway (r)	25 : 53	NP

Table 3.2: Demographic information of the ten interviews used. (r) and (l) refers to the political orientation of the speakers, *right*-leaning and *left*-leaning respectively. P indicates both speakers occupy the same political orientation, NP indicates non-partisanship.

3.4.2 Annotators

Eight human annotators volunteered to take part in this study. Annotators were known personally to the author of this thesis and were not paid or reimbursed for their efforts. All annotators were adults (18+) and spoke English as a first or second language. There was an even gender split, with four males and four females, with the ages of annotators ranging between 26 and 60. Annotators were naive to the purpose of the study and were only debriefed as to their role after participation.

3.4.3 Outcome measures

Determining the quality of politicians' responses during an interview is challenging. Unlike a job offer following an employment interview (Naim et al., 2018), or confession during a police interrogation (B. H. Richardson et al., 2014), political interviews are not easy to distinguish into successful and unsuccessful outcomes. Because of this, a quantitative measure was devised based on desired qualities of the interviewee's speech.

Literature in this area suggests that the completeness and truthfulness of a response (Rendle-Short, 2007) and clear articulation (Heritage, 1985) are desired linguistic features of political discourse. Politicians, however, are infamous for equivocation and evasiveness during interviews, with research suggesting fewer than half of the questions asked are answered satisfactorily (Waddle & Bull, 2020). Uncooperative politicians have also been shown to make superfluous comments (Pluss, 2010)

or rely on repetition of key phrases as a diversionary tactic (Young, 2008).

Based on the surrounding literature, a successful interview is expected to be characterised by a politician answering questions fully, directly, and clearly. Accordingly, this study introduces four simple linguistic behaviours intended to reflect this. These are broadly influenced by Gricean conversational maxims of quality, quantity, relation and manner (Grice, 1989). Each of these will now be described in turn.

Specificity

The first outcome variable defined was termed *specificity*. Specificity is intended to represent the antithesis of ambiguity and obfuscation of political speech. Its inclusion here has been influenced by best practices in police interrogations, where references to specific entities such as people, places, and motives contribute to evidence-gathering against a suspect (Alison et al., 2013; Collins & Carthy, 2018).

An automated measure of specificity was created by counting the number of named entities occurring within an interviewee's speech. Named entities are explicit references to real-world objects such as people, organisations, places, times, and dates, as illustrated in the example text below:

> *"When **Barack Obama** was **President of the United States** he met leaders from **around the world**, including **Queen Elizabeth**. **The president**, along with **his wife**, **Michelle Obama**, visited **Buckingham Palace** **three times** during his **eight years** in office, and it's safe to say that **everyone** got on perfectly."*

Here, references to people (magenta) and places (blue) are highlighted. Words in bold font are indirect references (i.e., noun phrases) that refer to a particular object but would not be classified as a named entity. Detection of named entities in interviewee speech was performed automatically using Named Entity Recognition (NER) via `spaCy` (Honnibal & Montani, 2017). Counts were based on unique named entities (repeat utterances of the same entity were ignored) and did not include the interviewer's name. To account for differences in interview length, a normalised measure was calculated by dividing the number of named entities by the number of noun phrases uttered by the interviewee.

Clarity

The second outcome variable was named *clarity*. Like specificity, it is intended to reflect an inverse of ambiguity and vagueness, but unlike specificity, it is less impacted by nouns. Clarity is based on a psycholinguistic norm known as word concreteness, which refers to the overall accessibility of language (Duran et al., 2010). Personality research has indicated that concreteness is positively associated with high degrees of extroversion (Walker et al., 2007). To measure clarity, each word uttered by the interviewee was compared to a word concreteness dictionary (see Wilson, 1988). The average concreteness score was then recorded as a measure of clarity.

Diversity

The third outcome variable was named *diversity* and is intended to capture the degree of openness expressed by the interviewee. An unwillingness to engage in conversation can be demonstrated by self-repetition. Similarly, reduced lexical diversity has been linked with deceptive behaviour in speech (Zhou et al., 2004). There is also tentative support for the claim that measures of lexical diversity predict misinformation (Hou et al., 2019). Consequently, diversity may reflect honesty and trustworthiness in political speech.

Following a measure of diversity described in Zhou et al. (2004), this study applies the Type-Token-Ratio (TTR) over all interviewee speech turns. TTR is the set count of unique words divided by the total word count.

Relevance

The fourth outcome variable this study explores is named *relevance*. Relevance reflects the extent an interviewee's response shares broad semantic similarity with the question they were asked. Prior research has indicated that politicians may strategically equivocate from a challenge posed by an interviewer to re-direct the line of questioning to something more favourable (Bull, 2003). Thus, providing an answer conceptually similar to the latest question may result in a more informative dialogue.

Relevance is measured by first converting all speech turns into message-level word embeddings. This step was performed using a pre-trained GLoVe language

model (Pennington et al., 2014). The output of GLoVe is a high-dimensional vector for each word, which is then averaged over all words within a speech turn (Flor & Andrews-Todd, 2022). The similarity of each question-answer pair is calculated by calculating the cosine of the angle between the two turn-level embeddings (i.e., cosine similarity). A single measure of relevance is obtained by taking the mean value over all question-answer pairs.

Overall quality score

An overall quality score was also devised based on the four measures described in this section. Overall quality was calculated by taking the mean value over all four outcomes.

3.4.4 Evaluation Survey

An accompanying evaluation survey was created to capture annotators' evaluations of each interview. This step was conducted online via SurveyMonkey[3]. The survey comprised ten pages, one page per interview. Each page included the interview as an embedded video file, five questions targeted at each outcome measure, and one overall quality rating. Answers were required on a 1:10 scale, where 1 indicated a low rating and 10 a high rating.

3.4.5 Annotation procedure

Human annotation procedure

Each annotator reviewed and provided evaluations for all ten interviews. At the beginning of the survey, annotators were presented with some preliminary instructions. These instructions required the annotator to watch each interview in its entirety in a single whilst trying to maintain a politically-neutral stance. Annotators could take breaks between interviews to prevent fatigue effects but were asked to complete the survey within one week. Full details regarding the survey are presented in the Appendix.

[3]Survey was administered online at https://www.surveymonkey.com/ under a free basic membership plan

Computational procedure

Transcripts of the ten interviews were tokenised, POS tagged and lemmatised using CoPyCAT. The five outcome measures described in Section 3.4.3 were calculated from interviewee speech turns. Post calculation, all measures were standardised on a 0-1 scale via `Min-Max` normalisation.

3.4.6 Evaluation results

The survey was completed in full by each of the eight annotators. Descriptive statistics of each interview and for each outcome measure are presented in Table 3.3. Overall, annotators tended to rate the interviews favourably, with an average of at least 5.93 out of 10 for each outcome measure. A clear preference can be observed for Interview F, which achieved the highest average score for each outcome. Equally, Interview J received the lowest score for each outcome. Inspection of this interview found that it was characterised by high levels of arguing between the interviewer and interviewee, with frequent interruptions by both speakers. Interview J was also the longest, which may have contributed to its low score. Indeed, exploring the relationship between interview length and annotator ratings indicated a sizeable negative correlation, with coefficients ranging between -0.52 and -0.61. This indicates that annotators reacted less favourably to more extended interviews, which may reflect signs of boredom and fatigue.

Interview	Specificity	Clarity	Diversity	Relevance	Overall
Interview A	6.50 (2.56)	7.75 (2.18)	5.25 (2.65)	2.62 (1.40)	4.25 (2.25)
Interview B	4.75 (2.76)	6.37 (2.97)	6.62 (2.44)	6.00 (2.97)	5.62 (3.06)
Interview C	7.12 (1.95)	7.87 (2.69)	6.5 (2.44)	7.25 (2.31)	7.37 (2.44)
Interview D	5.12 (2.69)	4.62 (3.37)	5.37 (1.92)	4.87 (2.53)	5.00 (1.85)
Interview E	7.87 (1.95)	8.62 (1.40)	7.00 (1.41)	7.25 (3.01)	7.75 (1.83)
Interview F	8.75 (1.48)	8.87 (1.24)	8.12 (1.12)	8.87 (1.12)	8.62 (1.18)
Interview G	7.00 (2.72)	8.50 (1.41)	6.87 (1.35)	7.62 (2.19)	7.50 (1.92)
Interview H	7.37 (2.19)	7.62 (2.32)	7.50 (1.92)	8.25 (2.05)	7.62 (1.84)
Interview I	4.75 (2.54)	5.37 (2.87)	5.37 (2.06)	4.50 (2.87)	5.00 (2.72)
Interview J	1.50 (1.41)	3.00 (1.92)	1.87 (1.72)	1.87 (1.35)	2.00 (1.77)
Total	6.08 (2.94)	6.86 (2.90)	6.05 (2.50)	5.93 (3.13)	6.08 (2.82)

Table 3.3: Mean (SD) ratings obtained from human annotators per interview for each outcome variable. Each interview was scored on an ascending 1:10 quality scale.

3.4.7 Intra-class correlation analysis

An Intra-Class Correlation (ICC) analysis was performed on the annotations to measure the consistency between annotators. ICC is a measure of inter-rater agreement, similar to the more commonly used Cohen's K. However, the latter is restricted to agreement between two raters on categorical variables. Like similar metrics, ICC calculates a reliability score between 0-1, where a score close to 0 indicates low inter-rater agreement, and a score close to 1 suggests high or perfect agreement. According to Koo and Li (2016), ICC scores greater than 0.75 can be considered evidence of good agreement.

A two-way random effects model was fitted to the ratings of each outcome measure. A good level of agreement (minimum ICC$>$ 0.8) was observed between annotators for each outcome. Table 3.4 presents each outcome variable's average agreement and confidence intervals.

Outcome	ICC	CI95%
Specificity	0.85**	$0.67 - 0.96$
Clarity	0.82**	$0.59 - 0.95$
Diversity	0.84**	$0.63 - 0.95$
Relevance	0.89**	$0.74 - 0.97$
Overall Quality	0.87**	$0.7 - 0.96$

Table 3.4: Intra-class correlation scores for each outcome measure, assuming a two-way random effects model. ** denotes $p<$ .001

Comparison with computationally-derived outcome measures

Having demonstrated that the human annotators tended to agree with each other, it is now important to evaluate the level of agreement with the computationally-derived measures. The human ratings were normalised between 0-1 to be placed on the same scale as the computational measures. Consistent with the computational measures, this step was performed using `Min-Max` normalisation.

Comparison between human and machine-derived measures are illustrated in Figure 3.1. For each outcome, the distribution of scores for each interview is represented as a separate box plot. The computationally-derived outcome measure is then overlaid with an X. Comparing the computational scores to the normalised distribution of human ratings, we find the following percentage of computational

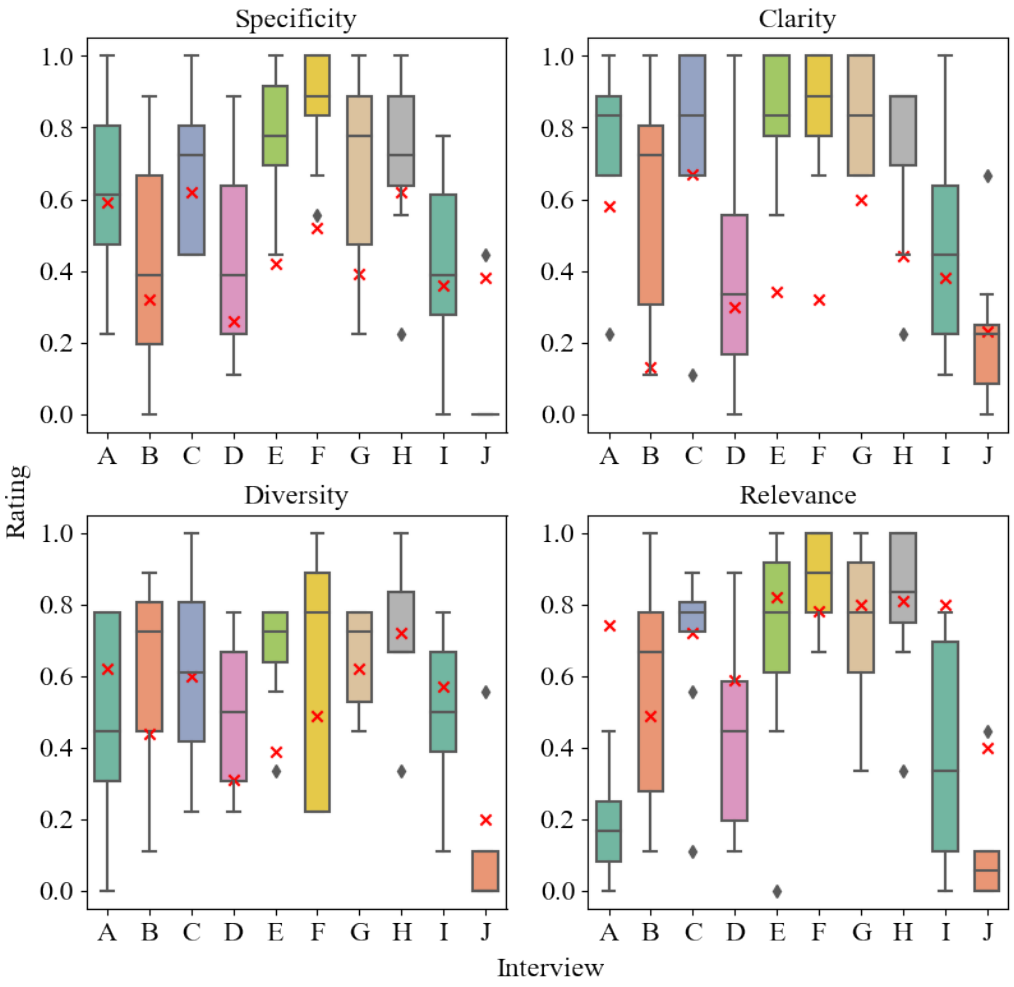

Figure 3.1: Normalised distribution of human ratings for specificity (top left), clarity (top right), diversity (bottom left), and relevance (bottom right). X refers to the normalised computational score predicted per interview, using all features

scores that fell within one standard deviation of the mean human score: specificity = 70%, clarity = 50%, diversity = 70%, relevance = 70%.

A linear regression model was used to measure how human annotators' overall rating was weighted by their ratings for each outcome. This approach was deemed a more suitable way to generate an overall score than simply taking the average, as it enabled a larger weighting for features deemed more critical.

Figure 3.2 shows the normalised distribution of overall success scores given by human annotators per interview. This new computationally-derived score marked with an X. A 90% agreement was observed with the mean human score – where computational scores are within one standard deviation of the average annotator score. Based on the average variance between the mean human score and the computational score, it would require four human raters to improve on the computational

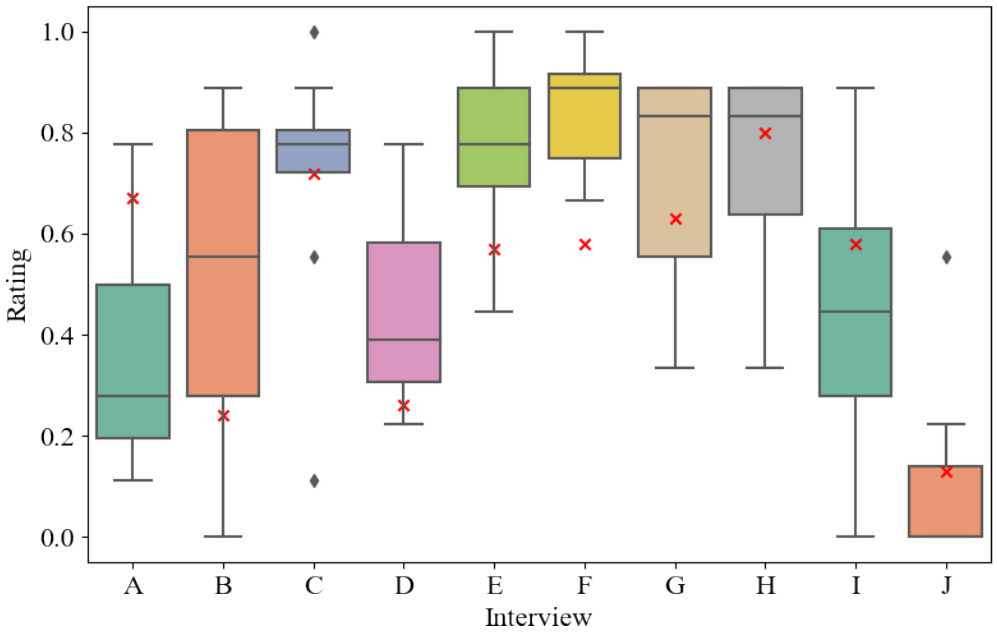

Figure 3.2: Normalised distribution of human ratings of overall interview quality. The overlaid X shows the corresponding normalised computational score.

predictions. Based on these scores, we can be confident that the human annotators' ratings capture meaningful interviewee behaviour within the context of a political interview. Furthermore, it has been shown that a computationally-derived measure broadly corresponds with the average human rating. In Section 3.5, supervised machine learning is used to predict each outcome score from mirroring-based features by the interviewer towards the interviewee.

3.5 Using supervised machine learning to predict interview success from interviewer mirroring behaviour

The previous section generated computational measures of success during a political interview based on interviewee speech properties. In this section, supervised machine learning is used to predict these values using features based on the extent the interviewer mirrors the verbal behaviour of the interviewee. As a reminder, mirroring is associated with a range of positive outcomes in conversation (Nenkova et al., 2008; Reitter & Moore, 2014). However, the extent that this behaviour extends to a political interview remains unknown.

3.5.1 Feature engineering

CoPyCAT was used to model the extent the interviewer mirrored the interviewee. Consistent with the claim that verbal mirroring should occur at multiple linguistic levels (see Garrod & Pickering, 2004; Pickering & Garrod, 2004), mirroring features were generated at the lexical, syntactic, stylistic, and semantic levels. Each of these will now be described briefly, although the reader is reminded that full details of how these features are generated are included in Chapter 2.

Micro-Level

Micro-level mirroring was calculated at the lexical, syntactic, and stylistic levels. All models were based on the approach introduced in Danescu-Niculescu-Mizil et al. (2011). This approach generates an individual feature value for each token (i.e., word or syntactic structure) produced by the interviewer. At the lexical level, mirroring was based on n-grams from unigrams (single words) to trigrams (three consecutive words). Syntactic features were based on dependency parse tags, and stylistic features were based on psycholinguistic word categories output from LIWC.

Macro-Level

At a macro level, features were calculated at the three levels discussed above in addition to a semantic measure. As highlighted in Chapter 2, lexical mirroring at the macro level was based on LLA (Fusaroli et al., 2012). Syntactic mirroring was based on the average negated branch factor difference (Murray & Oertel, 2018). Stylistic mirroring was based on the reciprocal LSM score introduced in Müller-Frommeyer et al. (2019). Finally, semantic mirroring was based on the average cosine similarity between adjacent speech turns.

Contextual Features

In addition to these features, contextual variables from the collected meta-data were included as one-hot encoded categorical variables. These included:

1. The interview length (measured in words as time-stamps were not included in all transcripts).

2. The host network.
3. The political orientation of the interviewer.
4. The political orientation of the interviewee.
5. Whether speakers shared the same political orientation.
6. The gender of the interviewer.
7. The gender of the interviewee.
8. Whether speakers shared a common gender identity.

The final output of these steps was a high-dimensional feature vector for each interview. These features were input into a supervised machine learning classifier to predict conversational quality.

3.5.2 Machine learning algorithms

This study experimented with four ensemble-based algorithms: (1) random forest (Breiman, 2001), (2) extremely randomised trees[4] (Geurts et al., 2006), (3) gradient boosting (Friedman, 2001), and (4) extreme gradient boosting[5] (T. Chen & Guestrin, 2016). The choice of the algorithm was based on suitability for a regression task, as each outcome variable comprises a real value. The inclusion of extra trees and XGBoost, in particular, was to minimise the impact of the high dimensional feature space relative to the number of data points (i.e., the high n low p problem). All models were initially used with default hyperparameters as specified in `scikit-learn`, except for the number of estimators, which was reduced from 100 to 40. Hyperparameters are adjustable settings the practitioner configures before executing the algorithm (Raschka, 2016).

All models were cross-validated using K-fold cross-validation, where K=10. A K-fold cross-validation strategy means the model is trained and tested iteratively K times (Yarkoni & Westfall, 2017). A model is trained on K-1 folds within each training cycle and tested on the holdout set. Rotating the dataset in this way ensures all data appears once in the test set. The decision to set K as 10, in this

[4] Also known as Extra Trees
[5] Also known as XGBoost

case, was influenced by similar research using high-dimensional features to predict social tension in online interactions (see Burnap et al., 2015).

For each cross-validated fold, the output of the machine learning algorithm is a numerical prediction for each interview in the holdout set. The model's ability to predict each outcome variable is measured by calculating the Root Mean Squared Error (RMSE) - the root of the average squared difference between the true and predicted values in the test set. We take the root to re-scale the error to reflect the units of the response variable. After cross-validation, we are left with K error values, which are then averaged as an overall measure of model performance.

As a measure of model error, lower RMSE indicates a more accurate model. Comparing performance to a baseline estimator indicates whether the model offers an improvement. In this chapter, models are compared to two baselines. Following Reitter and Moore (2014) and Espinoza et al. (2019), the first baseline (baseline 1) repeats the mean outcome variable in the training set for every prediction. Theoretically, it should be indistinguishable from chance. The second baseline (baseline 2) is a simple bag-of-words model based on interviewer word frequency counts. The reason for including this model as a baseline was to explore the relative difference in the performance of interviewer mirroring versus interviewer word use. Suppose mirroring contributes to the quality of the interviewee's response. In that case, we can expect an improvement in models that utilise mirroring-based features compared to one that only counts the frequency of individual words.

3.5.3 Prediction performance

Comparison of machine learning algorithms

The first experiment reported here compares model errors for the four algorithms. Figure 3.3 illustrates the average model error for each algorithm when predicting each outcome variable. Algorithms performed consistently within each outcome, with no apparent advantage from using any one algorithm over another. A series of one-way Analysis of Variance (ANOVA) tests were performed on the RMSE scores for each outcome variable to test for any statistical differences. In each case, the ANOVA was performed having first satisfied the appropriate statistical assumptions. Inspection of a box plot confirmed the absence of outliers, and the assumption of normality was satisfied via a non-significant Shapiro-Wilk's test ($p > .05$). Similarly,

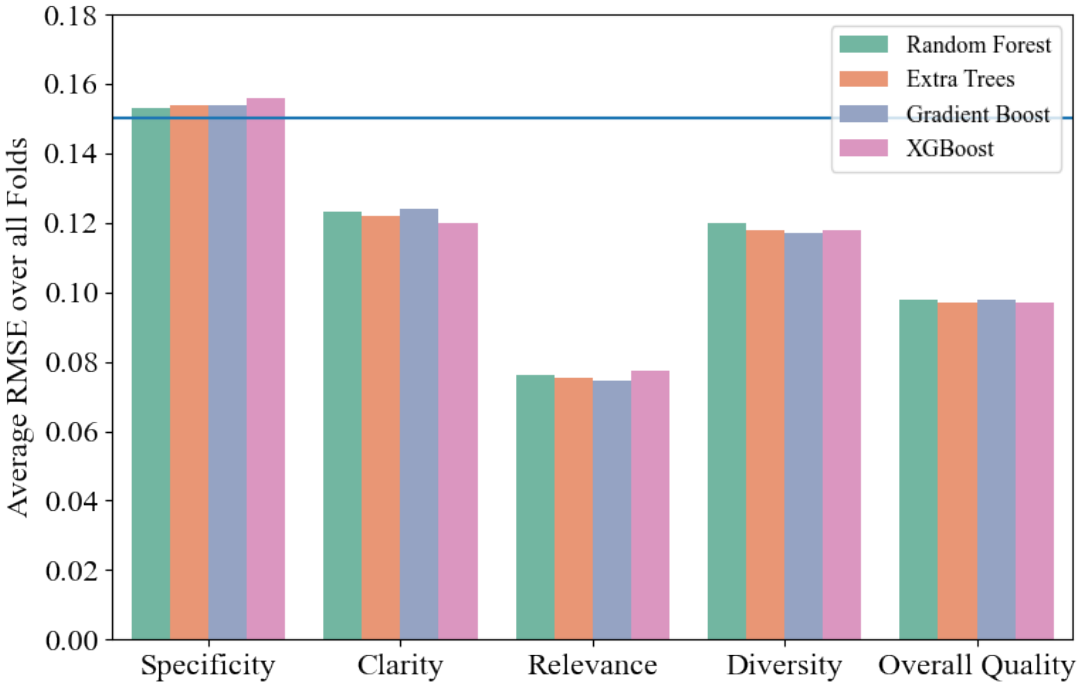

Figure 3.3: Prediction performance per algorithm on each of the outcome variables. Lower scores are better. The blue horizontal line indicates the average baseline estimator (repetition of training mean). Scores ***below*** this line indicate models that exceed baseline performance.

equality of the variance of differences between algorithms was assumed via a non-significant Mauchly's Test of Sphericity ($p > .05$). As expected, the output of the ANOVA revealed no statistically significant effect between the four algorithms for any of the outcome variables (all tests $p > .05$).

Comparison of outcome variables

That said, there were clear differences in performance between the outcome variables themselves - indicating that the models found some outcomes easier to predict than others. In particular, clarity, diversity, relevance, and overall quality consistently exceeded baseline one, irrespective of which algorithm was used. Error for diversity dropped by approximately half, while model error for clarity and relevance dropped by roughly 17% and 23%, respectively. Likewise, model error when predicting overall quality dropped by 46% over this baseline. A one-tailed Wilcoxon signed-rank test was performed on the prediction errors to explore the significance of each model's improvement over the baseline. The results of this test revealed that absolute model errors for clarity, diversity, relevance, and overall success were significantly lower than the baseline ($p < .001$). These findings suggest that the interviewer mirroring

the interviewee contributed to the interviewee's response for these variables. A comparison of each algorithm against baseline one is illustrated in Figure 3.3.

Conversely, the average model error for specificity (M=0.16) was more prominent than baseline 1, suggesting predictions were no different from chance. There are several reasons why this might be the case, including the notion that specificity may not be an essential characteristic to express in a political interview, unlike other domains (see Alison et al., 2013; Christiansen et al., 2018). These are explored in more detail in Section 3.6. For now, given this level of performance, specificity is dropped from the remainder of this chapter.

Hyperparameter tuning

The next step of the analysis involves attempting to improve performance by fine-tuning model hyperparameters. To simplify the process, and given the similarity in performance across the four algorithms, only a random forest is used in this section. Hyperparameters are tunable aspects of a machine-learning model set before model training. According to `scikit-learn` documentation (see Pedregosa et al., 2011), a random forest regressor has twelve hyperparameters. Commonly targeted hyperparameters include the number of estimators, the maximum depth of a decision tree, and the minimum number of examples allowed to generate a leaf node. In this section, six hyperparameters are tuned using a grid-search approach. Grid search is an optimisation technique that improves a model by exhaustively searching through a range of values for each hyperparameter. Table 3.5 presents the tuning strategy used for this section.

Description	Parameter Name	Range of Values
Number of trees	n_estimators	40, 80, 120, 160
Max. depth of trees	max_depth	No Max., 10, 15
Max. features to split node	max_features	n_estimators, $\sqrt{n\_estimators}$, $\log_2 n\_estimators$
Min. samples to split node	min_samples_split	2, 3
Min. samples at leaf node	min_samples_leaf	1, 2, 3
Min. impurity decrease to split node	min_impurity_decrease	0.0, 0.1

Table 3.5: Range of hyper-parameter settings used in grid-search

Nested cross-validation was used to tune hyperparameters without leaking information between the train and test sets. This approach required two sets of K-fold

loops called outer and inner loops. The outer loop operates identically to a standard K-fold, creating K different folds from the data. Then, we perform another, typically smaller K-fold within each outer fold to optimise parameters. This work used K=10 as the outer loop, with K=3 as the inner loop.

The hyperparameter tuning results found that the default random forest performance was equivalent to the optimised model for each outcome, indicating a point of diminishing returns. Results of a paired-samples t-test on the errors produced by each model confirmed that there were no statistical differences ($p > .05$) between the default and optimised models. As such, the default model was kept.

Feature comparison

The models included in this chapter have considered many low-level features using multiple linguistic representations at both macro and micro levels. Predictions were re-run for each outcome using a subset of the available features to examine whether the performance observed was associated with a particular feature set. Table 3.6 describes the average performance of feature subsets on each outcome variable. For comparison, both baseline estimators are included. As a reminder, models with an error **lower** than the baseline can be interpreted as a more accurate model.

Model	Clarity	Diversity	Relevance	Overall
Baseline 1	0.144±0.016	0.158±0.016	0.142±0.011	0.15±0.014
Baseline 2	0.138±0.017	0.123±0.012	0.129±0.014	0.117±0.013
Micro (all)	0.128±0.013	0.114±0.013	0.125±0.012	0.106±0.011
Lexical (micro)	0.129±0.013	0.115±0.013	0.126±0.013	0.105±0.01
Style (micro)	0.134±0.018	0.123±0.007	0.133±0.015	0.115±0.011
Syntax (micro)	0.134±0.015	0.121±0.008	0.137±0.014	0.111±0.011
Macro (all)	0.133±0.016	0.118±0.015	0.120±0.010	0.105±0.009
Contextual	0.160±0.016	**0.081**±0.008	0.153±0.008	0.12±0.007
Micro + Macro	**0.124**±0.010	0.108±0.014	**0.115**±0.010	**0.101**±0.01

Table 3.6: Mean ($\pm SD$) RMSE scores per feature subset when predicting each outcome variable. Lower is better. Number in **bold** indicates best performing feature subset per outcome.

Except for predicting diversity, the best-performing model for each outcome comprised both micro and macro mirroring features. Curiously, meta-features were highly predictive of linguistic diversity but not of any other outcome. This finding motivated a deeper examination of the unique features influencing each outcome's performance. Using permutation feature importance to rank the relative importance

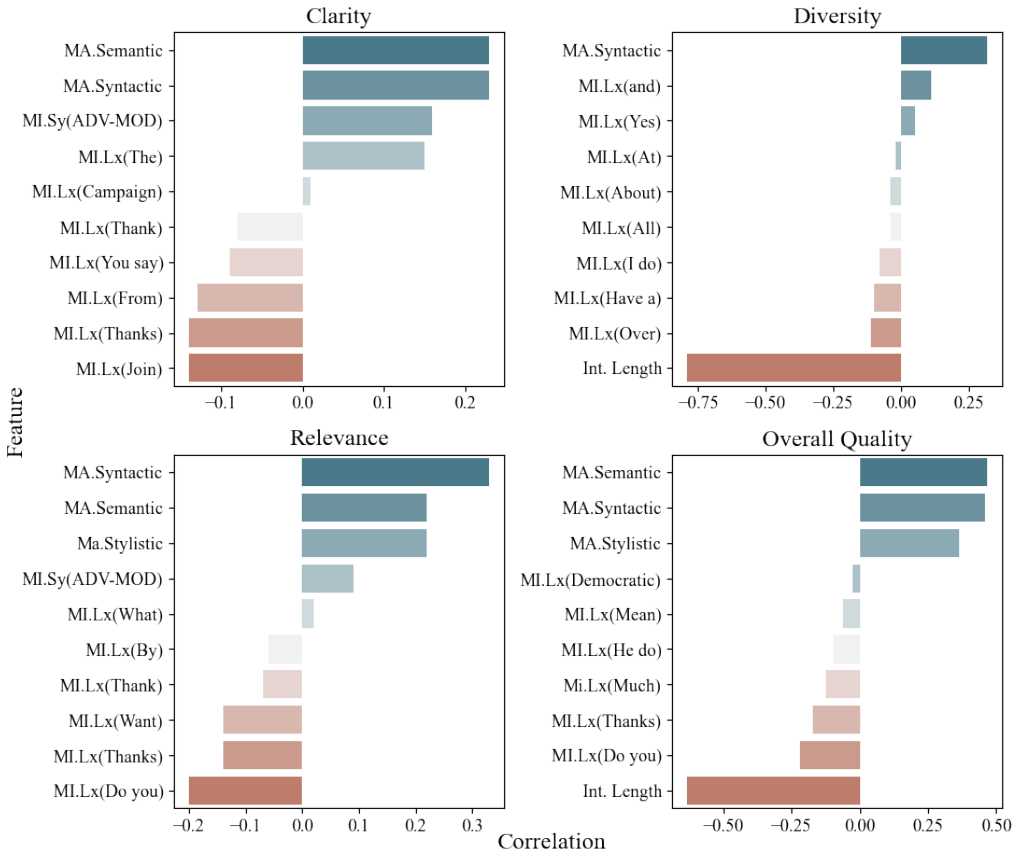

Figure 3.4: Correlations of the top ten features when predicting each outcome with the full complement of features. 'MI' refers to a micro-level feature, whereas 'MA' refers to macro. Similarly, 'Lx' indicates a lexical feature, whereas 'Sy', and 'St' refer to syntactic and stylistic respectively. For micro-level features, the precise feature value (i.e., the word or syntactic structure) is stated in parentheses.

of each feature (Breiman, 2001), Figure 3.4 describes the strength and direction of correlation between the top ten features per each outcome.

Notable observations here include the following: There was a substantial negative correlation between the length of the interview and the linguistic diversity of the interviewee. This relationship is a known consequence of using the TTR to measure language diversity. It is also indicative of *Zipf's law*, which states that the most common words will be inversely proportionate to their rank. The findings here suggest that interviews do not become more linguistically diverse by being longer. Given that televised political interviews occur within a dedicated time slot, it seems reasonable to posit that, in shorter interviews, politicians are communicating their crucial talking points more concisely than would otherwise be the case in longer-format discussions. Although TTR was used as the computational measure of diversity in this work, alternative approaches that may yield different results could be explored. Two such examples considered but ultimately discarded here are

content diversity and linguistic redundancy (Zhou et al., 2004).

A macro measure of semantic mirroring positively correlated with the clarity and relevance of interviewee responses. Similarly, macro-level stylistic mirroring positively correlated with relevance. Lastly, macro-level syntactic mirroring positively correlated with clarity and relevance. These patterns are consistent with communication theories that suggest that increased linguistic similarity improves understanding between speakers (Garrod & Pickering, 2004) and contribute to better conversational outcomes (Nenkova et al., 2008).

Generally, stronger correlations were observed between the macro-level features and the outcomes. These were exclusively in the positive direction, indicating that increased mirroring at a macro-level contributed to an increase in each outcome. Interestingly, the models also appeared to identify patterns of divergent interviewer behaviour – illustrated by negative correlations in the above plots. These were exclusively at a micro-level and suggested that some words and phrases were more specific to the interviewer. For instance, mirroring for phrases such as *do you* and *you say* negatively correlated with relevance and clarity. This hints that the divergence of these phrases by the interviewer aligned with clearer and more appropriate responses from the interviewee. Intuitively, this makes sense and may capture differences in language between speakers when one speaker is in the position of interviewer and another in the role of the interviewee. Whilst divergent behaviour tends to be overlooked within the broader mirroring literature, this finding highlights the potential impact of the speakers' roles in governing the extent of mirroring behaviour.

Error analysis

Lastly, a brief error analysis was performed on the overall conversation quality metric when predicted using the complete feature set. Analysis was performed by examining the relationship between rank-ordered predictions made by the model with the computational ground-truth scores. Based on Figure 3.5, it is possible to identify interviews where the model correctly predicted a high-quality interview (data points in the lower left corner) and where it incorrectly predicted a high-quality interview (lower right corner). Equally, the Figure can be used to identify those interviews that were correctly predicted to be of poorer quality (upper left corner) compared to incorrect predictions of poor quality (upper left corner). Whilst examining the differences between data points that fall within each of these regions was beyond the scope of the current thesis, such an approach might be helpful when a large cor-

pus such as this requires filtering before a more qualitatively-centric analysis. Such an approach might help inform a more detailed analysis exploring those instances where mirroring did not have the desired effect.

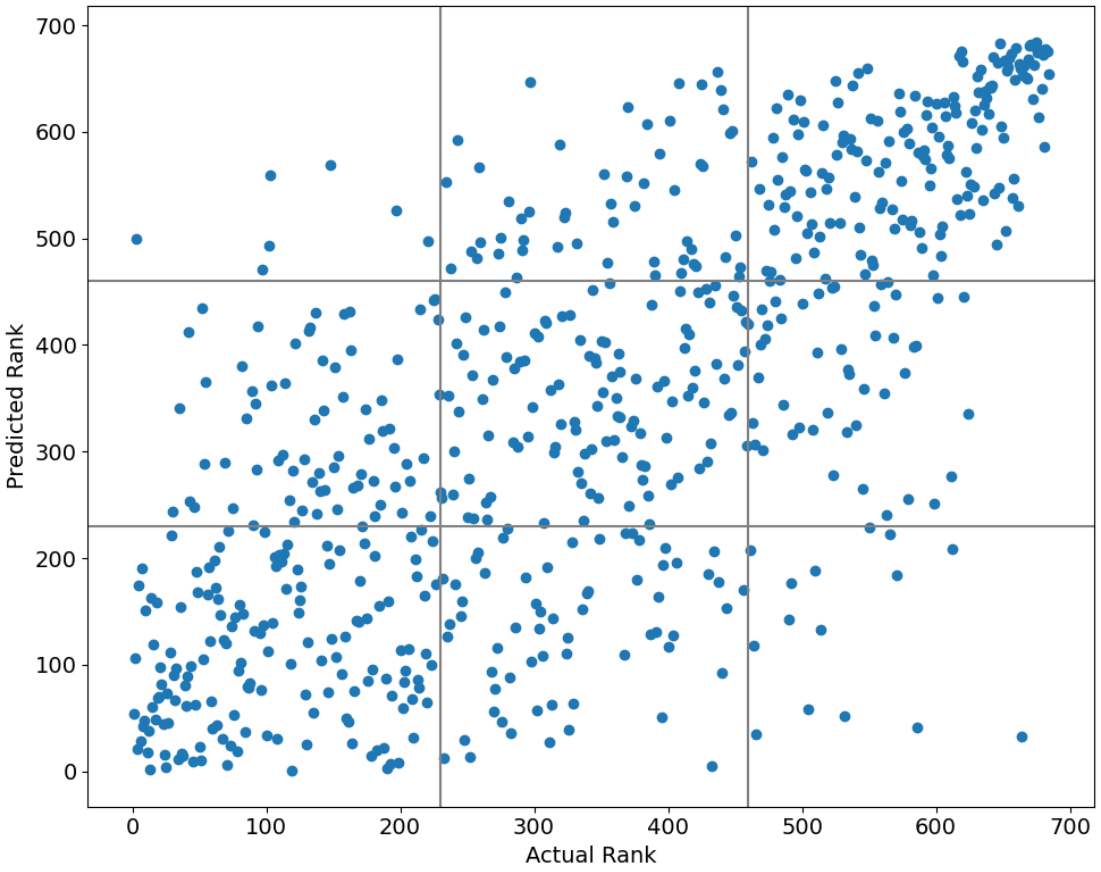

Figure 3.5: Relationship between actual and predicted conversational quality scores based on rank-ordering.

3.6 Discussion

This chapter has introduced an automated approach for analysing interview data and successfully demonstrated it on a corpus of publicly available political interviews. The results confirm that it is possible to successfully encode social scientific knowledge pertinent to interviewing into a computational analysis. Prudently, this can be harnessed both as a full analysis or as an initial mapping of a large corpus of conversational transcripts. The method offers an interpretable and reproducible alternative to time-consuming pre-labelling interview transcripts with behaviours of interest. This approach should offer an encouraging sign to both computer scientists and social scientists alike when seeking to analyse conversations at scale.

A series of psychologically-informed models inspired by theories of effective verbal communication significantly outperform a simple bag-of-words model across four-out-of-five related outcome variables and justify the inclusion of domain knowledge within computer science research. Using human-decipherable features has also meant that the analysis is helpful for future research within qualitative and quantitative domains. Although this work has modelled an array of interviewer behavioural features based on verbal mirroring, the choice of features is not exhaustive. Speakers in conversation have been shown to mirror each other at additional linguistic and paralinguistic levels such as vocal pitch and intensity (see R. Levitan et al., 2012; R. Levitan & Hirschberg, 2011; S. I. Levitan et al., 2018; Weise & Levitan, 2018). However, the extent to which these features contribute to the interview's outcome is yet to be known and will be the focus of an upcoming chapter.

Despite the close alignment between human and computer scores for specificity, the best-performing models did not successfully predict this measure based on the interviewer's behaviour. This result may be specific to the political interview domain as establishing specific information is an unlikely goal in and of itself within political interviewing. However, the construction of specificity as a computational score required the author to make several assumptions regarding which named entities would be relevant in a political interviewing context. Including different categories of named entities may have resulted in a different outcome. The decision to include specificity in this work was taken as it may be helpful in analysing interviews where the objective is more explicitly focused on information-gathering. Comparing a measure of specificity in a political interview with an alternative domain, for example, a police interrogation (see Alison et al., 2013), would be an interesting future step.

3.6.1 Study limitations

As discussed in Section 3.2, political interviewing, unlike other forms of social interaction, evades a clearly defined marker of success. This absence motivated the search for interviewee behaviours that were both accessible direct from the transcript and meaningfully related to the interviewer's goals. Considering that, in the context of a televised interview, the interviewer is essentially the voice of the audience (Heritage, 1985), it was concluded that specificity, clarity, diversity, and relevance were worthwhile behaviours to model. However, the importance of these behaviours was challenging to verify with the small number of annotators. Whilst comparison of human and computational measures indicated a reasonable level of alignment for

each outcome, it is acknowledged that other behaviours may also be of equal or greater importance. The lack of performance when predicting specificity highlights this, indicating that specificity may not be as crucial in a political interview as the author had initially assumed.

A general oversight with the human annotation task described in Section 3.4 is that interviews were presented in the same order to each annotator. Consequently, the tedium of the task may have impacted the integrity of the ratings provided to later interviews. It has already been noted that the worst-rated interview for each outcome variable was the longest and positioned as the last interview in the survey. The potential impact of this limitation has been somewhat alleviated by a manual inspection of the interview, which revealed it to be notably poor due to the high level of argumentation and hostility between the two speakers. Removing this interview, rotating the order in which the interviews were presented to the annotators, and creating a larger sample of both interviews and annotators, are all improvements that could be made to this aspect of the study.

Features derived from SSP are, by definition, low-level behaviours such as words. As such, it can be difficult for humans to interpret individual behaviours meaningfully. Mirroring is a subconscious behaviour; as such, we should be wary of overstating the power of merely repeating a particular word based on its prior use. It would be incorrect to assert that repeating a specific word or phrase has inherent value in encouraging greater cooperation from an interviewee. Instead, what CAT seeks to capture is a subtle shift in behaviour. Consciously trying to mirror another speaker's language style will undoubtedly result in a very staggered conversation. Instead, non-conscious mirroring serves as a proxy for behaviours such as attention or empathy – behaviours that would otherwise be labelled subjectively by experts.

This is not to say that conscious mirroring is ineffective. Intentional mirroring behaviours such as reflecting and paraphrasing are fundamental components in person-centred therapy (W. R. Miller & Rollnick, 2012). Like non-conscious mirroring, these techniques have been shown to elicit a prosocial response in social interactions (Alison et al., 2013). However, as conscious strategies require first identifying the intent to mirror an interlocutor, they are much harder to measure through automated methods.

Future reconstruction of this task should consider the additional factors that influence the quality of the interviewee's speech. A speaker's natural speaking rhythm and the topic of discussion are also likely to impact the outcome scores generated

here. It would have been sensible to monitor changes in response throughout the interaction. Alternatively, comparing a smaller number of interviewees across many interviewers may have also been helpful.

3.6.2 Recommendations

In light of the limitations discussed above, several recommendations should form the basis of future work in this area.

Using transcripts based on human effort was deemed necessary to ensure high-quality features. However, relying on manual transcription could cause a bottleneck in a real-world setting. This chapter uses secondary transcripts generated externally for this research, thus avoiding the lengthy time commitments required to generate a written record of the corpus (Moore, 2015). However, the availability of a pre-existing body of transcripts cannot be guaranteed in each new application. Instead, continuing to explore this work in novel domains will require considerable human effort if manual transcription remains the sole approach to generating data. Manual transcription will undoubtedly hinder the mobility of this work between different applications.

Instead, future work could explore the utility of transcripts derived from Automatic Speech Recognition (ASR) software. Recent comparisons between a manual and ASR transcription indicate that whilst the two methods produce drastically different transcriptions, performance on a downstream analytic task remained similar (S. J. Pentland et al., 2022). This indicates that ASR systems may be more capable than traditional evaluation metrics, such as word error rate (WER), would suggest. Indeed, a comparison of manual and ASR systems on dialogue act tagging found that ASR transcripts performed better in some settings (Malik et al., 2018). With the time-consuming nature of manual transcription, a similar comparison of the tasks reported in this chapter using automated transcription methods may increase confidence in automated transcription methods over isolated human effort.

A second recommendation involves exploring the variation in performance when different pre-processing decisions are considered. An area that could be particularly interesting is the treatment of stop-words. Guidelines in text mining routinely advocate the removal of stop-words as they offer little semantic value (Feldman & Sanger, 2006). This contrasts social psychological research highlighting the utility of these features when exploring social traits such as emotions (Tausczik & Pennebaker,

2010) and intent (Drouin et al., 2017). A systematic comparison of the different pre-processing decisions one can make rarely appear in the literature. Nevertheless, prior research suggests comparing performance with and without stop-words is worthwhile. For example, Silva and Ribeiro (2003) found that including stop-words in a document classification performance negatively impacted recall. However, an inverse effect was observed in Saif et al. (2014), who found that removal of a pre-compiled stop-word list negatively impacted performance in a sentiment classification task. Curiously, a dynamic list based on word frequency did improve performance by reducing data sparsity. The utility of a bag-of-stop-words approach has also been examined in Nenkova et al. (2008), where the most frequent words in a corpus were used to classify the naturalness of dialogues. In some sense, the efficacy of including stop-words has been indirectly examined throughout this chapter by including features based on linguistic *style*. Notwithstanding, the complete feature set's high dimensionality could contribute to overfitting the training data. As such, the removal of stop-words, alongside other potential dimensionality reduction techniques, could be worthwhile.

A further recommendation would be to compare the findings presented here to other forms of social interaction. As highlighted in Chapter 2, an overarching aim of `CoPyCAT` was to create a tool that could be easily applied to novel settings. One of the reasons behind verbal mirroring as a phenomenon of interest is its ubiquity as a social behaviour. It would, therefore, be interesting to examine its effects on other types of interviewing. Lastly, a final recommendation involves extending the scope beyond the effects of verbal mirroring to include other modalities. Interlocutors naturally mirror one another across various behaviours, including vocal behaviour (R. Levitan & Hirschberg, 2011) and physical gestures (Bergmann & Kopp, 2012). In particular, vocal analysis is a topic that could be highly relevant to effective interviewing, as it has been linked to rapport-building (Lubold & Pon-Barry, 2014) and empathy expression in dyads (Imel et al., 2014). Whilst prior research has indicated that non-verbal mirroring is present in political interviews (Gregory & Webster, 1996), its utility as an effective communication tool appears absent from the literature. Given the aims of this thesis, the following chapter will seek to address this gap.

3.6.3 Publication record

The work included within this chapter has been the subject of two co-authored submissions to conferences in computational linguistics. The first version, submitted in 2019 to the 27th International Conference on Computational Linguistics was ultimately unsuccessful. However, a revised version was submitted, accepted, and presented at Interspeech in 2021 (see Cook et al., 2021). The author of this thesis conducted data collection, analysis, and written aspects of this work. Collaborating authors supported the design of experiments and provided valuable feedback during the write-up.

3.6.4 Chapter summary

The objective of this chapter was to explore whether a computational model of interviewer behaviour could help evaluate the outcome of an interview. Automating the detection of interviewer behaviour through a linguistic analysis was intended to reduce reliance on traditional qualitative methods that rely heavily on human annotation. The work was conducted within a political interviewing domain due to sufficient quantities of available data. However, it is extendable to other types of dyadic interaction. The results presented in this chapter highlight the utility of an automated approach, relying on a limited amount of manual effort performed by a small number of human annotators. Automating both the behavioural features and outcome variables has increased the scale of the research, enabling in-depth analysis of a corpus larger than is conventional in typical conversation analysis within the social sciences. The findings of this chapter have generated new questions regarding the potential use of social signals as a predictor of interview outcomes and have contributed to existing social scientific theories of communication. Motivated by the findings presented here, in the next chapter, this work is extended by exploring the utility of acoustic (i.e., nonverbal) mirroring on the outcome of an interview.

Chapter 4

Multi-Modal Prediction of Interview Outcomes from Verbal and Non-Verbal Mirroring

4.1 Introduction

The previous two chapters have introduced a novel approach to analysing social interactions without heavy reliance on manual annotation. Results demonstrated that interviewer mirroring towards the interviewee could be quantified and that the extent of mirroring contributed to the conversational quality of the interview. Building models from text, however, ignore the myriad of other social signals humans exchange in everyday conversation. Alongside verbal behaviour, another instance is communication via non-verbal channels. How we speak, changes in pitch, cadence, or how loud or quiet our voice is all convey additional information about our emotions and motives. However, these effects are removed when working purely from a transcription.

In one of the main findings of the previous chapter, it was demonstrated that verbal mirroring of the interviewee by the interviewer could be used to predict the quality of the interviewee's response. Given the intricacy of human behaviour, however, a multi-modal approach may create a more robust picture of a social phenomenon (Alameda-Pineda et al., 2019). Alongside verbal mirroring, speakers have

been repeatedly shown to mirror each other's non-verbal behaviour (R. Levitan & Hirschberg, 2011). This is also affiliated with prosocial behaviour such as rapport-building (Lubold & Pon-Barry, 2014), which may contribute to success in dialogue-based games (Rahimi et al., 2017). Unlike measures based on verbal behaviour, non-verbal mirroring does not consider the content of speech. It is, therefore, less impacted by topicality, synonymy and homonymy effects. Moreover, the extent of non-verbal mirroring appears highly entwined with contextual factors such as gender (Bilous & Krauss, 1988).

However, researchers that study non-verbal mirroring in naturalistic settings face several challenges that are more problematic to address than verbal mirroring studies. Firstly, large text-based corpora are typically more accessible, with fewer computational storage requirements than high-quality audio. Second, transcripts are easier to parse into distinct analytic units (i.e., speech turns). Audio, by contrast, often requires dedicated techniques to achieve a similar effect. Due to storage issues, most publicly available audio tends to be in a format that makes it challenging to automatically identify who spoke and when. Third, as a form of social signal, there is greater availability and arguably a less-severe learning curve required to begin working with text than audio data. Lastly, there is a lack of a standardised approach concerning how non-verbal mirroring is measured. Collectively, these challenges have meant that comparatively fewer studies examine the effects of mirroring along a non-verbal channel than verbal. Very few studies also examine the effects of non-verbal mirroring in naturalistic settings. Consequently, the aims of this chapter will be the following:

- Describe the methodological literature as it relates to non-verbal mirroring
- Based on existing measures, determine whether interviewers mirror the non-verbal behaviour of the interviewee
- Explore the impact of contextual variables on interviewer mirroring behaviour. Namely, based on whether speakers share or differ on both political belief and gender
- As an extension of findings uncovered in the previous chapter, examine the predictive utility of a: (i) non-verbal, and (ii) multi-modal model of mirroring by combining verbal and non-verbal mirroring features

4.2 Automatic measures of acoustic-prosodic mirroring

As with mirroring based on verbal behaviour, studies measuring the effects of non-verbal similarity are scattered throughout the literature over the last fifty years. Although most commonly associated with linguistics, Communication Accommodation Theory CAT originated as a theory of accent mobility (see H. Giles, 1973). This work demonstrated that speakers adjust their accents to meet social goals. Referred to at the time as *speech* accommodation theory, individuals have been shown to converge their non-verbal behaviours for a variety of reasons, reflecting dynamics in social status (Gregory & Webster, 1996), social identities (Bourhis et al., 1975), and attraction (Michalsky et al., 2018).

Much of this early work was based on perceived similarity inferred by independent annotators (Pardo, 2006). Here, annotators are given audio samples in triplets (AXB testing). The audio is a short piece of dialogue between two speakers. Sample A serves as a control, reflecting a speaker's baseline prosodic behaviour. Sample X is audio by a second speaker, often a confederate or the researcher. Sample B is the first speaker's reply to Sample X. The task for annotators is to rate the extent they perceive audio sample B to be more similar to X than audio sample A. AXB testing suffers many of the same drawbacks as text-based coding. It is highly time-consuming to generate large bodies of data. Mirroring at a non-verbal level is also subtle, with considerable variation within and across speaker demographics (Pardo et al., 2018). This may increase the need for annotators with experience in AXB testing to ensure slight differences are detected.

In non-verbal mirroring research, there is a dichotomy between subjective and objective measurements, as discussed by (Weise et al., 2019). Subjective measures, like the AXB testing mentioned earlier, rely on human perception and offer a comprehensive evaluation of non-verbal mirroring. In contrast, objective measures are derived through computational methods based on quantifiable signal aspects, such as pitch or vocal intensity. Objective measures are characterised by their accuracy, as they can provide exact numerical values; however, they usually provide only a limited perspective on the phenomenon being studied.

The development of automated approaches to text-based mirroring has increased since the beginning of the 21st century, and similar advancements in non-verbal mirroring have been observed in the last decade. As with text-based methods, the

goal of automated approaches for non-verbal mirroring is to delegate the detection process to computational systems. This area of research has been referred to as "acoustic-prosodic entrainment" within the computer science literature (Beňuš et al., 2014; R. Levitan et al., 2012; R. Levitan & Hirschberg, 2011; S. I. Levitan et al., 2018; Lubold & Pon-Barry, 2014; Weise & Levitan, 2018; Weise et al., 2019). The methods used for detecting entrainment are less standardised than perceptual experiments such as AXB testing (Weise et al., 2019) and differ in the number of features they analyse and the level of detail at which mirroring is examined. Some frameworks assess behaviours at the conversation level (Ward & Litman, 2007), while others focus on changes in behaviour during turn-exchanges (i.e., alterations in vocal behaviour immediately upon re-occupying the conversational floor) (R. Levitan & Hirschberg, 2011). Some studies investigate each feature individually, comparing pitch, amplitude, speaking rate, and so forth on a feature-by-feature basis (R. Levitan et al., 2012; Lubold & Pon-Barry, 2014; Weise et al., 2019). Alternatively, some researchers, such as C. C. Lee et al. (2011), employ dimensionality reduction techniques to generate a single measure of vocal entrainment. Dimensionality reduction is a method used to simplify complex datasets by reducing the number of variables while preserving as much information as possible (Velliangiri, Alagumuthukrishnan, et al., 2019). This approach helps to streamline the analysis and facilitate the interpretation of the results.

Given the subtlety of non-verbal mirroring, several studies have explored whether vocal similarity occurs as a factor of the interaction or is merely an effect randomly produced between speakers. This is typically achieved by a data augmentation approach, whereby randomised data is created by re-pairing speakers who did not interact. Increased similarity for the authentic dialogues is evidence of entrainment in the corpus. This approach has been used to validate the presence of non-verbal mirroring in many settings and is further evidence of the ubiquity of mirroring more broadly. For instance, Willi et al. (2018) used this method to confirm the presence of non-verbal mirroring in a collaborative experimental task. Results of this study highlighted that not only did speakers adapt to each others' non-verbal behaviour during the task, but this predicted task success. Similar techniques have been deployed in more naturally-occurring settings. Ward and Litman (2007) found evidence of non-verbal mirroring in a corpus of student-tutor dialogues. Their results indicated a decay effect to mirroring, whereby the strength of the prime (i.e., the amount of mirroring taking place) reduces as the distance between utterance grows. In other words, mirroring is more prominent for the most recent behaviour than behaviour that appeared further back in the conversation. A study of interactions

during couples' therapy used the same augmentation approach to validate mirroring between married couples, finding evidence that the degree of entrainment predicted relationship affect (C. C. Lee et al., 2014). Moreover, an identical approach was used to confirm the presence of mirroring by a therapist towards their patient in (Imel et al., 2014). Correlation analysis revealed that increased mirroring of vocal intensity was positively related to increased empathy.

The possible impact of contextual factors on mirroring

The extent to which contextual factors contribute to how and when speakers mirror one another remains an open question. One of the factors most commonly explored in this regard is gender composition. In an experimental study of mixed and same-gender dyads, Bilous and Krauss (1988) found evidence indicating that both males and females converge on the length of their speech turns when speaking to a member of the opposite sex. The same study also observed that females were more inclined to mirror the extent to which they interrupted the other speaker when speaking with a male. However, the effect was not reciprocated by males towards females. Debatably, this finding can be contrasted by a novel study of movie dialogue that suggested that males are more likely to mirror the behaviour of females compared to other males (Danescu-Niculescu-Mizil & Lee, 2011). This study also observed that female-initiated dialogue encouraged greater mirroring than male-initiated dialogue independent of the gender of the non-initiating speaker. This finding is contrasted by previous research that found that male-male dyads displayed greater behavioural mirroring compared to mixed-gender dyads (Street, 1984), a general finding also supported by (Pardo, 2006). Given these varied results, one of the tasks explored in this chapter is to examine the impact of gender composition on mirroring in a political interview. Rarely does this line of research appear to have been applied in the current context.

A second potentially significant factor is the impact of partisanship on non-verbal mirroring. As illustrated previously in the thesis, one of the underlying causes of mirroring is affiliation and rapport-building (Lubold & Pon-Barry, 2014). We converge towards those we seek affiliation (Lakin et al., 2003), and diverge from those when we want to exemplify disagreement (Culpeper et al., 2003) or amplify personal identities (H. Giles et al., 1991). Studies of intergroup convergence indicate that we make greater linguistic and paralinguistic concessions when we like the person or group we are conversing with. For instance, bilingual speakers are likelier to switch

to the other speakers' native language when they like the other speaker (Palomares et al., 2016). The same effect may also be true within the landscape of the modern political interview. Political journalists are expected to uphold neutrality in their reporting. However, increased non-neutrality is becoming a more common form of discourse in televised political interviews (Hutchby, 2011). We may find differences in interviewer mirroring based on whether the interview involves a politician holding complementary or contrasting views to the host or network.

4.3 Multi-modal corpus of political interviews

This chapter utilises a subset (approximately 16%) of the corpus previously introduced in Chapter 3. From the original corpus, video footage of N=152 interviews was located online by searching YouTube. The author of this thesis examined each interview to ensure that the footage had not been edited and accurately reflected the entirety of the interview. This step was vital, as forty interviews ended up being removed due to various quality issues - the most common of which (19 interviews) was incomplete audio (i.e., evidence that the interview footage had been edited for production purposes). After inspecting each interview, n=112 interviews were used as the final sub-corpus.

The overall duration of the sub-corpus is 24.07 hours. Interviews ranged in length from 34.59 minutes at the longest to 4.45 minutes at the shortest (M=12.55 minutes, SD=7.04 minutes). Interview length was approximately balanced between the host networks (see Table 4.1). As a reminder, all interviews were dyadic - they comprised two speakers. Therefore, the corpus did not include panel interviews or political debates, which tend to be more widely available online in video format.

Network	Political Bias	Total Interviews	% of Corpus	% Non-Partisan	Duration
CNN	Left	43	38.4	60.47	13:28
Fox	Right	25	22.3	36.0	11:10
CBS	Left-Centre	24	21.4	58.33	13.:36
NBC	Left	10	8.93	70.0	11:31
ABC News	Left-Centre	10	8.93	70.0	14:38

Table 4.1: Demographic information of the news networks represented in the sub-corpus.

Speaker Demographics

Except for MSNBC, each news network featured in the previous chapter was represented here (see Table 4.1 for descriptive statistics of this sub-corpus). Network distribution was broadly similar to the whole corpus, with CNN (38%) and Fox (22%) being the most represented networks. There was also a roughly balanced number of partisan (44%) and non-partisan (56%) interviews. In total, 103 unique participants were featured in the corpus, of which 32 acted as interviewers and 71 acted as interviewees. The comparatively smaller number of interviewers can be explained by networks relying on a consistent host to conduct interviews regularly. As such, a substantial portion of the dataset (78%) features just ten interviewers: Jake Tapper (17 interviews), Margaret Brennan (12), Chuck Todd (10), John Dickerson (9), Chris Cuomo (9), Anderson Cooper (8), Chris Wallace (8), George Stephanopoulos (6), Tucker Carlson (5), and Wolf Blitzer (3). The remaining 22% of the corpus tended to feature stand-in hosts, with 60% of interviewers only conducting one interview.

Of the interviewees, 63% were Republican Party members, with the remaining 37% being members of the Democratic Party. Because of the opportunistic nature of the data collection, it was impossible to create a balanced corpus regarding the distribution of interviewees. Unsurprisingly, interviews with prominent politicians were more frequently uploaded to YouTube than others. Donald Trump (11 interviews), Andrew Yang (7 interviews), and Mike Pence (5 interviews) were the most frequent interviewees. As with the interviewer pool, most interviewees (71%) were only featured in a single interview. Each interviewer-interviewee combination was unique except for two dyads (Jake Tapper and Andrew Yang and Chris Cuomo and Rudy Giuliani) to avoid saturating the corpus with the same speaker combination. In these two outlying cases, each dyad appeared in the corpus twice.

4.4 Audio pre-processing

The audio was extracted from each video file using `ffmpeg` in Python with a sampling rate of 16KHz. The extracted audio was in monophonic (mono) format. That is, the utterances of both speakers were compressed into a single waveform. This poses an initial challenge for detecting non-verbal mirroring, as it is unclear which part of the waveform belongs to each speaker. Ideally, the speech attributed to each speaker would be isolated within a single audio channel via individual microphones.

However, this information is lost when the two audio channels are combined to create a single file.

The computational process of extracting and attributing audio segments to a local source is known as *diarization*. Given that non-verbal mirroring compares the acoustic behaviour of individual speakers, partitioning speech to a high degree of accuracy is essential. However, as automatic diarization techniques are less mature than other tasks within ASR such as word recognition, they have rarely been used with Conversation Analytic research (Moore, 2015). One of the main challenges to overcome when diarising everyday speech is the abundance of overlapping talk. In naturally-occurring conversations, speakers anticipate an exchange of the conversational floor (the shift from listening to talking). This anticipation is routinely characterised by a slight overlap in speech (Sacks et al., 1974). In more contentious interactions, overlapping speech in the form of interruption becomes increasingly common (Grimshaw, 1990).

4.4.1 Semi-mechanical speaker diarization

These factors motivated the decision to develop a semi-mechanical solution to speech segmentation. Here, audio files would be initially parsed using automatic speaker diarization before being validated and cleaned by human verifiers. This section will outline these two steps in more detail.

The initial diarization was performed using an acoustic model from IBM Watson[1] This tool receives a single audio file and outputs timings corresponding to each speech turn's beginning and end. These timings are then used as boundary markers to parse the audio into smaller, speaker-specific and time-ordered sub-files.

However, it was noted that this step resulted in a large amount of vocal overlap at the turn boundaries. Therefore, the next step required validating the purity of each audio sub-file through a manual verification step. This ensured that mirroring of non-verbal behaviour was not based on speech segments attributable to an external source (i.e., the other speaker in the interview). This step was performed in two stages. In the first stage, each sub-file was split into smaller units based on the presence of a vocal pause. This was performed automatically using the open-source audio processing tool PRAAT (Boersma & Weenink, 2007). A vocal pause was de-

[1]The speech-to-text model can be located here: https://www.ibm.com/watson/developercloud/speech-to-text.html.

fined as a period of at least 200 milliseconds where the average intensity (measured in decibels) did not exceed 49 dB. The next stage required three trained research assistants to inspect each vocal sub-unit manually and correct speech attributed to the incorrect speaker by setting new turn boundaries. Whilst this step took considerable time to perform and would undoubtedly limit scalability to larger datasets, it was preferable to a wholly manual approach. The three research assistants were paid for their time, which comprised approximately 60 hours of effort. As part of this step, the overlapping speech was filtered from the audio, as were any background noises, such as studio effects. The output of this effort was a series of time-ordered and speaker-specific speech turns, from which non-verbal mirroring could now be calculated.

4.4.2 Audio feature engineering

Non-verbal mirroring was calculated by comparing each interviewer's turn with the preceding interviewee's utterance across five acoustic dimensions: (i) pitch, (ii) intensity, (iii) jitter, (iv) shimmer, and (v) Harmonics to noise ratio (HNR). Pitch is a psychological attribute of sound wave frequency. It is measured in Hertz (Hz) and refers to how high or low humans perceive a sound. Intensity is measured in decibels (dB) and is a measure of amplitude (how loud or quiet a sound appears). Jitter is related to pitch. It measures pitch instability or the variation in sound wave frequency (R. Levitan et al., 2012). Shimmer is broadly equivalent to jitter, which measures amplitude instability within a given time frame. Lastly, HNR is commonly interpreted as a measure of vocal hoarseness.

All features were z-score normalised by sex ($z = (x - \mu)/\sigma$; where x = raw acoustic value, μ = the mean of the speaker's biological sex (male/female), and σ = the standard deviation of biological sex (male/female)). This step ensured that the calculation of non-verbal mirroring was not impacted by whether the two speakers are more similar to one another by sharing the same sex or gender (R. Levitan et al., 2012).

Rather than extracting features across an entire speech turn, features are based only on vocal behaviour at turn exchanges. Using PRAAT (Boersma & Weenink, 2007), each speech turn is split into Inter-pausal units (IPU) - defined in R. Levitan and Hirschberg (2011, p. 3082) as "pause-free units of speech separated by at least 50 milliseconds". Typically, an IPU is very short, with most being less than three seconds long. To capture different types of non-verbal mirroring from each acoustic

dimension, the mean, minimum and maximum frame values are extracted from the last IPU in each interviewer turn and the first IPU from each interviewee turn. Non-verbal mirroring performed by the interviewer is then based on comparing each interviewer's turn at time t with the preceding interviewee's turn at time t-1.

Features of non-verbal mirroring are based on three turn-level measures of prosodic entrainment: *proximity*, *convergence*, and *synchrony*. These features were first described in R. Levitan and Hirschberg (2011), and later adapted in other works (R. Levitan et al., 2012; S. I. Levitan et al., 2018; Lubold & Pon-Barry, 2014; Weise & Levitan, 2018). These are outlined in more detail in the following section.

Proximity

Proximity measures the absolute distance between the two speakers on a particular acoustic dimension. For instance, if the pitch of an interviewer at turn t is measured at $100Hz$ and the pitch of an interviewee at turn $t-1$ is $120Hz$, pitch proximity at turn t would be measured at $20Hz$[2]. A measure of interviewer proximity towards the interviewee is calculated by averaging proximity scores overall interviewer speech turns. This score is then inverted so that values closer to zero indicate greater proximity - in other words, if the average absolute difference in pitch is $40Hz$ for interview A and $60Hz$ for interview B, we would interpret interview A as having higher proximity (as $-60Hz < -40Hz$). Therefore, the lower the negated average absolute difference between the speakers, the higher the proximity. If the interviewer uses a similar pitch to the interviewee throughout the interview, this interview would be characterised by a higher pitch proximity. Conversely, if the interviewer's pitch was consistently high, whereas the interviewee's pitch was consistently low, a low proximity score would be recorded. See Figure 4.1 for an illustration of proximity.

Convergence

Convergence is a dynamic measure of non-verbal mirroring in that it describes a change in mirroring behaviour over time (Wynn & Borrie, n.d.). Convergence is an extension of the proximity measure. It reflects the change in proximity throughout the interview. An interviewer demonstrates an increase in convergence when the

[2]As pitch values have been normalised before calculating proximity, proximity scores will not be in Hertz but rather represent relative positions or ratios within the new scale.

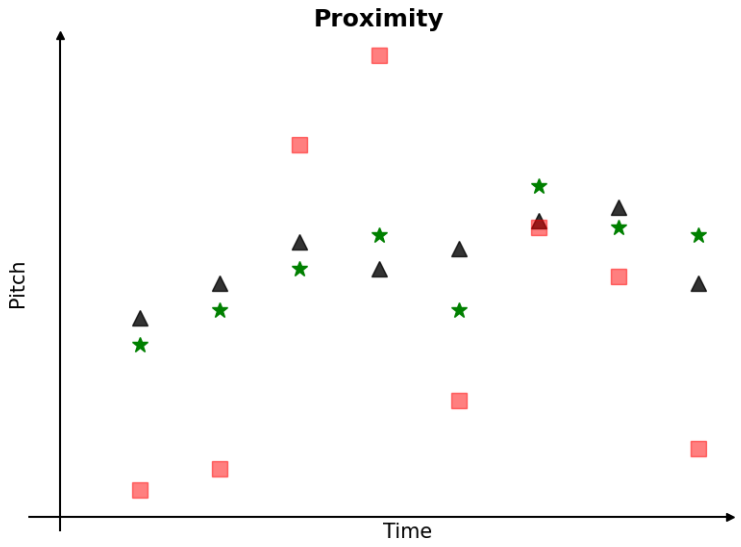

Figure 4.1: Hypothetical example of high (green) and low (red) proximity at each point in a conversation. Green points indicate high proximity as they are consistently close to a baseline (black). Conversely, red points indicate low proximity, as they are generally much further from the baseline.

proximity of their non-verbal behaviour becomes increasingly similar to the interviewee as the conversation progresses. Conversely, if an interviewer becomes consistently dissimilar or adjusts their behaviour irrespective of the interviewee, this would indicate divergent or maintaining behaviour (H. Giles et al., 1991). Pearson's correlation coefficient is calculated to measure convergence between. Correlation is based on turn-level proximity values and the count of interviewer speech turns at that point in the conversation. This calculation creates a single value per each interview (scores between -1 and 1, where higher positive values indicate greater convergence). An illustration of convergence is given in Figure 4.2.

Synchrony

Synchrony is a relative similarity measure between two speakers (Wynn & Borrie, n.d.). It is designed to capture the consistency with which two speakers maintain behavioural similarity. For example, if an interviewee frequently alters the pitch of their voice, and the interviewer responds by adjusting theirs, this would indicate higher levels of pitch synchrony. Low behavioural synchrony would occur if changes in the interviewer's behaviour bore little relationship to changes in the interviewee's. Like convergence, synchrony was also based on Pearson's correlation coefficient. It examines the correlation between the behaviour of the interviewer and interviewee at each speech turn. As such, synchrony generates a single value between -1 and 1

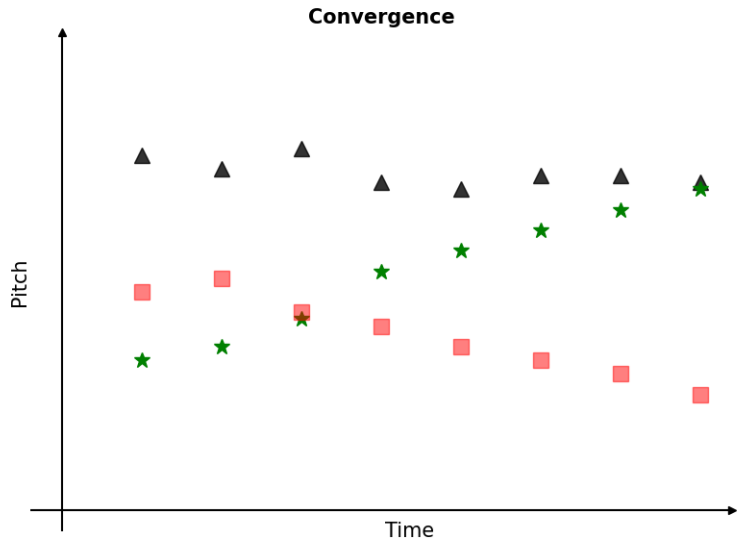

Figure 4.2: Hypothetical example of convergence (green) and divergence (red) relative to a baseline speaker (black). High levels of convergence indicate that the speaker is becoming increasingly similar to another speaker during the course of conversation. Conversely, low levels of convergence (red) are characterised by increased dissimilarity to another speaker.

for each behaviour, where higher positive values indicate greater speaker synchrony. See Figure 4.3 for an illustration.

4.5 Do interviewers mirror the non-verbal behaviour of the interviewee?

Before examining whether non-verbal mirroring by an interviewer contributes to an interview, it is worth first validating if the features described here are psychologically meaningful. Does pitch proximity, convergence of intensity, or synchrony of HNR describe anything meaningful about how the interviewer interacts with the interviewee? Non-verbal behaviours are subtle and can be hard to observe in real time. As such, low-level social signals such as non-verbal mirroring are not easily interpretable from a social scientific perspective. Therefore, its presence in a corpus can be difficult to distinguish from random noise. Assessing the validity of these behaviours is an essential prerequisite step before further analytic steps are performed (Lazer et al., 2014).

Consistent with prior research in this area (see Imel et al., 2014; C. C. Lee et al., 2014; R. Levitan & Hirschberg, 2011; Ward & Litman, 2007), a sensitivity check

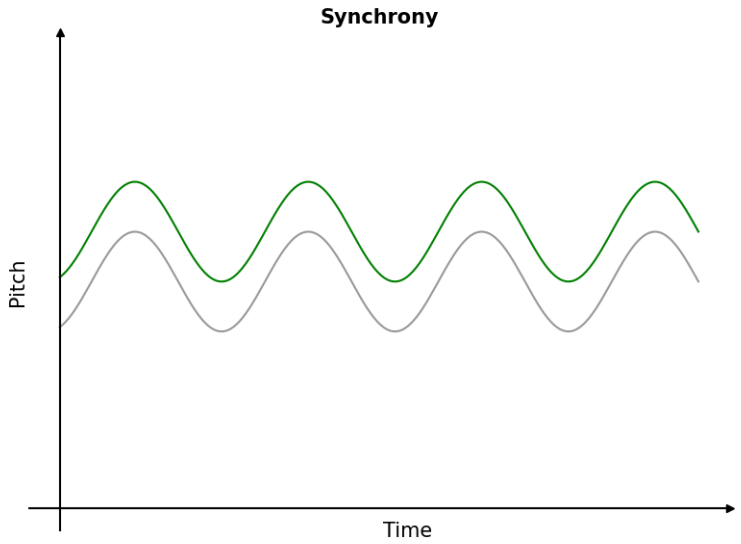

Figure 4.3: Hypothetical example of synchrony. High synchrony indicates the behaviour of the two speakers remain consistently similar throughout the conversation.

was performed by comparing non-verbal mirroring in real interviews with randomly generated dialogues. Theoretically, as interlocutors in the random dialogues did not converse, there should be minimal, if any, evidence of mirroring performed by the interviewer. By contrast, the values extracted from the actual interviews should be greater if the measures described above accurately detect behaviour changes by the interviewer in response to the interviewee.

Specifically, the following hypotheses were set for this section of the analysis:

- **H1 (proximity):** For each feature, there will be a larger negated mean absolute difference between speakers in the random dialogues compared to the real dialogues.

- **H2 (convergence):** For each feature, there will be a greater positive correlation between proximity and speech turns for the real dialogues versus the random dialogues.

- **H3: (synchrony)** For each feature, there will be a larger positive correlation between speakers for the real dialogues versus the random dialogues.

4.5.1 Dataset

Real interviews

The corpus of n=112 interviews that was described in Section 4.3 was used in this test. Features were generated for each interview as per Section 4.4.2.

Artificial Interviews

A corpus of random dialogues was generated by randomly reassigning each interviewer to a new interviewee with whom they did not converse. Excess speech turns by one speaker (caused by differences in interview length) was trimmed for each dialogue, ensuring that random dialogues maintained a sequential back-and-forth turn exchange pattern. To preserve any positional effects, the location of the turn within the interview was not randomised. That is, turns that appeared at the beginning of their actual interview maintained an approximate location in the random interviews. This step was performed due to an observation that interviewers tended to use a similar cadence to introduce and close an interview, particularly when addressing the camera. Randomly shuffling the order of speech turns would therefore exaggerate the differences between the real and random dialogues for reasons other than the rate of non-verbal mirroring and so were maintained. To ensure sufficient variation in the randomised corpus, each interviewer was randomly re-paired with a new interviewee ten times. The features generated by each interviewee were then averaged to create a single value per feature.

4.5.2 Analysis

A series of twenty-one paired t-tests were performed to compare the differences between the real and randomised interviews (once per each feature). Tests were performed in `R`. The normality assumption was confirmed by examining QQ plots, which were deemed satisfactory. Outliers were identified and inspected for potential errors. Tests were performed with and without outliers, with no meaningful difference in the overall interpretation of results. The results presented here *include* outlying data points.

4.5.3 Results

The outcome of the paired t-tests are reported in Table 4.2. A comparison of the mean values generated by the real and random dialogues indicates that, on the whole, the real interviews displayed more evidence of non-verbal mirroring compared to the random interviews. Increased similarity was observed in two-thirds of cases (as noted by a 'Y' in the 'Expected direction' column of Table 4.2). In most cases, the differences between the real and random interviews were too small for significant effects to be observed. Five behaviours were found to be statistically significant. *Proximity Intensity (Max.)*, and *Proximity Intensity (Mean)* were both found to be highly significant ($p < .001$ in both instances) with a moderate effect size. This suggests that the interviewer maintained proximity to the maximum and average vocal energy produced by the interviewee throughout the interaction. In other words, if the interviewee increased the volume of their voice, the interviewer was likely to reciprocate in the following utterance.

Less substantial but still significant effects ($p < .05$) were observed for both proximity and synchrony of pitch (mean) and synchrony of jitter. As a reminder, jitter is a measure of pitch instability. These findings indicate that interviewers match the average pitch of their interviewee on a turn-by-turn basis. Synchrony of jitter also suggests that the interviewer mirrors the pitch variation of the interviewee. Although, the application of a Bonferroni correction to account for potential Type 1 error will likely remove the significance of these effects due to the corrected p value ($p = .05 \div 21 = .002$). As such, it is necessary to be cautious when interpreting these findings literally. Future work exploring mirroring across a larger corpus would assist in this regard by providing enhanced statistical power.

Without a large significant effect, it is impossible to wholly reject the null hypothesis for either H1, H2, or H3. Partial support can be found for H1 (proximity). Three of the seven behaviours were statistically significant in the real versus random dialogues, and the real interview values for the remaining four behaviours were all at least equal to their random counterparts.

Minimal support was found for H2 (convergence). Correlations between proximity and turn count were slightly larger in real interviews for four of the seven behaviours, although differences between real and random interviews were slight. Curiously, convergence of jitter was considerably lower for real dialogues compared to random ones, with a t value indicating a potentially significant effect in the opposite direction. This would indicate that as the interviewee exhibited increased

jitter in their voice, the interviewer decreased their vocal jitter. Understanding the reasons behind this was deemed too time-consuming for the present thesis but would be worth exploring in future work.

Partial support was observed for H3 (synchrony), with five effects in the predicted direction (two of significance). Contrasting the pattern observed for H2, greater synchrony of jitter was observed in real interviews versus random ones. One interpretation of this contrast is that the interviewer matches the jitter of the interviewee early on in the interview but does not increase the level of mirroring over time. A similar pattern can be observed for synchrony of average pitch, where the interviewer fixes on an appropriate pitch distance, which is maintained throughout the interview.

Behaviour	Random Interview	Real Interview	t	df	Sig.	Exp. Direction
Convergence HNR	−.03	−.01	−0.45	111	$> .05$	Y
Convergence Jitter	−.04	−.08	−1.41	111	$> .05$	N
Convergence Intensity (Max.)	−.01	.06	−1.48	111	$> .05$	Y
Convergence Pitch (Max.)	.04	.04	.04	111	$> .05$	N
Convergence Intensity (Mean)	.0	.03	−0.75	111	$> .05$	Y
Convergence Pitch (Mean)	.04	.03	0.27	111	$> .05$	N
Convergence Shimmer	−.05	−.04	−0.48	111	$> .05$	Y
Proximity HNR	−3.7	−3.6	−0.92	111	$> .05$	Y
Proximity Jitter	−.02	−.02	−0.02	111	$> .05$	N
Proximity Intensity (Max.)**	−7.15	−4.84	−6.82	111	$< .001$	Y
Proximity Intensity (Mean.)**	−6.8	−4.37	−8.41	111	$< .001$	Y
Proximity Pitch (Max.)	−171.84	−164.28	−1.61	111	$> .05$	Y
Proximity Pitch (Mean)*	−73.09	−68.33	−1.73	111	$< .05$	Y
Proximity Shimmer	−.05	−.05	−0.19	111	$> .05$	N
Synchrony HNR	−.01	.04	−1.2	111	$> .05$	Y
Synchrony Jitter*	.01	.08	−1.75	111	$< .05$	Y
Synchrony Intensity (Max.)	.01	−.03	−0.97	111	$> .05$	N
Synchrony Intensity (Mean.)	−.01	−.02	0.37	111	$> .05$	N
Synchrony Pitch (Max.)	−.02	.04	−1.58	111	$> .05$	Y
Synchrony Pitch (Mean.)*	−.01	.06	−1.74	111	$< .05$	Y
Synchrony Shimmer	−.01	.01	−0.29	111	$> .05$	Y

Table 4.2: Scores and significance values of one-way paired t-tests. ** denotes a significance value $< .001$, * denotes $< .05$. p values are presented without Type 1 correction.

This section has used paired t-tests to explore whether interviewers engage in non-verbal mirroring towards their interviewee. Results are mixed, although there is some indication that interviewers are mirroring the interviewee on specific aspects of non-verbal behaviour. The measures utilised here reflect simple prosodic entrainment measures deployed in other settings (R. Levitan & Hirschberg, 2011). They do not represent the totality of methods in this space, however. As such, it would

be worth exploring whether more complex approaches yield more conclusive results (see C. C. Lee et al., 2014, for an example). In the next section of this chapter, the features described here are combined with supervised machine learning to examine individual differences in interviewer mirroring behaviour.

4.6 Detecting differences in gender-composition and partisanship

This section uses a combination of verbal and non-verbal mirroring behaviours to detect differences in gender composition and partisanship within the corpus. As noted at the beginning of this chapter, there are inconsistencies within the literature regarding whether intergroup differences should impact the degree and direction of mirroring an individual exhibits towards an interlocutor. A political interview is a novel domain to explore in this regard. As a public-facing form of discourse, the interviewer must balance the professionalism and neutrality expected of traditional journalism with building rapport and establishing a common ground with the interviewee.

To account for a large number of features when combining both verbal and non-verbal feature sets, a supervised classification task is constructed. Similar to the regression task described in the previous chapter, features based on interviewer mirroring are used as predictor variables. These are then used to classify whether a dyad comprises (a) a same-gender dyad and (b) a partisan dyad. This approach contrasts similar work that examines differences between groups by aggregating over all behaviours (i.e., *t*-tests).

4.6.1 Experimental set-up

Input features

Features of non-verbal mirroring were calculated for each interview in the same manner as those described in Section 4.4.2. In addition, several verbal mirroring behaviours from the previous chapter were included to examine whether performance improved when multiple modalities were included. Micro-level models were not included to minimise feature dimensionality and prevent non-verbal behaviours from

being washed out in a multi-modal model. Macro measures of verbal mirroring were included at the stylistic, syntactic, and semantic levels.

Target variables

Two binary target variables were constructed for this task based on: (i) gender composition and (ii) political partisanship. Table 4.3 includes summary statistics for each outcome. The dataset was broadly balanced for partisanship, with 49 (44%) partisan interviews and 63 (56%) non-partisan interviews. The dataset was imbalanced concerning gender composition, with almost two-thirds of the dataset (76) comprised interviews where speakers shared the same gender. Examining the same-gender interviews further revealed that 92% of these interviews featured male-male dyads. This reflects the comparatively few females that appeared in the dataset as either interviewers or interviewees. Indeed, only 30% of the interview participants in this corpus (interviewers and interviewees) were female.

Target	Outcome	Count of interviews in corpus (n=112)	% of corpus
Gender Composition	Same Gender	76	68%
	Mixed Gender	36	32%
Partisanship	Partisan	49	44%
	Non-Partisan	63	56%

Table 4.3: Summary of target variables.

Supervised machine learning

The problem was set as a supervised classification task to explore potential differences per each target variable. This structure departs from other studies in this area that use inferential statistics such as t-tests to compare differences between classes (Weise et al., 2019). The motivation for using machine learning here is two-fold. Firstly, based on the previous results, which used inferential statistics, differences between classes are likely to be subtle. Given the comparatively small number of data points (112) relative to a high-dimensional feature set, inferential tests are likely to be underpowered. Conversely, supervised machine learning, such as a random forest, help avoid this by splitting and combining random subsets of features to classify differences in the data (Breiman, 2001). Using a cross-validation approach

will also help avoid model over-fitting. The second motivation behind using a machine learning approach considers the complexity of mirroring as a social signal. A t-test assumes a linear relationship between the explanatory and target variable, the simplicity of which is unlikely to reflect differences in turn-taking behaviour adequately.

To remain consistent with the previous chapter, a forty-tree random forest was used to classify each target variable. The 21 non-verbal mirroring behaviours described in Section 4.4.2 were used as input features. Aside from the number of estimators ($n\_estimators$), all other hyperparameters were kept at default settings as specified in `scikit-learn`. To minimise potential over-fitting, models were cross-validated with k-fold cross-validation, where k=10. Models were evaluated by averaging performance over all folds.

4.6.2 Classification performance

Performance was evaluated for each target variable under three conditions: (i) using non-verbal features only, (ii) using verbal features only, and (iii) using a combination of verbal and non-verbal features. Models are evaluated using the F1 score to account for imbalanced classes. The positive label for gender was the classification of *same gender*, and the positive label for partisanship was *partisan*. If mirroring is impacted by either the gender composition or political affiliation of speakers, we can expect a machine-learning model to improve classification accuracy versus the baseline. However, if interviewers do not adjust their behaviour relative to these factors, we would not expect to exceed the baseline. Performance was compared to a simple estimator that either consistently predicted the positive label or each label with a 50:50 probability (whichever baseline was higher).

Gender Composition

Compared to the performance obtained via a majority-class baseline (68%), a performance improvement was observed for all three conditions, although only one differed from the baseline by a significant degree. Specifically, average performance over all folds using only non-verbal mirroring was $F1 = 0.79$ ($SD = 0.05$) – significantly above baseline as determined by t-test. Classification performance using only the verbal set was marginally over the baseline (although differences were not statistically significant from baseline) and was measured at $F1 = 0.73$ ($SD = 0.08$). A

combined model comprising both sets of features performed considerably better than the verbal model, marginally reducing the variation in scores versus the non-verbal model with a score of $F1 = 0.79$ ($SD = 0.04$). These results indicate that whilst verbal and non-verbal mirroring exceeded a random 50% baseline, performance only improved substantially when non-verbal features were included in the model.

Several alterations were then made to increase the classification accuracy further. First, classifications were re-run using Extreme Gradient Boosting (XGBoost) instead of a random forest. Hyperparameters were kept constant using the default values. Performance was broadly consistent with random forest for each feature set, with a maximum difference of $\pm 2\%$ observed across all models. These slight deviations did not alter the interpretations of the main findings and likely reflected noise due to the inherent randomness of machine learning algorithms.

In a separate attempt to improve model performance, micro-level linguistic features from Chapter 3 were also included in the combined model. This substantially increased the dimensionality of the data, as each word uttered by the interviewer (a set of 3978 n-grams) was considered an independent feature. To mitigate against potential over-fitting, micro models were restricted to 500 features[3] via Fisher scores (Li et al., 2017). Fisher scores is a filter-based feature selection method that ranks features according to their relationship to the outcome variable and returns only those features above a particular ranking position. Fisher scores were calculated within each fold rather than across the entire dataset. This step was to avoid leaking information between the train and test regions. The combined model was then re-run using a default random forest with the same cross-validation strategy. The results of this model improved slightly on the best-performing model, as well as maintaining a similar level of variation across folds $F1 = 0.8$ ($SD = 0.04$). Figure 4.4 illustrates model performance over each condition.

Feature Importance The relative importance of each behaviour was also calculated for the non-verbal model. As the analysis in this chapter has predominantly used random forest (unlike the previous chapter, which used four algorithms), the Gini Impurity method via `scikit-learn` is used to measure the importance of each non-verbal feature. The most important features and their average impurity decrease scores are reported in Figure 4.5.

Amongst the non-verbal features, average pitch proximity, even after normali-

[3]Other values that were explored were on a range between 10 and 500 features.

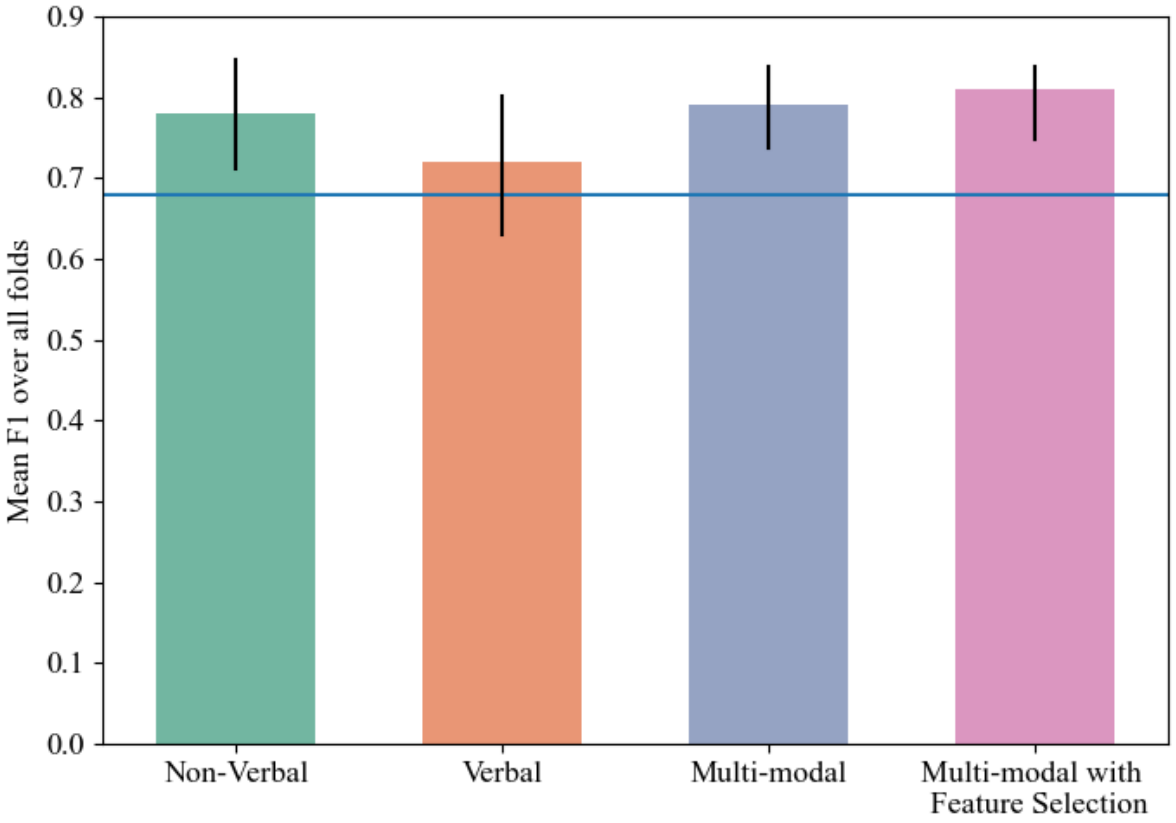

Figure 4.4: Model performance ($F1$) for random forest classifying same gender dyads based on interviewer mirroring. Bars represent the average score over 10-fold cross validation using non-verbal, verbal, multi-modal, and multi-modal with feature selection. Error bars represent 95% confidence intervals. Baseline (majority class estimator) is illustrated via the red dashed line. Bars above this value denote an improvement over the baseline.

sation, emerged as the most significant feature in the analysis. This observation aligns with expectations, as women generally have higher-pitched voices than men. Consequently, the pitch proximity for same-gender dyads is greater than that for mixed-gender dyads. Although it is possible to adjust one's pitch to match that of another speaker, the extent that this feature reflects actual non-verbal mirroring requires further investigation. An effort was taken to minimise this effect by normalising raw values by gender, as per previous studies (R. Levitan et al., 2012). To increase confidence in the model, the initial analysis was re-run without pitch proximity as an input feature. On average, results marginally degraded (-1%) and increased the standard deviation, but ultimately did not require re-interpreting the main findings. This is further supported by including pitch convergence and synchrony in the most important features. As both are temporal measures of mirroring, they indicate that the degree of change in pitch relative to the interviewee differed based on gender composition.

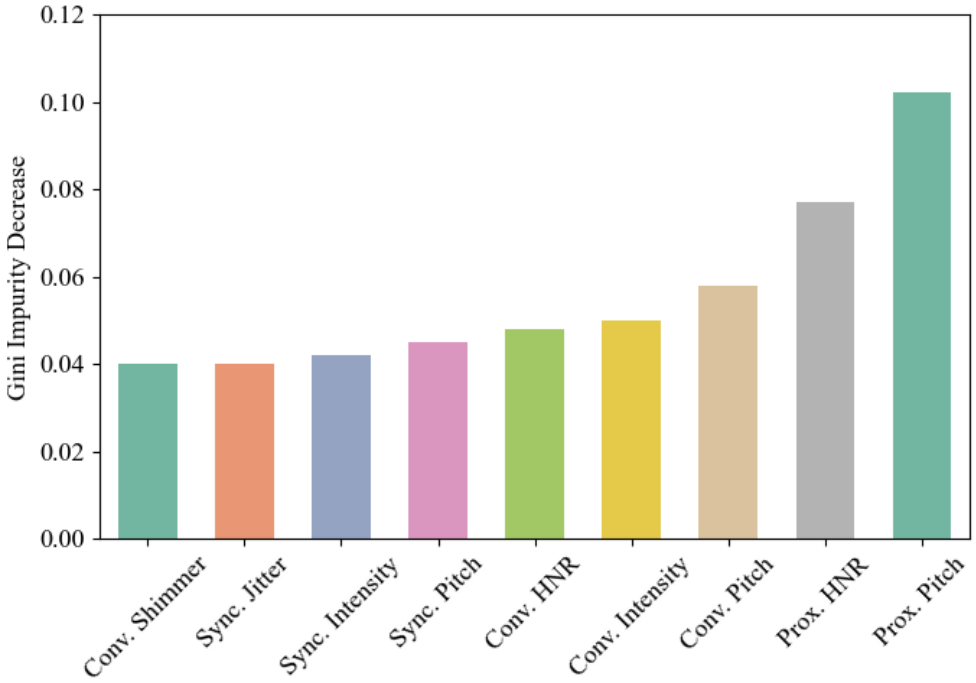

Figure 4.5: Top important features in the non-verbal model. Importance is based on average Gini Impurity over all cross-validated folds. Features with larger bars are more important.

Partisanship

The study was then re-created to detect differences in political partisanship. Given that interviews where speakers shared a common political orientation were a slight minority within the data (44%), a higher 50% random baseline was set. Performance exceeding this value can be interpreted as evidence of differences in interviewer mirroring based on whether the interviewee adopted a similar political ideology.

As before, classification was evaluated via F1 for the non-verbal, verbal, and combined feature sets. Using a random forest, the results of non-verbal mirroring were $F1 = 0.43$ ($SD = 0.26$). This indicates that interviewers did not appear to adjust their rates of non-verbal mirroring based on whether they were speaking to a member of an affiliated political party. The verbal feature set generated a worse score of $F1 = 0.39$ ($SD = 0.24$). The combined model performed better than both unimodal models, with a score of $F1 \approx 0.5$ ($SD = 0.21$), which fractionally exceeded the baseline, but not to a meaningful degree. The poor performance when classifying partisanship is further reflected in the large standard deviations obtained, indicating a sizable difference in $F1$ amongst the folds for each model. These results fail to exceed the 50% baseline, indicating that political partisanship

may not impact the degree to which an interviewer mirrors their interviewee. It may also be that the subtlety of mirroring is not adequately reflected in the measures adopted in this chapter. For completion, several steps were taken in an attempt to increase performance. First, as before, the analysis was re-run using XGBoost and with micro-features plus feature selection. XGBoost achieved a similar level of performance, falling within 3% of those obtained via a random forest. Each of these failed to exceed the baseline by any significant extent. The average $F1$ score via feature selection did exceed the baseline ($F1 = 0.54$, $SD = 0.17$), although due to the considerable variation in scores, this difference was not statistically significant from a 50% baseline.

Lastly, the impact of an alternative cross-validation strategy was examined. Leave-pair-Out Cross-Validation (LpOCV) is a technique whereby the model is iteratively trained on *n-2* rows, and tested on the 2-row holdout set. Because LpOCV uses almost all of the available data in training, it produces a model that is less impacted by bias between data splits. Compared to other techniques such as K-fold or Leave-One-Out cross-validation, which ensures each data-point features only once in the test set, LpOCV pairs each data point with every other data point in the test set. This step ensures that many training iterations can be performed from a comparatively small dataset. Specifically, the 112-row dataset produces 6216 splits, substantially larger than the ten folds used in the previous experiments. A potential advantage of this is that it may enable the detection of more subtle differences between classes. However, applying LpOCV to the best-performing model, in this case, did not exceed the original performance ($F1 < 0.5$). Moreover, the additional splits significantly inflated the run time from approximately 10 seconds to over three minutes.

Collectively, these results reflect mixed evidence as to the impact of contextual factors on verbal and non-verbal mirroring by an interviewer. Evidence was observed for a difference based on gender composition. This is consistent with prior research indicating that the degree of mirroring alters based on the gender of a conversational partner (Bilous & Krauss, 1988; R. Levitan et al., 2012). Results fail to demonstrate any observable impact of political partisanship on the mirroring behaviour of the interview. One possible explanation is that the interviewers, as professional journalists, exercised neutrality toward their interviewees. However, mirroring is generally considered beyond conscious awareness, so we may still expect to see subtle differences based on political orientation. This might indicate a potential flaw in the simplicity of the measures used. Whilst these results do not

necessarily prove a negative, they are at least congruent with existing research highlighting the difficulty in detecting such differences (Weise & Levitan, 2018; Weise et al., 2019). The significance of these findings is explored in more detail in the Discussion section of this chapter.

4.7 Does a multi-modal feature-set improve prediction of interview outcomes?

The final section in this chapter returns to the task of predicting the outcome of an interview from the behaviour of the interviewer. This task is an extension of the core task explored in the previous chapter but includes the additional non-verbal mirroring features introduced in this chapter.

Both non-verbal and verbal mirroring has been shown to contribute to task success in dyads (Friedberg et al., 2012; C.-C. Lee et al., 2010; R. Levitan & Hirschberg, 2011; Nenkova et al., 2008; Rahimi et al., 2017; Reitter & Moore, 2014), and promote a prosocial response (Kulesza et al., 2014; Lubold & Pon-Barry, 2014; van Baaren et al., 2003). As a reminder, the majority of work in this area is linked to theories of Communication Accommodation (CAT: H. Giles et al., 1991), and rapport-building (Tickle-Degnen & Rosenthal, 1990). Despite exploring these topics across various settings, however, the extent that such prosociality exists in the context of a political interview has yet to be explored. Building on the findings of the previous chapter, which provided evidence that the degree of verbal mirroring by an interviewer was related to conversational quality, this section explores this relationship when: (a) mirroring is based on non-verbal behaviour and (b) both feature sets are combined within a multi-modal model.

4.7.1 Input features

The non-verbal feature set used throughout this chapter is also used here. Two conditions are examined: (a) a unimodal model based on non-verbal features only and (b) a multi-modal model that combines the non-verbal features with the micro and macro-level verbal mirroring features.

4.7.2 Outcome measures

As in Chapter 3, features derived from the interviewer mirroring the interviewee were used to predict each of the five outcome measures. These are *specificity, clarity, diversity, relevance*, and an *overall quality* score based on a weighted average.

4.7.3 Experimental set-up

Consistent with the previous section, this task uses a forty-tree random forest with default hyperparameters. As the data used in this chapter is a subset of the whole corpus introduced in Chapter 3, a LpOCV cross-validation strategy was used. As before, models are evaluated via the RMSE metric and compared to a simple baseline model that repeats the mean outcome variable within the training set (Reitter & Moore, 2014). As RMSE is an error metric, models with a lower score should be interpreted as performing better.

4.7.4 Prediction performance

For each outcome, the average performance (RMSE) over every cross-validated fold is reported in Table 4.4. For comparison, the analysis of verbal mirroring is recalculated on this subset of data.

Table 4.4: Average performance of Non-Verbal, Verbal and Multi-Modal models per outcome indicator. Values reflect the mean RMSE over all cross-validated folds - **a lower score is better**. Values in parenthesis reflect the percentage improvement over the baseline.

Outcome	Non-Verbal	Verbal	Multi-Modal
Specificity	0.19	0.18 (4%)*	0.18 (3%)*
Clarity	0.17 (1%)$^{\triangle}$	0.175	0.174
Diversity	0.15 (8%)*	0.08 (47%)*	0.08 (47%)*
Relevance	0.18 (5%)*	0.16 (18%)*	0.16 (18%)*
Overall	0.15 (15%)*	0.11 (37%)*	0.11 (37%)*

1 Statistical significance between baseline and model is determined via paired t-test.
* indicates differences are significant to $p < .001$
$^{\triangle}$ indicates a non-significant improvement over the baseline.

Non-Verbal Model

An improvement in performance (as measured by a reduction in RMSE) versus the baseline was observed for all outcomes except specificity. This finding is identical to the previous chapter, providing further evidence that specificity, as measured here, may not be a suitable outcome measure for a political interview.

As RMSE values are inherently difficult to interpret, the percentage improvement over the baseline is given in parenthesis. This is a more readily-accessible metric, as it illustrates how much better the model is compared to the baseline. Accordingly, prediction performance was greatest for predicting the overall success of the interview, reducing RMSE by 15%. A statistically significant improvement (as measured by *t*-tests) was also observed for predicting the linguistic diversity of an interviewee's response (8%) and the relevance of the answer (5%). Clarity also exceeded a baseline estimator, although the average improvement over all cross-validated folds was not significant.

Verbal Model

In contrast to the previous chapter, a significant improvement *was* observed for predicting specificity via a verbal model. Verbal features exceeded the baseline RMSE by 4%. A possible explanation for these contrasting findings is the difference in cross-validation strategy. By increasing the number of folds during cross-validation, a more stable level of performance is obtained. In other words, it is possible that training over more folds detected a subtle effect that was not evident when only ten folds were used. Another contrast with the previous chapter is the lack of an improvement for predicting clarity. An alternative explanation for these findings is that the sub-corpus reflects a sampling discrepancy. This is explored further in the discussion section of this chapter.

Consistent with the previous chapter, a significant improvement was observed in predicting the linguistic diversity (47% reduction in RMSE) and relevance (18% reduction) of the interviewees' speech. A significant reduction was also observed in predicting the overall quality of the interview (37%). Interestingly, the percentage improvement is greater for verbal mirroring than for non-verbal mirroring. This would indicate that measuring the linguistic content of speech is more closely aligned to success compared to the vocal quality, although both are impactful behaviours to exhibit.

Multi-Modal Model

Combining non-verbal and verbal modalities into a single model did not improve the performance of the verbal model alone. Significant differences were observed for predicting specificity (3%), diversity (47%), relevance (18%), and overall quality (37%). The large feature imbalance between the two modalities may have resulted in a 'washing out' of the non-verbal features (Murray & Oertel, 2018). Restricting the number of available verbal features via Fisher scores was attempted but did not substantially improve performance. Instead, dropping the number of available verbal features in line with the number of non-verbal models decreased performance. Further experimentation with alternative feature selection algorithms was considered but deemed beyond the scope of the thesis.

4.8 Discussion

The focus of this chapter has been to extend the work covered in Chapter 3 by considering a non-verbal modality when predicting the quality of a political interview. Video footage of a sub-set (n=112) of the original corpus was collected, and a series of non-verbal mirroring features were generated from adjacent interviewee→interviewer turn exchanges. This chapter's results support the notion that interviewers engage in non-verbal mirroring towards the interviewee. However, this only appeared for certain behaviours. This was explored via a common data augmentation approach that compares mirroring in real versus artificial dialogue. In a second set of experiments, the ability of the non-verbal feature set to detect differences in (a) gender composition and (b) political partisanship was explored. These tests were motivated by prior experimental and computational work that has indicated speakers mirror differently based on contextual factors. Findings from this section indicate that non-verbal mirroring does appear to differ based on whether the interviewer is engaging with a member of the same or opposite gender. No such evidence was observed for partisanship, indicating that interviewers remain neutral in their behaviour irrespective of the political beliefs of the interviewee. The final experiment in this chapter explored whether a multi-modal feature set combining measures of both verbal and non-verbal mirroring by the interviewer predicted conversational quality. A broad interpretation of results suggested that verbal and non-verbal mirroring contributed to conversational quality; however, combining features in a multi-modal model did not increase performance over a unimodal feature set.

Summary of Findings

Section 4.5 indicated that interviewers, on average, tended to be more acoustically similar to the prior utterance of the interviewee compared to a random baseline. Greater similarity was observed in 66% (14 features) of the non-verbal features explored. However, only five of these were significantly different from chance. There are several potential interpretations of this. First, the methods used to measure non-verbal mirroring are overly broad. Studies have observed mixed effects that have utilised proximity, synchrony, and convergence as features since being introduced in R. Levitan and Hirschberg (2011). For example, the same feature set used here was used to successfully detect differences in deceptive and non-deceptive speech (S. I. Levitan et al., 2018). In Litman et al. (2016), the authors used a subset of these features to detect differences in within-group and between-group interactions. Moreover, Michalsky et al. (2018) found significant effects for attractiveness and conversational quality based on synchrony and proximity between romantic partners. Conversely, other studies have failed to identify any coherent structure using the same feature-set (Weise & Levitan, 2018; Weise et al., 2019). It is perhaps noteworthy that other studies that have explored non-verbal mirroring have often included additional acoustic features alongside those explored here (Willi et al., 2018). Other studies have used an entirely different approach, with success, based on features of mirroring derived from mel-frequency cepstral coefficients (MFCCs) (C. C. Lee et al., 2014; C. C. Lee et al., 2011; Murray & Oertel, 2018). MFCCs were initially considered for this work but were ultimately deemed beyond the current scope of the thesis.

Section 4.6 found evidence indicating that interviewers behaved differently based on whether they interacted with a member of the same or opposite gender. This finding broadly agrees with prior research suggesting that gender composition alters the extent speakers mirror one another in conversation (Bilous & Krauss, 1988; R. Levitan et al., 2012; Pardo, 2006; Street, 1984), although it contrasts other research in this area (Weise et al., 2019). Whilst not explored here, an area of future interest might be to contrast differences between the genders themselves. Prior research has suggested potential differences between, for instance, male-male and female-female dyads that may reflect social characteristics and status (Bilous & Krauss, 1988; R. Levitan et al., 2012). Since females were highly under-represented in the current corpus, exploring these distinctions further would be interesting.

Neither a model of non-verbal nor verbal behaviour detected differences in par-

tisanship. This contrasts initial assumptions made by the author, based on the link between non-verbal mirroring and rapport (Lubold & Pon-Barry, 2014). As before, it is possible that the feature set used was insufficient in detecting subtle differences between partisan and non-partisan interviews. It is, however, equally plausible that the interviewers were behaving the same based irrespective of the political orientation of the interviewee. As professional journalists, interviewers are expected to uphold a position of neutrality when conducting a televised interview (Heritage & Clayman, 2010). However, as noted in Hutchby (2011), a shift towards non-neutrality is becoming increasingly common, particularly in cable networks such as those explored here. Ultimately, the question of the interviewer's behaviour on issues of partisanship remains unresolved.

Lastly, in Section 4.7, non-verbal and multi-modal models of interviewer mirroring were used to predict the conversational quality of the interview. Results broadly support the view that mirroring contributes to a more successful interaction, although some findings contradict the results covered in the previous chapter. Specifically, the present chapter observed a significant performance improvement in predicting specificity. As highlighted in the results section, this may result from increased data splits performed during cross-validation. An alternative explanation may be due to a sampling discrepancy. Interviews within this sub-corpus were opportunistically sampled based on the availability of a suitable audio recording. The resulting subset may have introduced a hidden bias within the data that is not sufficiently accounted for.

Limitations

Several potential limitations within this approach are worth highlighting. One such weakness is whether validating the purity of speech turns via human effort justified the time it took to remove overlapping and incorrect speech from the audio. Even with a semi-automated approach, cleaning the data was incredibly time-consuming. Despite the participation of three research assistants, speech segmentation took more than twice the length of the entire sub-corpus ($\approx$ 60 hours in total). The time-consuming nature of this step would almost certainly be a barrier to exploring larger datasets, should they become available in future.

Manually preparing audio has been cited as time-consuming in similar studies (Imel et al., 2014). In this chapter, the RAs cited three main contributing factors in the prolonged pre-processing effort: (i) a general lack of familiarity with the

software (PRAAT), as well as audio processing more broadly, (ii) the intricacy of the data cleaning process, and (iii) interviews characterised by large amounts of speaker interruption. Efforts were made to address the first issue by training RAs in PRAAT before data cleaning occurred. The second and third issues, however, were much more challenging. An alternative approach to semi-mechanical diarization might have been to eliminate the data cleaning step entirely and instead rely solely on the turn boundaries generated from the automated output from IBM Watson. This decision was ultimately not taken as the author felt that the dataset was sufficiently small to justify a semi-automated approach and that performance would benefit from the increased purity of speech turns. However, the findings from this study could also be used in comparison with a wholly automated approach. Knowledge from such an effort would indicate whether human intervention in generating pure speech turns is beneficial in the current task.

A second limitation concerns the decision to focus on proximity, convergence, and synchrony as measures of non-verbal mirroring. As discussed in this chapter, prior research has utilised these methods with mixed success. For the most part, successful implementation appears to have occurred in studies under high levels of experimental control (R. Levitan & Hirschberg, 2011; Litman et al., 2016; Michalsky et al., 2018). The benefits of the feature set include their simplicity, which makes them simple to deploy and interpret. They are also relatively simple to explain to a non-expert audience. However, more complex methods based on MFCCs have arguably been more successfully applied in naturalistic settings (see C. C. Lee et al., 2014; C. C. Lee et al., 2011; Murray & Oertel, 2018, for examples of where such an approach has been used successfully). These studies have utilised the entirety of a speech turn rather than the segments at turn boundaries. Therefore, they may encapsulate a more appropriate representation of vocal behaviour.

4.8.1 Publication record

As with the previous chapter, the work included here is the subject of two co-authored manuscripts. A submission was initially made to the 24th ACM International Conference on Multi-modal Interaction. This submission was ultimately not accepted. However, reviewer feedback was valuable and helped shape both the current chapter and a revised manuscript currently in preparation (Cook et al., n.d.). As a joint submission with several co-authors, the distribution of analysis and written work was identical to the previous chapter. For both the conference submission

and the ongoing manuscript, the author of this thesis was responsible for collecting, preparing, and analysing the data. The three co-authors provided valuable feedback and support at key moments.

4.8.2 Chapter conclusions

The primary objective of this thesis phase has been to explore the utility of social signals as an alternative to traditional manual annotation. To this end, the results presented have highlighted how an automated approach can contribute to understanding interviewer behaviour and its effects on the interviewee. Automated techniques from computational fields such as text-mining and signal processing can scale the analysis of conversational data beyond that achievable through manual annotation alone. However, the current and previous chapters have highlighted the significance and difficulties of algorithmically modelling complex social phenomena. Interpersonal mirroring has been chosen as a behaviour of interest due to its theoretical backing as an effective conversational technique. It has also been subject to many algorithmic approaches that produce conflicting results. Chapter 2 sought to address this by aggregating a series of standard text-based algorithms into a single Python package. This was then explored in Chapter 3 and extended into the non-verbal domain in Chapter 4. However, the extent to which this approach can be applied to all behaviours seems unlikely. Social psychologists are often interested in latent behaviours such as rapport-building and empathy that evade an easily identifiable algorithmic interpretation.

An alternative framework for automating the annotation and analysis of conversation is based on supervised machine learning. Although this requires training a classifier on a small corpus of hand-labelled instances, once the machine has built a sufficient representation of the behaviour of interest, it can then be used to label a much larger corpus. The next phase of this thesis examines the suitability of this approach in a new conversational domain.

Chapter 5

Detecting Predatory Behaviour in Online Chats with Natural Language Inference

5.1 Introduction

Until this point in the thesis, there has been an emphasis on modelling low-level social signals directly from text and audio as an alternative to the hand-labelling activities routinely performed by social scientists. This approach is advantageous in many respects, as it avoids making the human annotator solely responsible for detecting behaviours of interest. Such an approach has the potential to aid both the scalability and consistency of annotations as it is less impacted by issues associated with annotator fatigue. However, avoiding human annotation can be problematic in cases where the complexity of a behaviour of interest makes it hard to define algorithmically. So far, this thesis has focused on behavioural mirroring, a well-understood social phenomenon for which various approaches have been applied in experimental and naturalistic settings. Mirroring is, at its core, a similarity task. Because of this, it is easy to imagine how comparing speakers to some pre-defined rules lends itself well to a computational solution. Social scientists, particularly psychologists, are often concerned with more abstract patterns of behaviour. Examples such as rapport-building (Alison et al., 2013) and empathy (Lord et al., 2015) are often

difficult to pin to a single definition, even amongst experts, and as a consequence, require a Gestalt 'felt sense' to identify in practice.

The last few years have seen deep learning technologies such as neural networks being increasingly applied to this task, with the idea of hand-labelling a small number of instances that a model can learn from. This approach minimises the need to define the behaviour of interest through explicit feature engineering. A deep learning methodology also affords considerable flexibility, as neural networks often utilise large pre-trained models trained on general-purpose corpora to acquire near-human levels of language comprehension. This is illustrated in the wide-ranging application of deep learning to different behaviours of interest. For instance, Ando et al. (2017) used a combination of approaches known as Long short-term memory (LSTM) (Hochreiter & Schmidhuber, 1997) and sequence-2-sequence (Seq2Seq) (Sutskever et al., 2014) models to automatically label a chat corpus with naturally-occurring conversational behaviours such as gratitude, agreement, questions, and greetings. In another example, J. Gibson et al. (2016) used an LSTM model to detect behaviours related to motivational interviewing in addiction therapy sessions. And lastly, Timoshenko and Hauser (2019) used a convolutional neural network approach to examine customer needs from a large corpus of online product reviews.

One area where such an approach could prove helpful is detecting online predatory behaviour. Internet predation is a growing problem on an international scale and one for which existing technologies are ill-equipped to address. Automatically detecting those behaviours that are most likely to lead to severe consequences for victims has significant ramifications for law enforcement. Accordingly, the current chapter explores the utility of a deep learning approach to label predatory behaviours in a corpus of known cases of online Child Sexual Exploitation (CSE). The aims of this chapter are as follows:

- Present the pertinent literature on common themes of online CSE from a psychological perspective.
- Describe an automated approach to online predatory detection based on Natural Language Inference (NLI).
- In an initial study, examine performance of an NLI model on classifying predatory behaviour when an increasing amount of training data is provided.
- Apply behaviours that demonstrate the best performance to a large corpus of unlabelled predatory chat-logs.

The approach covered in this chapter is novel in its use of deep learning methods to annotate a corpus of predatory conversations with domain-specific behaviours. The main findings indicate that a deep learning approach can be used for this purpose, with the best-performing behaviours achieving a performance consistent with expert annotation. However, considerable scope exists for improvement in some behaviours, particularly those that appear rarely in the corpus.

5.2 Theoretical background and related work

The growing popularity of social media amongst children and young adults raises serious concerns for their safety. The threat of online sexual grooming is an increasing problem in the digital age (Greene-Colozzi et al., 2020). In 2021 alone, UK police forces recorded over 5000 offences relating to sexual communications with a child, with social media apps such as Instagram and Snapchat popular amongst online predators (NSPCC, 2021). Minors who fall victim to offenders suffer considerably, with many abusers progressing from online to physical contact offline (Shelton et al., 2016). To assist law enforcement, academics have been trying to identify in advance those predators that steer the relationship towards physical encounters (Briggs et al., 2011; O'Connell, 2003; R. Williams et al., 2013; Winters et al., 2017).

Automatically detecting predatory behaviour, however, is challenging. Offenders use a variety of subtle and often idiosyncratic behaviours to manipulate the flow of the conversation. For instance, predators may use flattery to build trust (Barber & Bettez, 2021), make threats or bribe a potential victim as a coercion tactic (Joleby et al., 2021). Whilst human annotators, particularly experts, can identify these contextual psychological behaviours from chat logs, the implicitness of behaviours can be problematic for machines (Buckingham & Alali, 2020). As discussed earlier in this thesis, dictionary-based approaches (Tausczik & Pennebaker, 2010) have been the long-standing option used by computationally-focused social scientists to identify psychological characteristics such as neuroticism (Bogdanova et al., 2014). However, these methods heavily rely on the included vocabulary, causing a large number of false positives for some behaviours (Kaur et al., 2021) while overlooking others (Broome et al., 2020).

Prior work utilising machine learning has focused on identifying predators from a mixed corpus of illicit and everyday conversations (Ebrahimi et al., 2016; Gunawan et al., 2018; Inches & Crestani, 2012; Miah et al., 2011). This is often achieved by

creating a single corpus comprising predatory adult-child interactions, such as those in this chapter, with other forms of adult-adult dialogue. For instance, in Miah et al. (2011), the authors use adult-based dialogue where users are randomly paired with each other as a negative class. A similar approach was also used in Inches and Crestani (2012) PAN12 challenge, a dataset that has since become a staple in many studies seeking to automate the detection of online predatory behaviour (Escalante et al., 2013; Fauzi & Bours, 2020; Ngejane et al., 2021; Tomljanovic et al., 2016). While valuable in its own right, this line of research does not offer significant value to domain experts as it identifies grooming behaviour after-the-fact (Razi et al., 2021). It closes the stable door after the horse has bolted. As such, these studies do not substantially contribute to psychological insight regarding grooming or address real-world issues pertinent to online child safety.

A central issue with this line of research is that it tends to ignore the subtlety and sophistication of many online grooming techniques. Broadly, a classifier trained to detect conversational differences will tend to leverage those characteristics that appear disproportionately in one class over the other. At a surface level, it would appear reasonable to suggest that hyper-sexualised language might be an important distinction when classifying predatory and non-predatory dialogue. Indeed, conversations between the author of this thesis and practitioners indicate that hyper-sexualised keywords are often used to scan a large number of online interactions. However, this distinction is a gross oversimplification of the varied ways predators seek to lure vulnerable children online. van Gijn-Grosvenor and Lamb (2021) noted three other types of predators besides the hyper-sexualised variant. Their four-fold typology also included an intimacy-seeking type, a social type, and an opportunistic type. Predators not seeking immediate sexual gratification may withhold sexually explicit talk until the interaction has progressed to a level of familiarity. Indeed, Gottschalk and Hamerton (2022) note that some offenders sought to establish a relationship rather than an immediate sexual thrill.

To address this shortcoming, automated methods utilising aspects of machine learning are increasingly being used to detect behaviours that domain experts regard as problematic. This involves a deeper level of exploration than the work described above and instead seeks to tag clusters of messages. This can be useful in an applied context, highlighting particularly dangerous parts of the conversation. For instance, Gupta et al. (2012) used psycholinguistic features to detect differences in predatory behaviour throughout the conversation – linking LIWC categories to six phases of the conversation: *friendship forming, relationship forming, exclusivity,*

risk assessment, sexual and *conclusion*. Similarly, in Gunawan et al. (2018), the authors used supervised machine learning to pair each phase with more specific behaviours such as asking for a picture, talking about friends, discussing hobbies, and building mutual trust. Moreover, studies have also used a SSP approach, using many of the base-level features introduced earlier in the thesis, to detect behaviours related to different phases of the conversation (Cano et al., 2014). Other studies have used a combination of machine learning and dictionaries to detect qualitative differences in linguistic behaviour between the messages produced by predators and those generated by victims (Drouin et al., 2017) or quantify the level of predatory behaviour (Pendar, 2007).

Many of these studies indicate promising results with their chosen method, with an $F1$ score in the region of .75 not uncommon for what might be considered abstract behaviours such as building trust or assessing risk. However, little attention has been paid to evaluating the suitability of such methods for large-scale analysis. This chapter addresses this gap by training a machine learning algorithm with annotated chats before applying it to a larger corpus. Comparatively, the amount of manual annotation is noticeably smaller ($N = 24$ chat-logs) than related studies in this area – Gupta et al. (2012) and Gunawan et al. (2018) manually annotated 75 and 150 chats, respectively – a potential weakness in a methodological sense, but a factor that significantly aids scalability and application in the real world. Therefore, this chapter aims to identify the performance of a machine-learning classifier when a minimal amount of human annotation is provided.

5.3 Perverted Justice: a corpus of online CSE chat-logs

This chapter utilises a large corpus of online predatory conversations archived by a child-safety watchdog organisation known as The Perverted Justice Foundation[1]. These conversations occur in chat-based messaging forums such as MySpace and Yahoo Messenger, documenting the interaction between sexual predators and volunteers posing as children and teens. The outcome of these conversations often resulted in the predator and victim arranging to meet offline and forming part of an undercover police sting. According to the administrators of the Perverted Justice website, between 2003 and 2016, the activities of the undercover volunteers resulted

[1] http://perverted-justice.com/

in criminal convictions of over 600 predators in the US. As of 2019, Perverted Justice ceased all online operations but continues to make their archives publicly available.

Both the creation and usage of this data are controversial. The Perverted Justice model has faced accusations that using untrained members of the public as decoys encourage cyber-vigilantism (Zingerle, 2015). Moreover, the fact that offline meetings were routinely televised as part of NBC's *To Catch a Predator* series has resulted in claims of unnecessary humiliation towards individuals who, at that point, had been neither charged nor convicted of a crime (Adler, 2011). There is also some debate around whether the persistence of some volunteers constituted legal or moral entrapment (Butters et al., 2014; Egan et al., 2011).

However, it is equally valid that the predators featured in the chat logs were ultimately convicted of a grooming offence relating to a minor. Victims did not initiate contact with the predator and did not introduce sexual content but instead responded to messages sent by the predator. The conversations did not feature children but instead featured an adult volunteer playing the role of an underage victim. All volunteers playing the role of pseudo-victims were at least 18 years of age. Moreover, real-world data of this sort is rare, as law enforcement does not typically make such transcripts available for academic purposes (Kontostathis et al., 2010).

Notwithstanding the ethical challenges mentioned above, the difficulties associated with accessing either: (a) actual predator-victim dialogues or (b) predator-decoy (undercover law enforcement) dialogues have meant that the Perverted Justice archives have become a viable and popular alternative in the automatic detection of online predatory behaviour. However, the transcripts themselves are not pre-labelled in any meaningful way, contain minimal meta-data, and are not generally stored in a manner conducive to computational solutions. Because of this, the Perverted Justice corpus is often combined with another non-predatory corpus in a binary classification task (see the PAN12 competition described in Inches & Crestani, 2012). Detecting predatory from non-predatory chats has some practical utility in a real-world sense and, perhaps equally important, negates the requirement to explore (i.e., to read) the predatory messages in any particular depth.

An alternative use of this corpus has been to predict the level of deviance displayed by the offenders (Pendar, 2007). Visitors to the Perverted Justice website are encouraged to rate the 'sliminess' of the offenders on a 1-5 scale, tantamount to a crowd-sourced disgust metric. Other works have explored speaker differences in

sentiment (Bogdanova et al., 2014), as well as linguistic differences using tools such as LIWC (Drouin et al., 2017). These approaches exploit the potential of computational effort. Indeed, most studies in this area work at a scale infeasible via human effort alone. However, given that such works typically utilise surface-level features such as TF-IDF, they can be criticised for lacking domain knowledge. As such, the lack of expert knowledge in many approaches represents an ongoing weakness in this area.

5.4 Approach

The primary purpose of the work covered in this chapter is to explore how domain knowledge and computational effort can be combined to improve the detection of predatory behaviours in online CSE. Rather than classifying differences between predatory and non-predatory chats, or differences in predator and decoy messages, this chapter aims to identify those behaviours experts regard as indicative of increased risk towards a minor. A criticism of a more typical predatory versus non-predatory classification task is that the grooming has already come to fruition. Therefore the opportunity to reduce harm has been missed. Instead, this chapter focuses on detecting those behaviours that domain experts are trained to identify as potential risk markers for future offline contact. Given that child protection officers often monitor multiple chats simultaneously[2] the ability to automatically flag potentially harmful behaviours that can then receive expert attention represents a considerable opportunity.

An overview of the proposed approach in this chapter is given in Figure 5.1. This framework utilises a small number of expert-annotated chats, which are used as training input to fine-tune a pre-existing language model. The framework comprises eight primary steps, each of which will now be described in further detail.

(1) Coding framework

The first step involved the creation of a qualitative coding framework to generate a list of behaviours that appropriately describe the grooming process. The approach

[2]Conversations with a leading online child safety advocacy group revealed that many practitioners in this domain rely on a simple keyword approach to monitor potentially harmful behaviours. From these conversations, it was also revealed that each practitioner monitors dozens of conversations simultaneously.

is based on grounded theory (see Charmaz, 2006), which is a flexible methodology designed to extract descriptive (i.e., qualitative) patterns in data. An annotator with a background in forensic psychology (not the author of this thesis) examined a small selection of chat logs and applied various descriptive labels (codes henceforth) to the individual messages to the point of data saturation. That is, prospective codes were applied and amended until the annotator was satisfied they reliably described the interaction. As codes were derived inductively, steps were taken to increase confidence in the internal validity of the framework. This was achieved by comparing the identified codes with existing psychological theories of child grooming. In particular, the coding framework was compared to the Self-Regulation Model introduced by Elliott (2017), which states that online predation contains a *potentiality phase*, where the predator attempts to form a positive relationship with the victim, and a *disclosure phase*, where the predator becomes more explicitly goal-oriented. The framework was further validated by a collaborative re-coding exercise, whereby the original annotator and two additional researchers with a background in forensic psychology (including the author of this thesis) discussed and jointly coded a small sample of training chats. This enabled the refinement of several codes and proved essential in maintaining the reliability and integrity of the framework.

(2) Corpus annotation

Having created and validated the coding framework, messages from twenty-four randomly sampled chats were analysed. This step involved the annotator reading the chats and labelling messages corresponding to a particular code. An additional control variable was included corresponding to a null annotation. However, as the framework had been derived inductively from the same chat data, most messages corresponded to at least one behaviour. As has been highlighted previously in this thesis, the manual annotation of messages was highly time-consuming. Based on time logs provided by the annotator, it was estimated that manual annotation of the entire Perverted Justice corpus would take several years of effort to annotate in full.

(3) Natural Language Inference (NLI) preparation

The next step involved preparing the labelled chats for an NLI setup. As the task involved the detection of predatory behaviours at the message level, each predatory

message formed a single unit of analysis. Consequently, messages the decoy/victim sent were not included at this stage. For each message, labels identified in step 1 were treated as a binary encoding, where a label could either take the value of 1 if present and 0 otherwise. It should be noted, however, that messages could be labelled with multiple codes. Using the `transformers` library provided by Huggingface (Wolf et al., 2020), NLI requires a tuple comprised of two values; the premise and hypothesis. As a reminder, the objective of NLI is to infer whether there is a relationship between the two constituent parts. Consistent with previous research (Kecht et al., 2021; Sainz & Rigau, 2021), the content of each message was used as the premise, with the hypothesis formed by creating a sentence describing each label. For simplicity, the default setting specified in `transformers` was used – "This is an example of [LABEL]". Once premise-hypothesis pairs had been generated, the next step involved splitting the data into training and testing sub-regions. This was performed via `scikit- learn`, using a seed value to enable replication. The full-sized training set was 70% of the annotated corpus, with an additional 10% withheld as a validation set. The remaining 20% was used for testing.

(4) Variation in training set size

Despite the time and resources involved in annotating the chats, the labelled corpus represented less than 5% of all the available data. With this in mind, identifying the minimal amount of labelled instances needed to achieve an acceptable level of performance was deemed worthwhile. This was examined under three conditions by varying the training data available to the pre-trained language model during fine-tuning. The first condition (4a - full-shot) featured a model trained on all available training data. The second condition (4b - few-shot) used only X positive examples from the available training set, where $X = \{5, 25, 100, 150\}$. A stratified approach was used when sampling the positive classes. This ensured that the class distribution in the few-shot settings matched the distribution in the full labelled dataset. Lastly, the third condition (4c - zero-shot) featured no additional training beyond that already used to generate the language model.

(5) Classification of the test set

Once the model had been fine-tuned via one of the strategies detailed in the previous step, predictions were made on the messages within the test set. The classification output for each message and label combination are three probabilities relating

to whether the hypothesis entails, contradicts, or is neutral with respect to the premise. This step returns the probability of entailment, with the probability of non-entailment assumed by calculating $1 - p(entailment)$.

(6-7) Set entailment decision threshold

Given that the output of the previous step represents the probability that each message and label combination are related, the next step involved deriving an optimised cut-off from which positive and negative classifications could be based. To do this, the approach used here is influenced by the method introduced in (Kecht et al., 2021). First, a set of candidate thresholds was determined – $\{0.002, ...0.998\}$. For each candidate, probabilities above the specified cut-off would be considered positive class instances, while values below the threshold would be considered negative. Next, the predicted classes were split into five stratified folds using cross-validation. Using Matthews Correlation Coefficient (MCC) (Matthews, 1975), the correlation between true and predicted classes was calculated for each fold, with the optimal threshold being the candidate value with the highest MCC over all folds. This approach enabled a different threshold per behaviour code and was preferable to setting a pre-determined value (i.e., a 50% cut-off).

(8) Label remaining corpus

At this point, and with a suitable level of performance, the final step will involve labelling the remaining chats. If performance is deemed unsuitable, several amendments can be applied. First, if performance is poor for a particular label, that label can be dropped from the coding scheme. Alternatively, as explored in Kecht et al. (2021), applying different hypothesis sentence formulations might improve classifications for some tasks. A further option could be to leverage additional semantic information by using WordNet (see G. A. Miller, 1995), for instance, to provide an additional knowledge boost to the model (S. Wang et al., 2021). Lastly, combining machine and human approaches in a human-in-the-loop methodology may also improve performance, particularly if the model performs poorly regarding model recall. For the most part, these additional solutions were deemed too time-consuming in the present context, although a human-in-the-loop approach is experimented with in the following chapter.

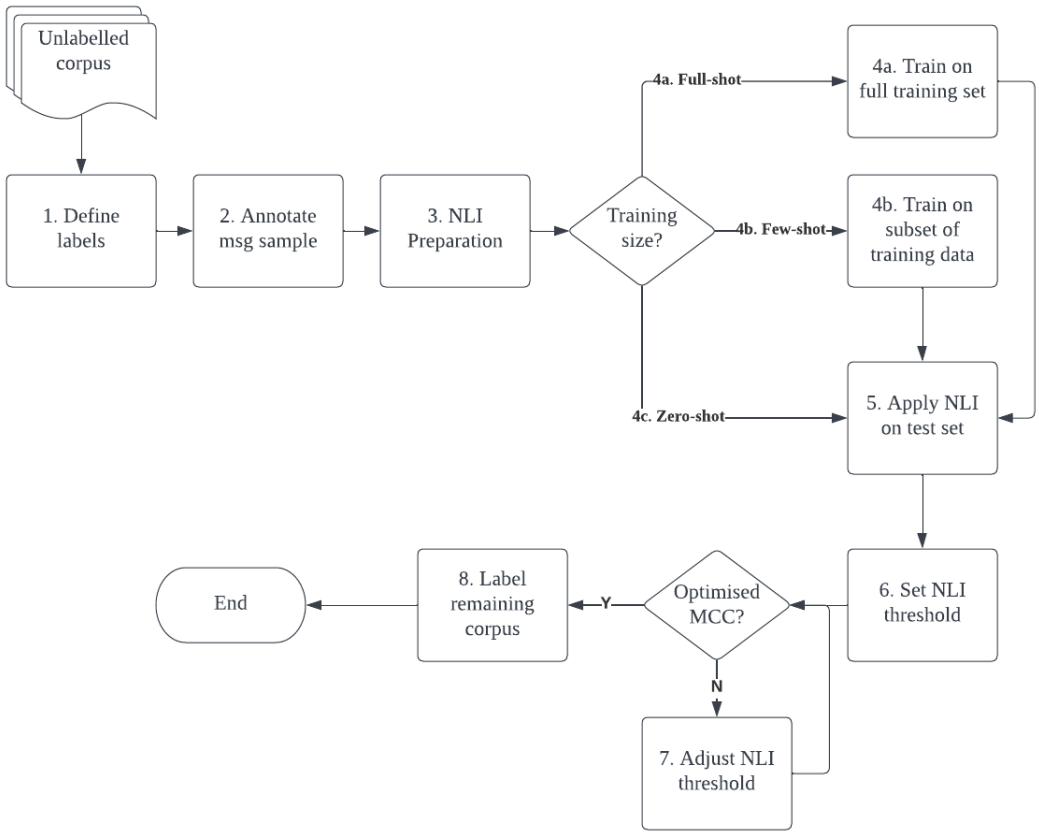

Figure 5.1: Overview of behaviour detection framework.

5.5 Experimental set-up

This section further outlines the approach used in this chapter and the specific experimental tests performed.

5.5.1 Pre-processing and labelling

Chat-logs of predatory chats were scraped from the Perverted Justice website using the `beautifulsoup` package in Python (L. Richardson, 2007). In total, N=623 chats were acquired, comprising approximately 1.35 million messages. A summary of the whole corpus is reported in Table 5.1.

Chats were inspected and cleaned using Regular Expressions (Aho, 1991). As the chats were not initially intended for text mining purposes, several inconsistencies in their structure and format required standardising. For instance, different formatting conventions were applied based on the origin of the chat. A different format was used if the conversation took place on Yahoo Messenger (i.e., `Username`

Region	Chats	Messages	% Corpus
Train	16	4712	3%
Test	5	1355	$< 1\%$
Validation	3	704	$< 1\%$
Unlabelled	599	1336086	96%
Total	623	1342857	-

Table 5.1: Summary statistics of the full CSE corpus.

`(hr:min:sec): message`) compared to MySpace (i.e., `Username - message (hr: min PM)`). In total, the author found eight different message formats throughout the corpus. In another example, some decoys included additional commentary and supplementary information that must be filtered from the message content. Regex was written to identify and remove these from the chats.

Moreover, changes in username mid-conversation were often not aligned within the chat log, or the associated meta-data scraped from the HTML due to moving from one messaging platform to another. This posed a problem for automatically allocating chat messages to a standardised speaker label. To ensure that all messages were attributed to the appropriate speaker role, the author created an exhaustive list of usernames within the corpus and semi-manually cross-referenced these with a list found on the Perverted Justice website. Each message was designated as belonging to a "Predator" or "Victim".

After pre-processing, one unlabelled chat was removed from the corpus due to the interaction featuring multiple predators engaging with a single victim.

Behaviour Codes

The inductive labelling approach described prior generated a coding framework comprising twelve predatory behaviours. The author would like to remind the reader that whilst this work was undertaken as part of a collaborative effort, the qualitative coding activities were performed external to this thesis. Because of this, one of the developed codes was deemed unsuitable for the aims of this chapter and so did not feature. The remaining eleven codes are briefly summarised below, alongside an example taken from the labelled sub-corpus:

Communication/Coordination The first code was named "Communication Coordination". It is used to start and maintain communication as offenders: (i)

exchange and clarify information with their intended victim, (ii) present reason/excuses, (iii) assess the level of engagement of the victim, (iv) find new ways to communicate (i.e., media exchange), (v) strategically use humour or linguistic fillers (i.e., "lol", "hehe"), and (vi) redirect the flow of conversation. One of the offenders' main purposes in this category is to maximise gain and potentially minimise time spent on non-compliant victims.

Rapport Building Offenders use positive behaviour to mimic romantic relationships, making it easier to introduce sexual topics (Elliott, 2017). Offenders use 'Rapport' to infiltrate victims' offline/online social and emotional life to create an illusion of exclusivity, reinforcing the offender as a trusted other. This is achieved through compliments/sweet talk, showing interest, and shared experiences. This special connection or bond is usually created in a short amount of time through excessive saturation and exposure to constant positive statements.

Control Control occurs when offenders use power to direct the flow of communication by influencing or directing the victim's behaviour. Controlling the conversation can occur through subtle (e.g., the illusion of control, rhetorical questions, checking for willingness to engage, or permissive behaviour) or direct strategies (e.g., making demands, persistence, use of coercion). Offenders may attempt to take control of the conversation through patronising language, persistence, frequently checking for engagement, making demands, or asking questions that give the illusion of consent - giving the impression that victims have control over what happens during an exchange.

Challenges An offender may challenge a victim when opposing motivations appear. As a result, confrontation ensues directly (e.g., offence, control, aggression) or indirectly (e.g., joke, mockery, irony). Offenders often challenge the victim as a way of authenticating identity or to exert more control.

Negotiation "Negotiation" can occur at any time during the exchange and is the process where offenders attempt to make decisions, compromise, incentivise continued interaction, or reach goal achievement (e.g., confirming a plan to meet). "Negotiations" can be brief or extensive depending on what goals the offender is trying to achieve. Incentives are particularly important when negotiating goals and can be financial or emotional.

Use of Emotion Offenders use emotive language to manipulate the victim's emotions to influence their behaviour. "Use of emotions" can be positive or negative, including sub-behaviours such as manipulation, expressing empathy, guilt-tripping, vilifying third parties, offering reassurance, or playing the victim. Offenders may employ positive strategies to isolate victims and use negative emotions to increase compliance.

Testing Boundaries "Testing Boundaries'' determines whether the conversation continues or ends. Offenders seek to test boundaries directly or indirectly to determine whether it is possible to desensitise victims through exposure to sexual topics (Elliott, 2017).

Use of Sexual Topics Offenders intentionally use sex to desensitise victims. This is done by directing the conversation toward the victim's prior sexual experiences, discussing fantasies, using explicit language, determining sexual preferences, suggesting media production, alluding to travelling for sex, and acting as a sexual mentor.

Mitigation "Mitigation" is a strategy that aims to soften or downgrade the intensity or seriousness of what is being expressed to convince the victim to participate. Offenders may use this technique to normalise the sexual exchange by lessening the idea of harm or criminality. Specific sub-behaviors include indirectly stating a sexual preference for children, implicating oneself in a previous crime, normalising sexual conversations, or discussing age differences. Normalisation occurs by talking about sex often without reservation and is the process of desensitising the victim to sexual topics or acts.

Encouraging Offenders use encouragement to comply with the victim's requests or to show support by acting as a mentor or trusted other.

Risk Management Lastly, "Risk Management" occurs when offenders assess risk and take steps to prevent discovery. This may be through incentivising secrecy using emotional manipulation, asking the decoy to delete messages/images, enquiring after third parties (i.e., the location of parents), acknowledging previous wrongdoing, and discussing the consequences of getting caught.

5.5.2 Training

In full-shot and few-shot settings, an existing language model was fine-tuned using examples taken from the training set. For training, `RoBERTa-large` (Liu et al., 2019) was used as a pre-trained language model with an implementation built-in `Pytorch` (Paszke et al., 2019). The specific implementation was held on the Huggingface (Wolf et al., 2020) model repository and has been fine-tuned for NLI tasks using the Multi-Genre Natural Language Inference corpus (A. Williams et al., 2018). The model card for this particular iteration of `RoBERTa-large` can be found at the link below[3]. As per the approach taken in S. Wang et al. (2021), models were trained for 10 epochs with batch size 32 and a learning rate of 10^{-5}. No additional training was performed in zero-shot settings beyond that already included in `RoBERTa-large-mnli`.

All experiments in this chapter were performed on Google Colaboratory under a "Colab+" membership tier, thus providing access to a higher-end Graphics Processing Unit (GPU). Both training and evaluation were performed using a Tesla P100 GPU, enabling a much faster processing time compared to training and evaluating models on a Central Processing Unit (CPU).

5.5.3 Evaluation

Models were evaluated on the predictions made on the test set using accuracy, precision, recall, and F1 metrics. Each of the 11 behaviour codes is highly imbalanced, and except for 'communication coordination', positive labels form the minority class. As such, relying solely on accuracy to evaluate model performance can be misleading, as performance will be biased towards the majority class. Precision and recall are, therefore, better suited to this dataset, with F1 (the harmonic mean of recall and precision) included to illustrate overall performance.

Baseline

Across all behaviour codes, the performance of each model was compared to a baseline estimator using traditional supervised machine learning on the full training set. This was intended to bring the current chapter in line with existing research

[3]Model card for the implementation of `RoBERTa-large` used in this chapter can be found at https://huggingface.co/roberta-large-mnli

(Gunawan et al., 2018; Gupta et al., 2012). Pre-processing of predator messages was performed by tokenising, POS tagging, and lemmatising message content using `spaCy`. Stop-words were also removed using the English-language set provided in NLTK. Tokenised messages were then transformed into TF-IDF vectors and used as features in four classification algorithms: Random Forest (Breiman, 2001), Support Vector Machine (SVM) (Cortes & Vapnik, 1995), Naive Bayes (Bayes, 1763), and Logistic Regression (Cox, 1958). For each algorithm, hyperparameters were optimised using GridSearch with a 3-fold cross-validation set-up (Pedregosa et al., 2011). Models are evaluated using F1, with the highest-performing model as the baseline.

5.6 Precision, Recall, and F1 on Hold-out Set

This section reports classification performance against messages with known ground truth (i.e., predatory messages included within the test set).

5.6.1 Coverage of behaviours in labelled corpus

Coverage statistics of each behaviour are reported in Table 5.2. For each behaviour, coverage is calculated by the percentage of positively labelled messages. For instance, if 10 out of 100 messages were labelled 'rapport', rapport would have a coverage of 10%. 'Communication coordination' was the most common behaviour in the corpus by some distance (73% coverage - the average coverage of the remaining ten behaviours was 14%). By comparison, the lowest represented category, 'challenge', was used in only 3% of messages. Manual inspection of messages labelled with the 'communication' behaviour revealed that predators were engaging in considerable amounts of information-sharing and information-gathering. This was particularly prominent at the beginning of conversations and characterised by targeted and directive questioning: "asl?"[4], "Are you there alone?", "do you want to give me your number?". There were also many attempts to use humour-related acronyms (i.e., "lol", "LMAO", "hehe") that explained the high coverage of 'communication coordination' in the labelled corpus.

Additionally, Table 5.2 provides the count of positive class instances for both the training and test regions. A stratified splitting approach ensured the distribution

[4]The phrase 'asl' is text-speak for "age, sex, location?"

of positive and negative class instances was the same between regions. Coverage of each behaviour in the test set was based on probabilities generated from the full-shot condition. As per steps 6-7 in Section 5.4, a message was deemed a positive instance of a given code if it exceeded the decision threshold with the highest MCC. For example, if a message was classified as 'rapport' with a .993 probability, and the threshold was set at .991, that message would be considered a positive instance of rapport. The optimised threshold for each behaviour code is also reported in Table 5.2.

Behaviour Code	Coverage (% of corpus)	Positive class (train)	Positive class (test)	Threshold
Communication	73%	3445	991	.001
Rapport	15%	718	206	.992
Control	21%	979	282	.004
Challenge	5%	211	60	.005
Negotiation	21%	986	283	.75
Use of emotion	16%	773	222	.71
Testing boundaries	31%	1470	423	.78
Use of sexual topics	18%	861	248	.994
Mitigation	3%	144	41	.7
Encouragement	8%	378	109	.004
Risk management	5%	217	62	.88

Table 5.2: Coverage statistics of each behaviour code

5.6.2 Evaluation on the test set

Zero-shot performance

The bar plots in Figure 5.2 summarise the performance of each behaviour in the zero-shot condition. Each sub-plot represents the accuracy (top-left), precision (top-right), recall (bottom-left), and f1 score (bottom-right) obtained from the test set. Red bars indicate that performance did not exceed the best-performing baseline algorithm, whereas green plots indicate an improvement over the baseline. All values are on a $0-1$ scale, where a score of 1.0 equates to 100% performance.

Evaluation of these plots has resulted in several observations. Broadly, all behaviours performed worse than the baseline regarding model accuracy and precision. 'Mitigation' and 'risk management' were the lowest-performing behaviours regarding precision. This resulted from an exceptionally low decision threshold ($< .01$), meaning positive instances of these behaviours were both highly over-predicted. This over-prediction also explains the high recall performance of both codes, as most

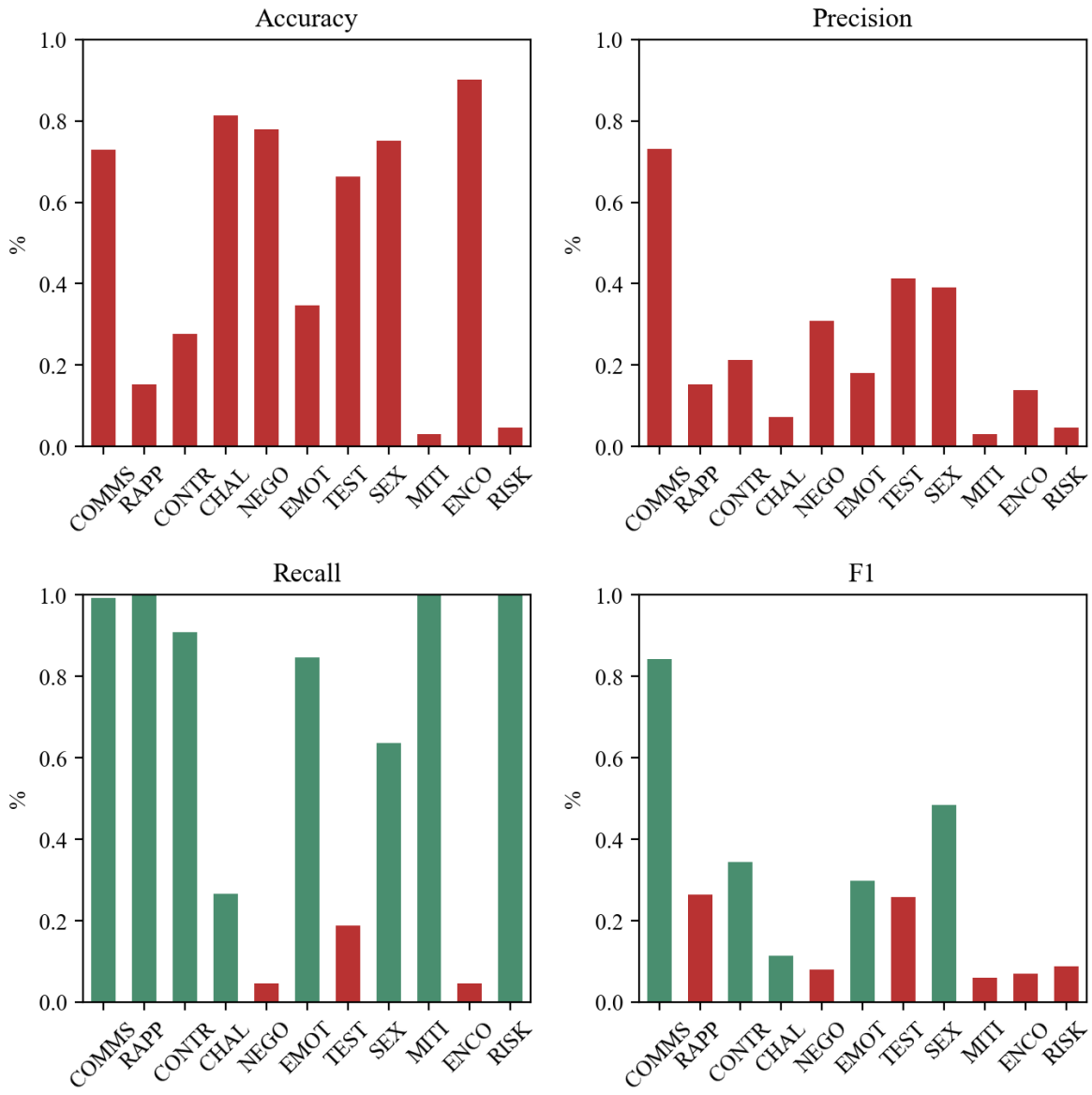

Figure 5.2: Zero-shot results on test set. Green bars indicate codes that exceeded the baseline.

messages were classified as positive instances. This factor, combined with the relative scarcity of both codes in the corpus (mitigation and risk management had 3% and 5% coverage, respectively), meant that each generated a low number of false negative classifications.

The generally high recall performance across most codes is an intriguing outcome but ultimately misleading in evaluating performance. As a reminder, each code appeared infrequently within the corpus except for 'communication coordination'. 'Mitigation', 'risk management', and 'rapport' all generated a recall score of 1, despite the combined coverage of these codes being less than 30%. This indicates a poor model, as it is likely that the positive class label was predicted for every

message, despite setting an optimised decision threshold. Indeed, the MCC for each of these codes was close to 0. This suggests that little to no relationship could be found between the predicted values and ground-truth classifications.

The abstractness of the behaviours likely influences the main reason for the lack of performance in the zero-shot condition. As noted by Heide (2013), concepts such as "rapport" are tough to define, even for humans. It, therefore, is not surprising that a machine fails at this task without any positive examples to learn from.

Regarding the F1 score, five of the eleven behaviours (communication coordination, control, challenge, use of emotions, and use of sexual topics) exceeded the equivalent baseline. This indicates that a pre-trained language model with no training input could better represent the concept than a more simplistic model trained on data comprising several thousand messages. 'Communication coordination' received the highest F1 score by a considerable margin (84%), followed by 'use of sexual topics' (48%), 'control' (34%), 'use of emotion' (30%), and 'challenge' (11%). However, as noted above, performance, in this case, is likely misleading due to an inflated recall score resulting from an over-classification effect.

Few-shot performance

In Figure 5.3, the change in the $F1$ score as the number of positive training examples increases is reported for each behaviour. A comparison to the baseline performance is also illustrated via a red line, with points above this line indicating that a few-shot approach improved over the baseline. Performance of most behaviours increases as the size of the training set grows. Nine of the eleven behaviours exceeded the baseline at some point within the few-shot condition. The two exceptions are 'communication coordination' and 'risk management'. Whilst neither behaviour exceeded their respective baseline, it should be noted that baseline performance for 'communication coordination' was already high ($F1 = 85\%$).

As demonstrated in Table 5.3, most behaviours attained a notable improvement with only a small number of positive training examples to train from. 'Control', 'challenge', and 'use of sexual topics' exceeded their respective baselines with as few as five labelled instances. On average, results indicate that a reasonable level of performance can be achieved by training on fewer than 50 positive examples.

A drop in performance was observed for several behaviours in a 25-shot set-up, notably 'Challenge', 'control', and 'emotion'. Similarly, a considerable drop in

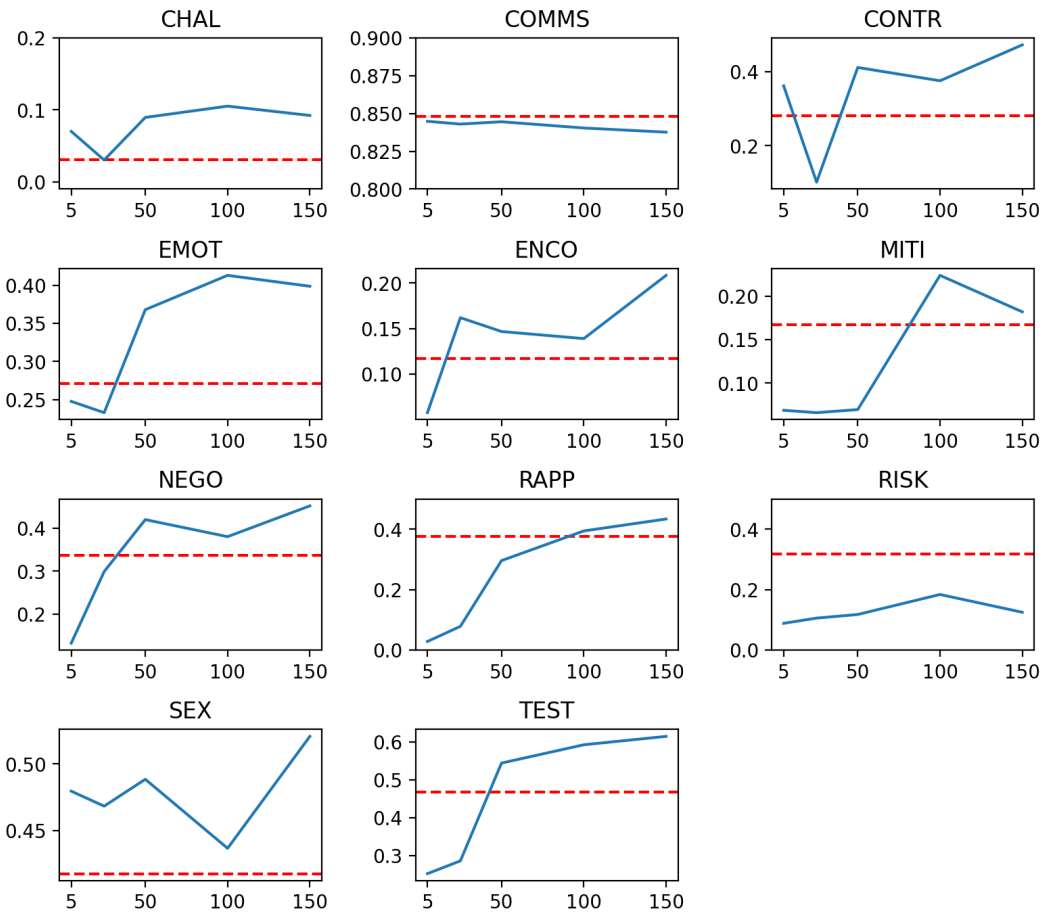

Figure 5.3: Change in $F1$ as the size of the training set increases. Value along the x-axis indicates the number of positive instances in the training set. Value at the y-axis indicates the $F1$ score. Baseline performance for each behaviour code is indicated by a red dashed line.

$F1$ was observed for 'use of sexual topics' at the 100-shot level. This may be a consequence of the split in the original data into train and test regions and could be resolved through cross-validation, where we expect to observe variation across folds. This step was, ultimately, not performed here for reasons of computational efficiency.

Full-shot performance

Having demonstrated a general increase in performance as the number of positive training examples grows, it was anticipated that the full-shot condition would generate a further boost in performance for all behaviours. Similar to the analysis of the zero-shot condition, an overview of performance on classifying predator messages in the full-shot condition is illustrated in Figure 5.4. Here, each of the eleven behaviours exceeded the respective baseline in $F1$ (compared to five behaviours in the zero-shot

Behaviour	Baseline $F1$	Min. examples to exceed baseline	$F1$ with min. examples	Max. $F1$
Communication	.85	n/a	n/a	.84
Rapport	.38	100	.39	.43
Control	.28	5	.36	.47
Challenge	.03	5	.07	.11
Negotiation	.34	50	.42	.45
Use of emotion	.27	50	.37	.41
Testing boundaries	.47	50	.54	.62
Use of sexual topics	.42	5	.48	.52
Mitigation	.17	100	.22	.22
Encouragement	.12	25	.16	.2
Risk management	.32	n/a	n/a	.18

Table 5.3: Number of positive training examples required to exceed baseline under **few-shot conditions**. n/a indicates performance did not exceed baseline

condition). Seven of the eleven behaviours also obtain an $F1$ score above 50%, with the best-performing behaviour being 'communication coordination' ($F1 = .87$), followed by 'testing boundaries' ($F1 = .62$), 'rapport' ($F1 = .61$), and 'use of sexual topics' ($F1 = .61$). Performance was poorest for 'challenge' ($F1 = .28$), followed by 'encouragement' ($F1 = .32$), and 'mitigation' ($F1 = .4$).

Comparing the differences in performance between zero-shot and full-shot conditions provides a further indication of the utility of training on larger datasets. On average, $F1$ performance increased between zero and full-shot conditions by 27%. Inspecting the precision and recall scores indicates why this is the case. Compared to zero-shot, the unrealistically large recall scores observed for 'mitigation' and 'encouragement' due to over-prediction have reduced considerably – recall for 'mitigation' having dropped from 1 to .32, and for 'encouragement' from 1 to .43. Whilst these are low, they do indicate that the model has at least learnt the ability to predict both positive and negative classes, even if this is likely to contribute to an undesirable number of false negatives. Equally, evaluating differences in model precision between zero and full-shot conditions indicates that the number of false positives has dropped across all behaviours. However, only four of the eleven behaviours exceeded the equivalent baseline. The false positive rate, in this case, remains an ongoing issue.

Qualitatively exploring classifications

To better understand performance, a brief qualitative exploration was performed on a random sample of $\approx$ 10% of the test set. This was performed theoretically

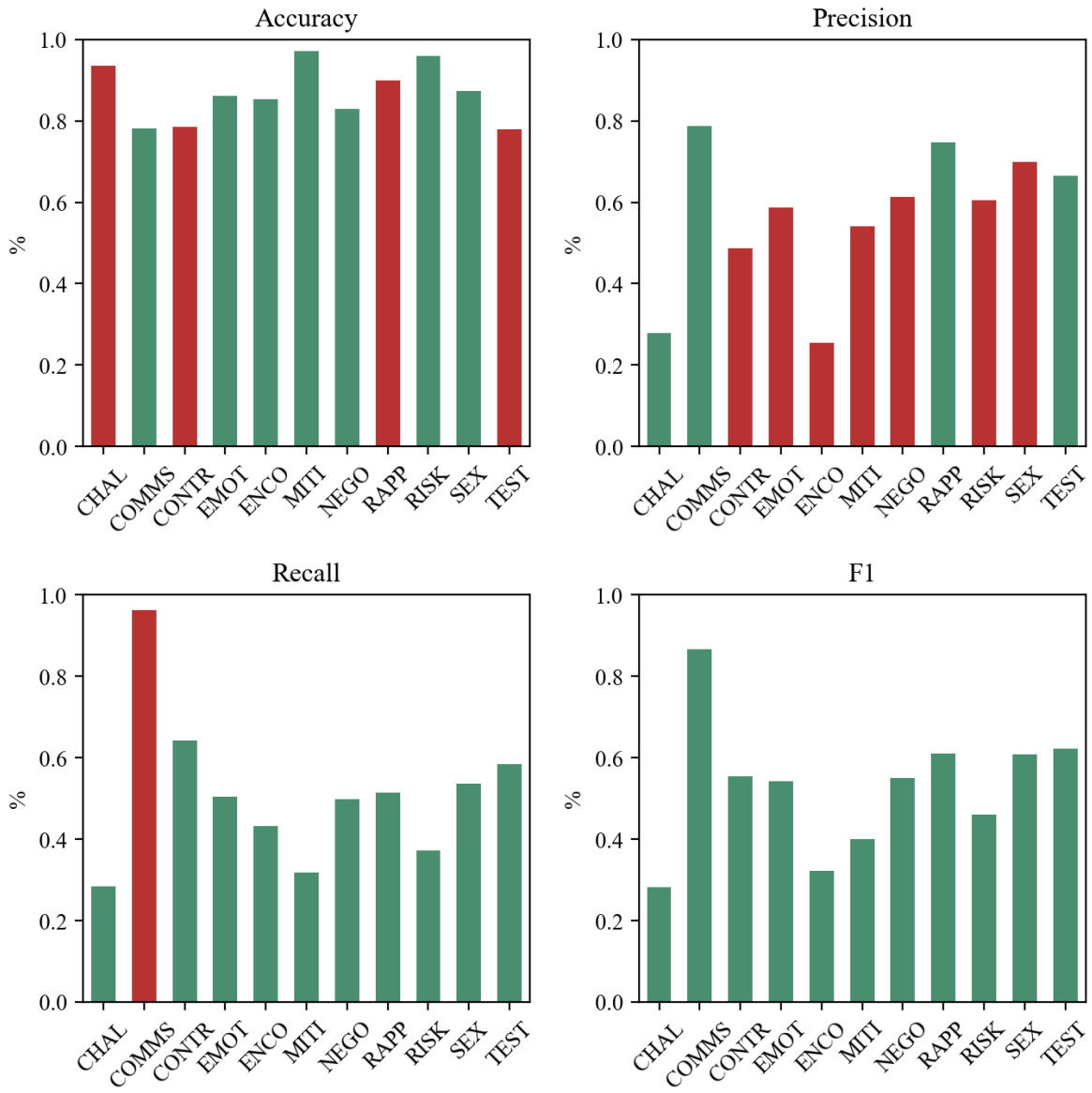

Figure 5.4: Full-shot results on test set. Green bars indicate codes that exceeded the baseline.

and intended to serve as a visual inspection of the classifications made in the full-shot condition, representing the best-performing models. It was observed that the NLI models had learned some rules contributing to several false positives. For instance, the transformer correctly anticipated several features emblematic of the *testing boundaries* code. Examples of this included assessing the victim's willingness to engage (e.g., "what are you looking for?"), as well as assessing their willingness to progress to offline contact (e.g., "How about I come by at 8 pm?").

The *rapport building* model correctly recognised complements and sweet talk as positive examples but missed more everyday examples of rapport building such as social greetings (e.g., "hi, how are you? asl?"). It also routinely failed to identify

general conversational patter as evidence of rapport (e.g., "how was your spring break?").

Some aspects of *control* appeared to take place over longer ranges than single messages. For example, persistently asking the same question was often misclassified, as each message was considered an independent event. In trying to predict *encouragement*, amongst the worst performing labels, the transformer appeared to overfit on short verbal nods (i.e., "kool" and "sure"). This appeared regularly in predator speech but was not always labelled as encouragement by the expert annotator. Over-reliance on these unigrams appeared to increase the false positive rate substantially.

Risk management appeared to perform better than other rare (coverage$< 5\%$) behaviours in the full-shot condition. Examination of the positive classifications indicates that this was largely a consequence of recognising attempts to establish the presence of a parent (e.g.,"Is ur dad gonna be home tomorrow?" and "When are they getting home?").

5.6.3 Does additional context enhance performance?

So far in the chapter, classification has been based solely on single messages. Whilst this appears to be an efficient strategy for classifying some behaviours, others perform poorly even in the full-shot condition. As discussed previously, one reason for this may be a lack of positive training examples. One could solve this problem by collecting more data. However, this approach is unlikely to be suitable for rare behaviours, given that collecting more positive examples will require an exponential increase in negative examples to maintain an existing distribution.

An alternative solution may be to leverage the additional context provided by the surrounding messages. When classifying messages individually, an implicit assumption is made that each message exists independently of others in the conversation. This is clearly not the case, as each message logically follows from, and has therefore been influenced by, those messages that have preceded it. It is not hard to imagine how several of the behaviours covered in this chapter might be improved by increasing the observable window. For instance, behaviours such as 'challenge', 'negotiation', and 'rapport' are all dynamic behaviours that, to varying degrees, require the involvement of an additional speaker. Conversely, behaviours such as 'use of sexual topics' are much more centred on the language a predator uses and less

dependent on prior utterances' content.

To explore whether the surrounding messages increase the contextual understanding of the language model, the full-shot condition was re-analysed over a multi-message window. This was performed by concatenating each predator message with the four previous messages as a contextual buffer[5]. Crucially, this included predator and decoy messages and reflected both sides of the conversation. Messages with fewer than the required number of prior messages, due to appearing earlier in the conversation, were concatenated up until the beginning of the conversation.

Each model was re-trained using the configuration described in Section 5.5.2. This analysis section was only performed on ten of the eleven behaviour codes. 'Communication coordination' was left out, as its high coverage in the labelled corpus (73%) would result in a positive classification for every message.

Single vs. Multi-Message Comparison

A per behaviour comparison of precision, recall, and $F1$ between single and multi-message input is illustrated in Figure 5.5. Regarding model precision, increases were marginal, with 'use of sexual topics' decreasing in precision by 5% (from 70% to 65%) when a multi-message window was used. The lowered precision indicates that the additional context increased the number of false positives.

A multi-message input increased the recall of several behaviours, including 'rapport' (increased from 52% to 81%), negotiation (from 50% to 67%), testing boundaries (from 58% to 76%), and risk management (from 37% to 48%). Increased recall suggests that the added context from the previous messages decreased the false negatives for these behaviours.

Summarising these findings, the multi-message input had a generally positive impact on model performance. Seven of the ten behaviours tested here generated an $F1 > .5$ in the multi-message condition. Whilst this would be insufficient in a real-world context, it does provide sufficient motivation to continue analysing the remainder of the corpus.

[5]The decision to opt for a 5-message window (current message + 2 prior predator + 2 prior victim) was based on additional experimentation with 3-message and 7-message windows. A 5-message window generally outperformed 3-messages, whereas 7-messages resulted in a high runtime during model training.

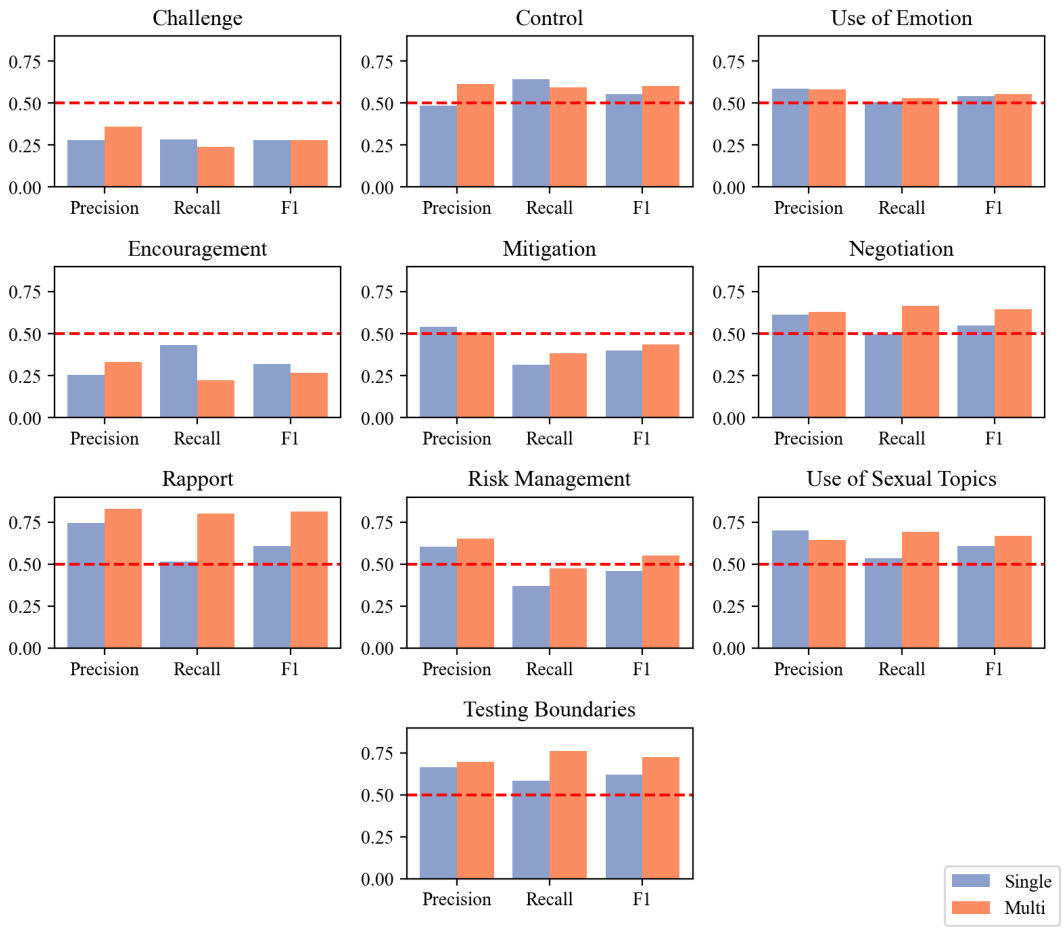

Figure 5.5: Comparison of classification performance between single and multi-message inputs. A 50% threshold was set as a simplistic cut-off point indicating a satisfactory level of performance – the precise value of a satisfactory model will be highly domain-specific.

5.6.4 Is the machine comparable to a human annotator?

The final experiment in this chapter explores an alternative way of evaluating model performance. Specifically, this section attempts to quantify how comparable a computational model is to a human annotator by examining the level of inter-rater agreement between human and machine classifications. This is subtly different from the analysis already conducted. Until this point, the labels generated through manual annotation have been considered equivalent to ground truth, implying some degree of objectivity in how labels have been distributed amongst the predatory messages. However, considerable variation often exists in human-annotated studies. When labelling highly abstract concepts such as those explored here, it is unlikely that two independent raters will come to the same decision about what constitutes an example of a given behaviour. Instead, within the social sciences, the level of agreement between raters is calculated to indicate a measure of objectivity within

coding efforts. The high agreement allows some flexibility and deviation in how annotators distribute labels. Whilst disagreement may exist, annotators ultimately converge on equivalent ratings in most cases.

One of the ways the level of agreement is assessed is through groups of raters (two or more) coding the entirety of a dataset independently before determining whether the level of exhibited agreement exceeds what would be expected through chance alone. This step is often impractical as it requires multiple passes of a dataset, which is both time-consuming and resource-intensive. To alleviate this, here, the predictions made by the best-performing model are used as a proxy for a second annotator (referred to as "Comp" below). To enhance the robustness of this approach, the author also performed a small manual validation step, which served as a second human annotator and third overall annotator (referenced as "R2" below). Post-validation involved comparing each message with the class predicted by the best-performing model. Each message was scored on a 1-3 scale, based on whether the author agreed with the classification (1=disagree, 2=uncertain, 3=agree). The author did not have access to the ground-truth labels during this stage. Messages were selected at random.

Inter-rater agreement

In total, the classifications of 645 messages were validated by the author of this thesis. A common agreement statistic known as Cohen's K (1960) was used to assess pairwise agreement, where a higher value of K indicates greater agreement between raters. Ironically, there is no universal agreement of how best to interpret K; however, many cite a subjective interpretation by Landis and Koch (1977) as acceptable (see Table 5.4). The total agreement between the three raters was assessed by averaging K values for each pairwise combination.

Kappa value	Interpretation of agreement
$< .0$	Poor / No agreement
$0.0 - 0.2$	Slight
$0.2 - 0.4$	Fair
$0.4 - 0.6$	Moderate
$0.6 - 0.8$	Substantial
$0.8 - 1.0$	Almost perfect

Table 5.4: Interpretation of Cohen's K indicating pairwise agreement, as per Landis and Koch (1977)

Figure 5.6 reports pairwise agreement scores for each behaviour. Across all behaviours, and for each combination of raters, values of K ranged between .46 and .95, indicating a minimum of 'moderate' agreement on any pairwise combination. Comparing R1 (initial annotations) with R2 (validations performed by the author), several of the behaviours received a K value above .8, indicating near-perfect levels of agreement. Interestingly, comparing these agreement scores with those generated by human-machine comparisons (R1 & Comp or R2 & Comp), values of K drop considerably. This finding suggests the two human raters found it easier to align with each other on certain behaviours than a computer was able to. For example, the average agreement between R1 and R2 for 'risk management', 'mitigation', 'negotiation', and 'challenge' was .91 – comfortably within the 'near perfect' range. Conversely, the average agreement between Comp and R1 or R2 on the same behaviours was .58.

Averaging values of K over every behaviour generates a single agreement score for each pairwise combination. Unsurprisingly, given the results already discussed, the average agreement between the two human raters was higher (M=.79, SD=.12) compared to the agreement of either rater with the computer (R1 & Comp: (M=.57, SD=.07); R2 & Comp: (M=.64, SD=.11)). The extent that this represents an unacceptable drop in rater agreement is likely domain specific. The level of human-human agreement found here is comparable to prior research (Shibata et al., 2017), with human-computer agreement close to that observed in (Cross et al., 2017; Espinoza et al., 2019; Flor & Andrews-Todd, 2022). A collective K score of .7 was deemed an "acceptable" level of group agreement in Mclaren et al. (2007).

5.7 Classification of remaining corpus

A primary advantage of having trained models to a satisfactory standard is that they can quickly classify the remaining 599 chat logs in the corpus. Eight of the eleven original behaviours are featured here: (i) communication coordination, (ii) rapport, (iii) control, (iv) negotiation, (v) use of emotions, (vi) testing boundaries, (vii) use of sexual topics, and (viii) risk management. The decision to drop *challenge*, *mitigation*, and *encouragement* was due to poor performance ($F1 < .5$) of these behaviours on the human-labelled holdout-set (see Section 5.6.2). Except for *communication coordination*, language models for seven behaviours are based on a multi-message window that includes both predator and victim speech turns. Due to high performance in the single-message condition, the language model for com-

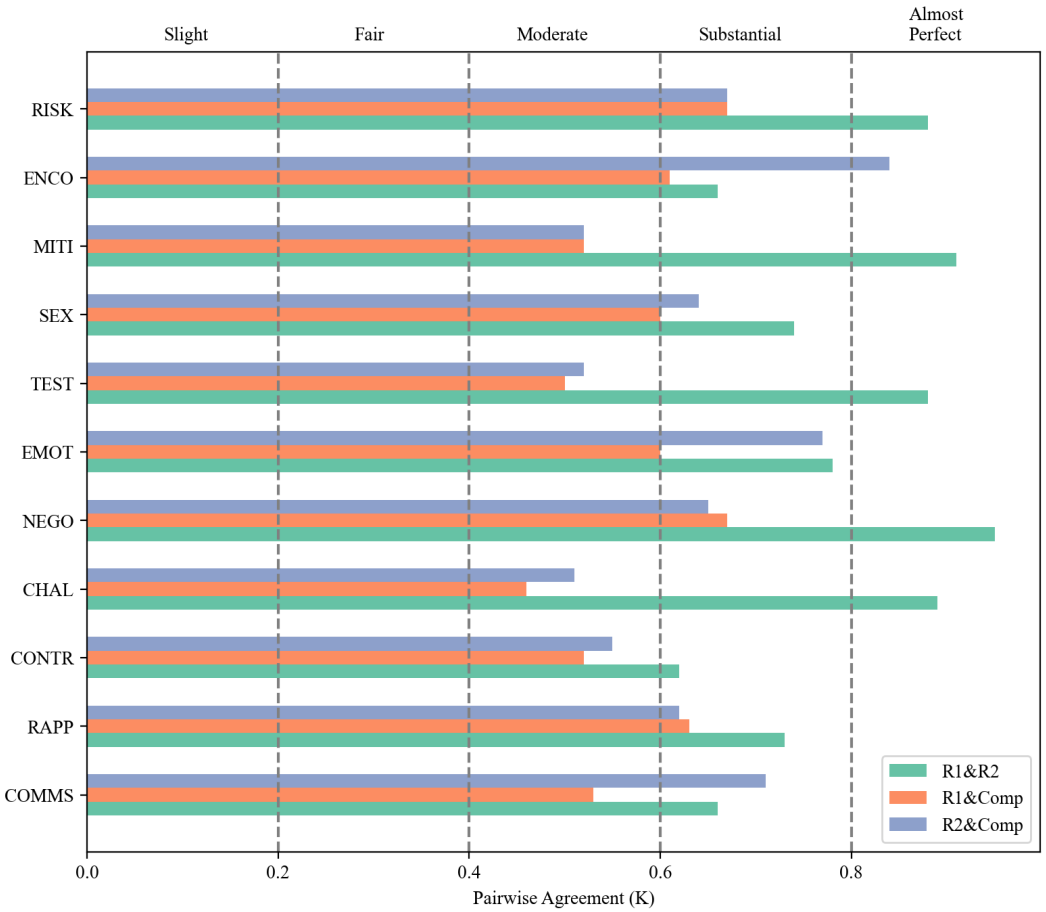

Figure 5.6: Pairwise agreement between initial human annotations (R1), human validations (R2), and machine classifications (Comp).

munication coordination is based on the content of a single predator message (i.e., additional content generated by prior messages was omitted).

5.7.1 Experimental set-up

Computationally, the unlabelled corpus was treated as a single, large test set. As before, labelling was performed within a Google Colab notebook with a Tesla P100 GPU. Classifications were performed via the zero-shot-classification pipeline using `Huggingface` (Wolf et al., 2020) with an increased batch size of 128 as no training was performed. The output was a data frame that included the message content, a chat ID, and a one-hot encoded marker for each label (where 1 indicates a positive class instance, and 0 is a negative class instance). See Figure 5.7 for an illustration.

message	chatID	COMM	RAPP	CONTR	NEGO	EMOT	TEST	SEX	RISK
iam just too old	pjchat_0	0	0	0	0	1	0	0	0
yes	pjchat_0	1	0	0	0	0	0	0	0
iam looking	pjchat_0	0	0	0	0	0	0	0	0
for a female	pjchat_0	1	0	0	0	0	0	0	0
what are you looking for?	pjchat_0	0	0	0	0	0	1	0	0
cool	pjchat_0	1	0	0	0	0	0	0	0
what do you like to do?	pjchat_0	1	1	0	0	0	0	0	0

Figure 5.7: Binary classification of each behaviour on predator messages

5.7.2 Coverage of machine-annotated behaviours

Frequency counts and coverage statistics of machine annotations are reported in Table 5.5. Machine annotations' coverage was compared to the expected coverage indicated by the human annotations. Comparing the coverage of each behaviour generated from machine annotations to the 95% confidence interval of human annotations indicates whether a behaviour is being over or under-predicted by the model. If coverage of machine annotations falls within the 95% range, it indicates that classifications are congruent with that expected by the human-annotated corpus.

Machine-annotated coverage of three of the behaviours – *testing boundaries*, *use of emotions*, and *risk management* – fall within the range expected by the human annotations, indicating the model performs well on these codes. Similarly, coverage for *rapport* and *negotiation* labels fall within 2% of the expected range. This finding suggests that both are slightly under-predicted by the model and are likely to contain an increased number of false negatives. *Use of sexual topics* was also under-predicted, although by a more considerable margin with only 11% coverage predicted despite a 17%-20% range expected by human annotations. Finally, the two most common behaviours in the human-annotated corpus (*communication coordination* and *control*) were each highly over-predicted by the model, indicating a more significant number of false positives. Coverage of *communication coordination* was 13% higher than the range expected by the human annotations (72% - 74%). Similarly, 'control' was 12% more populous in the machine annotations than the range of expected human annotations (19% - 23%).

5.7.3 Correlation between machine-annotated behaviours

An alternative way to test the robustness of the machine annotations is by exploring the correlations between pairs of behaviours and whether the strength of correlations

Behaviour	Pos. instances in unlabelled corpus	Coverage (% of corpus)	Expected range
Communication	61414	87%	72% - 74%
Rapport	77971	11%	13% - 17%
Control	248105	35%	19% - 23%
Negotiation	127031	18%	19% - 23%
Use of emotions	113951	16%	14% - 18%
Testing boundaries	207811	30%	29% - 33%
Use of sexual topics	79789	11%	17% - 20%
Risk management	22048	3%	3% - 7%

Table 5.5: Coverage statistics of machine-annotated behaviours in the unlabelled corpus. Expected range given by the 95% confidence interval of the human-annotations.

generated by the machine matches those in the human-annotated chats. Assuming human and machine datasets represent a truly random sample, we can expect models that have learnt a sufficient representation to mirror the pairwise relationships observed in the original human-annotated corpus. Pairwise correlations in both the human-annotated and machine-annotated corpora are referenced in Figure 5.8 and Figure 5.9 respectively, Comparisons between the two datasets are summarised fully in Table 5.6. The range of expected values indicated by human annotation was denoted by a z-transformed 95% confidence interval per Gorsuch and Lehmann (2011). The z-transformation (also referenced as a Fisher z-transform) corrects the negative skew in correlation data towards zero. Of the 28 pairwise correlations performed on the machine annotations, 17 (61%) fell within an expected range of human annotations. A further four pairs fell just outside this range (within .1 of the maximum or minimum expected value). In only one condition (*communication* and *use of sexual topics* did the polarity of the correlation invert from the expected range.

Several interesting correlations appear within the data. In particular, three behaviours ('control', 'testing boundaries', and 'use of sex') appear to cluster together, with a moderate average correlation of $r = .29$. This finding is consistent with existing literature, which argues that predators often create an illusion of control around sexual topics by appearing to only engage in behaviour to which the victim approves (Barber & Bettez, 2021). At the same time, predators are aware that interactions online are fleeting. As such, they will consistently test the victim's boundaries of acceptable conduct, knowing that failure will result in little to no consequence (Fransson et al., 2015). We see this illusion of control appears regularly in messages that were positively classified by the machine as representing the three behaviours:

Behaviour pair		Machine-annotated r	Expected value of r
Communication	Rapport	**−.18**	$-.2 < r < -.16$
Communication	Control	**−.08**	$-.13 < r < -.08$
Communication	Negotiation	**−.06**	$-.07 < r < -.02$
Communication	Use of emotion	−.34	$-.31 < r < -.27$
Communication	Testing boundaries	−.08	$-.16 < r < -.12$
Communication	Use of sexual topics	−.01	$.01 < r < .06$
Communication	Risk management	**.02**	$-.01 < r < .03$
Rapport	Control	−.02	$-.11 < r < -.06$
Rapport	Negotiation	.03	$-.04 < r < .01$
Rapport	Use of emotion	.11	$.05 < r < .1$
Rapport	Testing boundaries	**-.08**	$-.1 < r < -.05$
Rapport	Use of sexual topics	−.05	$-.13 < r < -.09$
Rapport	Risk management	**−.06**	$-.09 < r < -.04$
Control	Negotiation	.11	$.06 < r < .1$
Control	Use of emotion	**.02**	$-.02 < r < .03$
Control	Testing boundaries	**.26**	$.25 < r < .29$
Control	Use of sexual topics	.23	$.29 < r < .33$
Control	Risk management	**−.02**	$-.02 < r < .03$
Negotiation	Use of emotion	**.09**	$.06 < r < .11$
Negotiation	Testing boundaries	**.08**	$.06 < r < .11$
Negotiation	Use of sexual topics	**.06**	$.04 < r < .09$
Negotiation	Risk management	**.04**	$.04 < r < .08$
Use of emotion	Testing boundaries	**.05**	$.01 < r < .05$
Use of emotion	Use of sexual topics	**−.06**	$-.09 < r < -.04$
Use of emotion	Risk management	.04	$.08 < r < .12$
Testing boundaries	Use of sexual topics	.37	$.38 < r < .42$
Testing boundaries	Risk management	**.04**	$.04 < r < .09$
Use of sexual topics	Risk management	−.03	$-.02 < r < .03$

Table 5.6: Pairwise correlations between behaviours in machine-annotated data. Expected range represents z-transformed 95% confidence interval of human-annotations. Machine values that fall within the expected range are in **bold**.

> "dresses, skirts or jeans i like them all just want U to be comfortable in whet U wear"
>
> Predator in Chat 101.

> "are you finished looking at it my love. or do you want more"
>
> Predator in Chat 138.

> "so would you want to take my pic like that for you babe"
>
> Predator in Chat 45.

A moderate negative correlation was also observed between "communication coordination" and "use of emotion" ($r = -.34$). The size of the relationship is

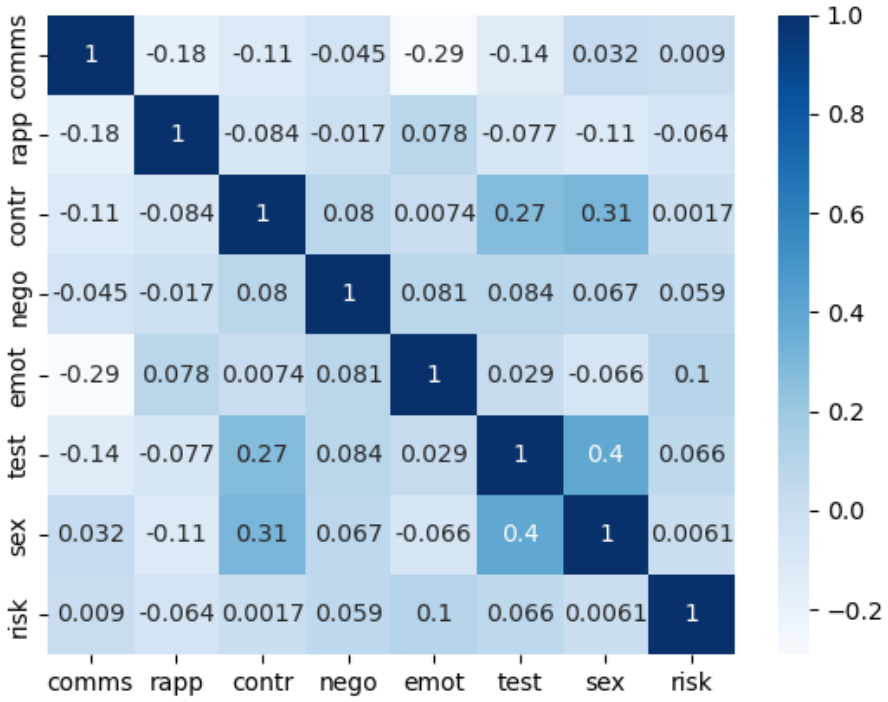

Figure 5.8: Correlations between behaviours on *human* annotations. 95% confidence interval is ±.027

particularly interesting given that "communication coordination" appeared in most messages and is likely too broad a category to be of particular significance. Exploring messages classified by the machine as positive instances of "use of emotions" but negative for "communication coordination", it appears the machine has learnt to detect several interesting sub-themes emblematic of the "use of emotions" code. First, "use of emotions" appears to be predicted when the predator intends to reassure the victim or is seeking reassurance themselves:

> "iam so sorry. iam too old. iam just too old"
>
> Predator in Chat 0.

> "I am sorry. its ok. we will talk tomorrow ok no pressure. no need to be nervous."
>
> Predator in Chat 103.

The results presented in this section emphasise the utility of an automated approach to behaviour coding as an alternative to traditional human annotations per-

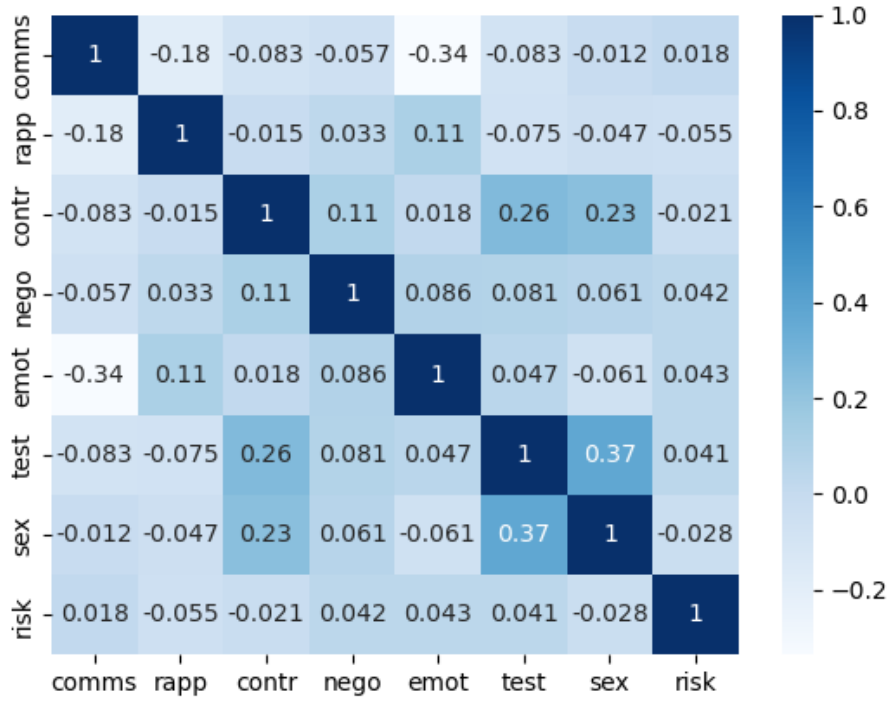

Figure 5.9: Correlations between behaviours on machine labelled annotations. 95% confidence interval is ±.002

formed manually. Using a small corpus of human annotations, labels for eight behaviours were automatically generated on a much larger corpus. Except for one study (see Fan & Ilk, 2020), to the best of the author's knowledge, the findings in this section represent one of the first attempts to extrapolate machine-generated behaviour codes based on social interactions to an unlabelled dataset several orders of magnitude larger. Given the absence of ground-truth labels beyond the initial training set, this approach is necessarily exploratory.

Human behaviour is highly idiosyncratic. Consequently, individual message-level behaviours may be too granular to be informative in a real-world setting. In online child safety, detecting single instances of rapport-building is unlikely to help identify one predator from a haystack of consenting conversations. As an alternative, future work could include models that track the changes and co-occurrences of multiple behaviours throughout the interaction. This approach was attempted in Fan and Ilk (2020), who used a directed graph approach to automatically detect behaviour patterns in a customer service setting. Such an approach could be useful here, as it could enable distinctions between when a behaviour such as rapport-building can be considered predatory and non-predatory. Perhaps predatory rapport-building co-

occurs with control? Perhaps risk management only occurs after the conversation has turned sexual? Such an analysis was considered to be beyond the scope of the present thesis but would potentially expand on the work covered here.

5.8 Discussion

This chapter has examined the utility of an automated approach to predatory behaviour detection based on a deep learning framework. The remainder of this section will now discuss how these findings contribute to the aims of the thesis and the prevention of online sexual exploitation more broadly.

5.8.1 Efficiency of automated behaviour detection

The primary aim of this thesis is to create scalable solutions to existing social scientific methods that rely on large amounts of manual annotation. Dynamic solutions enable the exploration of social behaviour on a large-scale and, as discussed in previous chapters, highlight subtle behaviours that might otherwise go undetected in smaller-scale studies. Manually annotating the twenty-four chat logs in the current chapter took over 600 hours to perform by a single annotator[6]. As a reminder, similar work in this area has typically elected to use a greater number of chats, as well as cover a broader range of predatory behaviours (see Gunawan et al., 2018; Gupta et al., 2012). The entire Perverted Justice corpus comprises 623 chat logs. Assuming a similar level of effort, manual annotation of the whole corpus would have taken 15,575 hours of continuous effort[7]. Equating that to a standard 38-hour working week would result in 410 weeks, or 8+ years to annotate the total corpus once. By contrast, model training took approximately 13 hours for all eleven behaviour labels and a further 7 hours to apply to the unlabelled portion of the corpus.

Differences between the two are further highlighted if we consider the added flexibility of an automated approach. Whilst the re-coding effort described in 5.4 did not result in fundamental changes to the core coding framework, any wholesale changes (i.e., the inclusion of another behaviour label) would have required the annotator to restart annotation from the beginning. This would have been undoubtedly frustrating at twenty-four chats, but ultimately workable given the comparatively fewer

[6]This includes the time taken to define the coding framework and verify inter-rater agreement in a joint re-coding effort with the author of this thesis and one other researcher.

[7]Based on the average chat taking 25 hours to annotate ($600 \div 24$)

data points. However, given the considerable time requirements, the same is not true if annotating the entire corpus was performed manually.

5.8.2 Accuracy of machine classifications

Identifying predatory behaviour is a large-scale problem that requires a high level of detection accuracy. Given the potential harm caused by misclassifying predatory behaviour as non-predatory, model recall ($TP \div (TP + FN)$) is arguably the more critical metric. Higher recall means fewer false negatives. At a surface level, a lower precision ($TP \div (TP + FP)$) represents less of a problem, as the increased number of false positives does not incorrectly eliminate predatory behaviours that should otherwise be detected. Arguably, this becomes more of a problem as the scale and complexity of the analysis increase as it creates additional unnecessary work to verify actual cases of predatory interactions. Because of this, in addition to a general class imbalance for each behaviour, this chapter has referred to $F1$ as the most relevant performance metric. Amongst the best-performing models, the average accuracy over all behaviour labels was $F1 = .55$, which is highly unlikely to be sufficient in a real-world setting. However, several behaviours performed well, particularly at a multi-message level. Classification for rapport-building ($F1 = .81$), negotiation ($F1 = .67$) and testing boundaries ($F1 = .76$) all achieved good levels of performance when an expanded window was used, and alongside 'communication/coordination' ($F1 = .87$) meant that over a third of the behaviours examined could be reliably predicted despite their infrequency within the corpus.

In their work examining behaviour in a sample of online Twitter interactions, Burnap et al. (2015) noted that groups of human coders, even if identically trained, are likely to express differences in coding due to the inherently subjective nature of the task. Such a degree of disagreement is a tolerance not factored into metrics such as precision and recall that assume a unified interpretation of ground truth (Razi et al., 2021). Debatably, a classification of predatory behaviour can be correct, even if not allocated as such by a single coder. In an ideal scenario, a closer approximation of ground truth could have been facilitated by conducting several passes of the data with multiple human annotators. Time and resource restraints prevented this in the current work. In place of this, an effort was made to verify the quality of the machine classifications through the post-validation step described in Section 5.6.4. This step created a pseudo-third annotator from which conventional agreement metrics could be calculated. The level of agreement between the original annotations and the post-

validations (an approximation for two human raters) tended to indicate high levels of agreement across behaviours. Whilst both variations of the human-computer agreement were lower, all behaviours fell comfortably within a range of 'moderate' agreement as per Landis and Koch (1977). At the very least, the machine can be interpreted as performing with a similar level of consistency as an individual human annotator.

5.8.3 Improvements with increased training

Comparison of the zero-to-full-shot classification task generally supports the view that increasing the data available for training improved model performance. Deep learning algorithms are notoriously greedy due to their complexity and tend to perform better with large amounts of training material (Venturott & Ciarelli, 2021). Whilst the language models used in the zero-shot condition have already been pre-trained on a large external dataset (i.e., English-language Wikipedia), these models struggle when no domain-specific fine-tuning is performed. Kecht et al. (2021) achieved greater success with a zero-shot approach when they classified Apple and Spotify customers' query types such as account, product or delivery queries. One of the biggest differences between their approach and the work performed here is the difference in label objectivity. There is arguably less debate regarding whether a complaint is about a delivery issue than there is about whether a message from a potential predator reflects an instance of rapport-building. According to Shibata et al. (2018), correctly inferring categories based on social dimensions requires considerable contextual awareness.

One solution to improving performance would be to conduct more manual annotation. Consistent with similar work (Flor & Andrews-Todd, 2022), the behaviour with the greatest number of positive training examples (communication/coordination) generated the best overall performance. Given the consistent increase in few-shot settings, as the amount of available data for training gets larger, it could be inferred that generating more data would improve some of the worst-performing categories. However, increasing the number of positive training samples would likely be unsuitable for the rarest behaviours examined here. Except for 'risk management', no behaviour with less than 10% coverage exceeded their respective baseline in either the single or multi-message task. Assuming the distribution of behaviours is a fair reflection of prevalence at a population level, it would take considerable effort to generate even a modest amount of positive training examples. An alternative

approach is to exploit data augmentation techniques to artificially boost the number of data points (see Wei & Zou, 2019). However, this was not considered for the current work.

5.8.4 Limitations

Several limitations of the approach taken in this work are worth considering here. First, the formulation of the hypothesis statement was kept at the default setting as specified in Wolf et al. (2020). This resulted in hypothesis statements that, whilst grammatically correct and portraying some degree of semantic meaning, are unlikely to exist in any applied setting. As a reminder, the core objective in natural language inference is determining the probability that a sentence logically entails another sentence. For example, given the premise "some people kick a ball", a high probability of entailment should be conferred to the sentence "two men are playing a sport", but not to "the people are sleeping". Relying on the default format, "This is an example of: [LABEL]" when crafting a hypothesis sentence has likely resulted in hypotheses with limited contextual significance. Take 'risk management' as an example. An alternative hypothesis statement might have been "the person is concerned about getting caught". Semantically, this bears a clear relationship to messages such as 'when are your parents coming home?' in a way that "this is an example of risk management" does not. Indeed, whilst the default value has been used in prior NLI research (see Jin et al., 2022), studies have shown that adjusting the structure of the hypothesis sentence can alter classification performance (Kecht et al., 2021).

A second limitation concerns the use of only a single annotator to perform the labelling. Human labelling is a naturally subjective task, and when paired with a task such as this, likely to require considerable interpretation of the surrounding context. Because of this, coders are likely to differ based on their expertise and prior experience in annotating conversational data (Burnap et al., 2015). Human-derived labels represent the closest to objective ground truth. However, absolute accuracy should not be assumed purely because a domain expert performed the coding. Fatigue and boredom can cause inconsistencies within the annotations (Maclin & Maclin, 2005). The complexity of the coding scheme can also result in spurious labelling (Snel et al., 2012). Where possible, multiple raters should perform several passes of the data to reach a suitable level of agreement. This will also enable edge cases and other sorts of disagreement (i.e., definitions of a particular label) to be discussed and

resolved. As highlighted previously, a single-rater approach was used here due to the time-consuming nature of labelling. Steps were taken to verify the quality of the annotations, and a reasonable level of agreement was observed during post-validation. However, the extent that this serves as a substitute for a secondary rater is debated. Consequently, it is prudent to speculate that re-training the model with annotations generated by a secondary coder will likely alter classification performance.

5.8.5 Next steps

The safety of children when they are online is of utmost concern in many countries. However, the threat posed by online predators is one for which governments relying on traditional methods are ill-equipped to manage. Whilst automated detection of predatory behaviour allows practitioners to monitor many chats on a large scale, it is, as has been demonstrated here, currently an imperfect solution. Given the results obtained throughout this chapter, several potential areas of future research are suggested. First, refinement of the eleven predatory behaviours through further qualitative analysis would ensure greater objectivity. As part of this, re-coding the data with secondary and tertiary raters would also be valuable.

In the following chapter, attempts are made to address the overall precision of the model by reformulating the task as an information extraction problem. Whilst the current work is of value to those seeking to study predatory behaviour throughout an interaction, flagging the most predatory aspect of a conversation is arguably of greater use to practitioners.

5.8.6 Publications and contributions associated with this chapter

The work included within this chapter has formed part of a larger collaborative effort to improve the detection of online predatory behaviour involving minors. As such, the author would like to note several key contributions made by others. First, the creation of the qualitative coding scheme and labelling of the manual corpus was performed by a University of Liverpool postgraduate enrolled on an Investigative and Forensic Psychology Master of Science (MSc) course. In addition to its use in this chapter, the coding scheme was used as a central component in a qualitative analysis as part of the postgraduate student's MSc dissertation. Additionally, a

co-authored conference submission is currently available on ArXiv (see Cook et al., 2022) that reflects much of the content covered in this chapter. In addition to the contributions already highlighted, specifically those around creating and validating the coding scheme, co-authors provided critical feedback during the analysis and writing process.

A psychology-focused manuscript combining both this and the following chapter has been submitted to the Sixth Conference on Artificial Intelligence, Ethics, and Society (AIES) (see Cook et al., 2023). The author wishes to acknowledge similar levels of contribution from co-authors.

Ethical considerations

Given the topic covered in this chapter, ethical approval was sought for the data collection and labelling of the Perverted Justice corpus. Ethical approval was granted by the University of Liverpool's Institute of Population Health Research Ethics Committee (REF: 9972 – Lead PI Dr Susan Giles).

Chapter 6

Improving Performance of Automated Systems through Collaborative Human-AI

6.1 Introduction

In the previous chapter, a transformer model was used to learn representations of eleven predatory behaviours in a corpus of chat logs featuring predators interacting with adult decoys posing as children. Given the risks posed by online child sexual exploitation Child Sexual Exploitation (CSE), the ability to automatically identify predators is of considerable value. However, the results presented in Chapter 5 indicated that whilst a deep learning solution offered optimal performance when sufficient training data was made available, model precision remained an issue for most behaviours. Consequently, it was observed that several behaviours were over-predicted by the model when applied to the full corpus. This can result in lost time in high-stakes settings where a high degree of precision is required. This is especially true in cases of CSE, with time misspent posing a considerable risk to victims' well-being. Whilst it is, undoubtedly, more tolerable to misidentify non-predatory behaviour as predatory (i.e., lower precision) than to identify predatory behaviour as non-predatory (i.e., lower recall), considerable resources will be sacrificed if precision is consistently poor. It is, therefore, vital that automated systems address this.

Humans, by contrast to machines, are particularly good at extracting salient information with high recall and precision. They are, however, constrained by cognitive limitations that make identifying and extracting information a slow and tedious task. As discussed previously, humans are prone to fatigue (Maclin & Maclin, 2005). Moreover, although we do not typically think of humans as overfitting, the subjectivity of manual annotation means humans are subject to various biases (Hasan et al., 2016). Machines, by contrast, are highly-efficient at extracting salient information from unstructured data. However, they regularly make mistakes where a human would not (Rosé et al., 2008). In a high-stakes setting such as CSE, a collaborative system that combines the accuracy of human cognition with the efficiency of cognitive resources offers the best of both worlds. This is the primary objective of this final research chapter.

Consequently, Chapter 6 examines the trade-off between precision and processing speed on detecting predatory behaviour in cases of online CSE – a reformulation of the previous chapter that seeks to improve model performance through a *human-in-the-loop* approach. In Butcher et al. (2023), it is claimed that such a system should satisfy the following criteria: (1) the ability to produce accurate predictions, (2) to be more efficient than can be expected from manual effort, (3) to demonstrate the capability to reproduce results, (4) be flexible in its application, and (5) reduce both human fatigue and limit potential harm. Such a system could be deployed in a real-world setting and would serve as an alternative to traditional methods that are over-reliant on simplistic technologies and large amounts of manual processing. Therefore, the main objectives of this chapter are as follows:

- Describe a model of human-AI cooperation with the intention of extracting information from CSE chat-logs.
- Assess the utility of a human-in-the-loop approach towards towards improving model precision whilst maintaining a suitably high level of model recall.
- Examine the trade-off between precision and processing speed.
- Discuss the potential merits of human-AI cooperation on reducing human annotator fatigue and promoting well-being.

The remainder of this chapter is structured accordingly; first, the framework used to extract salient information from the chat logs is presented. This work is based on a weak supervision approach that automatically recommends segments of

the chat log for human verification. Following this, the main experimental set-up is described, including an overview of how the problem formulation differs from that presented in the previous chapter. After this, the next section presents the main results, followed by a discussion and recommendations for future work.

6.2 Weak-supervision for information extraction

The central objective of this chapter is to combine the accuracy of human judgement with the processing power of computation through a human-in-the-loop approach. To this end, a weak supervision approach is suggested, whereby the machine makes an initial prediction before being validated by a human (the author).

For this task, the current chapter utilises a tool known as `ELICIT` to perform the initial extraction (Butcher et al., 2023). `ELICIT` is a recently developed tool designed to perform information extraction tasks with high precision (i.e., reducing the number of false positives). For a given information extraction problem, `ELICIT` proposes a series of candidate answers that a human annotator can verify. In principle, this should increase precision to near-perfect levels, as the human can reject any incorrect assertions made by the machine. It was initially developed as an information extraction tool for written documents such as press articles and sentencing remarks in legal and law enforcement settings. Here, the desired information centres on extracting factual information from the text – i.e., *was the victim male or female?* Consequently, its ability to classify a more interpretative set of labels with conversational data is unknown.

6.2.1 Machine phase: information extraction

`ELICIT` can be broadly split into two constituent phases: an automated machine phase, where for a given label, excerpts of text are extracted from the unstructured data – in this case, the chat logs – and a human validation phase, where a domain expert verifies the suitability of the extracted text and label. See Figure 6.1 for an outline of the framework.

In the machine phase, segments of an input text are automatically identified as potential evidence of a given behaviour label. To extract evidence, therefore, `ELICIT` requires both an input text and a predefined list of candidate labels. This phase aims

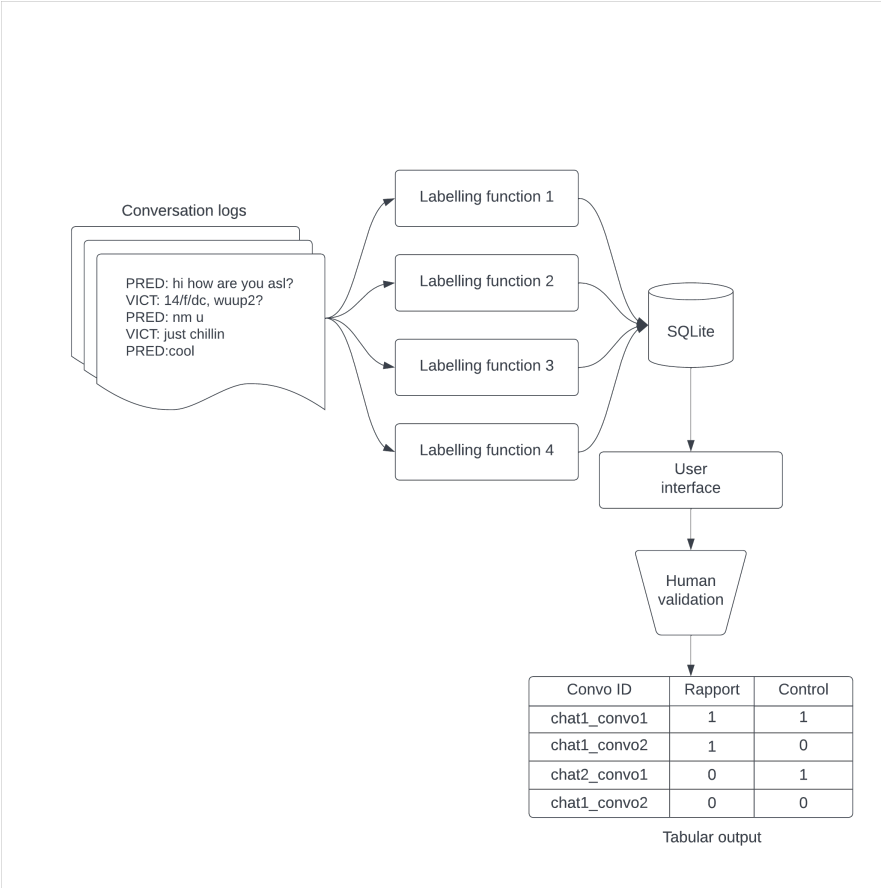

Figure 6.1: A high-level diagram of the ELICIT framework. For each variable extracted from the text, several labelling functions provide the user with possible answers and accompanying explanations. The user validates the correct option, creating a tabular dataset where each row corresponds to a conversation, and each column to an extracted behaviour label.

to identify potential evidence in the text that most suitably represents each label. To locate relevant evidence, `ELICIT` uses an ensemble of Labelling Function (LF). Each LF is pre-specified by the user. Examples of LF range from a simple keyword search, where the presence of a keyword in a text can be used as evidence, to sophisticated neural networks trained on large language models. Each LF extracts a segment of text with the greatest likelihood of representing each candidate label and stores it in a `SQLite` (Hipp, 2020) database for human review.

6.2.2 Human phase: verification

The second phase requires the human to decide whether the extracted evidence correctly represents a label (a true positive) or is incorrect (a false positive) and thus to be discarded. The user interacts with `ELICIT` through an API. For each

conversation, the user can inspect the evidence generated by each LF for every behaviour label (see Figure 6.2 illustrates this step). By selecting evidence for review, the user is provided with an expanded window containing a larger snippet of the conversation (see Figure 6.3). If the user deems the evidence a sufficient description of a behaviour label, they mark it as correct. Final ratings can then be extracted in tabular form, and precision, recall, and $F1$ can be analysed as usual.

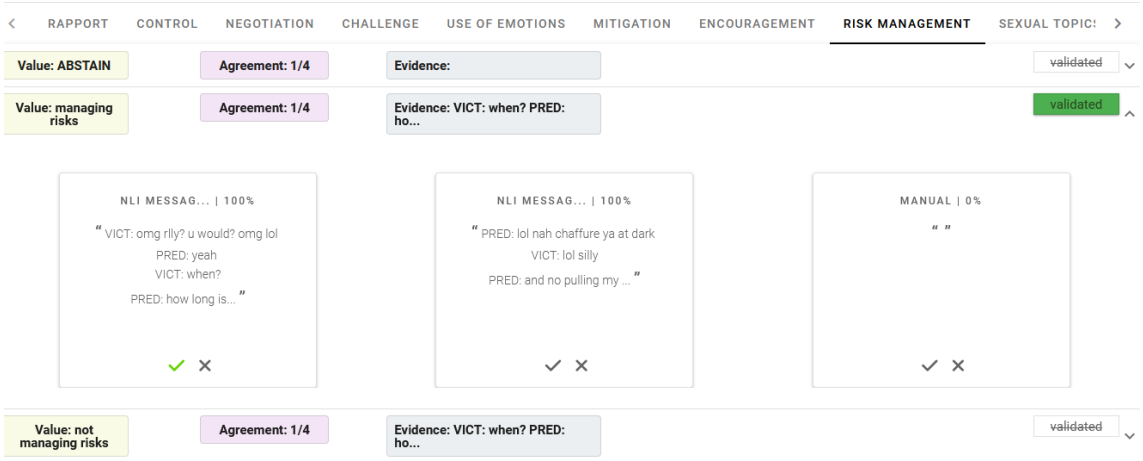

Figure 6.2: Example of ELICIT's user interface.

Figure 6.3: Example of the evidence generated by a labelling function.

6.3 Approach

6.3.1 Data source

This chapter uses the Perverted Justice corpus described previously in Chapter 5. As an information extraction task, the objective of `ELICIT` is to extract evidence relating to ten predatory behaviours[1]: (1) rapport-building, (2) control, (3) chal-

[1]Communication/coordination was removed due to its high frequency in the corpus.

lenge, (4) negotiation, (5) use of emotions, (6) testing boundaries, (7) use of sexual topics, (8) mitigation, (9) encouragement, (10) risk management. Whilst the previous chapter focused on a message-level classification, classification is performed at a conversational level. This was done to provide more information to the LF. The approach also possesses greater practical utility, as law enforcement is unlikely to analyse interactions on a message-by-message level and will instead seek to identify broad behaviour patterns. To parse chat-logs at a higher level, messages are grouped into *conversations* – a continuous sequence of messages where the maximum time between two messages did not exceed 60 minutes.

6.3.2 Labelling functions

An ensemble of five distinct LF was devised to meet the demands of this particular task. Text segments are extracted by each LF for each conversation and rank-ordered according to a confidence estimate. Segments of text with the highest confidence estimate are then returned to the user for verification. At this point, the user can either accept the text segment as representative of a given label or class or reject it as incorrect. This approach was influenced by previous work such as `SNORKEL` (Ratner et al., 2019). The primary difference here, however, is the inclusion of a human-centred verification step. Where methods such as `SNORKEL` fully automate the analytic pipeline, here, acceptance of a candidate text segment is performed by a human.

Labelling functions were designed to prioritise a high level of recall. Low thresholds were set to extract as much *potentially* useful information as possible. Consequently, the initial predictions made by each LF will be of low precision, at least before human verification. The motivation for this was because it was felt that correcting false positives made by the machine would be a more straightforward task compared to identifying missed opportunities (Yan et al., 2014). As such, it was expected that whilst precision on the unverified machine classifications would be low for all behaviours, performance would increase post-verification.

LF1: NLI sequence classifier

The fine-tuned version of `RoBERTa` from Chapter 5 was used for each behaviour label. As a reminder, the objective of NLI is to determine the likelihood that two sentences (a message and behaviour label in this case) are related. Each model

was trained on the development set of expert-annotated predatory messages. The wider, five-message window was used for each predatory message, as results from the previous chapter revealed this improved model performance ($F1$) over single-message classification for most behaviours. The classifier outputs ten confidence estimates for each five-message cluster, one for each label. Estimates higher than the established threshold are flagged as potential candidates for human verification. The decision threshold was set at 30%. After all message clusters have been assigned an estimate, they are ranked per behaviour, and the top K is returned to the user for verification.

LF2: Q&A → zero-shot classifier

An alternate version of `RoBERTa` fine-tuned using the Stanford Question and Answering Dataset (SQuAD2: Rajpurkar et al., 2018) was used to extract relevant sections of the conversation. For each behaviour label, a series of questions were generated. These questions were designed to extract a deeper understanding of the behaviour and were based on sub-categories generated during the initial qualitative coding effort described in Chapter 5. For example, the behaviour label "Control" comprised the sub-categories "patronising" and "confrontational", which could be formulated as a question – "was the predator patronising?" and "was the predator confrontational?" (A full list of questions devised per behaviour is included in the Appendix). Answers to these questions were short excerpts of the conversation the model deemed to best address the question asked, alongside a confidence estimate. Answers with an estimate below a certain threshold (30% in this case) were discarded, allowing the model to abstain. Conversely, those answers where the estimate was higher than the threshold were then submitted to a zero-shot NLI classifier (built on the same version of `RoBERTa-large-mnli` described in Chapter 5) to be compared with the appropriate behaviour label.

LF3: Q&A → cosine similarity

With a slight adjustment to the previous labelling function, answers are extracted from the same set of questions as described above. However, instead of a secondary transformer, `RoBERTa` embeddings for both the answer and the behaviour label are compared via cosine similarity. If the cosine similarity exceeds the stated threshold, it is returned to the user for verification. If not, the model abstains.

LF4: Message-cluster cosine similarity

A second cosine similarity measure is included based on embeddings between each five-message cluster and the behaviour label. Identical to the previous labelling function, a user-specified threshold is set. Message clusters with a similarity higher than the established threshold are returned to the user. If no similarity exceeds the threshold, the model abstains.

LF5: Keyword detection

An inductive approach was used to generate a suitable list of keywords for each behaviour. Keywords were derived from class-based TF-IDF trained using conversations within the development set. c-TF-IDF is calculated as follows:

$$c - TF - IDF_i = \frac{t_i}{w_i} \times log\frac{m}{\Sigma_j^n t_j} \tag{6.1}$$

Where t refers to each word in a corpus, w is the total number of words in the corpus, i indicates each class, n indicates all classes, and m refers to the number of conversation logs. There are two classes for each label that correspond to whether a conversation contains a message labelled with a behaviour or not (i.e., "rapport" and "not rapport"). As behaviour labels are considered independent, the same keyword may characterise multiple behaviours. To reduce noise, stop-words were removed before extracting keywords from conversation logs. The English set of stop-words in NLTK was used. See the Appendix for a full list of keywords per behaviour label.

To detect positive behaviours based on keywords, the following approach was taken. First, the set of all predatory words in a conversation was extracted. Second, the percentage of keywords in the predatory word set was calculated. If more than 1% of the predatory word set comprised a label's keyword list, it was deemed a positive classification for that label (with a confidence estimate of 100%).

6.3.3 Experimental setup

Dataset and Annotation

A test dataset was constructed based on $N = 67$ conversations obtained from the Perverted Justice corpus. In total, these conversations were taken from twenty-three randomly sampled chat-logs[2]. Predator messages were annotated in the same manner as before. However, in this chapter, annotation was performed by two post-graduate students enrolled on the University of Liverpool's MSc in Investigative and Forensic Psychology. In addition to this chapter, the annotations contributed to qualitative analyses as part of the students' respective MSc dissertations. Annotation was performed independently (i.e., not repeat coding) to boost the number of available data points. A short joint-coding exercise was performed to encourage a shared mental representation of each behaviour label to ensure consistency between annotators.

	Dataset		
	Annotator 1	Annotator 2	Total
Chat-logs	8	15	23
Conversations	19	48	67
Predator Messages	2328	10098	12426
Av. Msgs per Conversation	122.53	210.38	185.46

Table 6.1: Description of CSE conversation dataset.

In total, 12426 predatory messages were annotated for each of the ten behaviour labels. The average number of messages analysed per conversation was 185.46 ($SD = 188.77$). Conversations coded by annotator 1 were considerably shorter on average ($M = 122.53$) than by annotator 2 ($M = 210.38$). Inspection of the differences in message counts via a histogram revealed a large skew to the right, with several conversations approximately 2.5 standard deviations larger than the average. In all cases, annotation of longer conversations was performed by Annotator 2. Descriptive statistics for the dataset are presented in Table 6.1.

Evaluation

Top *k* Validation. To analyse the performance of the proposed approach, two conditions were defined based on the amount of evidence offered to the user: $k = 1$,

[2]Eight of the chat-logs also featured in the test region of the previous chapter. The remaining fifteen chat logs were randomly sampled from the unlabelled portion of the corpus.

and $k = 3$. As a reminder, `ELICIT` rank-orders extracted text segments based on its confidence level. By default, only the segment with the largest confidence score is offered as evidence ($k = 1$). For instance, if LF1 identifies a segment of text with a confidence score of 0.9, but LF2 identifies a separate segment with a higher confidence level, only LF2's evidence is output. An advantage of this approach is that it offers the user less material to verify, reducing the amount of human effort required. However, if the user deemed the evidence provided incorrect, accessing alternative candidates would not be possible. We might, therefore, expect a reduction in model recall in this case, as the absence of alternative segments of text to verify may increase the number of false negatives.

Alternatively, the second condition ($k = 3$) returns evidence for up to three labelling functions per behaviour label. This can mean a user can abstain from one text segment without rejecting the category entirely. Therefore, whilst $k = 3$ would increase the amount of human verification required by up to a factor of three, it could arguably afford the user greater flexibility. The optimum value of k is likely to be domain-specific and therefore reflects a user's willingness to accept a potentially inflated number of false negatives for improved efficiency. To examine the impact of this, performance metrics (precision and recall) are reported for each condition.

Time-Efficiency and Annotator Well-Being. In order to explore the time-efficiency of a collaborative human-AI approach, it was necessary to record the amount of time taken to perform the manual annotation. Unfortunately, this was not possible as part of the initial coding effort. Therefore, to measure the improved efficiency of the proposed approach, a segment of the dataset ($\approx 30\%$) was re-coded by the author of this thesis under timed conditions. Due to time constraints, this was not performed blind but with the visibility of the original annotations. This, undoubtedly, required less time and cognitive effort than would have been the case had the chat logs been re-coded from scratch. Thus, the timings generated from this exercise are likely much shorter than the case under optimal conditions. As an extension of this, the author also kept a diary of observations regarding mental fatigue and well-being whilst evaluating the conversations.

6.4 Results

6.4.1 Performance via top *k* validation

Precision

The first experiment in this chapter examines the difference in precision and recall when $k = 1$ compared to when $k = 3$. An illustration of performance over the two conditions is provided in Figure 6.4. Precision performance was generally high for both conditions, with two labels (Control and Negotiation) obtaining perfect precision. The average precision score over each behaviour was similar, with both conditions performing ≈ 0.94. The lowest performing behaviour was 'Challenge', which dropped by 13% (from 0.8 to 0.73) between the two conditions. This drop in performance is likely due to the $k = 3$ model incorrectly providing more information to the user to verify – thus, increasing the likelihood of incurring a false positive. Collectively, this finding suggests that a human-in-the-loop approach was able to consistently extract relevant segments of text for the user to review. High levels of precision also indicate a degree of robustness in the labels themselves. Only rarely did the user positively classify a text segment that had initially been negatively classified in the original annotation. Given the interpretive nature of the labels, imperfect precision indicates disagreement between annotators. A minimal difference was observed between the two values of k for precision, consistent with the author's initial assumptions.

Recall

However, poor performance was observed for most behaviour labels in terms of model recall. Low recall can be explained either as a consequence of inter-annotator disagreement (i.e., excerpts are provided but subsequently rejected by the user) or an inability of the model to identify salient information for a given category (i.e., the model fails to return anything for the user to verify). Supporting the latter explanation, providing more information for the user to review by increasing the value of k did improve performance for all categories. The average improvement in recall was 11%, from 0.52 when $k = 1$ to 0.63 when $k = 3$. The best-performing labels regarding precision were not necessarily the best-performing models regarding recall. For instance, whilst 'Negotiation' achieved perfect precision in both conditions,

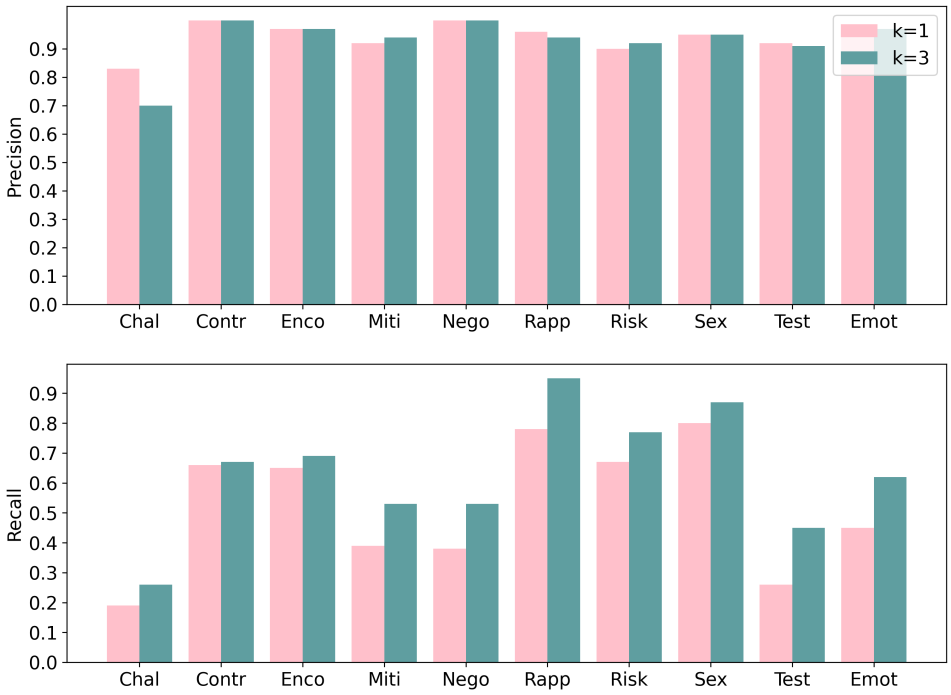

Figure 6.4: Precision and recall performance with either a $k = 1$ or $k = 3$ strategy for each behaviour label.

it only marginally exceeded a 0.5 chance effect when $k = 3$, and dropped to 0.38 when $k = 1$. Again, the lowest performing category was 'Challenge', with a recall score of 0.19 and 0.26 for $k = 1$ and $k = 3$, respectively. Coverage of 'Challenge' at the conversational level was 40% – making it an infrequent category compared to the other labels. A large number of false negatives (i.e., a model with low recall), particularly in cases where the true number of positives represents a minority class, suggest that the category is not suitably well-defined for this task.

Resolving Annotator and User Disagreement

Disagreement between the user and the original annotator is another reason for low recall. As a reminder, the user (the thesis author) was not responsible for the original annotations. A certain amount of disagreement is an expected consequence of subjective coding (Spooren & Degand, 2010). Consequently, expected disagreement between individuals may be masquerading as an inability of the model to extract salient information – artificially lowering recall as a result. Removing the effects of inter-rater disagreement would typically require manually re-annotating the entire

corpus to ensure consistency between the original annotations with those validated in ELICIT. Time restraints meant this was not possible in the present context. As an efficient alternative, the thesis author manually inspected the points of disagreement between the original annotations and the output from ELICIT. Cases where the author felt they would have disagreed with the original annotation were amended, and performance metrics were recalculated. Results on the amended annotations are provided in Figure 6.5.

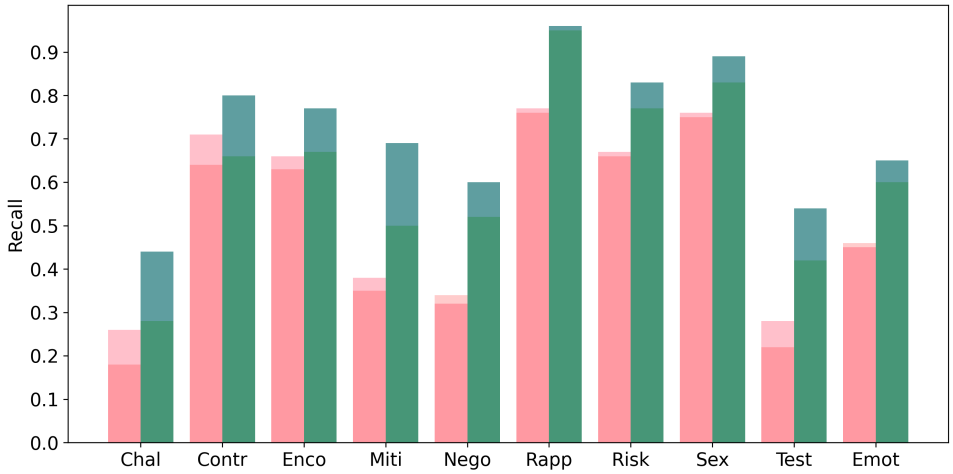

Figure 6.5: Annotator adjusted recall performance with either a $k = 1$ or $k = 3$ strategy for each behaviour label. The shaded area within each bar indicate non-adjusted recall

Consistent with expectation, adjusting points of disagreement between annotators improved model recall for most behaviour labels. Whilst it did not alter the overall interpretation of results, several behaviours attained high levels of recall in the $k = 3$ condition. In particular, recall for 'Mitigation' increased from 0.5 in the initial annotations to 0.69 in the adjusted version. Such a disparity in performance is further evidence that the labels lack objectivity and likely require refinement. Additional labels that performed well in the $k = 3$ condition include 'Control' (0.8), 'Encouragement' (0.77), 'Risk Management' (0.83), and 'Use of Sexual Topics' (0.89). As before, recall was highest for rapport-building, increasing marginally from 0.95 to 0.96. These findings broadly indicate that the model could extract salient text segments pertaining to the appropriate behaviour label. A substantial average increase was also observed between $k = 1$ and $k = 3$ from all conditions, increasing from 0.53 to 0.72. Furthermore, perfect precision was attained for all categories with the adjusted annotations. This indicates that a human-in-the-loop approach did not alter the user's perception of the behaviour labels – in other words, what

was interpreted by the user as 'rapport' using a wholly manual approach was equally identifiable via `ELICIT`.

Evaluating LF Performance

The large increase in recall between $k = 1$ and $k = 3$ highlights the utility of an ensemble approach. However, the size of the increase suggests that the labelling functions themselves may not be particularly well-calibrated to the current task. To examine this in further detail, Figure 6.6 reports confusion matrices for each LF irrespective of user input (i.e., accepting all suggestions made by the labelling functions as correct). Without user validation, a majority of true positives were produced by only two of the labelling functions: either the NLI sequence classifier or the keyword search – suggesting these LFs are the most useful in the current setting. These two labelling functions generated an average precision of .82 and an average recall of .86. This is perhaps unsurprising, given that these were the only labelling functions that received domain-specific training. Although the average precision of the remaining labelling functions was comparable (0.85), the average recall was substantially lower (0.14) – this suggests that, for most cases, no relevant information was detected via these methods.

6.4.2 Time-efficiency

The following experiment examined potential improvements in efficiency for the adopted method versus the time taken to generate annotations manually. Unsurprisingly, the time required to annotate a conversation manually depended on its length, with longer interactions (measured in words typed) taking considerably more time. On average, a conversation took just under 44 minutes to annotate (SD = 44 minutes 17 seconds). As a reminder, chat logs were re-annotated by the author using the original annotations as a guide. Based on the estimates provided by one of the original annotators, initial coding took considerably longer. The differences between manual and human-AI are likely a conservative estimate, with the actual difference likely to be much greater.

Timing differences between the manual and human-AI approach are illustrated in 6.7. Of the two values of k, validations took marginally longer to complete when k=3 versus when k=1. On average, validating a conversation took 2.45 minutes (SD = 1.08) where k=3, and 2.13 minutes (SD = 0.53) where k=1. This is to be

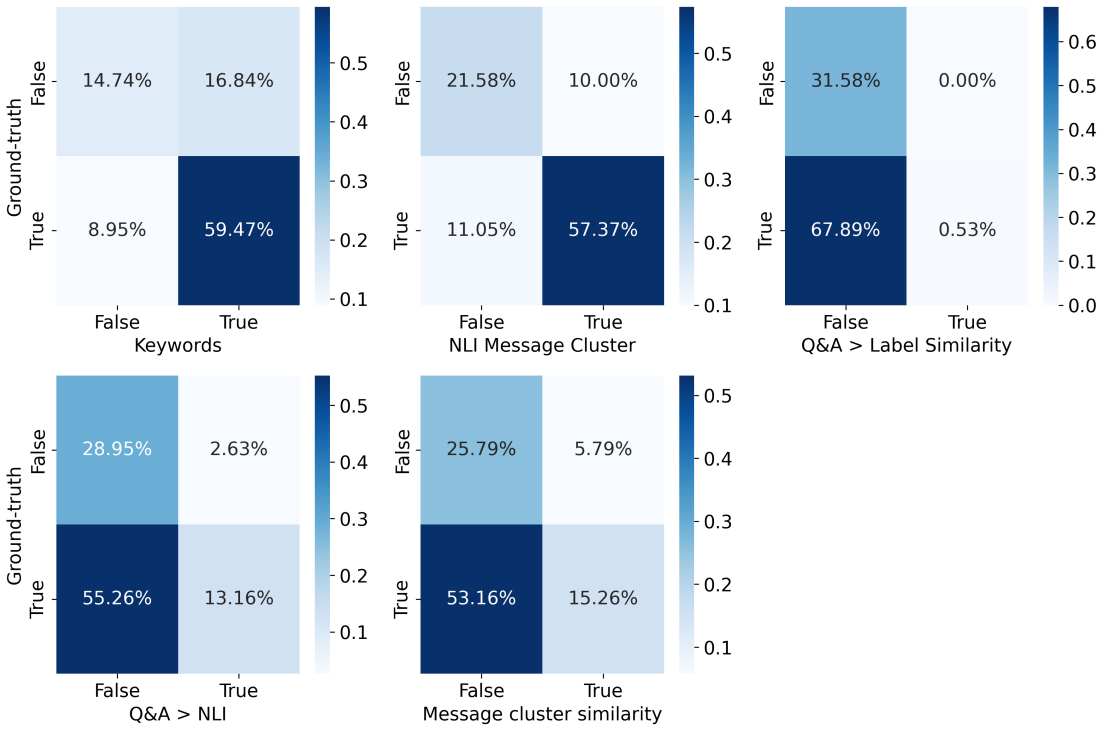

Figure 6.6: Confusion matrices on extracted segments for each LF where k=3. Based on the adjusted annotations. Extracted segments were unvalidated (i.e., they represent the raw output from `ELICIT` prior to user validation).

expected, given that k=3 increased the number of extracted segments the user could validate per conversation. The slightness of the difference also hints at a potential learning effect. As the user first undertook the $k = 1$ task, familiarity with both the software and the behaviour labels was reached. This likely increased the speed with which the user reviewed extracted segments in the $k = 3$ condition. Additionally, the length of time required when using `ELICIT` was independent of the length of the conversation. This suggests a human-AI solution may be preferable, particularly for more extended conversations.

6.4.3 Improving annotator fatigue and well-being

Finally, the user made several observations relating to annotator well-being throughout this work. One of the aims of the adopted human-AI approach is to minimise the fatigue often experienced by annotators and improve mental well-being by reducing the time spent with distressing content. Although no quantitative measures were collected to evaluate this formally, one of the main observations is worth highlighting.

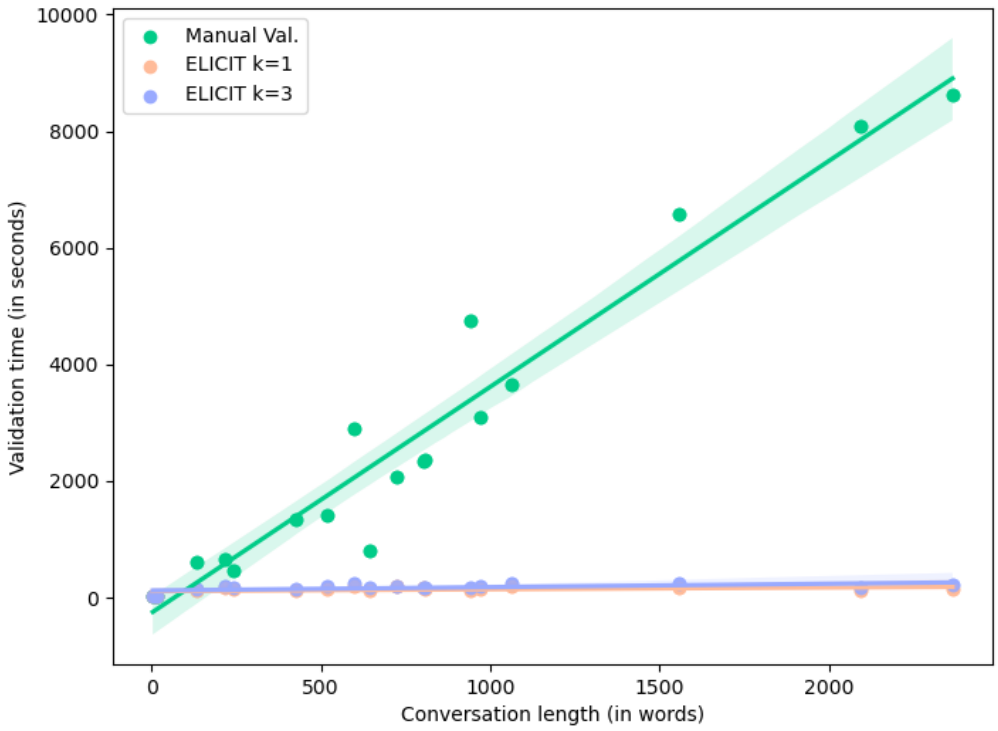

Figure 6.7: Comparison of time-efficiency for manual and human-AI approaches on the re-coded portion of the dataset.

Although the ELICIT interface does not prevent the user from having to read distressing content, it does significantly reduce the amount of immersion required. Manually annotating on a message-by-message basis requires a consistent focus on the flow of the conversation and, as such, requires considerable effort and deep concentration on the spoken content. As such, manually annotating becomes a highly intensive activity. The nature of the material also contributed to the degree of mental fatigue experienced when annotating manually. To maintain mental well-being, the author was limited to short periods of annotation, with more prolonged periods coinciding with a noticeable drop in mood.

By contrast, ELICIT offset much of this mental effort by providing only a snippet of the entire conversation. The impact of this was that the author could code at a more surface level by looking for generalised patterns within extracted segments. For example, the author found they become efficient at confirming the presence of 'risk management' due to the level of similarity between predators who engaged in this behaviour. Snippets of text such as "so when does your mom get home?" and "I could come round tonight if you're alone?" were regularly extracted by ELICIT and

proved easier to annotate over time. This is broadly consistent with research suggesting that validating machine-learnt positive classifications as incorrect is preferable to having to identify examples manually (Liew et al., 2015; Yan et al., 2014).

6.5 Discussion

Extending on the topic introduced in the previous chapter, this chapter has explored the utility of a human-in-the-loop framework for extracting salient information from chat logs. Grounding this work within the confines of the Perverted Justice corpus was intended to give real-world significance to the approach and address some of the performance limitations encountered in Chapter 5. This work was motivated by conversations between the author (as part of a wider research team) and domain experts. These conversations revealed that detecting online predatory behaviour remains a highly manual task conducted by a few experienced practitioners. The vast amounts of social interactions conducted online in the modern age, coupled with the subtlety of predatory techniques, make the likelihood of catching a single instance of predatory behaviour in real-time rare. This is highly problematic. At its most severe, predators seek to lure children into offline meetings, encourage them to engage in sexually explicit conversation online, and produce or share indecent images. Predators use a variety of behaviours to influence and groom their potential victims that are difficult even for trained experts to identify. Yet despite the inherent dangers, an absence of technological involvement in this area means many practitioners have come to rely on rudimentary tools such as monitoring a list of sexually explicit words or phrases[3]. Furthermore, practitioners typically monitor many conversations simultaneously, which increases cognitive load, and reduces both mental fatigue and well-being. To this end, the findings of the present chapter indicate that augmenting human effort through computation has the potential to reap considerable benefits.

The framework used in this chapter combines the efficiency of machines with the accuracy of human cognition. Given the considerable time and effort required to annotate a conversation manually, the results presented here suggest that human-AI collaboration may be particularly valuable when working with (a) very long texts, chat logs, or documents, (b) large corpora with limited manual resource, and (c) when there is a requirement for high-precision. The findings within this chapter

[3]The author's knowledge of the methods used in this area was based on a meeting in 2021 with the CEO of an online watchdog based in the UK (eSafe). This conversation helped shape the early groundwork in both this and the preceding chapter.

revealed that a human-AI approach could generate near human levels of precision at approximately $15x$ the speed of manual annotation[4]. Results suggest that this increases to closer to a $40x$ speed-up for longer conversations. Recall remained an issue for several behaviours. However, recall generally improved when more information was presented to the user. Determining how much information should be provided to the user is likely domain-specific and based on whether the main objective is efficiency or accuracy. However, the difference in time efficiency when $k = 1$ compared to $k = 3$ was minimal. This difference may become larger as higher values of k are considered or when more complex labelling functions are developed.

It was observed during the study that the adopted approach was considerably less cognitively demanding than manual annotation. An auxiliary reason this may have been the case is the focus on predatory behaviour at a conversational level instead of message-by-message. However, equally plausible is that reformulating the human role as confirmatory rather than exploratory lowers the level of immersion and consequently reduces annotator fatigue. A confirmatory approach, where the human confirms the prediction made by the machine, is likely to be particularly effective when a list of well-established behaviours is used. Conversely, an exploratory approach requires evaluating the entirety of the conversation.

Several limitations have been encountered throughout this work. First, of the five labelling functions developed for this task, only two appeared to perform capably. These were the only labelling functions that received any degree of domain-specific training, highlighting the task's difficulty. The considerable improvements in recall when $k = 3$ versus when $k = 1$ indicates that the user did not always deem the prediction with the highest confidence score relevant. As such, 'several bites of the cherry' were required from multiple labelling functions before relevant evidence was identified. This finding highlights the need to align the choice of LF with the domain of interest. A second limitation concerns the objectivity of the behaviour labels. Building on the work in Chapter 5, postgraduate students initially performed annotations rather than the user (author). However, the human-AI approach was found to perform poorly when predicting these annotations – this was deemed a consequence of disagreement between the user and annotator. Improvements in recall were observed when predictions were based on user-adjusted annotations. In future work, further validation of these behaviour labels is required to increase the level of annotations' agreement and subsequent objectivity. Performance improvement could be achieved through repeat coding or further refinement of the behaviour

[4]Based on the average timings for manual and k=3 annotations

labels.

6.5.1 Recommended changes

The findings from this chapter are an encouraging sign of the potential advantages of a human-AI approach to detect online predatory behaviour. Whilst several limitations have been identified, the general outcomes of this research support the fusion of technological capability with domain expertise. Several changes are recommended to improve the tool's utility based on the work presented here.

An advantage of the user verification step is that it can contribute to further model training. In Liew et al. (2015), active learning was used to attain incremental performance increases from user validations of an initial set of predictions. The validations performed in the current work could be used as training input in a secondary round of re-training. Over time, the amount of human-verified data would grow such that a considerable performance gain can be observed.

A second potential change could include human validation only during the initial training phase (see Boecking et al., 2020, for an example). This would enable expert knowledge to be refined and embedded within the labelling function, after which labelling is performed automatically. This affords further time savings compared to validating the classifications manually, as in the present chapter. Extending this idea could also involve outsourcing the entire pipeline to an automated solution. This would be equivalent to a $k = 0$ condition in the present work. Fully computational approaches to information extraction is an area of active research (Adnan & Akbar, 2019). As a tool, it affords the user maximum efficiency, as determining the correct classification by a series of labelling functions is performed automatically. However, even industry-leading solutions such as `SNORKEL` (Ratner et al., 2019) struggle when labels are not sufficiently defined.

6.6 Chapter summary

The development of human-AI solutions to analysing online conversational data affords considerable advantages to studying behaviour in a real-world context. Extending on the work explored in the previous chapter, the work presented here examines the trade-off between human accuracy and computational efficiency. The

main findings support a collaborative approach between experts and AI to address real-world problems.

6.6.1 Acknowledgements and associated publications

Part of the work contained within this chapter has been the subject of a collaborative manuscript submitted for publication (see Butcher et al., 2023) – of which the author of this thesis is a co-author. An additional manuscript based on this chapter is also being prepared in collaboration with colleagues from the University of Liverpool's School of Psychology. The author wishes to acknowledge the contribution of Brad Butcher for creating the `ELICIT` API and defining several labelling functions used in this work. The author also acknowledges the contribution of the two postgraduate students at the University of Liverpool who performed the initial annotation.

Chapter 7

General Discussion

The central aim of this thesis was to examine the suitability of a computational approach to behaviour coding within the social sciences. The techniques used have been adopted from computer science fields, including signal processing, machine learning, and natural language processing, with domain knowledge from social science areas such as social and forensic psychology. The main findings obtained in the thesis support the inclusion of computational effort to scale social research but emphasise the value of domain knowledge throughout the analytic pipeline.

Following an overview of modern approaches to this problem, presented in Chapter 1, the thesis is broadly split into two parts. The first part explored automated detection of behaviour mirroring through a Social Signal Processing (SSP) framework (see Vinciarelli, Pantic, et al., 2009; Vinciarelli, Salamin, et al., 2009). This involved defining behaviours of interest directly from the source data (i.e., text or audio) without reference to human-based coding. Chapter 2 discusses the strengths and drawbacks of several algorithms that have been applied to this task and explores the social scientific literature that posits a pro-social response to mirroring in everyday conversation. The utility of verbal mirroring as a communicative strategy is then formally tested in Chapter 3. The findings from this study motivated Chapter 4, which applied a similar approach from a non-verbal (i.e., acoustic) perspective. The results obtained from these studies support the use of an SSP approach in behaviour coding. However, social scientists are typically interested in behaviours at a higher level of abstraction. Consequently, reliance on low-level features to model behaviour would be difficult to operationalise in many contexts.

To address this limitation, the second part of the thesis takes an alternative approach based on deep learning. An advantage of this is that the expert can state the behaviours of interest without explicitly defining them algorithmically. Instead, a small annotation sample is used as training data, which can be applied to a larger corpus. The utility of this approach was examined throughout Chapters 5 and 6. The results obtained during this part of the thesis suggest that deep learning can be used to detect latent behaviours in text, although imbalanced data can lead to over-prediction in rare events. The balance between accuracy and efficiency was then explored in Chapter 6, which emphasised the utility of human deferral in high-stakes settings. In summary, the contributions offered by the thesis include: (i) understanding whether automated labelling techniques could be used to code social behaviours in naturalistic settings reliably, (ii) identifying and examining the trade-off between accuracy and efficiency when working with large conversational datasets[1]. Additionally, `CoPyCAT`, the package used to detect verbal mirroring between two speakers automatically, has been made publicly available[2]. The remainder of this concluding chapter is structured accordingly:

1. Summarise each of the previous chapters covered during the thesis.
2. Discuss the main findings of the four research chapters and their implications for future cross-disciplinary study of social interactions.
3. Critically evaluate the merits and limitations of the thesis.
4. Describe future steps in this area based on the main findings.

7.1 Chapter overview

Chapter 1

Chapter 1 provided an overview of the challenges associated with traditional social scientific study of conversation and how a cross-disciplinary approach can help overcome some of these issues by adopting an array of computational technologies. The chapter also highlighted the promise of automation and the potential barriers to increased cross-disciplinary effort between the social and computational sciences.

[1]Large, in this sense, is interpreted from a social science perspective. The author acknowledges that typical computer science research tends to use datasets several orders of magnitude larger than the corpora used here.

[2]https://www.github.com/cookie1986/CoPyCAT

Chapter 2

Having contrasted the advantages and barriers to increased automation in social research within the previous chapter, Chapter 2 describes a Social Signal Processing approach to analysing conversational data. To address a lack of standardised methods in this area, a Python package named `CoPyCAT` was introduced. Using low-level linguistic behaviour, the package is designed to extract features pertinent to effective dialogue based on a social psychological theory known as Communication Accommodation (CAT: H. Giles et al., 1991).

Chapter 3

Building on the previous chapter, Chapter 3 explored how a Social Signal Processing approach could be used to predict the conversational quality of an interview. `CoPyCAT` was used to extract verbal mirroring behaviour from interviewers in a corpus of dyadic political interviews. These features were then used to predict four verbal success measures based on the interviewee. This chapter's findings justify using computational techniques to analyse interview data at scale and motivated exploration of behaviours beyond a linguistic dimension.

Chapter 4

Encouraged by the performance of the previous chapter, Chapter 4 explored interviewer mirroring from a non-verbal perspective. Three types of behaviour (proximity, synchrony, and convergence) were modelled from the acoustic behaviour of the interviewer in a semi-automated approach. Three experiments were then performed. The first experiment validated the presence of non-verbal mirroring towards the interviewee. In the second experiment, differences in interviewer mirroring were examined based on differences in (a) gender composition and (b) partisanship. The final experiment in this section explored the performance of a non-verbal and multimodal model on predicting the interview's conversational quality, extending the work undertaken in the previous chapter. These findings indicate that the interviewer's non-verbal mirroring also influences the interview's conversational quality. Despite these findings, the limited interpretability of low-level behaviours means the knowledge gained from such studies is difficult to operationalise.

Chapter 5

Chapter 5 used an alternative approach based on a deep learning architecture to improve feature interpretability. The chapter also transitioned from effective interviewing as the domain of interest to detecting predatory behaviour online. Natural Language Inference was used to automatically learn a representation of eleven predatory behaviours using a small training set of expert-annotated chats between known child groomers and decoys posing as underage victims. Performance on a hold-out set indicated that several behaviours could be reliably predicted. However, the imbalanced distribution of infrequent behaviours resulted in over-prediction.

Chapter 6

Seeking to address some of the limitations encountered in the previous chapter, Chapter 6 explored the trade-off between human accuracy and computational efficiency through a human-in-the-loop approach. In this chapter, the problem of online predator detection is reformulated as an information extraction task. Weak supervision was used to recommend segments of text for human verification automatically. The findings of this section illustrate that a combined human-AI approach can offer the best of both worlds, leading to near-human levels of accuracy with an approximately 15x speedup compared to manual methods alone.

7.2 Summary of main research findings

The outset of this thesis posed several research questions that the intervening chapters have sought to address. In this section, the main findings of each of these questions will be presented.

7.2.1 Can computational techniques be used to overcome limitations of expert labelling?

The main contribution made during this thesis has been examining how computational approaches can be used to overcome the limitations associated with wholly manual processing. One of these major limitations of traditional qualitative coding is the minimal *scalability* of research when behaviour coding is performed by

hand. In Chapter 1, it was observed that while a scarcity of data has historically characterised social science, the rise of digital media has presented opportunities for researchers to work at previously unfathomable scales (Grimmer et al., 2021). However, the considerable time and effort required to perform traditional coding suggest that much of this newly accessible data has remained underutilised.

The manner in which computational effort has been leveraged throughout this thesis has meant that the datasets used are considerably larger than is typical in social science. In Chapter 3, the $\approx$ 700 interviews are far in excess of similar research that has sought to explore behaviour during a political interview. For instance, Huls and Varwijk (2011) used a corpus of just 12 transcripts to explore political bias in talk-show format interviews. Another study sought to classify the behaviour of interlocutors in a corpus of 26 interviews with UK political party leaders (Waddle & Bull, 2020) – less than 4% of the corpus explored in Chapter 3. Whilst these studies differ in their approach from that explored here, both used a formalised coding framework. As such, the techniques deployed throughout this thesis indicate how similar approaches could have been used to increase the amount of data explored. An advantage of working at scale is detecting increasingly subtle yet potentially meaningful behaviours. This was illustrated by the general drop in prediction performance between Chapter 3, and Chapter 4 – the latter representing $\approx$ 15% of the transcripts included in the former.

Similarly, Chapter 5 generated labels for the entire catalogue of chat logs on the Perverted Justice website by fine-tuning language models with domain-specific content labels. Coincidentally, the number of chat logs in the entire corpus was almost identical to the number of interviews at $\approx$ 700. As before, qualitatively-focused social scientific studies of online predatory behaviour have typically utilised only a fraction of the available data. For instance, Egan et al. (2011) used 20 randomly selected chat logs in a language analysis of predatory behaviour. A mixed-method approach, combining expert content analysis and an automated LIWC based psycholinguistic analysis, used 65 chat-logs (Broome et al., 2020). Whilst these studies are broadly equivalent to the development set deployed in Chapter 5, the ability to quickly scale this work to the remaining $\approx$ 600 unlabelled chats represents a clear advantage of an automated approach. Even amongst studies that utilise automated text mining to detect predatory behaviour, it is more common to extract only a subset of the available data (Chiang et al., 2021; Gunawan et al., 2018; Kontostathis et al., 2010). To the author's knowledge, no prior study has sought to identify predatory behaviour across the entire Perverted Justice dataset.

A related limitation associated with a manual approach is the reduced mobility to apply analysis to novel domains. For the most part, formalised coding schemes are highly context-dependent. This means they are tightly entwined with the domain in which they originated. As a consequence of this, they possess limited utility outside of their original setting. Even simplistic computational approaches, such as dictionaries, have been criticised as offering minimal use beyond their initial application (Scharkow, 2013). Moreover, due to the inefficiency of manual approaches, coding frameworks are often tied to the decisions made at their inception. In other words, they offer limited refinement opportunities as new theories or contrasting viewpoints emerge. Imagine a situation whereby a new code is identified only toward the end of the coding effort. Researchers, in this case, are faced with one of two choices: drop the code entirely, or re-start the annotation effort. The inverse of this problem was encountered during Chapter 5, whereby the high frequency of the 'communication/coordination' label arguably justifies breaking it down into several distinct behaviours. However, given the time-intensive nature of the manual coding effort, this was deemed unfeasible.

Contrasting a wholly manual approach, the framework defined in Chapter 2 offers considerable re-usability and is intended to be flexible to new domains. Although it was applied to political interviewing in Chapter 3, the full feature set would be equally appropriate across many other dyadic settings. It would be interesting, for instance, to examine how well these features perform in other domains such as police interrogations (see Alison et al., 2013). Additionally, the choice of verbal mirroring algorithms is not intended to be an exhaustive account of all methods in this area. Other approaches that have been used with success in this area are described in Doyle et al. (2016), Carrick et al. (2016), and Boghrati et al. (2018). The work conducted in this thesis section can be easily re-aligned with these alternate methods by leveraging computational effort that does not rely on first performing manual annotation. In other words, it does not require a domain-specific training set.

Another limitation of a manual approach is the cognitive demands placed on annotators. Coding unstructured data such as a transcript is a time-consuming and laborious task (Schneider & Pea, 2015). It also requires a high concentration level, leading to fatigue and frustration. These factors combined can threaten the accuracy of coding efforts, with studies suggesting that prolonged coding can lead to spurious labelling (Snel et al., 2012). Fatigue effects can also lead to within-annotator inconsistencies (Karlgren et al., 2020). This presents a problem for particularly long-form

interactions, such as those explored in this thesis, as inconsistent allocation of codes can result in reduced accuracy when examining more extended interactions (McGhee et al., 2011).

By contrast, a Social Signal Processing approach, such as that used in Chapters 3 and 4, can generate more accurate and less biased data on a large-scale (Liebregts et al., 2019). This is because the allocation of a code to a particular data point is determined by a set of predefined rules. Using an SSP approach can also identify more complex behaviours than would be possible by human effort. For example, several studies examining behaviour mirroring have often used perception-based judgements to infer similarity (H. Giles, 1973; Pardo, 2006). This, however, overlooks speakers' tendencies to align on some behaviours whilst diverging on others (Danescu-Niculescu-Mizil et al., 2011). It would not be possible for human annotators to simultaneously monitor multiple behaviours in a mirroring study, with research instead relying on a Gestalt 'felt sense' to capture perceived similarity (Alison et al., 2013; Dunbar et al., 2014). By contrast, the high-dimensionality feature-set described in Chapter 2, and deployed in Chapters 3 and 4 enabled detection of mirroring at an increasingly granular level.

Lastly, the reliability of the generated annotations is another area improved by an automated approach. This is partly influenced by the ability to perform analysis on a larger scale, enabling the detection of subtle differences that would be impractical through manual methods alone (Yarkoni & Westfall, 2017). Defining behaviours of interest through feature engineering, as described in Chapter 2, ensures a level of unmatched consistency when using human annotators impacted by fatigue. Similarly, whilst the machine learning algorithms utilised in Chapter 5 rely on a degree of randomness, using seed values ensures a reproducible output.

Therefore, the main findings throughout this thesis generally support the claim that an automated solution can substantially improve manual effort in several ways. Scalability, complexity, and adaptability are all advantages enabled by offsetting the modelling and detecting behaviours to a machine. However, a claim that all behaviours can be equally automated is misleading, with problems encountered throughout this thesis, particularly in aspects of Chapter 4 and Chapter 5. The potential limitations encountered by automated effort are the focus of the next section.

7.2.2 What tasks are more or less suited to a computational solution?

In Lazer et al. (2014), the authors warn against big data hubris – readily accepting that a big data approach will outperform traditional methods simply due to the difference in scale. Indeed, findings throughout this thesis indicate that one cannot assume that the annotation of every behaviour can be reliably offset to a machine. For instance, the findings from Chapter 5 demonstrated considerable difficulties in detecting infrequent behaviours. This contrasts the performance obtained for behaviours that appeared more frequently in the corpus, which was generally better quality. A likely reason for this disparity is the number of positive examples available during training. In a stratified data split into training and testing regions, behaviours rarely appearing in the corpus will also appear rarely during training. In the case of abstract or latent behaviours, insufficient information may be available to form an adequate representation within the chosen language model.

Zero-shot learning was used in an attempt to overcome the lack of labelled data (see Huang et al., 2021). However, performance for most behaviours was substantially worse when no training data was provided. This finding is at odds with previous work that observed a high level of performance when classifying sentences with topical labels such as *science, art,* and *travel* (Sainz & Rigau, 2021). Likely, the latent nature of the behaviours used in Chapter 5 is the reason for the poor performance. In previous research, abstract concepts have often proven difficult to detect under zero-shot conditions (Srivatsa et al., 2022). Thus, the findings from this section suggest that it is essential to generate a sufficient body of positive training examples, mainly when labels are not well-defined.

Addressing a significant class imbalance for rare events is non-trivial via supervised machine learning (Chandola et al., 2009). The extensive time and effort required to generate labelled data for large-scale datasets mean creating larger training sets without incurring high costs is usually impossible. As an alternative, methods such as oversampling or undersampling have been used to reduce the imbalance by adding or deleting data as appropriate. Unfortunately, these methods have been shown to increase the risk of model overfitting and alter the underlying distribution of infrequent labels (Yap et al., 2014).

One approach to address class imbalance without oversampling or undersampling is cost-sensitive learning methods. Cost-sensitive learning assigns different

misclassification costs to different classes, making the classifier more sensitive to rare events. This approach can help the learning algorithm to focus on the minority class without the need to manipulate the original dataset (Thai-Nghe et al., 2010).

Another viable option is ensemble methods, such as bagging and boosting. Ensemble methods combine multiple learning models to improve overall performance. Bagging, for instance, generates multiple bootstrapped samples from the original dataset and trains a classifier on each of these samples (Breiman, 1996). The final classification is then based on the majority vote of these classifiers. Boosting, conversely, iteratively trains a series of classifiers by focusing on the instances misclassified by the previous classifier, thereby increasing the weight of misclassified instances (Freund & Schapire, 1997). This iterative process allows the classifier to pay more attention to the rare events, improving the classification of the minority class (Feng et al., 2022).

However, it is essential to note that some classification tasks may have fuzzy boundaries between positive and negative classes, as observed in those behaviours identified in Chapter 5. Disagreement between human annotators can indicate that the classification problem is more complex than initially anticipated. As the primary purpose of the coding framework introduced in this thesis was qualitative research, finding that the identified behaviours did not immediately lend themselves to automation is not surprising. Instead, it highlights the distance still to go before language models reach human levels of understanding. It also serves as a cautionary tale of assuming a "plug in and play" mindset irrespective of context (Lazer et al., 2014). Further refinement of the eleven predatory behaviours is likely required before the performance in an automated analysis can be significantly improved. The next steps in this area could include other qualitative exploration or statistical approaches such as principal components analysis.

Whilst the ultimate goal of this thesis was to automate behaviour coding, Chapter 6 highlighted the value of including expert domain knowledge in cases that call for high levels of precision. Detection of predatory behaviour is one such domain due to the considerable risk posed to victims. Despite advances in NLP over the last three decades, computers continue to fall short of humans in accuracy on many tasks that call for interpretation of language (Grishman, 2019). The change in focus from message-level classification to conversation-level information extraction prevents a direct comparison between the performance metrics in Chapters 5 and 6. That notwithstanding, a general interpretation of the two chapters indicates an improvement in the precision metric when the potential for human deferral was

incorporated.

7.2.3 Do machines perform comparably with humans?

A third research question explored by this thesis considers whether the quality of automatically generated annotations approximates the performance of a human annotator. Usually, in supervised machine learning, the quality of a classifier is based on how closely it performs to 'ground truth' (N. C. Chen et al., 2018). Ground truth is assumed to represent a degree of objectivity. By training models on portions of this data, performance can be evaluated by standard metrics such as accuracy, precision, and recall. A good model is one where predictions overlap considerably with ground truth, whereas a poor model predicts inconsistently or with little apparent relation to the gold standard labels. In other words, the validity of ground-truth labels is not debated. This assumption, however, becomes less stable when deriving ground truth from features based on human judgement, which is both subjective and inconsistent (Chuang et al., 2015). Two human annotators, equally trained but with different backgrounds, attitudes, and experience levels, are likely to perform coding differently. Consequently, training models on each annotator will likely generate different performance levels.

This issue raises questions regarding the validity of annotations generated by a machine. As noted by Rosé et al. (2008), machines are not replicating the cognitive processes performed by humans when classifying data. This means the machine will likely make mistakes where humans would not. In supervised machine learning, models can overfit the training data (Grimmer et al., 2021). This was observed in several of the predictions made in Chapter 5, where the classifier incorrectly leveraged simple linguistic rules. An example of this was predators' frequent use of the stylised 'kool', routinely misidentified as an indicator of 'Encouragement'.

Complimentary to conventional metrics, a supplementary method for assessing the quality of automatically generated annotations is considering the machine as an independent annotator. From here, it is possible to measure the level of agreement between human and machine annotations (Anjewierden & Gijlers, 2008). Inter-rater reliability measures can also be used to measure the level of agreement between different algorithms as long as it is tested on the same data (Araya et al., 2012). Additional reliability metrics not explored in this thesis include examining *within-rater* consistency through intra-rater measures (Espinoza et al., 2019). Ideally, measuring agreement requires several passes of the data by different human

annotators to determine the between-rater agreement level. This can subsequently serve as a benchmark for pairwise human-computer comparisons and give an average agreement score across all pairwise comparisons. An acceptable level of performance would be inferred in cases where minimal discrepancies between humans and machines are observed when the agreement between multiple humans is high (Rosé et al., 2008). Measures of human-computer agreement were tested in Chapter 5 between the original ground truth labels, post validations performed by the author, and the automatic annotations. The main findings support the claim that machine annotations were broadly consistent with a human annotator, achieving at least a 'moderate' agreement score for most behaviour codes. However, the level of observed human-computer agreement was typically not as high as the agreement between the two human annotators.

7.2.4 Can an automated solution extend domain knowledge and advance theories of social behaviour?

A fundamental question asked at the outset of this thesis was how automated technologies could advance knowledge gained from social research in a conversational setting. Recent reviews by Yarkoni et al. (2021), Adjerid and Kelley (2018), and Lazer et al. (2020) have highlighted the opportunities enabled by increased computational effort within the social sciences. These included: (1) standardisation of methods, (2) open access to data, (3) addressing real-world problems, and (4) increased complexity of research. The main findings presented throughout this thesis have sought to contribute towards each of these. For instance, Chapter 2 sought to combine several prominent approaches to verbal mirroring into a single Python package. A growing body of verbal mirroring work has been characterised by disarray, with overlapping and contrasting methods leading to conflicting research findings (Healey et al., 2010; Nenkova et al., 2008; Rahimi et al., 2017; Reitter & Moore, 2007). Combining multiple methods into a single, easily accessible package was intended to allow future researchers to compare and contrast different verbal mirroring methods on the same data.

As has been highlighted elsewhere in this chapter, access to data is something that until recently has presented a significant challenge to many within the social sciences (Grimmer et al., 2021). Whilst the internet and digital media represent vast opportunities to those with the appropriate skill set, it is generally the case that those within qualitatively-focused fields lack the necessary training to take advan-

tage of such computational resources (N. C. Chen et al., 2018). As such, providing easy-to-use tools and appropriately formatted datasets can lower the barrier to access (Yarkoni et al., 2021). With this in mind, whilst the unstructured transcripts that formed the basis of both Chapter 3 and Chapter 5 remain publicly available, extensive cleaning and formatting were required to transform them into an organised tabular structure. Both datasets have been made freely available alongside supporting meta-data to support and encourage future research in this area.

Each use-case explored throughout this thesis was selected because they represent essential real-world problems. In the case of Chapter 3 and Chapter 4, accurately measuring interviewer performance through automation presents opportunities to improve evaluation and training in many professional interviewing settings. Although the focus here was on political interviewing, the intention was to develop a flexible approach that could be readily applied to new domains. One of the primary motivations of this section of the thesis was Alison et al. (2013), who, as part of ongoing work with UK and US law enforcement, continues to train professional police interrogators in effective communication techniques. However, heavy reliance on manual annotation remains a significant barrier to continuous learning for participants. Similar concerns have been raised in other domains, indicating that this is a widespread problem that extends to many professional settings (Flemotomos et al., 2018; J. Gibson et al., 2015; Naim et al., 2018). In many cases, once the initial training has been completed, there is little opportunity to receive ongoing feedback on these newly acquired skills.

By contrast, the automated evaluation performed in Chapters 3 and 4 could, theoretically, form part of a tool that monitors and evaluates interviewer performance for continuous skill development (Hirsch et al., 2018b). However, several amendments would first be required to automate the pipeline fully. In particular, manual transcription and audio diarization would need to be reliably offset to a computational solution, as the existing manual approach would serve as a considerable bottleneck in practice. The prospect of both automatic speech recognition (ASR) and automatic diarization (see Giannakopoulos, 2015) was explored in the development stages of Chapter 4. However, it was found to perform poorly on the extracted audio recordings. The successful fusion of ASR with conversation analysis in prior research illustrates the potential in this area (Moore, 2015). However, generating a transcript of sufficient quality irrespective of the audio quality represents a challenging problem within the signal processing space (Narayanan & Georgiou, 2013).

Equally, the second use-case featured in this thesis, detecting online predatory behaviour, is also an area where a computational approach addresses a real-world need. Given the potential risks posed to victims manipulated into an offline meeting with a predator, detecting known risk markers in real time is of immense value. Prior research has used traditional supervised machine learning to detect distinct phases of the conversation (Gupta et al., 2012; Kontostathis et al., 2010; McGhee et al., 2011). However, as these approaches rely on acquiring many labelled instances as training input, it is debatable how effective they would be in a fast-paced environment. Again, generating a sufficient body of manually annotated training data is a considerable limitation to real-world deployment. The main findings from Chapter 5 suggest that state-of-the-art deep learning solutions could exceed traditional approaches with as few as 50 labelled instances – potentially a considerable cost saving. However, given the overall level of performance obtained even when a large amount of training data was provided, further refinement would likely be required before the proposed approach could be utilised to detect predatory behaviour in real time. The extended findings from Chapter 6 go some way to improving performance and highlight that sacrificing efficiency by including a human deferral step can lead to improvements in accuracy. The extent that this trade-off is deemed acceptable in practice is highly likely to be domain-specific.

Finally, the work comprised within this thesis highlights the additional complexity afforded when the detection of behaviour is outsourced to a computational solution. The desiderata outlined in Chapter 2 would be infeasible to adopt via manual methods alone. As mentioned, early methods for detecting mirroring were based on human judgement (H. Giles, 1973; H. Giles et al., 1991; Pardo, 2006). Moreover, whilst text-based approaches have been used increasingly over the last decade, most research has tended to regard mirroring as unidirectional and not account for differences between speakers. Conversely, the findings presented in Chapter 3 revealed the importance of divergent behaviours in addition to convergence. The inclusion of computational technologies, therefore, enabled a more granular level of analysis than is conventional in prior studies of political interviewing (see Heritage, 1985; Pluss, 2010; Waddle & Bull, 2020, for some examples cited previously in this thesis).

7.3 Recommendations and limitations

The current section of the discussion describes some of the methodological limitations encountered during this thesis and presents recommendations on how these

might be resolved in future work. For clarity, this section is divided between the two case studies.

7.3.1 Chapter 3 and Chapter 4

This section of the thesis focused on automatically predicting the conversational quality of a political interview from the interviewer's behaviour. The motivation for this case study was the considerable effort required to analyse conversational data through human effort alone. As Chapter 3 focused on verbal behaviour, interview transcripts were the primary data source. Transcription was performed manually, however, was undertaken independently of this thesis. While this represented a considerable time saving for the author, using secondary transcripts brings several limitations. First, it was not possible to quantify the time and effort that would have been required to generate the large corpus of political interviews. As such, the claim of increased scalability through automation is somewhat weakened without estimating the time required to perform the analysis manually.

Second, secondary transcripts meant the author had very little control over what was included in the transcription. Details included within a transcript vary based on the type of transcription performed. A transcript generated by conversation analysts typically includes much more detail than that intended for record-keeping purposes (Moore, 2015). The corpus generated in this thesis comprised orthographic transcripts, which are human-readable records of *what was said and by whom*. Orthographic transcriptions are formatted with appropriate capitalisation, punctuation, and correct spelling (Bonsignori, 2009). Conversely, prosodic transcription includes false starts, repetitions, fillers, disfluencies, filled pauses, and back channels. They are difficult for humans to read but are more in keeping with conventional conversation analysis as described by Sacks et al. (1974). Prior computational work has sought to automatically detect some of these prosodic behaviours (see Dideriksen et al., 2019; Fusaroli et al., 2017). However, the absence of a suitably annotated dataset rendered this line of research beyond the scope of the present thesis.

The substantial time and effort required to generate a suitably sized corpus mean that assessing the feasibility of automatic speech recognition (ASR) software is a promising next step. Until recently, ASR has been regarded as a hard problem within NLP research (Narayanan & Georgiou, 2013; Tang, 2009). ASR performance is conventionally measured by a Word-Error-Rate (WER) metric, which reflects the number of substitutions, deletions, and insertions required to transform

a hypothesised transcript into a gold standard. High WER for particularly noisy data presents an ongoing challenge for ASR, with errors over 50% cited in prior research (Kim et al., 2019). ASR transcripts can also be difficult for humans to read (Moore, 2015) and may vary significantly from human-generated transcripts on the same source data (S. J. Pentland et al., 2022). Additionally, audio such as that used in Chapter 4 would require additional processing to allocate speech to the appropriate speaker. In such cases, automatic diarization would continue to be a requirement (Giannakopoulos & Petridis, 2012). Given these challenges, a concern when using ASR is the presumption of reduced accuracy compared to human-derived transcripts. Studies have suggested that significant errors generated during transcription can significantly affect downstream performance (Biel et al., 2013). Similarly, research has suggested that mistrust in predictions increases in cases where users had concerns about the quality of the transcription (Hirsch et al., 2018a). Addressing these concerns, however, general improvements in audio quality and improved software capabilities have reduced WER to levels closely approximating human performance ($WER \approx 5\%$) (Thomas et al., 2019). Encouragingly, a direct comparison of ASR and manual transcripts found that ASR transcripts can lead to more stable performance and improved accuracy on a downstream classification task (Malik et al., 2018). The authors of this study highlight the cumbersome nature of manual transcription as a motivator for using ASR. Research has also suggested that despite the differences in individual features, classification tasks perform similarly regardless of whether manual or automatic transcription is used (S. J. Pentland et al., 2022). This indicates that relying solely on WER to guide whether ASR is suitable may be misguided. Indeed, preliminary experiments in Chapter 4 explored the feasibility of ASR. However, large errors ($WER > 60\%$) meant that resulting transcripts bore minimal resemblance to the actual audio. More concerning was the substantial between-speaker deviations, which varied considerably. It was ultimately decided that ASR would reduce the opportunity to detect subtle mirroring between speakers, so a manual approach was pursued as an alternative. Therefore, a recommendation for future work would be to test the utility of a fully automated approach using a combination of ASR and automatic diarization techniques.

7.3.2 Chapter 5 and Chapter 6

The latter part of the thesis sought to detect predatory behaviour in online adult-child interactions automatically. The primary motivation behind this section of the

thesis is the real-world need to efficiently extract a variety of known risk markers as a means of improving child safety. As highlighted in this chapter, generating a sufficiently large number of positive class labels was a considerable problem for most behaviours examined. Deep learning models are notoriously data-hungry as they require substantial training input to build an adequate representation of a particular concept. Whilst latent, abstract or subjective behaviours are relatively straightforward for humans to detect, they represent a considerable problem for machines. Given the poor performance in the zero-shot setting, it would appear prudent to regard the behaviour labels as insufficiently defined. Pre-trained language models should be able to leverage the vast amounts of generic data they were trained on. However, with noisy labels such as those used here, the model will likely overfit simple linguistic rules.

Several solutions to this problem have already been described in Section 7.1. In addition, an active learning approach could be utilised to increase the data available for training. Active learning can increase the amount of training data by manually correcting misclassified data and re-training. It is, therefore, an effective solution in cases where the amount of positive class labels is scarce (Liew et al., 2015), and may be preferable to more computationally demanding approaches such as semi-supervised learning.

Another limitation of the approach used in this section is the opaqueness of the deep learning algorithm. Whereas the first part of the thesis relied primarily on random forest, which offers a degree of interpretability through techniques such as feature importance (Breiman, 2001), deep learning methods are considered black boxes (Zini & Awad, 2022). This can make them difficult to debug, as it is not always clear why a particular prediction has been made. Whilst statistical approaches such as LIME and SHAP offer a degree of interpretability, they are often unsuitable for a non-technical audience (Jeyakumar et al., 2020). Big data approaches with limited interpretability have been treated with scepticism within the social sciences and criticised as being *atheoretical* (Adjerid & Kelley, 2018). Furthermore, whilst models with low explainability can harm the users' trust in the quality of the predictions made (Rudin, 2019), the ability to explain a prediction increases confidence and trust in the system (Yang et al., 2020). This can make the difference between a community embracing a technical solution and rejecting it. For instance, a study of worker attitudes towards AI solutions found that individuals with greater domain experience were more likely to challenge the output of a computational analysis (Hirsch et al., 2018b). It is, therefore, vital that solutions intended for real-world

use possess the capability to explain why a particular prediction has been made.

7.4 Conclusions

The detection of complex social behaviours is becoming increasingly important in a society characterised by widespread communication on a worldwide scale. 24-hour news and the increase in online text, audio, and video communication, such as social media, represent just two exciting domains in which social scientists can explore social interactions in naturally-occurring settings. However, whilst the increased availability of affordable computational resources presents the opportunity to study social phenomena at a previously unfathomable scale, a reticence to embrace technology hinders the mobility and application of social research to new and exciting domains.

Throughout this thesis, the principal aim has been to help bridge the gap between social science questions and computer science technologies through a cross-disciplinary approach. Specifically, the main aim of the thesis has been to examine the suitability of computational effort toward the task of automated conversation analysis and behaviour coding. Such work is a highly intensive task routinely performed via manual effort alone. To this end, the thesis has contributed to existing literature within a growing area of necessary cross-disciplinary research.

The thesis presents the advantages of cross-disciplinary research in this space and highlights some of the pitfalls researchers face when seeking to scale conventional social scientific practice. The main findings of the thesis support the inclusion of automation into social scientific research but warn against the dangers of accepting a 'plug-in-and-play' mindset. Neither should the computer be considered a replacement for the value offered by true domain expertise. Instead, the author calls for greater collaboration between social and computational disciplines to tackle important real-world issues. Concluding this thesis, the true value of computational effort within the social sciences is one that augments rather than replaces the human expert.

References

Adjerid, I., & Kelley, K. (2018). Big data in psychology: A framework for research advancement. *American Psychologist*, *73*(7), 899.

Adler, A. (2011). To catch a predator. *Colum. J. Gender & L.*, *21*, 130.

Adnan, K., & Akbar, R. (2019). Limitations of information extraction methods and techniques for heterogeneous unstructured big data. *International Journal of Engineering Business Management*, *11*, 1847979019890771.

Aho, A. V. (1991). Algorithms for finding patterns in strings, handbook of theoretical computer science (vol. a): Algorithms and complexity.

Alameda-Pineda, X., Ricci, E., & Sebe, N. (2019). Multimodal Behavior Analysis in the Wild: An Introduction. In X. Alameda-Pineda, E. Ricci, & N. Sebe (Eds.), *Multimodal behavior analysis in the wild: Advances and challenges* (pp. 1–8). Academic Press. https://doi.org/10.1016/c2017-0-01387-3

Alison, L. J., Alison, E. E., Noone, G., Elntib, S., & Christiansen, P. (2013). Why tough tactics fail and rapport gets results: Observing rapport-based interpersonal techniques (ORBIT) to generate useful information from terrorists. *Psychology, Public Policy, and Law*, *19*(4), 411–431. https://doi.org/10.1037/a0034564

Allen, J. A., Fisher, C., Chetouani, M., Chiu, M. M., Gunes, H., Mehu, M., & Hung, H. (2017). Comparing Social Science and Computer Science Workflow Processes for Studying Group Interactions. *Small Group Research*, *48*(5), 568–590. https://doi.org/10.1177/1046496417721747

Ameisen, E. (2020). *Building machine learning powered applications: Going from idea to product.* " O'Reilly Media, Inc."

Ando, K., Shibata, C., & Inaba, T. (2017). Coding Collaboration Process Automatically : Coding Methods Using Deep Learning Technology. *International Journal on Advances in Intelligent Systems*, *10*(3), 345–354. https://doi.org/10.1109/tale.2018.8615295

Anjewierden, A., & Gijlers, H. (2008). An exploration of tool support for categorical coding.

Araya, R., Plana, F., Dartnell, P., Soto-Andrade, J., Luci, G., Salinas, E., & Araya, M. (2012). Estimation of teacher practices based on text transcripts of teacher speech using a support vector machine algorithm. *British Journal of Educational Technology*, *43*(6), 837–846.

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational linguistics*, *34*(4), 555–596.

Ashktorab, Z., Desmond, M., Andres, J., Muller, M., Joshi, N. N., Brachman, M., Sharma, A., Brimijoin, K., Pan, Q., Wolf, C. T., et al. (2021). Ai-assisted human labeling: Batching for efficiency without overreliance. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 1–27.

Aspinwall, E. R., & VerBurg, A. L. (2007). The use of liwc to understand how undergraduate students learn the apa style.

Bahgat, M., Wilson, S., & Magdy, W. (2022). Liwc-ud: Classifying online slang terms into liwc categories. *14th ACM Web Science Conference 2022*, 422–432.

Barber, C. S., & Bettez, S. C. (2021). Exposing patterns of adult solicitor behaviour: Towards a theory of control within the cybersexual abuse of youth. *European Journal of Information Systems*, *30*(6), 591–622. https://doi.org/10.1080/0960085x.2020.1816146

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and machine learning: Limitations and opportunities* [http://www.fairmlbook.org]. fairmlbook.org.

Bass, L., & Semetko, H. A. (2021). Content analysis. *Research Methods in the Social Sciences: an AZ of Key Concepts*, *54*(1), 56.

Baum, M. A. (2005). The Talk Show Circuit. *Political Science*, *49*(2), 213–234.

Bayes, T. (1763). Lii. an essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, frs communicated by mr. price, in a letter to john canton, amfr s. *Philosophical transactions of the Royal Society of London*, (53), 370–418.

Bayram, A. B., & Ta, V. P. (2018). Diplomatic chameleons: Language style matching and agreement in international diplomatic negotiations. *Negotiation and Conflict Management Research*, *12*(1), 23–40. https://doi.org/http://dx.doi.org/10.1111/ncmr.12142

Behera, S. K., & Dash, R. (2022). Fine-tuning of a bert-based uncased model for unbalanced text classification. In *Advances in intelligent computing and communication: Proceedings of icac 2021* (pp. 377–384). Springer.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, *35*(8), 1798–1828.

Beňuš, Š., Gravano, A., Levitan, R., Levitan, S. I., Willson, L., & Hirschberg, J. (2014). Entrainment, dominance and alliance in supreme court hearings. *Knowledge-Based Systems*, *71*, 3–14. https://doi.org/10.1016/j.knosys.2014.05.020

Bergmann, K., & Kopp, S. (2012). Proceedings of the Annual Meeting of the Cognitive Science Gestural Alignment in Natural Dialogue. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *34*(34), 1326–1331.

Biel, J. I., Tsiminaki, V., Dines, J., & Gatica-Perez, D. (2013). Hi YouTube! Personality impressions and verbal content in social video. *ICMI 2013 - Proceedings of the 2013 ACM International Conference on Multimodal Interaction*, 119–126. https://doi.org/10.1145/2522848.2522877

Biggiogera, J., Boateng, G., Hilpert, P., Vowels, M., Bodenmann, G., Neysari, M., Nussbeck, F., & Kowatsch, T. (2021). Bert meets liwc: Exploring state-of-the-art language models for predicting communication behavior in couples' conflict interactions. *Companion Publication of the 2021 International Conference on Multimodal Interaction*, 385–389.

Bilous, F. R., & Krauss, R. M. (1988). Dominance and accommodation in the conversational behaviours of same and mixed-gender dyads. *Language & Communication*, *8*, 183–194. https://doi.org/10.1016/0271-5309(88)90016-x

Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4). Springer.

Boecking, B., Neiswanger, W., Xing, E., & Dubrawski, A. (2020). Interactive weak supervision: Learning useful heuristics for data labeling. *arXiv preprint*.

Boersma, P., & Weenink, D. (2007). Praat: Doing phonetics by computer. https://www.fon.hum.uva.nl/praat/

Bogdanova, D., Rosso, P., & Solorio, T. (2014). Exploring high-level features for detecting cyberpedophilia. *Computer speech & language*, *28*(1), 108–120. https://doi.org/10.1016/j.csl.2013.04.007

Boghrati, R., Hoover, J., Johnson, K. M., Garten, J., & Dehghani, M. (2018). Conversation level syntax similarity metric. *Behavior Research Methods*, *50*(3), 1055–1073. https://doi.org/10.3758/s13428-017-0926-2

Bonin, F., De Looze, C., Ghosh, S., Gilmartin, E., Vogel, C., Polychroniou, A., Salamin, H., Vinciarelli, A., & Campbell, N. (2013). Investigating fine temporal dynamics of prosodic and lexical accommodation. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, (January), 539–543.

Bonsignori, V. (2009). Transcribing film dialogue: from orthographic to prosodic transcription. In M. Freddi & M. Pavesi (Eds.), *Analysing audiovisual dialogue. linguistic and translation insights* (pp. 185–200). Clueb.

Bourhis, R. Y., Giles, H., & Lambert, W. E. (1975). Social consequences of accommodating one's style of speech: A cross-national investigation. https://doi.org/10.1515/ling.1975.13.166.55

Bowling, A., & Ebrahim, S. (2005). Quantitative social science: The survey. *Handbook of health research methods: Investigation, measurement and analysis*, 190–214.

Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. (2022). The development and psychometric properties of liwc-22. *Austin, TX: University of Texas at Austin*, 1–47.

Brandt, P. M., & Herzberg, P. Y. (2020). Is a cover letter still needed? using liwc to predict application success. *International Journal of Selection and Assessment*, *28*(4), 417–429.

Braun, M. T., Van Swol, L. M., & Vang, L. (2015). His Lips Are Moving: Pinocchio Effect and Other Lexical Indicators of Political Deceptions. *Discourse Processes*, *52*(1), 1–20. https://doi.org/10.1080/0163853X.2014.942833

Braun, V., & Clarke, V. (2012). *Thematic analysis.* American Psychological Association.

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*, 123–140.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1017/CBO9781107415324.004

Brennan, S. E. (1996). Lexical entrainment in spontaneous dialog. *Proceedings of the 2nd International Symposium on Spoken Dialogue*, 41–44. http://www.cs.columbia.edu/$%5Csim$julia/papers/brennan96.pdf

Briggs, P., Simon, W. T., & Simonsen, S. (2011). An exploratory study of internet-initiated sexual offenses and the chat room sex offender: Has the internet enabled a new typology of sex offender? *Sexual Abuse*, *23*(1), 72–91. https://doi.org/10.1177/1079063210384275

Brönnimann, R., Herlihy, J., Müller, J., & Ehlert, U. (2013). Do testimonies of traumatic events differ depending on the interviewer? *European Journal of Psychology Applied to Legal Context*, *5*(1), 97–121. https://doi.org/10.5167/uzh-69183

Broome, L. J., Izura, C., & Davies, J. (2020). A psycho-linguistic profile of online grooming conversations: A comparative study of prison and police staff considerations. *Child Abuse & Neglect*, *109*, 104647. https://doi.org/10.1016/j.chiabu.2020.104647

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877–1901.

Buckingham, L., & Alali, N. (2020). Extreme parallels: A corpus-driven analysis of isis and far-right discourse. *Kōtuitui: New Zealand Journal of Social Sciences Online*, *15*(2), 310–331. https://doi.org/10.1080/1177083x.2019.1698623

Bull, P. (2003). The analysis of equivocation in political interviews. *Doing social psychology research*, 205–228.

Burgoon, J. K., Dunbar, N. E., & Giles, H. (2017). Interaction Coordination and Adaptation. In J. K. Burgoon, N. Magnenat-Thalmann, M. Pantic, & A. Vinciarelli (Eds.), *Social signal processing* (pp. 78–96). Cambridge University Press. https://doi.org/10.1017/9781316676202.008

Burnap, P., Rana, O. F., Avis, N., Williams, M., Housley, W., Edwards, A., Morgan, J., & Sloan, L. (2015). Detecting tension in online communities with computational Twitter analysis. *Technological Forecasting and Social Change*, *95*, 96–108. https://doi.org/10.1016/j.techfore.2013.04.013

Butcher, B., Zilka, M., Cook, D., & Weller, A. (2023). Optimising Human-Machine Cooperation for Efficient High-Precision Information Extraction from Text [Submitted to]. *Sixth ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT)*.

Butters, R. R., Kendall, T., & Carter, P. (2014). Internet traps and the creation of linguistic crimes: Perverted justice as broadcast entertainment. *Internet Traps and the Creation of Linguistic Crimes: Perverted Justice as Broadcast Entertainment*, 223–240.

Cano, A. E., Fernandez, M., & Alani, H. (2014). Detecting child grooming behaviour patterns on social media. *International conference on social informatics*, 412–427.

Carrick, T., Rashid, A., & Taylor, P. J. (2016). Mimicry in online conversations: An exploratory study of linguistic analysis techniques. In J. Rokne & V. Subrahmanian (Eds.), *Proceedings of the ieee/acm international conference on advances in social networks analysis and mining* (pp. 732–736). ASANOM. https://doi.org/10.1109/ASONAM.2016.7752318

Celis, L. E., Keswani, V., & Vishnoi, N. (2020). Data preprocessing to mitigate bias: A maximum entropy based approach. *International conference on machine learning*, 1349–1359.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR)*, *41*(3), 1–58.

Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative analysis*. sage.

Chartrand, T. L., & Bargh, J. A. (1999). The chameleon effect: The perception–behavior link and social interaction. *Journal of personality and social psychology*, *76*(6), 893–910. https://doi.org/10.1037/0022-3514.76.6.893

Chen, L., Feng, G., Martin-Raugh, M. P., Leong, C. W., Kitchen, C., Yoon, S.-Y., Lehman, B., Kell, H., & Lee, C. M. (2016). Automatic scoring of monologue video interviews using multimodal cues. *INTERSPEECH*, 32–36.

Chen, N. C., Drouhard, M., Kocielnik, R., Suh, J., & Aragon, C. R. (2018). Using machine learning to support qualitative coding in social science: Shifting the focus to ambiguity. *ACM Transactions on Interactive Intelligent Systems*, *8*(2). https://doi.org/10.1145/3185515

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794.

Chi, M. T. (1997). Quantifying Qualitative Analyses of Verbal Data: A Practical Guide. *Journal of the Learning Sciences*, *6*(3), 271–315. https://doi.org/10.1207/s15327809jls0603_1

Chiang, E., Nguyen, D., Towler, A., Haas, M., & Grieve, J. (2021). Linguistic analysis of suspected child sexual offenders' interactions in a dark web image exchange chatroom. *International Journal of Speech Language and the Law*, *0*(0). https://doi.org/10.1558/ijsll.41446

Christiansen, P., Alison, L., & Alison, E. (2018). Well begun is half done: Interpersonal behaviours in distinct field interrogations with high-value detainees. *Legal and Criminological Psychology*, *23*(1), 68–84.

Chuang, J., Wilkerson, J. D., Stewart, B. M., & Roberts, M. E. (2015). Computer-Assisted Content Analysis : Topic Models for Exploring Multiple Subjective Interpretations. *NIPS Workshop on Human-Propelled Machine Learnin*, 1–9.

Church, K. W. (2000). Empirical Estimates of Adaptation: The Chance of Two Noriegas is Closer to p/2 than p. In M. Kay (Ed.), *Proceedings of the 18th conference on computational linguistics* (pp. 180–186). Association for Computational Linguistics. https://doi.org/10.3115/990820.990847

Cioffi-Revilla, C. (2014). Introduction to computational social science. *London and Heidelberg: Springer*.

Clark, H. H. (1996). *Using language*. Cambridge University Press.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, *20*(1), 37–46.

Collins, K., & Carthy, N. (2018). No rapport, no comment: The relationship between rapport and communication during investigative interviews with suspects. *Journal of Investigative Psychology and Offender Profiling*, *16*(1), 18–31. https://doi.org/10.1002/jip.1517

Cook, D., & Humann, M. (n.d.). Casualty perceptions of first response communication during a major incident [In progress].

Cook, D., Zilka, M., Alison, L., & Maskell, S. (n.d.). Does multimodality improve unimodal performance in predicting the outcome of a political interview? [In progress].

Cook, D., Zilka, M., de Sandre, H., Giles, S., Weller, A., & Maskell, S. (2022). The Mechanical Psychologist: Can We Automate the Analysis of Online Child Sexual Exploitation Discourse? *arXiv preprint*.

Cook, D., Zilka, M., DeSandre, H., Giles, S., & Maskell, S. (2023). Protecting Children from Online Exploitation: Can a trained model detect harmful communication strategies? [Submitted to]. *Sixth AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society*.

Cook, D., Zilka, M., Maskell, S., & Alison, L. (2021). A Psychology-Driven Computational Analysis of Political Interviews. *21st Annual Conference of the International Speech Communication Association*, 1942–1946. https://doi.org/10.21437/Interspeech.2021-2249

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273–297.

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, *20*(2), 215–232.

Cross, S., Waters, Z., Kitto, K., & Zuccon, G. (2017). Classifying help seeking behaviour in online communities. *ACM International Conference Proceeding Series*, 419–423. https://doi.org/10.1145/3027385.3027442

Croteau, D. (1998). Examining the 'liberal media'claim. *Fairness & Accuracy In Reporting: New York, NY (June)*.

Culpeper, J., Bousfield, D., & Wichmann, A. (2003). Impoliteness revisited: With special reference to dynamic and prosodic aspects. *Journal of pragmatics*, *35*(10-11), 1545–1579.

Cypress, B. S. (2019). Data analysis software in qualitative research: Preconceptions, expectations, and adoption. *Dimensions of critical care nursing*, *38*(4), 213–220.

Dalati, S., & Marx Gómez, J. (2018). Surveys and questionnaires. In *Modernizing the academic teaching and research environment* (pp. 175–186). Springer.

Danescu-Niculescu-Mizil, C., Gamon, M., & Dumais, S. (2011). Mark My Words! Linguistic Style Accommodation in Social Media. In S. Sadagopan, K. Ramamritham, A. Kumar, & M. Ravindra (Eds.), *Proceedings of the 20th international conference on world wide web* (pp. 745–754). Association for Computing Machinery. https://doi.org/10.1145/1963405.1963509

Danescu-Niculescu-Mizil, C., & Lee, L. (2011). Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In F. Keller & D. Reitter (Eds.), *Proceedings of the 2nd workshop on cognitive modeling and computational linguistics* (pp. 76–87). Association for Computational Linguistics. http://arxiv.org/abs/1106.3077

Daniel, C. (2023). Facebook messenger revenue and growth statistics (2023) [Accessed: 2023-04-15]. https://www.usesignhouse.com/blog/facebook-messenger-stats

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dideriksen, C., Fusaroli, R., Tylén, K., Dingemanse, M., & Christiansen, M. (2019). Contextualizing Conversational Strategies: Backchannel, Repair and Linguistic Alignment in Spontaneous and Task-Oriented Conversations. In A. Goel, C. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st annual meeting of the cognitive science society* (pp. 261–267). https://doi.org/10.31234/osf.io/fd8y9

Dixon, S. (2023). Number of monthly active facebook users worldwide as of 4th quarter 2022 [Accessed: 2023-04-15]. https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, *55*(10), 78–87. https://doi.org/10.1145/2347736.2347755

Dönmez, P., Rosé, C., Stegmann, K., Weinberger, A., & Fischer, F. (2005). Supporting CSCL with automatic corpus analysis technology. *Computer Supported Collaborative Learning 2005: The Next 10 Years - Proceedings of the International Conference on Computer Supported Collaborative Learning 2005, CSCL 2005*, 125–134. https://doi.org/10.3115/1149293.1149310

Doyle, G., Yurovsky, D., & Frank, M. C. (2016). A Robust Framework for Estimating Linguistic Alignment in Twitter Conversations. In J. Bourdeau, J. A. Hendler, R. Nkambou Nkambou, I. Horrocks, & B. Y. Zhao (Eds.), *Proceedings of the 25th international conference on world wide web* (pp. 637–648). International World Wide Web Conferences Steering Committee. https://doi.org/10.1145/2872427.2883091

Drouin, M., Boyd, R. L., Hancock, J. T., & James, A. (2017). Linguistic analysis of chat transcripts from child predator undercover sex stings. *Journal of Forensic Psychiatry and Psychology*, *28*(4), 437–457. https://doi.org/10.1080/14789949.2017.1291707

Du, M., Liu, N., Yang, F., Ji, S., & Hu, X. (2019). On attribution of recurrent neural network predictions via additive decomposition. *The World Wide Web Conference*, 383–393.

Dunbar, N. E., Jensen, M. L., Tower, D. C., & Burgoon, J. K. (2014). Synchronization of nonverbal behaviors in detecting mediated and non-mediated deception. *Journal of Nonverbal Behavior*, *38*(3), 355–376.

Duran, N. D., Hall, C., Mccarthy, P. M., & Mcnamara, D. S. (2010). The linguistic correlates of conversational deception : Comparing natural language processing technologies. *Applied Psycholinguistics*, *31*, 439–462.

Duran, N. D., Paxton, A., & Fusaroli, R. (2019). ALIGN: Analyzing Linguistic Interactions With Generalizable techNiques—A Python Library. *Psychological Methods*, *24*(4), 419–438. https://doi.org/10.1037/met0000206

Dutta, S., Arunachalam, A., & Misailovic, S. (2022). To seed or not to seed? an empirical analysis of usage of seeds for testing in machine learning projects. *2022 IEEE Conference on Software Testing, Verification and Validation (ICST)*, 151–161. https://doi.org/10.1109/ICST53961.2022.00026

Ebrahimi, M., Suen, C. Y., & Ormandjieva, O. (2016). Detecting predatory conversations in social media by deep convolutional neural networks. *Digital Investigation*, *18*, 33–49. https://doi.org/https://doi.org/10.1016/j.diin.2016.07.001

Egan, V., Hoskinson, J., & Shewan, D. (2011). Perverted Justice: A Content Analysis of the Language Used vt Offenders Detected Attempting to Solicit Children for Sex. *Antisocial Behavior: Causes, Correlations and Treatments*, (44).

Elliott, I. A. (2017). A self-regulation model of sexual grooming. *Trauma, Violence, & Abuse*, *18*(1), 83–97. https://doi.org/10.1177/1524838015591573

Endo, K. (2017). Public sphere and social capital in the age of intermediality: Approach from computational social science. *Reconstruction of the Public Sphere in the Socially Mediated Age*, 1–6.

Ericsson, K. A., Hoffman, R. R., Kozbelt, A., & Williams, A. M. (2018). *The cambridge handbook of expertise and expert performance*. Cambridge University Press.

Escalante, H. J., Villatoro-Tello, E., Juárez, A., Montes, M., & Villaseñor-Pineda, L. (2013). Sexual predator detection in chats with chained classifiers. *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 46–54.

Espinoza, C., Lamsa, J., Araya, R., Hamalainen, R., Jimenez, A., Gormazx, R., & Viiri, J. (2019). Automatic content analysis in collaborative inquiry-based learning. In O. Levrini & G. Tasquier (Eds.), *Proceedings of esera 2019: The beauty and pleasure of understanding : Engaging with contemporary challenges through science education* (pp. 2041–2050). https://www.esera.org/publications/esera-conference-proceedings/esera-2019

Fan, S., & Ilk, N. (2020). A text analytics framework for automated communication pattern analysis. *Information and Management*, *57*(4), 103219. https://doi.org/10.1016/j.im.2019.103219

Fauzi, M. A., & Bours, P. (2020). Ensemble method for sexual predators identification in online chats. *2020 8th international workshop on biometrics and forensics (IWBF)*, 1–6. https://doi.org/10.1109/iwbf49977.2020.9107945

Feldman, R., & Sanger, J. (2006). *The Text Mining Handbook*. https://doi.org/10.1017/cbo9780511546914

Feng, H., Qin, W., Wang, H., Li, Y., & Hu, G. (2022). A combination of resampling and ensemble method for text classification on imbalanced data. *Big Data–BigData 2021: 10th International Conference, Held as Part of the Services Conference Federation, SCF 2021, Virtual Event, December 10–14, 2021, Proceedings*, 3–16.

Fiore, V. (2019). Automatic Identification of Political Ideology in Online News Articles. *RAIS Journal for Social Sciences*, *3*(2), 50–54. https://doi.org/10.5281/zenodo.3549475

Flemotomos, N., Martinez, V. R., Gibson, J., Atkins, D. C., Creed, T. A., & Narayanan, S. S. (2018). Language Features for Automated Evaluation of Cognitive Behavior Psychotherapy Sessions. In C. Sekhar, P. Rao, P. Ghosh, H. Murthy, B. Yegnanarayana, S. Umesh, P. Alku, S. Prasanna, & S. Narayanan (Eds.), *19th annual conference of the international speech communication association* (pp. 1908–1912). International Speech Communication Association. https://doi.org/10.21437/Interspeech.2018-1518

Flor, M., & Andrews-Todd, J. (2022). Towards Automatic Annotation of Collaborative Problem-Solving Skills in Technology-Enhanced Environments. *Journal*

of Computer Assisted Learning. https://doi.org/https://doi.org/10.1111/jcal.12689

Formosa, R. (2023). 25 fun facts about facebook every user should know [Accessed: 2023-04-15]. https://tech25s.com/25-fun-facts-about-facebook/

Forsberg, L., Forsberg, L. G., Lindqvist, H., & Helgason, A. R. (2010). Clinician acquisition and retention of motivational interviewing skills: A two-and-a-half-year exploratory study. *Substance Abuse Treatment, Prevention, and Policy*, *5*(1), 1–14.

Fransson, E., Martinsen, T., & Staksund, E. (2015). *Rape in the age of the Internet* (Vol. 7). Routledge. https://www.researchgate.net/publication/269107473_What_is_governance/link/548173090cf22525dcb61443/download%0Ahttp://www.econ.upf.edu/$%5Csim$reynal/Civil%20wars_12December2010.pdf%0Ahttps://think-asia.org/handle/11540/8282%0Ahttps://www.jstor.org/stable/41857625

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, *55*(1), 119–139.

Friedberg, H., Litman, D., & Paletz, S. B. F. (2012). Lexical Entrainment and Success in Student Engineering Groups, 404–409.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of statistics*, 1189–1232.

Fusaroli, R., Bahrami, B., Olsen, K., Roepstorff, A., Rees, G., Frith, C., & Tylén, K. (2012). Coming to terms: Quantifying the benefits of linguistic coordination. *Psychological science*, *23*(8), 931–939.

Fusaroli, R., Tylén, K., Garly, K., Steensig, J., Christiansen, M. H., & Dingemanse, M. (2017). Measures and mechanisms of common ground: Backchannels, conversational repair, and interactive alignment in free and task-oriented social interactions.

Garcia, A. C. (2018). Presidential campaign talk: Question-answering in 'Neutral Informational Interviews'. *Discourse and Society*, *29*(3), 256–278. https://doi.org/10.1177/0957926517734662
Includes some useful information around what a good IE would do.

Garrod, S., & Pickering, M. J. (2004). Why is conversation so easy? *Trends in Cognitive Sciences*, *8*(1), 8–11. https://doi.org/10.1016/j.tics.2003.10.016

Gaston, J., Narayanan, M., Dozier, G., Cothran, D. L., Arms-Chavez, C., Rossi, M., King, M. C., & Xu, J. (2018). Authorship attribution via evolutionary hybridization of sentiment analysis, liwc, and topic modeling features. *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, 933–940.

Gaume, J., Hallgren, K. A., Clair, C., Mast, M. S., & Atkins, D. C. (2019). Modeling Empathy as Synchrony in Clinician and Patient Vocally-encoded Emotional Arousal: A Failure to Replicate. *Journal of Counseling Psychology*, *66*(3), 341–350. https://doi.org/10.1037/cou0000322

Georgiou, P. G., Black, M. P., & Narayanan, S. S. (2011). Behavioral signal processing for understanding (distressed) dyadic interactions: Some recent developments. *Proceedings of the 3rd International Workshop on Social Signal Processing*, 7–12. https://doi.org/10.1145/2072572.2072576

Gervais, M. M., Kline, M., Ludmer, M., George, R., & Manson, J. H. (2013). The strategy of psychopathy: Primary psychopathic traits predict defection on low-value relationships. *Proceedings of the Royal Society B: Biological Sciences*, *280*(1757), 20122773.

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, *63*, 3–42.

Giannakopoulos, T. (2015). PyAudioAnalysis: An open-source python library for audio signal analysis. *PLoS ONE*, *10*(12), 1–17. https://doi.org/10.1371/journal.pone.0144610

Giannakopoulos, T., & Petridis, S. (2012). Fisher linear semi-discriminant analysis for speaker diarization. *IEEE Transactions on Audio, Speech and Language Processing*, *20*(7), 1913–1922. https://doi.org/10.1109/tasl.2012.2191285

Gibson, J., Can, D., Xiao, B., Imel, Z. E., Atkins, D. C., Georgiou, P., & Narayanan, S. S. (2016). A Deep Learning Approach to Modeling Empathy in Addiction Counseling. *Interspeech*. http://www.itopf.com/knowledge-resources/countries-regions/countries/netherlands/

Gibson, J., Malandrakis, N., Romero, F., Atkins, D. C., & Narayanan, S. S. (2015). Predicting Therapist Empathy in Motivational Interviews using Language Features Inspired by Psycholinguistic Norms. *Proceedings of the 16th Annual Conference of the International Speech Communication Association*, 1947–1951.

Gibson, W., & Brown, A. (2009). *Working with qualitative data*. Sage.

Giles, D., Stommel, W., Paulus, T., Lester, J., & Reed, D. (2015). Microanalysis Of Online Data: The methodological development of "digital CA". *Discourse, Context and Media*, *7*, 45–51. https://doi.org/10.1016/j.dcm.2014.12.002

Giles, H. (1973). Accent Mobility: A Model and Some Data. *Anthropological Linguistics*, *15*(2), 87–105. http://www.jstor.com/stable/30029508

Giles, H., Coupland, N., & Coupland, J. (1991). Accommodation theory: Communication, context, and consequence. In H. Giles, N. Coupland, & J. Coupland (Eds.), *Contexts of accommodation* (pp. 1–68). Cambridge University Press. https://doi.org/10.1017/cbo9780511663673.001

Gorsuch, R. L., & Lehmann, C. S. (2011). Correlation Coefficients: Mean Bias and Confidence Interval Distortions. *Journal of Methods and Measurement in the Social Sciences*, *1*(2), 52. https://doi.org/10.2458/jmm.v1i2.114

Gottschalk, P., & Hamerton, C. (2022). Online grooming. In *White-collar crime online* (pp. 219–243). Springer.

Greene-Colozzi, E. A., Winters, G. M., Blasko, B., & Jeglic, E. L. (2020). Experiences and perceptions of online sexual solicitation and grooming of minors: A retrospective report. *Journal of child sexual abuse*, *29*(7), 836–854. https://doi.org/10.1080/10538712.2020.1801938

Gregory, S. W., & Webster, S. (1996). A Nonverbal Signal in Voices of Interview Partners Effectively Predicts Communication Accommodation and Social Status Perceptions. *Journal of Personality and Social Psychology*, *70*(6), 1231–1240. https://doi.org/10.1037/0022-3514.70.6.1231

Grice, P. (1989). *Studies in the way of words*. Harvard University Press.

Grimmer, J., Roberts, M. E., & Stewart, B. M. (2021). Machine Learning for Social Science: An Agnostic Approach. *Annual Review of Political Science*, *24*, 395–419. https://doi.org/10.1146/annurev-polisci-053119-015921

Grimshaw, A. D. (1990). 13 research on conflict talk: Antecedents, resources, findings, directions. *Conflict talk: Sociolinguistic investigations of arguments in conversations*, 280.

Grishman, R. (2019). Twenty-five years of information extraction. *Natural Language Engineering*, *25*(6), 677–692.

Groseclose, T., & Milyo, J. (2005). A measure of media bias. *The quarterly journal of economics*, *120*(4), 1191–1237.

Gubrium, J. F., & Holstein, J. A. (2003). *Postmodern interviewing*. Sage.

Gunawan, F. E., Ashianti, L., & Sekishita, N. (2018). A simple classifier for detecting online child grooming conversation. *Telkomnika (Telecommunication*

Computing Electronics and Control), *16*(3), 1239–1248. https://doi.org/10.12928/TELKOMNIKA.v16i3.6745

Guo, L., Vargo, C. J., Pan, Z., Ding, W., & Ishwar, P. (2016). Big social data analytics in journalism and mass communication: Comparing dictionary-based text analysis and unsupervised topic modeling. *Journalism and Mass Communication Quarterly*, *93*(2), 322–359. https://doi.org/10.1177/1077699016639231

Gupta, A., Kumaraguru, P., & Sureka, A. (2012). Characterizing pedophile conversations on the internet using online grooming. *arXiv preprint arXiv:1208.4324*. https://doi.org/https://doi.org/10.48550/arXiv.1208.4324

Hasan, M., Kotov, A., Carcone, A. I., Dong, M., Naar, S., & Hartlieb, K. B. (2016). A study of the effectiveness of machine learning methods for classification of clinical interview fragments into a large number of categories. *Journal of biomedical informatics*, *62*, 21–31.

Haslinger, A. (1997). Automatic coding and text processing using ngrams. *Conference on European Statisticians Statistical Standards and Studies*, 199–209.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (Vol. 2). Springer.

Haug, S., Rietz, T., & Maedche, A. (2021). Accelerating deductive coding of qualitative data: An experimental study on the applicability of crowdsourcing. In *Proceedings of mensch und computer 2021* (pp. 432–443).

Healey, P. G. T., Howes, C., & Purver, M. (2010). Does structural priming occur in ordinary conversation? *Proceedings of Linguistic Evidence*.

Heide, F. J. (2013). "easy to sense but hard to define": Charismatic nonverbal communication and the psychotherapist. *Journal of Psychotherapy Integration*, *23*(3), 305.

Heritage, J. (1985). Analyzing News Interviews: Aspects of the Production of Talk for an Overhearing Audience. In T. A. van Dijk (Ed.), *Handbook of discourse analysis* (pp. 95–117). Academic Press London.

Heritage, J., & Clayman, S. (2010). *Talk in Action*. John Wiley & Sons. https://doi.org/10.1002/9781444318135

Heritage, J., & Greatbatch, D. (1986). Generating applause: A study of rhetoric and response at party political conferences. *American journal of sociology*, *92*(1), 110–157.

Higashinaka, R., & Isozaki, H. (2008). Corpus-Based Question Answering for Why-Questions. In J.-H. Lee (Ed.), *Proceedings of the third international joint conference on natural language processing: Volume-i* (pp. 418–425). Association for Computational Linguistics.

Hipp, R. D. (2020). SQLite. https://www.sqlite.org/index.html

Hirsch, T., Soma, C., Merced, K., Kuo, P., Dembe, A., Caperton, D. D., Atkins, D. C., & Imel, Z. E. (2018a). " it's hard to argue with a computer" investigating psychotherapists' attitudes towards automated evaluation. *Proceedings of the 2018 Designing Interactive Systems Conference*, 559–571.

Hirsch, T., Soma, C., Merced, K., Kuo, P., Dembe, A., Caperton, D. D., Atkins, D. C., & Imel, Z. E. (2018b). "It's hard to argue with a computer:" Investigating psychotherapists' attitudes towards automated evaluation. *DIS 2018 - Proceedings of the 2018 Designing Interactive Systems Conference*, 559–572. https://doi.org/10.1145/3196709.3196776e

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Holtgraves, T., & Perdew, A. (2016). Politeness and the communication of uncertainty. *Cognition*, *154*, 1–10.

Honnibal, M., & Montani, I. (2017). Spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, *7*(1), 411–420.

Hopkins, D. J., & King, G. (2010). A Method of Automated Nonparametric Content Analysis for Social Science. *American Journal of Political Science*, *54*(1), 229–247.

Hou, R., Pérez-Rosas, V., Loeb, S., & Mihalcea, R. (2019). Towards automatic detection of misinformation in online medical videos. *2019 International conference on multimodal interaction*, 235–243.

Housley, W., Albert, S., & Stokoe, E. (2019). Interactional data and conversation analysis. *HTTF 2019, November 19–20, 2019, Nottingham, United Kingdom*, 1–7.

Hu, Q., He, Q., Huang, H., Chiew, K., & Liu, Z. (2016). A formalized framework for incorporating expert labels in crowdsourcing environment. *Journal of Intelligent Information Systems*, *47*(3), 403–425. https://doi.org/10.1007/s10844-015-0371-6

Huang, Y., Jin, L., & Peng, D. (2021). Zero-shot chinese text recognition via matching class embedding. *International Conference on Document Analysis and Recognition*, 127–141.

Huls, E., & Varwijk, J. (2011). Political bias in TV interviews. *Discourse and Society*, *22*(1), 48–65. https://doi.org/10.1177/0957926510382836

Hutchby, I. (2011). Non-neutrality and argument in the hybrid political interview. *Discourse Studies*, *13*(3), 349–365. https : / / doi . org / 10 . 1177 / 1461445611400665

Hutchby, I., & Wooffitt, R. (2008). *Conversation analysis*. Polity.

Imel, Z. E., Barco, J. S., Brown, H. J., Baucom, B. R., Kircher, J. C., Baer, J. S., & Atkins, D. C. (2014). The Association of Therapist Empathy and Synchrony in Vocally Encoded Arousal. *Journal of Counseling Psychology*, *61*(1), 146–153. https://doi.org/10.1037/a0034943

Inches, G., & Crestani, F. (2012). Overview of the international sexual predator identification competition at pan-2012. *CLEF (Online working notes/labs/workshop)*, *30*.

Ireland, M. E., Slatcher, R. B., Eastwick, P. W., Scissors, L. E., Finkel, E. J., & Pennebaker, J. W. (2011). Language style matching predicts relationship initiation and stability. *Psychological Science*, *22*(1), 39–44. https://doi.org/10.1177/0956797610392928

Jacob, C., Guéguen, N., Martin, A., & Boulbry, G. (2011). Retail salespeople's mimicry of customers: Effects on consumer behavior. *Journal of Retailing and Consumer Services*, *18*(5), 381–388.

Jager, W., Abramczuk, K., Komendant-Brodowska, A., Baczko-Dombi, A., Fecher, B., Sokolovska, N., & Spits, T. (2020). Looking into the Educational Mirror: Why Computation Is Hardly Being Taught in the Social Sciences, and What to Do About It. *Springer Proceedings in Complexity*, 239–245. https://doi.org/10.1007/978-3-030-34127-5_22

Jeyakumar, J. V., Noor, J., Cheng, Y.-H., Garcia, L., & Srivastava, M. (2020). How can i explain this to you? an empirical study of deep neural network explanation methods. *Advances in Neural Information Processing Systems*, *33*, 4211–4222.

Jiang, J. A., Wade, K., Fiesler, C., & Brubaker, J. R. (2021). Supporting serendipity: Opportunities and challenges for human-ai collaboration in qualitative analysis. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 1–23.

Jin, Z., Lalwani, A., Vaidhya, T., Shen, X., Ding, Y., Lyu, Z., Sachan, M., Mihalcea, R., & Schölkopf, B. (2022). Logical fallacy detection. *arXiv preprint arXiv:2202.13758*.

Joleby, M., Lunde, C., Landström, S., & Jonsson, L. S. (2021). Offender strategies for engaging children in online sexual activity. *Child Abuse & Neglect*, *120*, 105214. https://doi.org/10.1016/j.chiabu.2021.105214

Jordan, W., & Miller, S. R. (2003). Inter-rater agreement in analysis of open-ended responses: Lessons from a mixed methods study of principals. *Knowledge Creation Diffusion Utilization*, *15*.

Jurafsky, D., & Martin, J. H. (2019). Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.

Karamshuk, D., Shaw, F., Brownlie, J., & Sastry, N. (2017). Bridging big data and qualitative methods in the social sciences: A case study of twitter responses to high profile deaths by suicide. *Online Social Networks and Media*, *1*, 33–43.

Karlgren, J., Li, R., & Milgrom, E. M. M. (2020). Text Mining for Processing Interview Data in Computational Social Science, 1–9. http://arxiv.org/abs/2011.14037

Kaur, S., Singh, S., & Kaushal, S. (2021). Abusive content detection in online user-generated data: A survey. *Procedia Computer Science*, *189*, 274–281. https://doi.org/10.1016/j.procs.2021.05.098

Kecht, C., Eggert, A., Kratsch, W., & Roglinger, M. (2021). Event Log Construction from Customer Service Conversations Using Natural Language Inference. *Proceedings - 2021 3rd International Conference on Process Mining, ICPM 2021*, 144–151. https://doi.org/10.1109/ICPM53251.2021.9576869

Kim, J. Y., Liu, C., Calvo, R. A., Mccabe, K., Taylor, S. C. R., Schuller, B. W., & Wu, K. (2019). A Comparison of Online Automatic Speech Recognition Systems and the Nonverbal Responses to Unintelligible Speech. *ArXiv*, *abs/1904.1*, 1–13.

Kochanowicz, J., Tan, A.-H., & Thalmann, D. (2016). Modeling human-like non-rationality for social agents. *Proceedings of the 29th International Conference on Computer Animation and Social Agents*, 11–20.

Kontostathis, A., Edwards, L., & Leatherman, A. (2010). Text mining and cybercrime. In: Text Mining: Applications and Theory, 1–14.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, *15*(2), 155–163.

Kosinski, M., Matz, S. C., Gosling, S. D., Popov, V., & Stillwell, D. (2015). Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American Psychologist*, *70*(6), 543–556. https://doi.org/10.1037/a0039210

Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. *Information*, *10*(4), 150.

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. Sage publications.

Kulesza, W., Dolinski, D., Huisman, A., & Majewski, R. (2014). The Echo Effect: The Power of Verbal Mimicry to Influence Prosocial Behavior. *Journal of Language and Social Psychology*, *33*(2), 183–201. https://doi.org/10.1177/0261927X13506906

Lakin, J. L., & Chartrand, T. L. (2003). Using nonconscious behavioral mimicry to create affiliation and rapport. *Psychological science*, *14*(4), 334–339.

Lakin, J. L., Chartrand, T. L., & Arkin, R. M. (2008). I am too just like you: Nonconscious mimicry as an automatic behavioral response to social exclusion. *Psychological science*, *19*(8), 816–822.

Lakin, J. L., Jefferis, V. E., Cheng, C. M., & Chartrand, T. L. (2003). The chameleon effect as social glue: Evidence for the evolutionary significance of nonconscious mimicry. *Journal of nonverbal behavior*, *27*(3), 145–162.

Lamb, M. E., Sternberg, K. J., Orbach, Y., Esplin, P. W., & Mitchell, S. (2002). Is ongoing feedback necessary to maintain the quality of investigative interviews with allegedly abused children? *Applied Developmental Science*, *6*(1), 35–41.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, 159–174.

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of google flu: Traps in big data analysis. *Science*, *343*(6176), 1203–1205. https://doi.org/10.1126/science.1248506

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A. L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., & Van Alstyne, M. (2009). Social science: Computational social science. *Science*, *323*(5915), 721–723. https://doi.org/10.1126/science.1167742

Lazer, D., Pentland, A., Watts, D. J., Aral, S., Athey, S., Contractor, N., Freelon, D., Gonzalez-Bailon, S., King, G., Margetts, H., Nelson, A., Salganik, M. J., Strohmaier, M., Vespignani, A., & Wagner, C. (2020). Computational social science: Obstacles and opportunities. *Science*, *369*(6507), 1060–1062. https://doi.org/10.1126/science.aaz8170

Learning, M. (1997). Series in computer science.

Lee, C. C., Katsamanis, A., Black, M. P., Baucom, B. R., Christensen, A., Georgiou, P. G., & Narayanan, S. S. (2014). Computing vocal entrainment: A signal-derived PCA-based quantification scheme with application to affect analysis in married couple interactions. *Computer Speech and Language*, *28*(2), 518–539. https://doi.org/10.1016/j.csl.2012.06.006

Lee, C. C., Katsamanis, A., Black, M. P., Baucom, B. R., Georgiou, P. G., & Narayanan, S. S. (2011). An analysis of PCA-based vocal entrainment measures in married couples' affective spoken interactions. In P. Cosi, R. De Mori, G. Di Fabbrizio, & R. Pieraccini (Eds.), *Proceedings of the 12th annual conference of the international speech communication association* (pp. 3101–3104). ISCA Archive. http://www.isca-speech.org/archive/interspeech_2011

Lee, C.-C., Black, M., Katsamanis, A., Lammert, A. C., Baucom, B. R., Christensen, A., Georgiou, P. G., & Narayanan, S. S. (2010). Quantification of prosodic entrainment in affective spontaneous spoken interactions of married couples. *Eleventh Annual Conference of the International Speech Communication Association*.

Lee, J., & Hamilton, J. T. (2022). Anchoring in the past, tweeting from the present: Cognitive bias in journalists' word choices. *Plos one*, *17*(3), e0263730.

Leech, N. L., & Onwuegbuzie, A. J. (2008). Qualitative Data Analysis: A Compendium of Techniques and a Framework for Selection for School Psychology Research and Beyond. *School Psychology Quarterly*, *23*(4), 587–604. https://doi.org/10.1037/1045-3830.23.4.587

Levitan, R., Gravano, A., Willson, L., Beňuš, Š., Hirschberg, J., & Nenkova, A. (2012). Acoustic-prosodic entrainment and social behavior. *NAACL HLT 2012 - 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, 11–19.

Levitan, R., & Hirschberg, J. (2011). Measuring Acoustic-Prosodic Entrainment with Respect to Multiple Levels and Dimensions. In P. Cosi, R. De Mori, G. Di Fabbrizio, & R. Pieraccini (Eds.), *Proceedings of the 12th annual conference of the international speech communication association* (pp. 3081–

3084). International Speech Communication Association. https://www.isca-speech.org/archive/interspeech_2011/i11_3081.html

Levitan, S. I., Maredia, A., & Hirschberg, J. (2018). Acoustic-prosodic indicators of deception and trust in interview dialogues. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *2018-Septe*, 416–420. https://doi.org/10.21437/Interspeech.2018-2443

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature Selection: A Data Perspective. *ACM Computing Surveys*, *50*(6), 1–45. https://doi.org/https://doi.org/10.1145/3136625

Liebman, N., & Gergle, D. (2016). Capturing turn-by-turn lexical similarity in text-based communication. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*, *27*, 553–559. https://doi.org/10.1145/2818048.2820062

Liebregts, W., Darnihamedani, P., Postma, E., & Atzmueller, M. (2019). The promise of social signal processing for research on decision-making in entrepreneurial contexts. *Small Business Economics*. https://doi.org/10.1007/s11187-019-00205-1

Liew, J. S. Y., McCracken, N., Zhou, S., & Crowston, K. (2015). Optimizing Features in Active Machine Learning for Complex Qualitative Content Analysis. (1050), 44–48. https://doi.org/10.3115/v1/w14-2513

Lin, G., & Walker, M. (2017). Stylistic variation in television dialogue for natural language generation. *Proceedings of the Workshop on Stylistic Variation*, 85–93.

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, *16*(3), 31–57.

Litman, D., Paletz, S., Rahimi, Z., Allegretti, S., & Rice, C. (2016). The teams corpus and entrainment in multi-party spoken dialogues. *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, 1421–1431. https://doi.org/10.18653/v1/d16-1149

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*. https://doi.org/10.48550/arXiv.1907.11692

Loper, E., & Bird, S. (2002). Nltk: The natural language toolkit. *arXiv preprint cs/0205028*.

Loper, E., & Bird, S. (2006). NLTK: The natural language toolkit. *COLING/ACL 2006 - 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Interactive Presentation Sessions*, 69–72.

Lord, S. P., Sheng, E., Imel, Z. E., Baer, J., & Atkins, D. C. (2015). More Than Reflections: Empathy in Motivational Interviewing Includes Language Style Synchrony Between Therapist and Client. *Behavior Therapy*, *46*(3), 296–303. https://doi.org/10.1016/j.beth.2014.11.002

Lubold, N., & Pon-Barry, H. (2014). Acoustic-prosodic entrainment and rapport in collaborative learning dialogues. *MLA 2014 - Proceedings of the 2014 ACM Multimodal Learning Analytics Workshop and Grand Challenge, Co-located with ICMI 2014*, 5–12. https://doi.org/10.1145/2666633.2666635

Lumontod III, R. Z. (2020). Seeing the invisible: Extracting signs of depression and suicidal ideation from college students' writing using liwc a computerized text analysis. *International Journal of Research*, *9*(4), 31–44.

Lyu, S., Ren, X., Du, Y., & Zhao, N. (2023). Detecting depression of chinese microblog users via text analysis: Combining linguistic inquiry word count (liwc) with culture and suicide related lexicons. *Frontiers in psychiatry*, *14*.

Maclin, O. H., & Maclin, M. K. (2005). Coding observational data: A software solution. *Behavior Research Methods*, *37*(2), 224–231. https://doi.org/10.3758/BF03192690

Maddux, W. W., Mullen, E., & Galinsky, A. D. (2008). Chameleons bake bigger pies and take bigger pieces: Strategic behavioral mimicry facilitates negotiation outcomes. *Journal of Experimental Social Psychology*, *44*(2), 461–468.

Mahmoodi, J., Leckelt, M., van Zalk, M. W., Geukes, K., & Back, M. D. (2017). Big data approaches in social and behavioral science: Four key trade-offs and a call for integration. *Current Opinion in Behavioral Sciences*, *18*, 57–62.

Malik, U., Barange, M., Saunier, J., & Pauchet, A. (2018). Performance Comparison of Machine Learning Models Trained on Manual vs ASR Transcriptions for Dialogue Act Annotation. *2018 IEEE 30th International Conference on Tools with Artificial Intelligence Performance*, 1013–1017. https://doi.org/10.1109/ICTAI.2018.00156

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, *405*(2), 442–451.

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis?: What does 'failure to replicate' really mean? *American Psychologist*, *70*(6), 487–498. https://doi.org/10.1037/a0039400

McGhee, I., Bayzick, J., Kontostathis, A., Edwards, L., McBride, A., & Jakubowski, E. (2011). Learning to identify Internet sexual predation. *International Journal of Electronic Commerce*, *15*(3), 103–122. https://doi.org/10.2753/JEC1086-4415150305

Mclaren, B. M., Scheuer, O., Laat, M. D. E., Hever, R., Groot, R. D. E., & Rosé, C. P. (2007). Using Machine Learning Techniques to Analyze and Support Mediation of Student E-Discussions, 331–338.

Mehu, M. (2011). Smiling and laughter in naturally occurring dyadic interactions: Relationship to conversation, body contacts, and displacement activities. *Human Ethology Bulletin*, *26*(1), 10–28.

Metzler, K., Kim, D. A., Allum, N., & DenMan, A. (2016). Who Is Doing Computational Social Science? Trends in Big Data Research. *Sage*, 1–21. https://doi.org/10.4135/wp160926.

Miah, M. W. R., Yearwood, J., & Kulkarni, S. (2011). Detection of child exploiting chats from a mixed chat dataset as a text classification task. *Proceedings of the Australasian Language Technology Association Workshop 2011*, 157–165.

Michalsky, J., Schoormann, H., & Niebuhr, O. (2018). Conversational quality is affected by and reflected in prosodic entrainment. *Proceedings of the International Conference on Speech Prosody*, *2018-June*, 389–392. https://doi.org/10.21437/SpeechProsody.2018-79

Miller, G. A. (1995). Wordnet: A lexical database for english. *Communications of the ACM*, *38*(11), 39–41.

Miller, W. R., & Rollnick, S. (2012). *Motivational interviewing: Helping people change*. Guilford press.

Mills, K. A. (2019). *Big data for qualitative research*. Taylor & Francis.

Mohajan, H. K., et al. (2018). Qualitative research methodology in social sciences and related subjects. *Journal of economic development, environment and people*, *7*(1), 23–48.

Moore, R. J. (2015). Automated Transcription and Conversation Analysis. *Research on Language and Social Interaction*, *48*(3), 253–270. https://doi.org/10.1080/08351813.2015.1058600

Moreno, A., van Delden, R., Poppe, R., & Reidsma, D. (2013). Socially aware interactive playgrounds. *IEEE Pervasive Computing*, *12*(3), 40–47. https://doi.org/10.1109/MPRV.2013.40

Moyers, T. B., Rowell, L. N., Manuel, J. K., Ernst, D., & Houck, J. M. (2016). The Motivational Interviewing Treatment Integrity Code (MITI 4): Rationale, Preliminary Reliability and Validity. *Journal of Substance Abuse Treatment*, *65*, 36–42. https://doi.org/10.1016/j.jsat.2016.01.001

Müller-Frommeyer, L. C., Frommeyer, N. A., & Kauffeld, S. (2019). Introducing rLSM: An integrated metric assessing temporal reciprocity in language style matching. *Behavior Research Methods*, *51*(3), 1343–1359. https://doi.org/10.3758/s13428-018-1078-8

Murray, G., & Oertel, C. (2018). Predicting group performance in task-based interaction. In S. D'Mello, S. Scherer, & P. G. Georgiou (Eds.), *Proceedings of the 20th annual conference on multimodal interaction* (pp. 14–20). Association for Computing Machinery. https://doi.org/10.1145/3242969.3243027

Naim, I., Tanveer, M. I., Gildea, D., Mohammed, & Hoque. (2018). Automated Analysis and Prediction of Job Interview Performance. *IEEE Transactions of Affective Computing*, *9*(2), 191–204. https://doi.org/10.1109/TAFFC.2016.2614299

Narayanan, S., & Georgiou, P. G. (2013). Behavioral Signal Processing: Deriving Human Behavioral Informatics From Speech and Language. *Proceedings of the IEEE*, *101*(5), 1203–1233. https://doi.org/10.1109/JPROC.2012.2236291

Nenkova, A., Gravano, A., & Hirschberg, J. (2008). High frequency word entrainment in spoken dialogue. In D. Lin, Y. Matsumoto, & R. Mihalcea (Eds.), *Proceedings of the 46th annual meeting of the association for computational linguistics on human language technologies short papers* (pp. 169–172). Association for Computational Linguistics. https://doi.org/10.3115/1557690.1557737

Netemeyer, R. G., Bearden, W. O., & Sharma, S. (2003). *Scaling procedures: Issues and applications*. sage publications.

Ngejane, C. H., Eloff, J. H., Sefara, T. J., & Marivate, V. N. (2021). Digital forensics supported by machine learning for the detection of online sexual predatory chats. *Forensic science international: Digital investigation*, *36*, 301109.

Nguyen, L. S., & Gatica-Perez, D. (2015). I would hire you in a minute: Thin slices of nonverbal behavior in job interviews. In P. R. Cohen, Z. Zhang, D. Bohus, R. Horaud, & H. Meng (Eds.), *Proceedings of the 2015 association for computing machinary international conference on multimodal interaction* (pp. 51–58).

Niederhoffer, K. G., & Pennebaker, J. W. (2002). Linguistic style matching in social interaction. *Journal of Language and Social Psychology*, *21*(4), 337–360. https://doi.org/10.1177/026192702237953

Norvig, P. (2016). *How to write a spell checker*. Retrieved September 30, 2019, from https://norvig.com/spell-correct.html

NSPCC. (2021). Record high number of recorded grooming crimes lead to calls for stronger online safety legislation. Retrieved May 11, 2022, from https://www.nspcc.org.uk/about-us/news-opinion/2021/online-grooming-record-high

O'Connell, R. (2003). A typology of child cybersexploitation and online grooming practices. http://image.guardian.co.uk/sys-files/Society/documents/2003/07/17/Groomingreport.pdf

Oliveira, M., Bitencourt, C., Teixeira, E., & Santos, A. C. (2013). Thematic content analysis: Is there a difference between the support provided by the maxqda®

and nvivo® software packages. *Proceedings of the 12th European Conference on Research Methods for Business and Management Studies*, 304–314.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

OpenAI. (2023). Gpt-4 technical report.

Palomares, N. A., Giles, H., Soliz, J., & Gallois, C. (2016). Intergroup accommodation, social categories, and identities. *Communication accommodation theory: Negotiating personal relationships and social identities across contexts*, 123–151.

Pamungkas, E. W., Basile, V., & Patti, V. (2020). Do you really want to hurt me? predicting abusive swearing in social media. *Proceedings of the 12th language resources and evaluation conference*, 6237–6246.

Papadopoulou, A., Ineson, E., & Williams, D. (1996). The graduate management trainee preselection interview: Candidates' perceptions of the influence of interpersonal and communication factors on the interview outcomes. *Personnel Review*.

Pardo, J. S. (2006). On Phonetic Convergence during Conversational Interaction. *The Journal of the Acoustical Society of America*, *119*(4), 2382–2393. https://doi.org/10.1121/1.2178720

Pardo, J. S., Urmanche, A., Wilman, S., Wiener, J., Mason, N., Francis, K., & Ward, M. (2018). A comparison of phonetic convergence in conversational interaction and speech shadowing. *Journal of Phonetics*, *69*, 1–11. https://doi.org/10.1016/j.wocn.2018.04.001

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems 32* (pp. 8024–8035). Curran Associates, Inc. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*, *12*(Oct), 2825–2830.

Pendar, N. (2007). Toward spotting the pedophile telling victim from predator in text chats. *International Conference on Semantic Computing (ICSC 2007)*, 235–241. https://doi.org/10.1109/icsc.2007.32

Pennebaker, J. W. (1993). Putting stress into words: Health, linguistic, and therapeutic implications. *Behaviour Research and Therapy*, *31*(6), 539–548. https://doi.org/10.1016/0005-7967(93)90105-4

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The Development and Psychometric Properties of LIWC2015*. Pennebaker Conglomerates.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Linguistic inquiry and word count [computer software and manual].

Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, *77*(6), 1296–1312. https://doi.org/10.1037/0022-3514.77.6.1296

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological Aspects of Natural Language Use: Our Words, Our Selves. *Annual Review of Psychology*, *54*, 547–577. https://doi.org/10.1146/annurev.psych.54.101601.145041

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. In A. Moschitti, B. Pang, & W. Daelemans (Eds.),

Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp). Association for Computational Linguistics. https://doi.org/10.3115/v1/d14-1162

Pentland, A. (2007). Social Signal Processing. *IEEE Signal Processing Magazine*, *24*(4), 108–111. https://doi.org/10.1109/msp.2007.4286569

Pentland, S. J., Fuller, C. M., Spitzley, L. A., & Twitchell, D. P. (2022). Does accuracy matter? Methodological considerations when using automated speech-to-text for social science research. *International Journal of Social Research Methodology*, *00*(00), 1–17. https://doi.org/10.1080/13645579.2022.2087849

Pérez-Rosas, V., Catley, D., Singh, S., Mihalcea, R., Ann, L., Resnicow, K., Goggin, K. J., Singh, S., Ann, L., Goggin, K. J., Catley, D., An, L., Goggin, K. J., & Catley, D. (2017). Predicting Counselor Behaviors in Motivational Interviewing Encounters. In M. Lapata, P. Blunsom, & A. Koller (Eds.), *Proceedings of the 15th conference of the european chapter of the association for computational linguistics* (pp. 1128–1137). Association for Computational Linguistics. https://doi.org/10.18653/v1/e17-1106

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*(2), 169–190. https://doi.org/10.1017/s0140525x04000056

Pidgeon, N. F., Turner, B. A., & Blockley, D. I. (1991). The use of grounded theory for conceptual analysis in knowledge elicitation. *International journal of Man-machine studies*, *35*(2), 151–173.

Pluss, B. (2010). Non-Cooperation in Dialogue. In S. Demir, J. Raab, N. Reiter, M. Lopatkova, & T. Strzalkowski (Eds.), *Proceedings of the acl 2010 student research workshop* (pp. 1–6). Association for Computational Linguistics. https://www.aclweb.org/anthology/P10-3000

Potter, J., & Wetherell, M. (1987). *Discourse and social psychology: Beyond attitudes and behaviour*. Sage.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza : A Python natural language processing toolkit for many human languages. In A. Celikyilmaz & T.-H. Wen (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics: System demonstrations* (pp. 101–108). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-demos.14

Qiu, L., Chan, S. H. M., & Chan, D. (2018). Big data in social and psychological science: theoretical and methodological issues. *Journal of Computational Social Science*, *1*(1), 59–66. https://doi.org/10.1007/s42001-017-0013-6

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Rahimi, Z., Kumar, A., Litman, D., Paletz, S., & Yu, M. (2017). Entrainment in multi-party spoken dialogues at multiple linguistic levels. In F. Lacerda, D. House, M. Heldner, J. Gustafson, S. Strombergsson, & M. Wlodarczak (Eds.), *Proceedings of the 18th annual conference of the international speech communication association (interspeech)* (pp. 1696–1700). International Speech Communication Association. https://doi.org/10.21437/interspeech.2017-1568

Rajpurkar, P., Jia, R., & Liang, P. (2018). Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*.

Ramos, J., et al. (2003). Using tf-idf to determine word relevance in document queries. *Proceedings of the first instructional conference on machine learning*, *242*(1), 29–48.

Raschka, S. (2016). Model evaluation, model selection, and algorithm selection in machine learning, part iii—cross-validation and hyperparameter tuning.

Ratner, A., Hancock, B., Dunnmon, J., Sala, F., Pandey, S., & Ré, C. (2019). Training complex models with multi-task weak supervision. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 4763–4771.

Razi, A., Kim, S., Alsoubai, A., Stringhini, G., Solorio, T., De Choudhury, M., & Wisniewski, P. J. (2021). A human-centered systematic literature review of the computational approaches for online sexual risk detection. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 1–38. https://doi.org/10.1145/3479609

Reitter, D., & Moore, J. D. (2007). Predicting Success in Dialogue. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL)*, (June), 808–815.

Reitter, D., & Moore, J. D. (2014). Alignment and task success in spoken dialogue. *Journal of Memory and Language*, *76*, 29–46. https://doi.org/10.1016/j.jml.2014.05.008

Rendle-Short, J. (2007). Neutralism and adversarial challenges in the political news interview. *Discourse & Communication*, *1*(4), 387–406. https://doi.org/10.1177/1750481307082205

Reverdy, J., & Vogel, C. (2017). Linguistic repetitions, task-based experience and a proxy measure of mutual understanding. *8th IEEE International Conference on Cognitive Infocommunications, CogInfoCom 2017 - Proceedings*, *2018-Janua*(CogInfoCom), 395–400. https://doi.org/10.1109/CogInfoCom.2017.8268278

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144.

Richardson, B. H., McCulloch, K. C., Taylor, P. J., & Wall, H. J. (2019). The Cooperation Link: Power and Context Moderate Verbal Mimicry. *Journal of Experimental Psychology: Applied*, *25*(1), 62–76. https://doi.org/10.1037/xap0000200

Richardson, B. H., Taylor, P. J., Snook, B., Conchie, S. M., & Bennell, C. (2014). Language style matching and police interrogation outcomes. *Law and human behavior*, *38*(4), 357.

Richardson, L. (2007). Beautiful soup documentation. *April*.

Rogers, E. M., & Bhowmik, D. K. (1970). Homophily - Heterophily: Relational Concepts for Communication Research. *Public Opinion Quarterly*, *34*(4), 523–538. https://doi.org/10.1086/267838

Romero, D. M., Swaab, R. I., Uzzi, B., & Galinsky, A. D. (2015). Mimicry Is Presidential: Linguistic Style Matching in Presidential Debates and Improved Polling Numbers. *Personality and Social Psychology Bulletin*, *41*(10), 1311–1319. https://doi.org/10.1177/0146167215591168

Rose, L. T., & Fischer, K. W. (2011). Garbage in, garbage out: Having useful data is everything. *Measurement: Interdisciplinary Research & Perspective*, *9*(4), 222–226.

Rosé, C., Wang, Y. C., Cui, Y., Arguello, J., Stegmann, K., Weinberger, A., & Fischer, F. (2008). Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. *International Journal of Computer-Supported Collaborative Learning*, *3*(3), 237–271. https://doi.org/10.1007/s11412-007-9034-0

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Rytsarev, I. (2020). Text data mining using conversation analysis. *CEUR Workshop Proceedings*, *2667*, 159–161.

Sacks, H., Schegloff, E., & Jefferson, G. (1974). A Simplest Systematics for the Organization of Turn-Taking for Conversation. *Language*, *50*(4), 696–735. https://doi.org/10.1353/lan.1974.0010

Saif, H., Fernandez, M., He, Y., & Alani, H. (2014). On stopwords, filtering and data sparsity for sentiment analysis of twitter. *Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014*, 810–817.

Saini, J. R., & Rakholia, R. M. (2016a). On Continent and Script-Wise Divisions-Based Statistical Measures for Stop-words Lists of International Languages. *Procedia Computer Science*, *89*, 313–319. https://doi.org/10.1016/j.procs.2016.06.076

Saini, J. R., & Rakholia, R. M. (2016b). On continent and script-wise divisions-based statistical measures for stop-words lists of international languages. *Procedia Computer Science*, *89*, 313–319.

Sainz, O., & Rigau, G. (2021). Ask2transformers: Zero-shot domain labelling with pre-trained language models. *arXiv preprint arXiv:2101.02661*.

Scharkow, M. (2013). Thematic content analysis using supervised machine learning: An empirical evaluation using German online news. *Quality and Quantity*, *47*(2), 761–773. https://doi.org/10.1007/s11135-011-9545-7

Schegloff, E. A. (1993). Reflections on Quantification in the Study of Conversation. *Research on Language and Social Interaction*, *26*(1), 99–128. https://doi.org/10.1207/s15327973rlsi2601_5

Schneider, B., & Pea, R. (2015). Does Seeing One Another's Gaze Affect Group Dialogue? A Computational Approach. *Journal of Learning Analytics*, *2*(2), 107–133. https://doi.org/10.18608/jla.2015.22.9

Schultz, B. G., O'Brien, I., Phillips, N., McFarland, D. H., Titone, D., & Palmer, C. (2016). Speech rates converge in scripted turn-taking conversations. *Applied Psycholinguistics*, *37*(5), 1201–1220.

Seger, C. (2018). An investigation of categorical variable encoding techniques in machine learning: Binary versus one-hot and feature hashing.

Shah, F. P., & Patel, V. (2016). A review on feature selection and feature extraction for text classification. *2016 international conference on wireless communications, signal processing and networking (WiSPNET)*, 2264–2268.

Sharpley, C. F., Munro, D. M., & Elly, M. J. (2005). Silence and rapport during initial interviews. *Counselling Psychology Quarterly*, *18*(2), 149–159.

Shaw, A. D., Horton, J. J., & Chen, D. L. (2011). Designing incentives for inexpert human raters. *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, 275–284.

Shelton, J., Eakin, J., Hoffer, T., Muirhead, Y., & Owens, J. (2016). Online child sexual exploitation: An investigative analysis of offender characteristics and offending behavior. *Aggression and violent behavior*, *30*, 15–23. https://doi.org/10.1016/j.avb.2016.07.002

Shepherd, J. (2023). 22 essential twitter statistics you need to know in 2023 [Accessed: 2023-04-15]. https://thesocialshepherd.com/blog/twitter-statistics

Shibata, C., Ando, K., & Inaba, T. (2017). Towards Automatic Coding of Collaborative Learning Data with Deep Learning Technology. In L. A. Ludovico & A. M. F. Yousef (Eds.), *Elml 2017 : The ninth international conference on mobile, hybrid, and on-line learning* (pp. 65–70). https://doi.org/10.1109/tale.2018.8615295

Shibata, C., Ando, K., & Inaba, T. (2018). Evaluation of Automatic Collaborative Learning Process Coding Using Deep Learning Methods Based on Multi - Dimensional Coding Scheme. *2018 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*, (December), 251–258.

Shu, X., & Ye, Y. (2023). Knowledge discovery: Methods from data mining and machine learning. *Social Science Research*, *110*, 102817.

Sileo, D., Muller, P., Van de Cruys, T., & Pradel, C. (2022). A pragmatics-centered evaluation framework for natural language understanding. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2382–2394.

Silva, C., & Ribeiro, B. (2003). The Importance of Stop Word Removal on Recall Values in Text Categorization. *Proceedings of the International Joint Conference on Neural Networks*, *3*, 1661–1666. https://doi.org/10.1109/ijcnn.2003.1223656

Sinha, P., Verma, A., Shah, P., Singh, J., & Panwar, U. (2020). Prediction for the 2020 United States Presidential Election using Machine Learning Algorithm: Lasso Regression. (103889). https://mpra.ub.uni-muenchen.de/id/eprint/103889

Snel, J., Tarasov, A., Cullen, C., & Delany, S. (2012). A Crowdsourcing Approach to Labelling a Mood Induced Speech Corpus. *4th International Workshop on Corpora for Research on Emotion Sentiment & Social Signals (ES3 2012)*. http://arrow.dit.ie/dmccon/97/

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (2008). Cheap and fast - But is it good? Evaluating non-expert annotations for natural language tasks. *EMNLP 2008 - 2008 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference: A Meeting of SIGDAT, a Special Interest Group of the ACL*, (October), 254–263.

Spooren, W., & Degand, L. (2010). Coding coherence relations: Reliability and validity. *Corpus Linguistics and Linguistic Theory*, *6*(2), 241–266. https://doi.org/10.1515/CLLT.2010.009

Srivatsa, S., Mohan, T., Neha, K., Malakar, N., Kumaraguru, P., & Srinivasa, S. (2022). Zero-shot entity and tweet characterization with designed conditional prompts and contexts. *arXiv preprint arXiv:2204.08405*.

Stivers, T. (2015). Coding Social Interaction: A Heretical Approach in Conversation Analysis? *Research on Language and Social Interaction*, *48*(1), 1–19. https://doi.org/10.1080/08351813.2015.993837

Street, R. L. (1984). Speech Convergence and Speech Evaluation in Fact-Finding Interviews. *Human Communication Research*, *11*(2), 139–169. https://doi.org/10.1111/j.1468-2958.1984.tb00043.x

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in nlp. *arXiv preprint arXiv:1906.02243*.

Sun, J., Schwartz, H. A., Son, Y., Kern, M. L., & Vazire, S. (2020). The language of well-being: Tracking fluctuations in emotion experience through everyday speech. *Journal of Personality and Social Psychology*, *118*(2), 364.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, *27*.

Suvorova, A. (2022). Interpretable Machine Learning in Social Sciences: Use Cases and Limitations. *Communications in Computer and Information Science*, *1503 CCIS*, 319–331. https://doi.org/10.1007/978-3-030-93715-7_23

Suzuki, N., & Katagiri, Y. (2004). Alignment of human prosodic patterns for spoken dialogue systems. *8th International Conference on Spoken Language Processing, ICSLP 2004*, (100), 2989–2992. https://doi.org/10.21437/interspeech.2004-750

Tabani, H., Balasubramaniam, A., Marzban, S., Arani, E., & Zonooz, B. (2021). Improving the efficiency of transformers for resource-constrained devices. *2021 24th Euromicro Conference on Digital System Design (DSD)*, 449–456.

Tang, X. (2009). Hybrid hidden markov model and artificial neural network for automatic speech recognition. *2009 Pacific-Asia Conference on Circuits, Communications and Systems*, 682–685.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, *29*(1), 24–54. https://doi.org/10.1177/0261927x09351676

Taylor, P. J., & Thomas, S. (2008). Linguistic Style Matching and Negotiation Outcome. *Negotiation and Conflict Management Research*, *1*(3), 263–281. https://doi.org/10.1111/j.1750-4716.2008.00016.x

Thai-Nghe, N., Gantner, Z., & Schmidt-Thieme, L. (2010). Cost-sensitive learning methods for imbalanced data. *The 2010 International joint conference on neural networks (IJCNN)*, 1–8.

Thakerar, J. N., Giles, H., & Cheshire, J. (1982). Psychological and linguistic parameters of speech accommodation theory. *Advances in the social psychology of language*, *205*, 205–255.

Thomas, S., Suzuki, M., Huang, Y., Kurata, G., Tuske, Z., Saon, G., Kingsbury, B., Picheny, M., Dibert, T., Kaiser-Schatzlein, A., & Samko, B. (2019). English Broadcast News Speech Recognition by Humans and Machines. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, *2019-May*, 6455–6459. https://doi.org/10.1109/ICASSP.2019.8683211

Tickle-Degnen, L., & Rosenthal, R. (1990). The Nature of Rapport and Its Nonverbal Correlates. *Psychological Inquiry*, *1*(4), 285–293. https://doi.org/10.1207/s15327965pli0104_1

Timoshenko, A., & Hauser, J. R. (2019). Identifying customer needs from user-generated content. *Marketing Science*, *38*(1), 1–20. https://doi.org/10.1287/mksc.2018.1123

Tomljanovic, J., Zuanovic, L., & Šebrek, T. (2016). Sexual predator identification using word2vec features. *Text Analysis and Retrieval 2016 Course Project Reports*, 70.

Trotta, D., Tonelli, S., Aprosio, A. P., & Elia, A. (2019). Annotation and analysis of the polimodal corpus of political interviews. In R. Bernardi, R. Navigli, & G. Semeraro (Eds.), *Proceedings of the sixth italian conference on computational linguistics*. http://ceur-ws.org/

Tsamardinos, I., Greasidou, E., & Borboudakis, G. (2018). Bootstrapping the out-of-sample predictions for efficient and accurate cross-validation. *Machine learning*, *107*, 1895–1922.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, *9*(11).

van Baaren, R. B., Holland, R. W., Steenaert, B., & van Knippenberg, A. (2003). Mimicry for money: Behavioral consequences of imitation. *Journal of Experimental Social Psychology*, *39*(4), 393–398. https://doi.org/10.1016/S0022-1031(03)00014-3

van Gijn-Grosvenor, E. L., & Lamb, M. E. (2021). Online groomer typology scheme. *Psychology, Crime & Law*, *27*(10), 973–987.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Velliangiri, S., Alagumuthukrishnan, S., et al. (2019). A review of dimensionality reduction techniques for efficient computation. *Procedia Computer Science*, *165*, 104–111.

Venturott, L. I., & Ciarelli, P. M. (2021). Application of data augmentation techniques for hate speech detection with deep learning. *EPIA Conference on Artificial Intelligence*, 778–787.

Vijayan, V. K., Bindu, K., & Parameswaran, L. (2017). A comprehensive study of text classification algorithms. *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 1109–1113.

Vinciarelli, A. (2017). Social signal processing for automatic role recognition. *Social Signal Processing*, 225.

Vinciarelli, A., Pantic, M., & Bourlard, H. (2009). Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, *27*(12), 1743–1759. https://doi.org/10.1016/j.imavis.2008.11.007

Vinciarelli, A., Salamin, H., & Pantic, M. (2009). Social Signal Processing: Understanding social interactions through nonverbal behavior analysis. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 42–49. https://doi.org/10.1109/cvprw.2009.5204290

Vraga, E. K., Edgerly, S., Bode, L., Carr, D. J., Bard, M., Johnson, C. N., Kim, Y. M., & Shah, D. V. (2012). The correspondent, the comic, and the combatant: The consequences of host style in political talk shows. *Journalism and Mass Communication Quarterly*, *89*(1), 5–22. https://doi.org/10.1177/1077699011428575

Waddle, M., & Bull, P. (2020). You're Important, Jeremy, but not that Important": Personalized Responses and Equivocation in Political Interviews. *Journal of Social and Political Psychology*, *8*(2), 560–581. https://doi.org/10.5964/jspp.v8i2.1095

Waldman, B. P., & James, D. (1998). Photographs the 1996 Presidential Election: The Question of Bias. *Journalism and Mass Communication Quarterly*, *75*(2), 302–311. https://doi.org/10.1177/107769909807500206

Walker, M. A., Mehl, M. R., Moore, R. K., & Mairesse, F. (2007). Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. *Journal of Artificial Intelligence Research*, *30*(40), 457–500. https://www.jair.org/index.php/jair/article/view/10520zz%0Ahttps://pdfs.semanticscholar.org/7a5e/a7c513aac77880d37bd98527f4d24dda44f8.pdf

Wang, S., Fang, H., Khabsa, M., Mao, H., & Ma, H. (2021). Entailment as few-shot learner. *arXiv preprint arXiv:2104.14690*. https://doi.org/10.48550/arXiv.2104.14690

Wang, Y., Yen, J., & Reitter, D. (2015). Pragmatic Alignment on Social Support Type in Health Forum Conversations. In T. O'Donnell & M. van Schijndel (Eds.), *Proceedings of the 6th workshop on cognitive modeling and computational linguistics* (pp. 9–18). Association for Computational Linguistics. https://doi.org/10.3115/v1/w15-1102

Ward, A., & Litman, D. (2007). Automatically measuring lexical and acoustic/prosodic convergence in tutorial dialog corpora. *Proceedings of the SLaTE Workshop on Speech and Language Technology in Education*, 57–60. http://www.isca-speech.org/archive/slate_2007

Watts, D. J. (2007). A twenty-first century science. *Nature*, *445*(7127), 489–489.

Wei, J., & Zou, K. (2019). EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the conference on empirical methods in natural language processing and 9th international joint conference on natural language processing* (pp. 6382–6388). Association for Computational Linguistics. https://doi.org/10.18653/v1/d19-1670

Weise, A., & Levitan, R. (2018). Looking for Structure in Lexical and Acoustic-Prosodic Entrainment Behaviors. In M. Walker, H. Ji, & A. Stent (Eds.), *Proceedings of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 297–302). Association for Computational Linguistics. https://doi.org/10.18653/v1/n18-2048

Weise, A., Levitan, S. I., Hirschberg, J., & Levitan, R. (2019). Individual differences in acoustic-prosodic entrainment in spoken dialogue. *Speech Communication*, *115*, 78–87. https://doi.org/10.1016/j.specom.2019.10.007

Wen, T.-H., Gasic, M., Kim, D., Mrksic, N., Su, P.-H., Vandyke, D., & Young, S. (2015). Stochastic language generation in dialogue using recurrent neural networks with convolutional sentence reranking. *arXiv preprint arXiv:1508.01755*.

Wiedemann, G. (2013). Opening up to Big Data : Computer-Assisted Analysis of Textual Data in Social Sciences. *38*(4), 332–357.

Willi, M. M., Borrie, S. A., Barrett, T. S., Tu, M., & Berisha, V. (2018). A discriminative acoustic-prosodic approach for measuring local entrainment. In C. Sekhar, P. Rao, P. Ghosh, H. Murthy, B. Yegnanarayana, S. Umesh, P. Alku, S. Prasanna, & S. Narayanan (Eds.), *19th annual conference of the international speech communication association: Speech research for emerging markets in multilingual societies* (pp. 581–585). https://doi.org/10.21437/Interspeech.2018-1419

Williams, A., Nangia, N., & Bowman, S. (2018). A broad-coverage challenge corpus for sentence understanding through inference. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1112–1122. https://doi.org/10.18653/v1/n18-1101

Williams, R., Elliott, I. A., & Beech, A. R. (2013). Identifying sexual grooming themes used by internet sex offenders. *Deviant Behavior*, *34*(2), 135–152. https://doi.org/10.1080/01639625.2012.707550

Wilson, M. (1988). MRC Psycholinguistic Database : Machine Usable Dictionary. *Behavior Research Methods, Instruments, & Computers*, *20*(1), 6–10. https://doi.org/10.3758/bf03202594

Winters, G. M., Kaylor, L. E., & Jeglic, E. L. (2017). Sexual offenders contacting children online: An examination of transcripts of sexual grooming. *Journal of sexual aggression*, *23*(1), 62–76. https://doi.org/10.1080/13552600.2016.1271146

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (2020). Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 38–45.

Wolfman, M., Brown, D., & Jose, P. (2016). Talking past each other: Interviewer and child verbal exchanges in forensic interviews. *Law and Human Behavior*, *40*(2), 107–117. https://doi.org/10.1037/lhb0000171

Wynn, C. J., & Borrie, S. A. (n.d.). *Classifying Conversational Entrainment of Speech Behavior: An Updated Framework and Review*.

Xu, Y., & Reitter, D. (2015). An Evaluation and Comparison of Linguistic Alignment Measures. In T. O'Donnell & M. van Schijndel (Eds.), *Proceedings of the 6th workshop on cognitive modelling and computational linguistics* (pp. 58–67). Association for Computational Linguistics. https://doi.org/10.3115/v1/w15-1102

Xu, Y., Chen, L., Fang, M., Wang, Y., & Zhang, C. (2020). Deep reinforcement learning with transformers for text adventure games. *2020 IEEE Conference on Games (CoG)*, 65–72.

Yan, J. L. S., McCracken, N., & Crowston, K. (2014). Semi-automatic content analysis of qualitative data. *IConference 2014 Proceedings*.

Yang, F., Huang, Z., Scholtz, J., & Arendt, D. L. (2020). How do visual explanations foster end users' appropriate trust in machine learning? *International Conference on Intelligent User Interfaces, Proceedings IUI*, 189–201. https://doi.org/10.1145/3377325.3377480

Yap, B. W., Rani, K. A., Rahman, H. A. A., Fong, S., Khairudin, Z., & Abdullah, N. N. (2014). An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets. *Proceedings of the first inter-*

national conference on advanced data and information engineering (DaEng-2013), 13–22.

Yarkoni, T., Eckles, D., Heathers, J. A. J., Levenstein, M. C., Smaldino, P. E., & Lane, J. (2021). Enhancing and Accelerating Social Science Via Automation: Challenges and Opportunities. *Harvard Data Science Review*, (3). https://doi.org/10.1162/99608f92.df2262f5

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122.

Yin, M., Vaughan, J. W., & Wallach, H. (2019). Understanding the effect of accuracy on trust in machine learning models. *Conference on Human Factors in Computing Systems - Proceedings*, 1–12. https://doi.org/10.1145/3290605.3300509

Young, S. (2008). The broadcast political interview and strategies used by politicians: How the Australian prime minister promoted the Iraq War. *Media, Culture and Society*, *30*(5), 623–640. https://doi.org/10.1177/0163443708094012

Zheng, Q., Tang, Y., Liu, Y., Liu, W., & Huang, Y. (2022). UX Research on Conversational Human-AI Interaction: A Literature Review of the ACM Digital Library. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3491102.3501855

Zhou, L., Burgoon, J. K., Nunamaker, J. F., & Twitchell, D. (2004). Automating Linguistics-Based Cues for detecting deception in text-based asynchronous computer-mediated communication. *Group Decision and Negotiation*, *13*(1), 81–106. https://doi.org/10.1023/B:GRUP.0000011944.62889.6f

Zingerle, A. (2015). Scambaiters, human flesh search engine, perverted justice, and internet haganah: Villains, avengers, or saviors on the internet. *ISEA Conference*.

Zini, J. E., & Awad, M. (2022). On the explainability of natural language processing deep models. *ACM Computing Surveys (CSUR)*.

Appendix

A.1 Chapter 3: Instructions to human annotators during interview evaluation survey

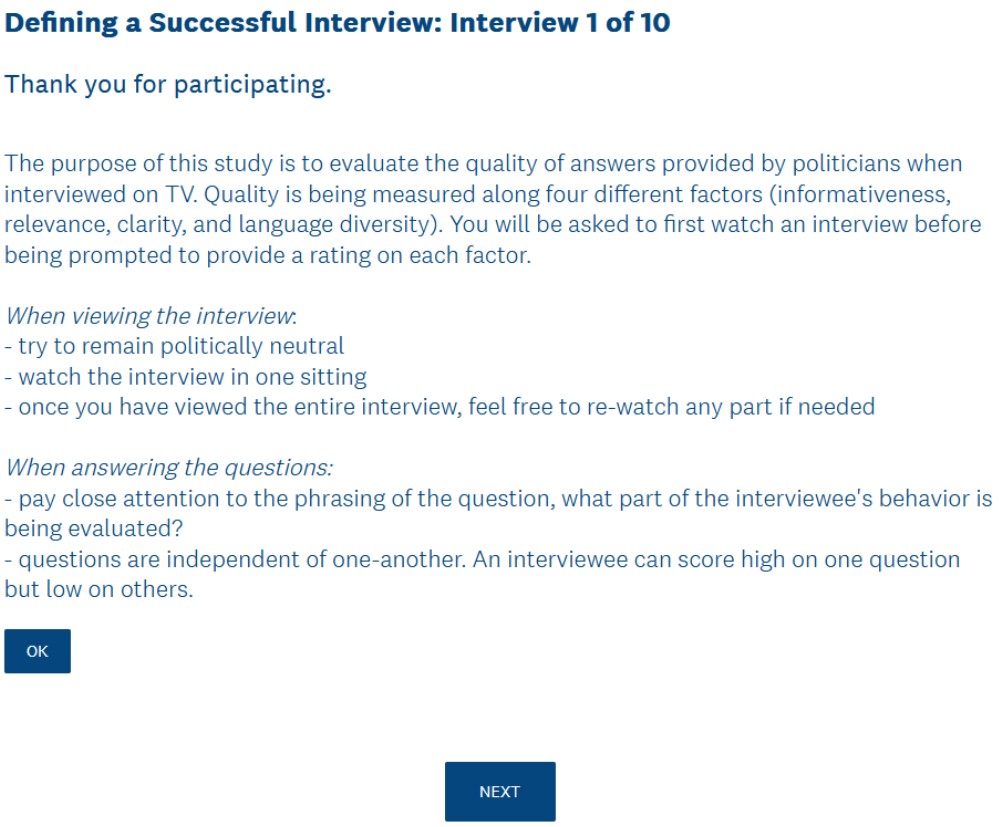

Figure S1: Survey instructions provided to human annotators as described in Section 3.4.5

Figure S2: Example of interview embedded into survey as part of annotation task described in described in Section 3.4.5

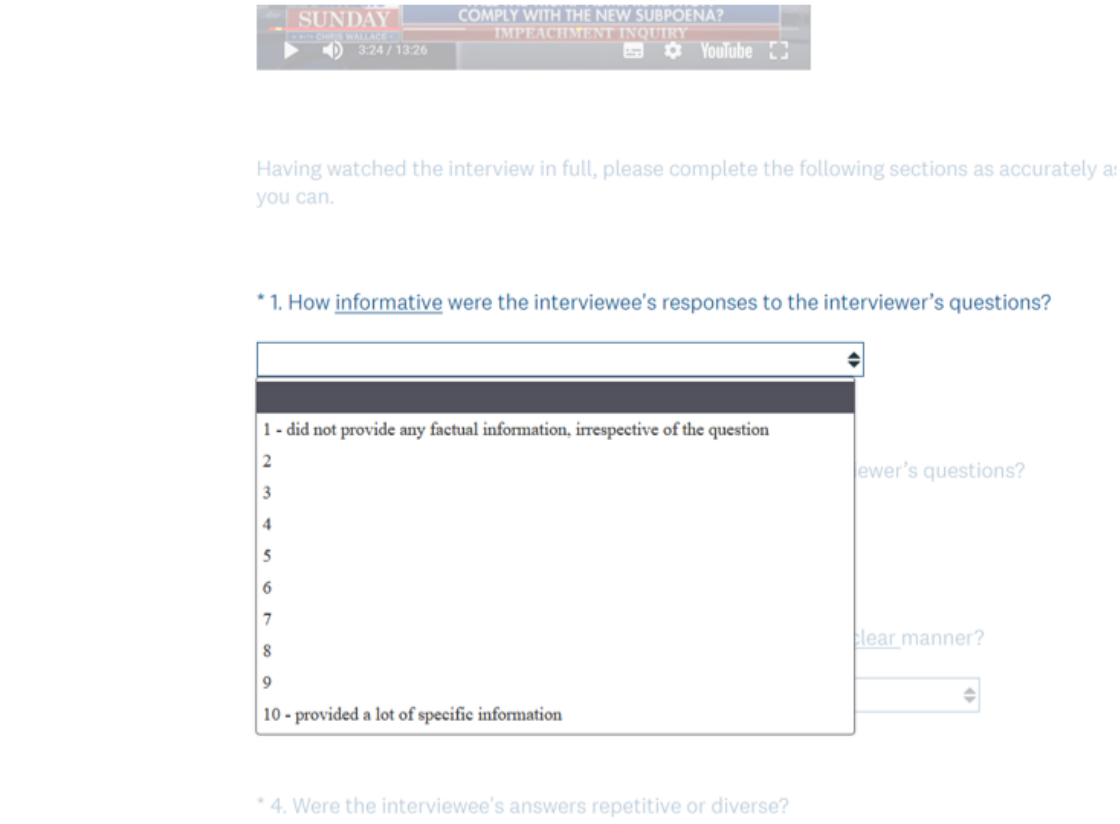

Figure S3: Example of question layout given to annotators during Section 3.4.5

A.2 Chapter 6: Questions included as part of LF2

Rapport

- is the offender giving a compliment?
- is the offender accepting a complement?
- is the offender building a special bond?
- is the offender being romantic?
- is the offender showing interest?
- is the offender talking about personality?
- is the offender talking about personal similarities?

Control

- is the offender being persistent?
- is the offender talking about consent?
- is the offender trying to please the victim?
- is the offender complying with requests?
- is the offender jealous?
- is the offender being compliant?
- is the offender being assertive?
- is the offender asking a rhetorical question?
- is the offender being patronising?
- is the offender asking for permission?
- is the offender checking for engagement?
- is the offender seeking permission?

Challenge

- is the offender mocking the victim?
- is the offender insulting the victim?
- is the offender confronting the victim?
- is the offender rejecting the victim?
- does the victim trust the offender?

Negotiation

- is the offender offering incentives?
- is the offender making plans to meet?
- is the offender persuading the victim?
- is the offender defensive?
- is the offender talking about alcohol?
- is the offender talking about drugs?
- is the offender arranging plans?

Use of Emotions

- is the offender showing concern?
- is the offender looking for validation?
- is the offender shocked?
- is the offender angry?
- is the offender sad?
- is the offender confused?
- is the offender embarrassed?
- is the offender happy?

- does the offender reassure the victim?
- does the offender ask for reassurance?
- does the offender reassure acknowledge wrongdoing?
- does the offender blackmail the victim?
- does the offender apologise?

Testing Boundaries

- does the offender set boundaries?
- does the offender check the victim's willingness to engage?
- does the offender talk about sex?
- does the offender talk about relationships?
- does the offender talk about sharing pictures?
- does the offender talk about meeting offline?
- does the offender talk about fantasies?
- does the offender talk about sharing pictures?
- is the offender being secretive?
- is the offender bored?

Use of Sexual Topics

- is the offender talking about sexual topics?
- is the offender talking about fantasies?
- is the offender talking about sexual preferences?
- is the offender talking about pornography?
- is the offender talking about sexual acts?
- is the offender talking about relationships?
- is the offender talking about age differences?

Mitigation

- does the offender implicate themselves in a crime?
- does the offender have a sexual preference for children?

Encouragement

- does the offender express willingness to engage?
- does the offender encourage the victim?
- does the offender comply with the victim?
- does the offender flirt with the victim?
- does the offender request a picture of the victim?

Risk Management

- does the offender ask if the victim is real?
- does the offender ask if the victim is a cop?
- does the offender ask about the victim's mom?
- does the offender ask about the victim's dad?
- does the offender ask about the victim's family?
- does the offender talk about the dangers on the internet?
- does the offender ask about meeting the victim?

A.3 Chapter 6: Keywords included as part of LF5

Rapport

- would
- time
- good

- cute
- baby
- ur
- really
- want
- love
- like

Control

- good
- feel
- dont
- know
- see
- wanna
- tell
- would
- want
- like

Challenge

- well
- late
- cause
- right

- would
- ur
- really
- know
- yea
- dont

Negotiation

- ur
- call
- love
- ll
- time
- well
- get
- would
- like
- want

Use of Emotions

- ur
- call
- love
- ll
- time

- well
- get
- would
- like
- want

Testing Boundaries

- think
- wanna
- ur
- well
- see
- know
- get
- would
- want
- like

Use of Sexual Topics

- sex
- yea
- dick
- cum
- get
- well

- ur
- want
- would
- like

Mitigation

- phone
- dont
- ex
- call
- one
- want
- boys
- like
- he
- really

Encouragement

- yea
- sure
- really
- good
- yes
- like
- nice

- want
- kool
- cool

Risk Management

- get
- alone
- could
- know
- home
- long
- parents
- mom
- dad
- ur