2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)

FB-SEC-1: A Social Emotion Cause Dataset

Abdullah Alsaedi *Department of Computer Science*
University of Liverpool & Taif University
Liverpool, UK & Taif, Saudi Arabia
a.alsaedi@tu.edu.sa

Stuart Thomason *Department of Computer Science*
University of Liverpool
Liverpool, UK
s.thomason@liverpool.ac.uk

Floriana Grasso *Department of Computer Science*
University of Liverpool
Liverpool, UK
floriana@liverpool.ac.uk

Phillip Brooker *Department of Sociology, Social Policy and Criminology*
University of Liverpool
Liverpool, UK
p.d.brooker@liverpool.ac.uk

October 4, 2023

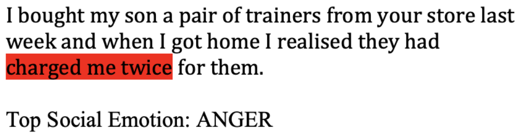
Abstract

"Social emotion" in text mining refers to the emotion experienced by the reader exposed to a text, as opposed to the emotion conveyed from the author's perspective. Mining the cause of social emotion from text is a new challenging task with a wide range of applications, but its progress is hindered by the lack of annotated datasets. In this paper, we release the first English dataset for social emotion cause. The dataset is based on a well-established corpus of Facebook posts, and it was annotated through a crowdsourcing experiment. Together with the dataset, we provide two baseline models to be used as benchmarks for future studies.

social emotion cause extraction, SECE, emotion analysis

1 Introduction

The analysis and detection of emotions in text, especially social media content, with the aim to provide a better understanding of the text possibly based on



I bought my son a pair of trainers from your store last week and when I got home I realised they had charged me twice for them.

Top Social Emotion: ANGER

Figure 1: An example of the social emotion extraction task. The highlighted span is the cause of anger, which is the top social emotion.

psychological emotion models, is an ever-growing area with an established corpus of methods and techniques [1]. The vast majority of this work focuses on determining which emotions are in fact 'expressed' in a text, or, in other words, which emotions could be attributed to the text *from the writer's perspective*. Gaining more and more attention, however, is work aimed at establishing the emotions evoked *in the readers* when they are exposed to the text, with the aim to analysing or predicting the effect of the text on the reader, with various applications in human-computer interaction, psychology, media, and so on [2,3]. The term that has established itself in the field to refer to this phenomenon, borrowed and specialised from the concept in behavioural sciences, is *social emotion* [4].

Mining social emotion in text helps predict how reading the textual content will affect the user's emotion, but it does not reveal what, specifically, in the text triggered such emotion, something that could have many applications, particularly on social media, such as exploring emotional reactions to events and overall improving user experience [5]. Finding out which specific part of the text reveals the emotion of the writer is a well understood task, going under the name of Emotion Cause Extraction (ECE) [6,7], and in the same way the emerging field of Social Emotion Cause Extraction (SECE) seeks to extract either the cause or causes that are most related to the dominant emotion(s) that the readers experienced [8]. Figure 1 depicts a typical example of an output from this task, where the social emotion of Anger was attributed to the "charged me twice" piece of text.

The difficulty of SECE as a task lies in the fact that, unlike ECE tasks, the clue to the elicited emotion may not explicitly appear in the text, as well as being potentially different from reader to reader due to their experience and background.

Progress in the SECE task is seriously impacted by the lack of relevant datasets [8]. In this paper, we contribute to filling this gap by releasing what is, to the best of our knowledge, the first English dataset specifically annotated for social emotion cause. The dataset is based on the well-established FacebookR corpus [9] which was further annotated through a crowdsourcing experiment.

The paper is structured as follows: we first report on related work, especially commenting on existing datasets that support such work. Then we describe how we constructed our dataset[1], starting from FacebookR, through the experiment.

[1] In this submission, we present some samples of the content of the dataset, the full URL

Then, as a way to provide baselines for future work, we provide two models for span labelling using our dataset, before our concluding remarks.

2 Related Work

Understanding how SECE might work as a task is underpinned by understanding how generally ECE works, therefore we will start our report on related work with a selection of relevant ECE research, before moving into what is available to scholars attempting SECE.

2.1 Emotion Cause Extraction Task

The Emotion Cause Extraction task, firstly introduced by Lee et al. [6], is an established task that is concerned with extracting the emotion cause represented in a text from the writer's point of view. The majority of early work in emotion cause extraction relied on rule-based methods. Chen et al. [10] presented a rule-based model for the emotion cause extraction based on the work of Lee et al. [6] with the ultimate goal to develop linguistic rules for detecting the emotion cause in the text.

Gui et al. [11] approached the task as a clause-level classification problem, in which the text is divided into clauses, and the goal is to find the clause that contains the cause.

Li et al. [12] combined a Bi-directional Long Short-Term Memory network, Co-Attention mechanism, and Convolutional neural network in a unified framework to extract the emotion cause.

Bostan et al. [13] used Conditional Random Field (CRF) model and Bi-directional Long Short-Term Memory network with a CRF layer (BiLSTM-CRF) models in the evaluation for the automatic extraction of emotion cause, and demonstrated the challenges of both the emotion cause annotation and extraction tasks.

Recently, most works aim to extract pairs of clauses or spans from the text that represent the emotion and the cause, in a task known as Emotion-Cause Pair Extraction (ECPE) [7, 14–17].

2.2 Datasets for Emotion Cause Extraction

The first dataset for emotion cause extraction [6] was built from the Academia Sinica Chinese Corpus, labelled with Turner emotion model [18], using predefined emotion keywords, and then manually annotated with the emotion cause.

Neviarouskaya [19] manually annotated 500 English sentences with emotion, experiencer, and cause, aiming to extract the linguistic relation between the emotion and its cause.

link to the data was eliminated to preserve anonymity but will be included in the final version.

ElectoralTweets [20] is a dataset for semantic roles in tweets that includes the emotion cause and is part of an automatic system for emotion cause extraction. The data was collected from Twitter hashtags regarding the 2012 US presidential elections. Interestingly, the cause which is labelled in the dataset is taken from a set of pre-defined entities, rather than being an actual span in the original text.

Ghazi et al. [21] manually annotated emotion cause in sentences using FrameNet [22] emotion framework.

Gui et al. [11] released a dataset for emotion cause extraction collected from SINA city news and used explicit emotion keywords presented in the text for the emotion labelling. The emotion cause annotation process was carried out manually by two annotators, with the third annotator only intervening when there were conflicts.

Gao et al. [23] manually annotated a corpus of English and Chinese texts with emotion and emotion cause for the evaluation of the emotion cause analysis task.

REMAN [24] is a relational emotion annotation for entities/events in the emotion interaction.

Finally on our list, GoodNewsEveryone [13] is an English news headlines dataset annotated through crowdsourcing with semantic roles in addition to the emotion cause. The dataset consists of 5000 news headlines, annotated with spans that represent the emotion cause in the text.

2.3 Social Emotion Cause Datasets

When it comes to Social Emotion Cause Dataset, the selection is very limited. Xiao et al. [8, 25] established the social emotion cause extraction as a task, and constructed a dataset from news articles collected from Weibo, the widely popular Chinese micro-blogging platform. The dataset annotation was performed manually for each post, by selecting the comments with highest number of likes, and considering the emotions expressed by the writer of such comments as as the social emotion related to the post. Such social emotion cause is then annotated manually back in the dataset. The dataset consists of 1000 articles, labelled with the following emotions: Happy, Sad, Gratitude, and Anger.

This dataset is the only published dataset for social emotion cause extraction (SECE), to the best of our knowledge, and none exists in English.

3 The first English SECE dataset: FB-SEC-1

The starting point for FB-SEC-1 is FacebookR [9], a dataset scraped from Facebook posts, with the reactions of their readers representing the social emotion. Posts are taken from the customer service page of 12 major retailers from UK and US, mostly written by customers. The reactions collected belong to the initial post (not to the replies). We follow the sampling procedures of [9] that filters out posts without any reaction, and posts with only *"like"* reaction, as it is deemed ambiguous, because it is used both to show support, and to simply show that the post was read. The final dataset we used for our experiment consists of 8103 posts, distributed across the top emotion suggested by the Facebook reaction, as shown in the following table:

Top emotion	**Number of posts**
ANGRY	2276
HAHA	2253
LOVE	1648
WOW	1237
SAD	689

3.1 Annotation Approach

For the annotation, we used Amazon's Mechanical Turk (MTurk) service, which has reportedly been used with success for analogous tasks [26]. In this service, a *Requester* can submit and organise a diverse range of so called Human Intelligence Tasks (*HITs*), and users (known as *Workers*) can search *HITs* using keywords related to the tasks they are willing to participate in, and receive a (small) sum of money for each task they complete.

We used FacebookR to run an experiment where participants *(Workers)* on MTurk were provided a survey and asked to annotate posts with the emotion cause.

In order to perform this task, we first needed to decide the annotation mechanism. There are three typical approaches:

1. Categorising the emotion cause into predefined entities [20]
2. Clause-level binary labelling, i.e. identifying whether or not the clause contains the cause [11], and
3. Span labelling, identifying the tokens of the text that represent the emotion cause [13, 21, 24].

The first approach is rarely used, and it is only effective when the causes are limited and well defined. Oberländer et al. [27] compared the clause-level labelling approach with the span labelling approach for emotion cause extraction in English over four datasets, demonstrating that the span labelling approach outperforms the clause-level labelling approach in three out of the four. This prompted us to use the span labelling approach in the construction of our dataset.

3.2 Annotation task: Challenges and Solutions

Although there are certain advantages to employing MTurk, including low cost, low complexity, and fast response time, there are also obvious issues. The most significant issue is quality control. The generated reward might encourage cheaters who submit random answers as well as harmful annotators who could intentionally submit incorrect data. Furthermore, we have no influence over the workers' educational background, and we cannot assume the typical *Worker* will fully read and understand the given instructions. Another issue is obtaining enough *Workers* who are willing to do the task. If the task does not necessitate any specific expertise, more *Workers* will be assigned to it. Moreover, *Workers*' engagement with the task is also affected by how desirable the activity is to them as well as how appealing the reward is.

Building a dataset for emotion in text is challenging due to the nature of emotion and how we understand it [13, 28]. However, annotating posts with social emotion and their cause is even more challenging, as the emotion can be different from one person to another depending on their background and experiences. For example, the sentence *"France won the World Cup with a 4-2 win over Croatia"*, will provoke happiness in France supporters, but not in Croatia supporters. Moreover, in the same sentence, the main cause of happiness for some people could be winning the World Cup, while for others, it could be the score difference.

One of the challenges that annotators can face is deciding whether to label the emotion that the reader had or the emotion that the author of the post experienced. For instance, in the post *"I was surprised that your store uses CCTVs."*, the customer is clearly expressing surprise, but the reader is unlikely to be surprised by this information.

In order to mitigate the huge potential variation due to the personal experience of the annotators and with the goal of obtaining annotations that represent as many people as possible, we asked the annotators to see the text from the point of view of the "typical reader". This would have, we hoped, also the added benefit of minimising the number of "neutral" responses, when the annotators themselves have no interest in or emotions towards the text.

Low-effort responses increase mistakes and necessitate extensive manual review [29]. The inadequate effort is a key challenge, especially in crowdsourcing, where participants are uncontrolled and are more encouraged to finish surveys as quickly as possible in order to maximise their reward [30]. Participants who do not read or follow instructions, overlook critical elements of the experiment, or create random or irrelevant responses are considered negligent and rejected from participating in any further assignments. The instructions ask the annotator to write *"No Cause"* if the text does not contain any emotion-provoking causes. That, we believe, will push the *Worker* to focus and limit the number of easy-to-submit empty responses.

With this in mind, and given that we decided to use a span labelling approach, the annotation task for a post consisted in revealing, together with the post, what emotion label was associated with the post (from the FacebookR dataset) and asking the annotator to identify which span is best associated with such an emotion. The question was as follows (if for example the post was labelled with ANGER):

> *"This post is labelled with the emotion of ANGER. Please COPY and PASTE the shortest amount of text that provokes that emotion."*.

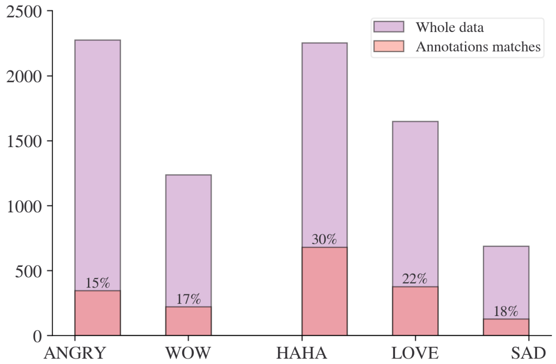


Figure 2: The percentage of annotation matches for each social emotion label.

3.3 Annotation Experiment: Setup, Collection, and Refinement

In order to obtain large-scale annotations for social emotion cause, we administered the experiment using crowdsourcing, through Amazon's Mechanical Turk platform (MTurk). The task has been clearly defined in simple and brief language in order to achieve high-quality annotations, as it is noted that when brief and simple instructions are given, the MTurk platform has proven to provide high inter-annotator agreements [26].

Moreover, rules have been established to filter the *Workers* who can see and contribute to the submission. We only accepted annotators with a reliable history (at least a thousand approved annotations in their history and approval received for at least 90% of their annotations).

For each post, we requested three annotations. We divided the data into batches of 100 posts each, to make it easier for the annotator to follow and finish. The MTurk platform allows us to specify the desired criteria for the targeted annotators of the survey. As the FacebookR dataset is made up of posts from popular supermarkets in the United States and the United Kingdom, we were able to send data from US supermarkets to annotators based in the United States, and we did the same with data from supermarkets in the UK. The targeted country for MTurk Workers for each supermarket in FacebookR is as follows:

Country	**Supermarket**
United States	Amazon Walmart Target Bestbuy Costco Macys Publix Safeway Walgreens
United Kingdom	Tesco Sainsburys AldiUK

We manually reviewed for approval all of the collected annotations, starting with the post-processing step for annotations that had common mistakes, like dots at the end, or new line markers. We did not accept spans of social emotion causes that were not part of the original text. We also rejected empty submissions that did not specify whether a cause existed. Multiple manual reviews were carried out to make sure that wrong or accidental annotations were caught, rejected, and resubmitted for another round of annotation.

A complete HIT sample for one of the posts is provided in the Sample in Table 2.2. It should be noted that the *Workers* must fill in every text field with an answer and must not leave any blank fields. The survey was approved by out The University of Liverpool ethics committee (Ethic Approval No. 7800), as were the participant's consent form and instructions that was presented to the *Workers* at the beginning of each task.

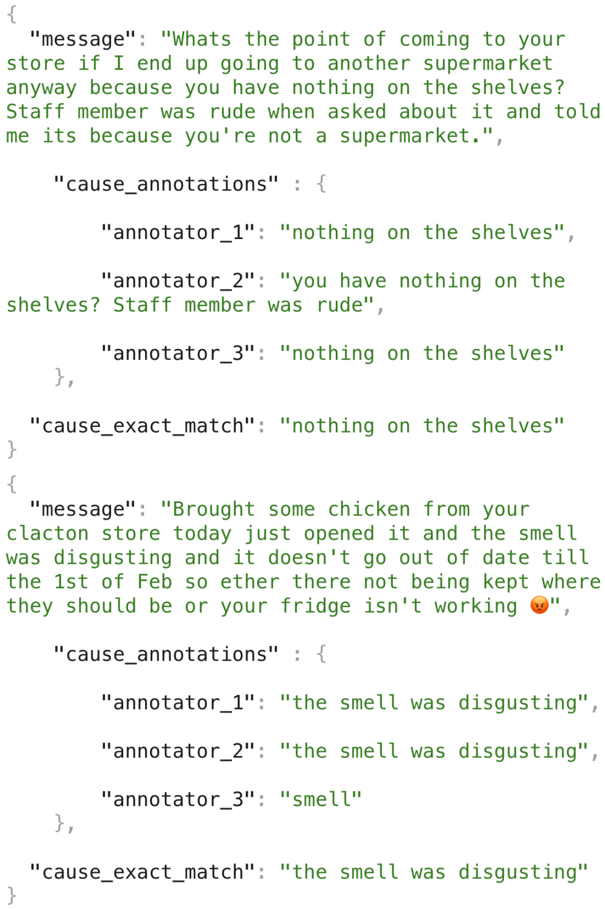
3.4 Inter-annotator Agreements

As previously stated, the experiment intended to collect three annotations of social emotion cause for each post from FacebookR. To measure how frequently the annotators agreed with one another, we used Fleiss' Kappa (κ) [31]. The inter-annotator agreements for the annotated dataset, with the average accuracy of matches between all pairs of annotators (%) is as follows:

Fleiss' Kappa (κ)	**Accuracy (%)**
0.07	8.95

while a detailed comparison between the annotators using the Cohen's Kappa (κ) metric, with the relative accuracy is as follows:

Measure	**a vs b**	**b vs c**	**a vs c**
Cohen Kappa's (κ)	0.06	0.07	0.09
Accuracy (%)	8.0	8.3	10.4



```
{
  "message": "Whats the point of coming to your
store if I end up going to another supermarket
anyway because you have nothing on the shelves?
Staff member was rude when asked about it and told
me its because you're not a supermarket.",

    "cause_annotations" : {

        "annotator_1": "nothing on the shelves",

        "annotator_2": "you have nothing on the
shelves? Staff member was rude",

        "annotator_3": "nothing on the shelves"
    },

  "cause_exact_match": "nothing on the shelves"
}
{
  "message": "Brought some chicken from your
clacton store today just opened it and the smell
was disgusting and it doesn't go out of date till
the 1st of Feb so ether there not being kept where
they should be or your fridge isn't working 😡",

    "cause_annotations" : {

        "annotator_1": "the smell was disgusting",

        "annotator_2": "the smell was disgusting",

        "annotator_3": "smell"
    },

  "cause_exact_match": "the smell was disgusting"
}
```

Figure 3: Examples of the social emotion cause annotation fields in the dataset.

We can see that the agreement is low, and we attribute this to a range of factors. First, as many prior works have emphasised [13, 28], the difficulty of emotion annotation task itself. Second, there is a high number of posts that are not emotionally charged, making them difficult for the annotator to identify the emotion as well as the emotion cause. Moreover, the exact match agreement is strict. Annotators may disagree on minor details such as including names, letters, commas, and so on into the annotation.

For many posts, the annotators have not agreed on any spans. To investigate how annotators' agreements may differ depending on the emotion class, we analysed posts with at least two annotation agreement for the following social emotions labels: ANGRY, WOW, HAHA, LOVE, and SAD. Figure 2 shows the percentage of annotation agreements for each social emotion label.

Positive emotions labels such as HAHA and LOVE appear to have more agreement than negative ones such as ANGRY and SAD, while WOW may refer to both positive and negative surprise.

3.5 Dataset Description

FacebookR was released as a MongoDB database that stores documents in a JSON-like format. FB-SEC-1 is released in JSON format and adds the annotations of the social emotion cause to posts in the FacebookR dataset (see Figure 3). For the 8013 posts, we have added three social emotion cause annotations. Furthermore, we have added a field for exact annotation matches that have a majority agreements between the annotators. The resulting dataset includes annotations for 8103 posts, as well as 1749 exact annotation matches for posts with a majority of annotator agreements, 1034 of which are agreements on exact span matches and 715 of which are labelled as "No Cause". Figure 3 shows examples of the annotated posts.

4 Span Labelling Baselines

Baseline models that work using intuitive features have been built to evaluate the challenge of identifying the causes of social emotion. The dataset contains posts annotated with social emotion and cause; therefore, we can investigate the opportunity to train a machine learning model to extract the social emotion cause. To estimate the consistency of the FB-SEC-1 and hopefully boost the task of social emotion cause extraction in English, we adopted the span labelling approach following recent research [32] and built two baseline models for span labelling.

The first model is a Conditional Random Field (CRF) [33] that has been trained on the following features: word and part-of-speech tags for each word, as well as previous and succeeding words. The part-of-speech tags were extracted using the SpaCy library[2]. With LBFGS [34], the model was trained for 200 iterations.

The second model employs Bi-directional Long Short-Term Memory with a Conditional Random Field layer (BiLSTM-CRF) that was trained on GloVe [35] embedding of 200 dimensions, 30 epochs, and 100 batch size.

Because there are many posts with no annotation agreements, we only considered posts with at least two exact annotation matches in order to train the models on gold standard data. the inter-annotator agreements for the selected subset of posts, both including and excluding those with the *"No Cause"* label match is as follows

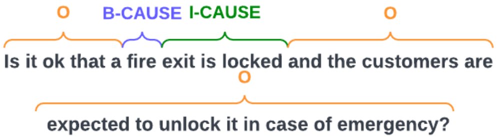
Data Samples	**Fleiss' Kappa**	**Accuracy**
Incl. *"No Cause"*	0.33	40.38
Excl. *"No Cause"*	0.37	37.52

[2] https://spacy.io/

A detailed comparison of the inter-annotator agreement using the Cohen's Kappa (κ) metric is as follows:

Data Samples	**a vs b**	**b vs c**	**a vs c**
Including "No Cause"	0.28	0.30	0.41
Without "No Cause"	0.33	0.31	0.46

As a sequence labelling problem, the data was encoded using the inside-outside-beginning (IOB) scheme, which allowed the models to capture the relationship between the words. This is an example of how the cause "fire exit is locked" is encoded:


O
B-CAUSE
I-CAUSE
O
Is it ok that a fire exit is locked and the customers are
O
expected to unlock it in case of emergency?

We evaluated the exact sequence matches using the Accuracy metric, as well as Precision, Recall, and F1-score for the cause span only, which we consider the most relevant. The performance of the baselines is shown in Table 2. As shown in the table, the BiLSTM-CRF model outperforms the CRF model in terms of the token partial match: Precision, Recall, and F1-score. However, the CRF model is slightly superior in terms of exact match Accuracy, with a nearly 2% increase.

The performance is low for both models, particularly for the exact sequence match. However, considering the difficulty of detecting the exact match [13] and the fact that this work is still in its early stages, this performance represents a step toward the development of more accurate models. Furthermore, model performance in detecting the cause span is promising, particularly for *BiLSTM-CRF*, where all precision, recall, and F1-score were significantly higher than *CRF*, with a precision of 0.59, a recall of 0.52, and an F1-score of 0.55.

5 Conclusion

In this paper, we released FB-SEC-1, the first English dataset for social emotion cause, and discussed the challenges of labelling text with social emotion cause and how we overcame them using a crowdsourcing platform, Amazon's Mechanical Turk. Moreover, we created baseline models on the dataset to be used in the evaluation of future models. This is, we believe, a significant step towards the development of established social emotion cause mining techniques and will hopefully lead to future research on how annotators' partial agreement can be used to increase the dataset size and improve training performance. Social emotion cause extraction is an exciting avenue of research on social communication, and we hope that joint learning approaches based on social emotions can be used to better identify their causes.

6 Ethical Impact Statement

The annotation experiment received ethical approval from our own institutions, and much care was taken not only with the specification of the experiment itself but also with determining a fair fare to reimburse participants for their efforts.

Participants received a consent form for the data collection process, as approved by our institution, a participant information sheet, where the purpose of the study was explained, and the way in which data was going to be stored, and were clearly communicated they could withdraw from the experiment at any time.

In order to determine whether the fare and the allocated time was fair, we conducted a preliminary, small collection, where we monitored the response time for calibration in the full experiment.

Besides the data collection task, we think it is important also to consider the high level implications of our research.

Identifying emotions that a reader of a text might experience poses a significant privacy concern as personal data might be used to identify and extract the social emotion cause, therefore, any practical implementation of such a tool must ensure that readers are informed about their collected data. While a valuable research tool, it would be concerning if such a facility were to be introduced for instance as an automatic feature of social media posts unbeknown to the readers, and more importantly so if readers are potentially vulnerable.

Accuracy and bias in SECE models is also a concern, emotion analysis models, just as any data analysis models, have been shown to perpetuate and in fact magnify bias towards specific characteristics.

Generally, a tool that can potentially detect what emotions a text evokes in the reader is open to a high degree of potentially malicious use. Such a tool would fall into what for instance the EU Artificial Intelligence Act [36] considers high risk application, and developers including such a feature would be liable to measure their objectives against standards [37].

References

[1] P. Nandwani and R. Verma, "A review on sentiment analysis and emotion detection from text," *Social Network Analysis and Mining*, vol. 11, no. 1, pp. 1–19, 2021.

[2] M. De Choudhury and E. Kiciman, "The language of social support in social media and its effect on suicidal ideation risk," in *Eleventh International AAAI Conference on Web and Social Media*, 2017.

[3] K. Park, H. Kwak, J. An, and S. Chawla, "How-to present news on social media: A causal analysis of editing news headlines for boosting user engagement.," in *ICWSM*, pp. 491–502, 2021.

[4] X. Li, Q. Peng, Z. Sun, L. Chai, and Y. Wang, "Predicting social emotions from readers' perspective," *IEEE Transactions on Affective Computing*, vol. 10, no. 2, pp. 255–264, 2017.

[5] J. P. Jokinen, "Emotional user experience: Traits, events, and states," *International Journal of Human-Computer Studies*, vol. 76, pp. 67–77, 2015.

[6] S. Y. M. Lee, Y. Chen, S. Li, and C. R. Huang, "Emotion cause events: Corpus construction and analysis," *Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC 2010*, pp. 1121–1128, 2010.

[7] R. Xia and Z. Ding, "Emotion-cause pair extraction: A new task to emotion analysis in texts," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1003–1012, 2019.

[8] X. Xiao, L. Wang, Q. Kong, and W. Mao, "Social emotion cause extraction from online texts," in *2020 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pp. 1–6, IEEE, 2020.

[9] F. Krebs, B. Lubascher, T. Moers, P. Schaap, and G. Spanakis, "Social emotion mining techniques for facebook posts reaction prediction," in *Proceedings of the 10th International Conference on Agents and Artificial Intelligence*, pp. 211–220, 2018.

[10] Y. Chen, S. Y. M. Lee, S. Li, and C. R. Huang, "Emotion cause detection with linguistic constructions," *Coling 2010 - 23rd International Conference on Computational Linguistics, Proceedings of the Conference*, vol. 2, no. August, pp. 179–187, 2010.

[11] L. Gui, D. Wu, R. Xu, Q. Lu, and Y. Zhou, "Event-driven emotion cause extraction with corpus construction," *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, pp. 1639–1649, 2016.

[12] X. Li, K. Song, S. Feng, D. Wang, and Y. Zhang, "A Co-Attention Neural Network Model for Emotion Cause Analysis with Emotional Context Awareness," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4752–4757, 2018.

[13] L. A. M. Bostan, E. Kim, and R. Klinger, "GoodNewsEveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception," in *Proceedings of the 12th Language Resources and Evaluation Conference*, (Marseille, France), pp. 1554–1566, European Language Resources Association, May 2020.

[14] A. Singh, S. Hingane, S. Wani, and A. Modi, "An End-to-End Network for Emotion-Cause Pair Extraction," in *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pp. 84–91, 2021.

[15] J. Yu, W. Liu, Y. He, and C. Zhang, "A Mutually Auxiliary Multitask Model with Self-Distillation for Emotion-Cause Pair Extraction," *IEEE Access*, vol. 9, pp. 26811–26821, 2021.

[16] Y. Chen, W. Hou, S. Li, C. Wu, and X. Zhang, "End-to-End Emotion-Cause Pair Extraction with Graph Convolutional Network," in *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 198–207, Dec. 2021.

[17] C. Fan, C. Yuan, J. Du, L. Gui, M. Yang, and R. Xu, "Transition-based Directed Graph Construction for Emotion-Cause Pair Extraction," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3707–3717, July 2020.

[18] J. Turner, *On the origins of human emotions: A sociological inquiry into the evolution of human affect.* Stanford University Press, 2000.

[19] A. Neviarouskaya and M. Aono, "Extracting causes of emotions from text," in *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, (Nagoya, Japan), pp. 932–936, Asian Federation of Natural Language Processing, Oct. 2013.

[20] S. Mohammad, X. Zhu, and J. Martin, "Semantic role labeling of emotions in tweets," in *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, (Baltimore, Maryland), pp. 32–41, Association for Computational Linguistics, June 2014.

[21] D. Ghazi, D. Inkpen, and S. Szpakowicz, "Detecting emotion stimuli in emotion-bearing sentences," in *International Conference on Intelligent Text Processing and Computational Linguistics*, pp. 152–165, Springer, 2015.

[22] C. F. Baker, C. J. Fillmore, and J. B. Lowe, "The Berkeley FrameNet Project," in *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98/COLING '98, p. 86–90, 1998.

[23] Q. Gao, G. Lin, Y. He, J. Hu, Q. Lu, R. Xu, and K.-F. Wong, "Overview of NTCIR-13 ECA Task," in *Proceedings of the 13th NTCIR Conference on Evaluation of Information Access Technologies*, pp. 361–366, 2017.

[24] E. Kim and R. Klinger, "Who Feels What and Why? Annotation of a Literature Corpus with Semantic Roles of Emotions," *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 1345–1359, 2018.

[25] X. Xiao, W. Mao, Y. Sun, and D. Zeng, "A cognitive emotion model enhanced sequential method for social emotion cause identification," *Information Processing & Management*, vol. 60, no. 3, p. 103305, 2023.

[26] S. M. Mohammad and P. D. Turney, "Crowdsourcing a word-emotion association lexicon," *Computational Intelligence*, vol. 29, no. 3, pp. 436–465, 2013.

[27] L. A. M. Oberländer and R. Klinger, "Token sequence labeling vs. clause classification for English emotion stimulus detection," in *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, (Barcelona, Spain (Online)), pp. 58–70, Association for Computational Linguistics, Dec. 2020.

[28] S. Mohammad and P. Turney, "Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon," in *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, (Los Angeles, CA), pp. 26–34, Association for Computational Linguistics, June 2010.

[29] J. C. Lind and B. D. Zumbo, "The continuity principle in psychological research: An introduction to robust statistics.," *Canadian Psychology/Psychologie canadienne*, vol. 34, no. 4, p. 407, 1993.

[30] J. B. Ford, "Amazon's mechanical turk: A comment," *Journal of Advertising*, vol. 46, no. 1, pp. 156–158, 2017.

[31] J. L. Fleiss, B. Levin, and M. C. Paik, *Statistical methods for rates and proportions*. John Wiley & Sons, 2013.

[32] M. Li, H. Zhao, H. Su, Y. R. Qian, and P. Li, "Emotion-cause span extraction: a new task to emotion cause identification in texts," *Applied Intelligence*, vol. 51, no. 10, pp. 7109–7121, 2021.

[33] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the Eighteenth International Conference on Machine Learning*, p. 282–289, 2001.

[34] D. C. Liu and J. Nocedal, "On the limited memory bfgs method for large scale optimization," *Mathematical programming*, vol. 45, no. 1, pp. 503–528, 1989.

[35] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

[36] E. Commission", "Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. com/2021/206 final." https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206, 2021.

[37] L. Floridi, M. Holweg, M. Taddeo, J. Amaya Silva, J. Mökander, and Y. Wen, "CapAI-A Procedure for Conducting Conformity Assessment of AI Systems in Line with the EU Artificial Intelligence Act," *Available at SSRN 4064091*, 2022.

Sample 1: A HIT from the published survey

Title: Reader's Emotion Cause Extraction

Description: Read social media posts and select what provokes the readers' emotion from them.

Keywords: emotion, emotion cause, sentiment, text, labelling

Task purpose:
Your responses will be used to train an automatic system to extract the emotion cause from emotion arousing posts. Posts are taken from customer service pages of popular supermarkets.

Instructions:

1. Your responses are confidential. Your specific responses will not be used in any future work, only aggregate information from many contributors. We will not ask any information that can be used to identify who you are.
2. Complete these questions only if you are familiar with social media.
3. Complete these questions only if you understand the meaning of given post.
4. You will be asked to define the cause that provoked the emotion from the post. The cause is usually an action, person, object, place, or group of people. The cause might be more than one word. If there is one, please COPY and PASTE the cause from the post. Don't paraphrase or change the words order. If there is nothing type "No cause".

The beginning of the survey..

Post:
"Can someone please contact me and advise how I can report a delivery driver for dangerous driving"

This post is labelled with the emotion **ANGRY**. Please **COPY and PASTE** the shortest amount of text that provokes that emotion.

Table 1: Complete HIT sample for one of the posts.

Model	**Accuracy** (exact match)	**Precision**	**Recall**	**F1**
CRF	13.50	0.39	0.18	0.25
BiLSTM-CRF	11.25	0.59	0.52	0.55

Table 2: Results of the baseline models.