PLOS ONE

RESEARCH ARTICLE

Twitter mining using semi-supervised classification for relevance filtering in syndromic surveillance

Oduwa Edo-Osagie[1]*, Gillian Smith[2], Iain Lake[4], Obaghe Edeghere[3], Beatriz De La Iglesia[1]

1 School of Computing Science, University of East Anglia, Norwich, Norfolk, United Kingdom, **2** Real-time Syndromic Surveillance Team, National Infection Service, Public Health England, Birmingham, United Kingdom, **3** Epidemiology West Midlands, Field Service, National Infection Service, Public Health England, Birmingham, United Kingdom, **4** School of Environmental Sciences, University of East Anglia, Norwich, Norfolk, United Kingdom

* o.edo-osagie@uea.ac.uk

Citation: Edo-Osagie O, Smith G, Lake I, Edeghere O, De La Iglesia B (2019) Twitter mining using semi-supervised classification for relevance filtering in syndromic surveillance. PLoS ONE 14 (7): e0210689. https://doi.org/10.1371/journal.pone.0210689

Editor: Olalekan Uthman, The University of Warwick, UNITED KINGDOM

Received: December 6, 2018

Accepted: June 13, 2019

Published: July 18, 2019

Data Availability Statement: The Twitter data was collected via the Twitter Application Programmer Interface and cannot be shared, as they are third-party data and are restricted by Twitter's terms of service. However, while we cannot provide this data directly, we can and have provided details of the search parameters used to construct this dataset in the Methods section as well as in the Supporting Information. Additionally, due to the sensitive nature of the public health data, and the fact that the authors do not have ownership of it, parties interested in that data may privately contact

Abstract

We investigate the use of Twitter data to deliver signals for syndromic surveillance in order to assess its ability to augment existing syndromic surveillance efforts and give a better understanding of symptomatic people who do not seek healthcare advice directly. We focus on a specific syndrome—asthma/difficulty breathing. We outline data collection using the Twitter streaming API as well as analysis and pre-processing of the collected data. Even with keyword-based data collection, many of the tweets collected are not be relevant because they represent chatter, or talk of awareness instead of an individual suffering a particular condition. In light of this, we set out to identify relevant tweets to collect a strong and reliable signal. For this, we investigate text classification techniques, and in particular we focus on semi-supervised classification techniques since they enable us to use more of the Twitter data collected while only doing very minimal labelling. In this paper, we propose a semi-supervised approach to symptomatic tweet classification and relevance filtering. We also propose alternative techniques to popular deep learning approaches. Additionally, we highlight the use of emojis and other special features capturing the tweet's tone to improve the classification performance. Our results show that negative emojis and those that denote laughter provide the best classification performance in conjunction with a simple word-level *n*-gram approach. We obtain good performance in classifying symptomatic tweets with both supervised and semi-supervised algorithms and found that the proposed semi-supervised algorithms preserve more of the relevant tweets and may be advantageous in the context of a weak signal. Finally, we found some correlation ($r = 0.414$, $p = 0.0004$) between the Twitter signal generated with the semi-supervised system and data from consultations for related health conditions.

Public Health England for access at syndromic.surveillance@phe.gov.uk. Finally, the code for the project can be found at https://github.com/oduwa/semisupervised-twitter-asthma.

Funding: We acknowledge support from NHS 111 and NHS Digital for their assistance and support with the NHS 111 system; Out-of-Hours providers submitting data to the GPOOH syndromic surveillance and Advanced Heath & Care. The authors also acknowledge support from the Public Health England Real-time Syndromic Surveillance Team. Iain Lake and Beatriz De La Iglesia received support from the National Institute for Health Research Health Protection Research Unit (NIHR HPRU) in Emergency Preparedness and Response. Beatriz De La Iglesia received support from grant ES/L011859/1, from The Business and Local Government Data Research Centre, funded by the Economic and Social Research Council.

Competing interests: The authors have declared that no competing interests exist.

1 Introduction

Surveillance, described by the World Health Organisation (WHO) as "the cornerstone of public health security" [1], is aimed at the detection of elevated disease and death rates, implementation of control measures and reporting to the WHO of any event that may constitute a public health emergency or international concern. Disease surveillance systems often rely on laboratory reports. More recently some countries such as the UK and USA have implemented a novel approach called "syndromic surveillance", which uses pre-diagnosis data and statistical algorithms to detect health events earlier than traditional surveillance [2]. Syndromic surveillance can be described as the real-time (or near real-time) collection, analysis, interpretation, and dissemination of health-related data, to enable the early identification of the impact (or absence of impact) of potential human or veterinary public health threats that require effective public health action [3]. For example, they use emergency department attendances or general practitioner (GP, family doctor) consultations to track specific syndromes such as influenza-like illnesses (ILI). The digital world is generating data at an astonishing rate. This generated big data has unlocked novel solutions to the area of public health [4, 5]. The expansion in digital technology and increasing access to online user-generated content, like Twitter, has provided another potential source of health data for syndromic surveillance purposes. Expanding access to communications and technology makes it increasingly feasible to implement syndromic surveillance systems in low and middle income countries (LMIC) too and some early examples in Indonesia and Peru have indicated reason for optimism [6].

The use of data from microblogging sites, such as Twitter, for disease surveillance has been gaining momentum (e.g. [7–11]). This may not only complement existing surveillance systems, but may also support the precise monitoring of disease activity in sub-groups of the population that do not routinely seek medical help via existing healthcare services. The real-time streaming nature of Twitter data could provide a time advantage for syndromic surveillance activities aimed at early detection of disease outbreaks. In addition to this, the low cost of utilisation of this data means that in LMIC where access to medical services may be restricted but where the of use digital technology and social media is becoming more common, such data may support the development of cost-effective and sustainable disease surveillance systems.

It is in light of this that we develop our work. Our ultimate aim is to establish the utility of social media data, and specifically, Twitter data for syndromic surveillance.

Our first objective is to extract a reliable signal from the Twitter stream for different syndromes and health conditions of interest. To achieve this, we must be able to effectively identify and extract tweets expressing discomfort or concern related to a syndrome of interest and reflecting current events. Such symptomatic tweets are considered "relevant" for our purpose of syndromic surveillance. In this paper, we look at asthma/difficulty breathing as our syndrome of interest, which has received less attention in studies using social media data than other syndromes (e.g. ILI). Keywords such as "asthma" or other asthma-related keywords can capture tweets such as "*oh I used to have asthma but I managed to control it with will power*" or "*Does your asthma get worse when you exercise?*" which we consider as not relevant. On the other hand, tweets such as "*having an asthma attack atm*" or "*why is my asthma so bad today?*" express a person currently affected, which we would like to consider as relevant. Hence, the problem becomes one of text classification. Other authors [11, 12] have already identified that much of the data captured on Twitter represents chatter, concern, or awareness instead of actual infection or suffering from symptoms of a disease, talk of past events or a reflection on news content. Such data is therefore irrelevant as a signal for syndromic surveillance. This irrelevant content may greatly magnify the signal and lead to incorrect results and over-

estimation [13]. Once the signal has been extracted, we then compare it to real world syndromic surveillance data to understand how well Twitter works for monitoring our syndromes.

When extracting the Twitter signal, we focus on two novel aspects of tweet classification. Firstly, we investigate emojis in tweet classification, and show their worth in a syndromic surveillance context. While there is published literature making use of emoticons in text classification [11], there are few studies describing the use of emojis. Secondly,we compare both supervised and semi-supervised approaches to text classification. We consider semi-supervised methods because they enable us to use a small amount of labelled data, thereby reducing the initial labelling effort required to build a classifier. Finally, we compare the signal we extracted using our methods, to syndromic surveillance data from Public Health England (PHE) to investigate the utility of Twitter for the syndromic surveillance of asthma/difficulty breathing.

2 Related work

In a survey carried out in 2015, Charles-Smith et al. [8] identified 33 articles that reported on the integration of social media into disease surveillance with varying degrees of success. However, they reported that there is still a lack of application in practice, despite the potential identified by various studies. Many studies are retrospective as it is relatively easy to predict a disease post-outbreak, but practical application would need to be prospective. Uses of social media data vary from global models of disease [14] to the prediction of an individual's health and when they may fall ill [15].

The most commonly studied disease is influenza (or ILI) [16]. Ginsberg et al. [17] put forward an approach for estimating influenza trends using the relative frequency of certain Google search terms as an indicator for physician visits related to influenza-like symptoms. They found that there was a correlation between the volume of specific Google searches related to ILI and the recorded ILI physician visits reported by CDC [17]. De Quincey and Kostkova [9] introduced the potential of Twitter in detecting influenza outbreaks. They posited that the amount of real-time information present on Twitter, either with regards to users reporting their own illness, the illness of others or reporting confirmed outbreaks from the media, is both rich and highly accessible. Achrekar et al. [18] also investigated the use of Twitter for detecting and predicting seasonal influenza outbreaks, and observed that Twitter data is highly correlated with the ILI rates across different regions within USA. They concluded that Twitter data can act as supplementary indicator to gauge influenza within a population and could be useful in discovering influenza trends ahead of CDC.

In this study, our objective is to collect relevant tweets for our given syndrome. We notice that a majority of tweets are not relevant as they do not express the required sentiment (i.e. a person suffering from the particular ailment at the current time). We view this as a text (or tweet) classification problem, and build models to filter relevant tweets. Several papers have looked at the tweet classification problem using supervised learning for different applications. Sriram et al. [19] classified tweets to a predefined set of generic classes, such as news, events, opinions, deals, and private messages, based on information on the tweets' authors and domain-specific features extracted from tweets. Dilrukshi et al. [20] applied a Support Vector Machine (SVM) to classify tweets to different news categories. Some other papers have also investigated tweet classification in the context of syndromic surveillance to varying degrees. As early as 2013, Lamb et al. [11] used Twitter data to investigate influenza surveillance. They argued that for accurate social media surveillance, it is essential to be able to distinguish between tweets that report infection and those that express concern or awareness. To accomplish this, they made use of a log-linear model with word-level n-gram features extracted from the tweets. Over the years, more sophisticated approaches to Twitter surveillance have arisen.

Most recently, the rise in popularity of deep learning has seen it applied to social media data and health research, yielding positive results. Hu et al. [21] used a Convolutional Neural Network (CNN) to classify drug abuse behaviour in tweets. Lee et al. [22] applied a CNN to the classification of tweets as being related to adverse drug effects or not. Dai et al. [23] made use of deep learned distributed representations of words [24] to characterise tweets which were then separated into two clusters—related or unrelated to a topic (e.g. influenza). Researchers in Australia made use of Recurrent Neural Networks (RNNs) for the classification of chief complaints to the Australian Emergency Department [25]. Chinese researchers applied RNNs to spatio-temporal influenza data from the Shenzen Centre for Disease Control and Prevention to predict flu trends [26]. SENTINEL [10] is a deep learning powered syndromic surveillance system which utilises RSS news feeds in addition to Twitter, and makes use of supervised deep learning classifiers, comparing them to Naive Bayes and SVM. It is set in the US and looks at Influenza-like illnesses, similar to the majority of work on syndromic surveillance. We have also previously carried out some work on the syndromic surveillance of asthma/difficulty breathing using deep learning, making use of a deep Long Short Term Memory (LSTM) RNN after implementing other deep architectures and comparing them [27].

One problem with the above approaches is that they rely on having a set of labelled data for learning, i.e. a sufficient set of tweets must first be labelled as, say, relevant/irrelevant for the learning to take place. Such labelling can be very time consuming, so it often means that researchers do not use all of the data available, but instead use a subset of labelled data to develop their classifiers. Since the syndromes/events we wish to study may not be mentioned frequently in a Twitter feed, we wish to use as many tweets as possible to build our models. Semi-supervised classification approaches try to produce models using a small set of labelled data whilst also taking into account the larger set of unlabelled data. As such, we investigate them next. A number of papers have looked at using semi-supervised learning for sentiment analysis, and in particular, self-training [28, 29]. Baugh [30] proposed a hierarchical classification system with self-training incorporated, with the goal of classifying tweets as *positive*, *negative* or *neutral*. Liu et al. [31] proposed a semi-supervised framework for sentiment classification of tweets, based on co-training. They converted tweets into two kinds of distinct features: textual and non-textual. The deep learning approaches highlighted above can also be considered as semi-supervised, as they use unlabelled data to build vector feature representations which are subsequently used for classification. However, the classification process is still entirely supervised. In fact, Oliver et al. [32] argue that while deep neural network tasks have proven successful on standard benchmark tasks, these tasks are not comparable to the scarce data scenarios for which widely-used semi-supervised classification techniques are implemented. Lim et al. [33] proposed a completely unsupervised model for identifying latent infectious diseases on social media. They made use of the unsupervised sentiment analysis system *SentiStrength* [34] to identify positive and negative tweets. They then tried to check if the negative tweets contained mention of a relevant body part. The fully unsupervised system was found to perform well without any human labelling at all and achieved an $F1$-score of 0.724.

In this paper, we build classification models for tweets based on the relevance in the context of a specific syndrome/event which perform well in an environment with little labelled data. In parallel work, we have been looking at deep learning methods [27]. There is an advantage in developing methods with minimal labelled sets because they simplify the establishing of Twitter data classification for new syndromes. As part of our investigation, we look into feature representation and feature selection for text, which is an important part of text classification. We experiment with different types of features, taking into consideration suggestions from previous work. We consider the addition of emojis for tweet classification, and show their worth for improving classification in a syndromic surveillance context. We compare both

supervised and semi-supervised approaches to text classification in order to understand if, and how, we can utilise more of the data that we collect.

3 Methods

We discuss the data collection, pre-processing and analysis of tweets in order to extract a relevant signal for a given syndrome. We narrow our efforts to asthma/air pollution incidents in this paper.

3.1 Data collection and pre-processing

We collected tweets in different periods to account for the seasonality of the syndromes under study and to have a better chance of observing a significant episode, which is unpredictable. This collection strategy also enables us to observe periods with no significant episodes, as a form of control. Different periods also enable us to monitor changes in the use of Twitter as well as in the language used on Twitter over time. We started with an Autumn period (September 2015 to November 2015), followed by a summer period (June 2016 to August 2016) and a winter through to mid-summer period (January 2017 to July 2017).

Tweets were collected using the official Twitter streaming Application Programmer's Interface (API). The Twitter streaming API provides a subset of the Twitter stream free of charge. The whole stream can be accessed on a commercial basis. Studies have estimated that using the Twitter streaming API, users can expect to receive anywhere from 1% of the tweets to 40% of tweets in near real-time [35]. The streaming API has a number of parameters that can be used to restrict the tweets obtained. We extracted tweets in the English language with specific terms that may be relevant to a particular syndrome. For this, in conjunction with experts from Public Health England (PHE), we created a set of terms that may be connected to the specific syndrome under scrutiny, in this case asthma/difficulty breathing. We then expanded on this initial list using various synonyms from regular thesauri as well as from the urban dictionary (https://www.urbandictionary.com) as those may capture some of the more colloquial language used on Twitter. Examples of our keywords are "asthma", "wheezing", "couldn't breathe" etc. A full list of keywords used is provided in S1 List.

The restriction on specific terms can be implemented in the Twitter API by by using the parameter "track" followed by a comma-separated list of phrases which will be used to determine which tweets will be delivered from the stream. A phrase may be one or more terms separated by spaces, and a phrase will match if all of the terms in the phrase are present in the tweet, regardless of order and ignoring case. Hence in this model, commas act as logical ORs and spaces are equivalent to logical ANDs. The tracked terms are matched against a number of attributes of the tweet including the *text* attribute of the tweet, *expanded_url* and *display_url* for links and media and *screen_name* for user.

We collected 10 million tweets obtained over the three collection periods. The general characteristics of the collected tweets are reported in Table 1.

Table 1. Information on the data corpus collected before cleaning.

	Counts
Tweets	10,702,063
URLs	2,225,155
Hashtags	177,506
Emojis	3,103,598
Number of unique users	5,861,247
Number of tweets per user	4.1

https://doi.org/10.1371/journal.pone.0210689.t001

The anatomy of a tweet is presented in the Status Map in Fig 1. There are several attributes associated with a tweet that are available to our analysis. We did not consider all the available tweet attributes to be useful for our experiments, so we collected those that could help us in our task. More specifically, we collected "tweet_Id", "text", "created_at", "user_id", "source" as well as information that may help us establish location such as "coordinates", "time_zone" and "place.country". We stored the collected tweets using MongoDB, which is an open source no-SQL database whose associative document-store architecture is well suited to the easy storage of the JSON Twitter responses.

3.1.1 Location filtering. Because the aim of this project is to assess the utility of Twitter data for syndromic surveillance systems in England, we would like to exclude tweets originating from outside England. Doing this will give a realistic signal, however, inferring the location of Twitter users is notoriously difficult. Fewer than 14% of Twitter users disclose city-level information for their accounts and even then, some of those may be false or fictitious locations [36]. Less than 0.5% turn on the location function which would give accurate GPS coordinate information from mobile devices. *time_zone*, *coordinates* and *place* attributes, which we collected, can help in the geolocation of a tweet but are not always present or even correct as is shown in Table 2. The most reliable source of location information at the time of tweeting, *coordinates*, is only present in a very small percentage of tweets.

For building a relevance classifier, accurate location is of relative importance. In this work, we are not overly concerned with accurate location filtering. For the purpose of symptomatic tweet classification for relevance filtering, location is of no importance. We collect tweets from the whole of the UK. We employ all three geolocation fields, filtering out tweets that do not have a UK timezone, a place in the UK or coordinates in the UK. We acknowledge that the location filtering is not entirely accurate and may have a disruptive effect when we compare our signal with public health data collected within England. However, we leave the task of improving on location filtering for future work where we will extend our signal comparisons to include longer periods of time and other syndromes.

3.1.2 Cleaning the data. The Twitter dataset contained retweets (sometimes abbreviated to RT) which are the re-posting of a tweet; other duplicate tweets not marked as RT but containing exactly the same text with different URLs appended; and users tweeting multiple times on the same day. We removed all duplication from the final dataset in order to minimise the detection of false signals.

In addition, we removed URLs, which are often associated with news items and blogs, and replaced them with the token "<URL>". This helped with identification of duplication but also identification of "bot" posting and news items. A "bot" is the term used when a computer program interacts with web services appearing as a human user. Tweets from bots, news and web blogs are not relevant to syndromic surveillance so we developed algorithms to identify them and remove them. An overview of the data after cleaning, showing a considerable reduction in volume, is shown in Table 3.

3.1.3 Labelling. 3,500 tweets from the first data collection period were manually labelled as "relevant" or "not relevant". A tweet was labelled as relevant if it announced or hinted at an individual displaying symptoms pertaining to the syndrome of choice. The labelling was done by three volunteers. A first person initially labelled the tweets. This took approximately 1 hour per 1,000 tweets. A second person checked the labels and flagged up any tweets with labels that they did not agree with. These flagged tweets were then sent to a third person who made the decision on which label to use. 23% of the labelled tweets were labelled as "relevant" while 77% were labelled as "irrelevant". A second set of 2,000 tweets, selected at random, were later labelled following the same procedure from the last data collection period. 32% of these tweets were labelled as relevant and 68% were labelled as irrelevant. From the second sample of 2000

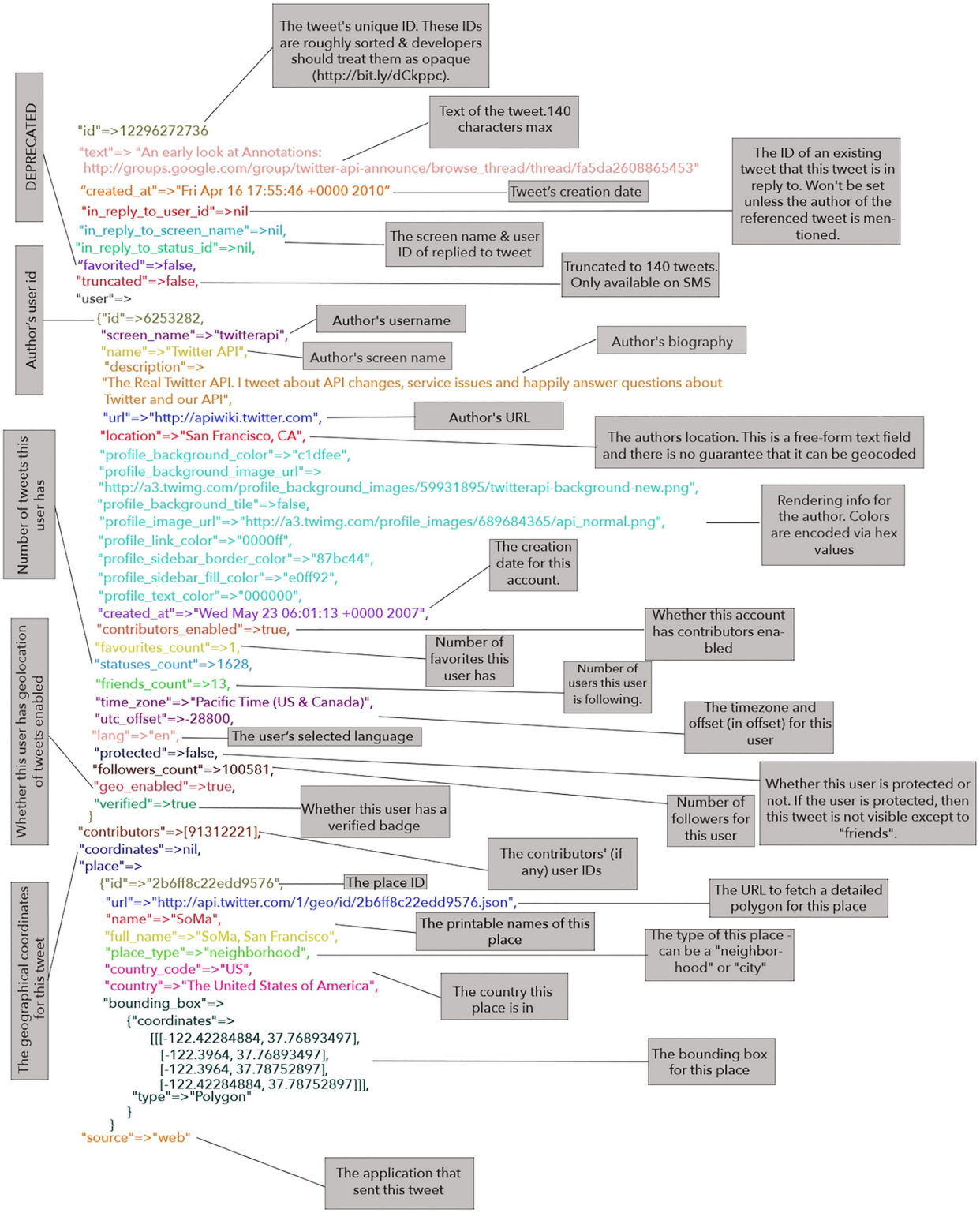


Fig 1. Map of a tweet from the Twitter API.

https://doi.org/10.1371/journal.pone.0210689.g001

Table 2. Availability of geolocation attribute in collected Twitter Data.

Data Collection Period	Percentage of tweets Containing attributes		
	Coordinates	Timezone	Place
September 23, 2015—November 30, 2015	0.30%	57.90%	2.17%
June 15, 2016—August 30, 2016	0.29%	61.12%	2.10%
January 27, 2017—July 31, 2017	0.21%	59.21%	1.61%

https://doi.org/10.1371/journal.pone.0210689.t002

tweets, the Inter-Rater Agreement was computed using Fleiss' Kappa [37] which is given by the following equation:

$$\kappa = \frac{\bar{P} - P_e}{1 - P_e}$$

where $\bar{P}$ is the mean agreement between raters and P_e is the probability of agreement by chance calculated from the observed data using the probabilities of each rater randomly labelling a tweet as relevant or irrelevant. The Fleiss' kappa was chosen over other kappas due to the fact that it is intended to be used when assessing the agreement between three or more raters which is appropriate for this scenario. A value of 1 suggests complete agreement while a value of 0 suggests complete disagreement. We obtained a value of **0.906** for κ.

3.1.4 Basic text classification features. We acknowledge that deep learned word vectors are an effective avenue for text feature representation. However, as was described in section 2, in this work, our focus is not on deep learning, but semi-supervised approaches that can be used in an environment where labelled data is scarce. In addition, the training and deployment of high-performant industry-grade deep learning systems with hundreds of layers can be intensive and require considerable hardware resources such as dedicated GPUs and TPUs [38]. A system such as ours will require no such special hardware or resources, so it will be easier for low and middle income countries (LMIC) to incorporate such systems at whatever scale.

Classification of tweets may be challenging as they are very short and in our scenario, target classes may share common vocabularies. That is, both relevant and irrelevant tweets could contain the same words. Twitter has specific language and styles of communication that people use. In particular, we found that *emojis and emoticons* are promising additional tokens that we could exploit for classification:

- An emoticon is a pictorial representation of a facial expression using punctuation marks, numbers and letters, usually written to express a person's feelings or mood. :-) is an example of an emoticon.
- Emojis on the other hand are miniature graphics of various objects and concepts including facial expressions. 😀 is an example of an emoji.

Table 3. Information on the data corpus collected after cleaning.

	Counts
Tweets	127,145
URLs	147,102
Hashtags	23,189
Emojis	36,872
Number of unique users	115,583
Number of tweets per user	5.3

https://doi.org/10.1371/journal.pone.0210689.t003

Emoticons have been used successfully as features to improve tweet classification performance in sentiment analysis as well as syndromic surveillance [11]. Practically speaking, emojis can be used for the same purposes as emoticons. However, emojis have seen a recent surge in popularity, presumably due to the fact that emojis provide colourful graphical representations as well as a richer selection of symbols [39]. In fact, as Table 1 shows, there were a large number of emojis in our corpus. Because of this, we extend existing emoticon feature techniques with the inclusion of emojis. A further advantage is that while emoticon use can differ around the world, emoji features are less variant. Take for example the crying emoticon. In Western countries, it can be represented by:'(or:'-(. In Asian countries, "kaomoji", which refers to emoticons depicting faces and made from a combination of Western punctuation characters and CJK (Chinese-Japanese-Korean) characters, are more popular than regular emoticons [40]. An example of a popular kaomoji is "¯_(ツ)_/¯". Using the earlier example of the crying face, we could now also expect to see (T_T) for the same concept. Emojis on the other hand are a predefined set of unicode characters. Even though they may be rendered differently on different devices, the underlying mapping between a concept and an emoji remains the same. In this sense, emojis may transcend language barriers.

We believe that emoticons and emojis can help with assessing the tone of a tweet. Tweets we are interested in will most likely have a negative tone as they reflect people expressing that they are unwell or suffer some symptoms. This means they may contain one or more emojis/emoticons denoting sadness, anger or tiredness, for example. On the other hand, the presence of emojis/emoticons denoting happiness and laughter in a tweet may be an indication that the tweet is not relevant to our context of syndromic surveillance. We also investigate more complex features derived from our words or additional tokens.

3.1.5 Feature construction.

3.1.5.1 Word classes We extend our *n*-gram feature set with further syntactical features in order to make up for the shortcomings words may present when applied to Twitter data. Word classes are labels that Lamb et al. [11] found useful in the context of analysing tweets to categorise them as related to infection or awareness. The idea is that many words can behave similarly with regard to a class label. A list of words is created for different categories such as "*possessive words*" or "*infection words*". Word classes function similarly to bag of word features in that the presence of a word from a word class in a tweet triggers a count based feature. We manually curated a list of words and classes which are shown in Table 4. As we applied lemmatisation, we did not include multiple inflections of the words in our word classes.

3.1.5.2 Positive and negative word counts: We constructed two dictionaries of positive and negative words respectively. These dictionaries are shown in S2 and S3 Lists. This feature computes for every tweet, the number of positive words and negative words it contains. Words that do not appear in either of our dictionaries are not counted. The classifier should then infer a matching between ratios of positive to negative counts and tweet relevance.

3.1.5.3 Denotes laughter: This is a simple binary feature which measures the presence of a token (emoji and/or emoticon) that might suggest laughter or positivity. We manually curated

Table 4. Our list of word classes with their member words.

Word Class	Member Words
Infection	sick, down, ill, infect, caught, recover
Possession	have, contain, contaminated, my
Concern	awful, worried, scared, afraid, terrified, fear, sad, unhappy, feel
Humour	laugh, ha, haha, hahaha, lol, lmao, rofl, funny, hilarious, amused
Symptomatic	runny nose, cough, spray, shots, wheezing, mucus, cold

https://doi.org/10.1371/journal.pone.0210689.t004

Table 5. Distribution of some constructed features and classes across the dataset.

Feature	Value Distribution		Class Distribution	
			Relevant	Not Relevant
Denotes Laughter	TRUE	3.9%	31.8%	68.2%
	FALSE	96.3%	24.2%	75.8%
Negative Emojis/Emoticons	TRUE	5.5%	74.8%	25.2%
	FALSE	94.5%	21.6%	78.4%

https://doi.org/10.1371/journal.pone.0210689.t005

and saved a list of positive emojis/emoticons for this. The usefulness of this feature was augmented by also checking for the presence of a small list of more established and popular internet and slang for laughter or humour such as "lol" or "lmao" which stand for "Laughing Out Loud" and "Laughing My Ass Off" respectively. Table 5 shows this feature's distribution over the data.

3.1.5.4 Negative emojis/emoticons: This is similar to the *Denotes Laughter* feature but this time looking at the presence of an emoji or emoticon that can be associated with an illness or the symptoms that it may bring., i.e. negative emotions. We decided to include this feature because we discovered the ubiquity of emojis on Twitter and wanted to investigate their potential. Table 5 shows this feature's distribution over the data. We find that this feature may be the most discriminative of the two emoji-based features. Of the instances with a positive value, a high percentage belong to the "relevant" class and of the instances with a negative value, a high percentage belong to the "not relevant" class.

We experimented with two other features—*Contains Asthma-Verb Conjugate* and *Indicates Personal Asthma Report* but found that they underperformed compared to the other features so we do not report on them. We also constructed features from the tweets collected in the latest time period in order to see how the features generalised across time periods. The distributions of the non-continuous features from the latest time period are shown in Table 6.

For each tweet, we appended all of the above features together to form one feature vector. Each tweet T_i is therefore represented by an f dimensional vector, where f is a sum of the number of terms, n, in the constructed vocabulary, and the dimensionality of our custom features C (*Word Classes*, *Positive and Negative Word Counts*, *Denotes Laughter* and *Negative Emojis/Emoticons*). This gives us

$$T_i = \{t_{i1}, t_{i2}, ..., t_{in}\} \cup \{C_i^1\} \cup \{C_i^2\} \cup \{C_i^3\} \cup \{C_i^4\}$$

where t_{ij} represents the weight of the j-th vocabulary term in the i-th tweet and C_i^k represents the value of the k-th custom feature in the i-th tweet. The feature vectors are represented in code by dictionary (or hashmap) objects which allows them to contain different types of values (i.e.. binary, continuous and categorical).

Table 6. Distribution of constructed features and classes across tweets from a different time period 2 years apart from our that of our investigated dataset.

Feature	Value Distribution		Class Distribution	
			Relevant	Not Relevant
Denotes Laughter	TRUE	4.0%	13.9%	86.1%
	FALSE	96.0%	32.5%	67.5%
Negative Emojis/Emoticons	TRUE	14.4%	41.3%	58.7%
	FALSE	85.6%	30.2%	69.8%

https://doi.org/10.1371/journal.pone.0210689.t006

3.2 Text classification

A classification algorithm for text can be used to automatically classify tweets, in this case, to the categories of relevant/not relevant. We first applied a variety of popular and powerful supervised classification algorithms to the data namely—Naive Bayes, Decision Trees, Logistic Regression, Support Vector Machines (SVMs) and Multilayer Perceptron (MLP) neural networks. We used the Python implementations found in the Natural Language ToolKit (NLTK) and Sci-Kit Learn [41].

As per our objective, we also implemented a semi-supervised approach which is suited to small to medium sized datasets [32]. Semi-supervised learning attempts to make use of the combined information from labelled and unlabelled data to exceed the classification performance that would be obtained either by discarding the unlabelled data and applying supervised learning or by discarding the labels and applying unsupervised learning. Our intention is to extend the labelling in a semi-supervised fashion. We make use of the heuristic approach to semi-supervised learning and employ a ***self-training iterative labelling algorithm***. We then extend this work by using a form of ***co-training***.

3.2.1 Self-training model. We adopted an *Iterative Labelling Algorithm* for semi-supervised learning [42]. Iterative labelling algorithms are closely related to and are essentially extensions of the Expectation-Maximization (EM) algorithm put forward by Dempster et al. [43]. The iterative labelling algorithm is a sort of *meta-algorithm* which uses a data set S of labelled instances L, unlabelled instances U, and a supervised learning algorithm A with

$$S = \{L \cup U\}$$

An iterative learning algorithm aims to derive a function f which provides a mapping from S to a new dataset S':

$$f(S, A) = S' \leftrightarrow \{L' \cup U' \} \mid |U'| \leq |U|, |L'| \geq |L|$$

Such an algorithm can be defined simplistically as an iterative execution of three functions: *Choose-Label-Set*(U, L, A) selects and returns a new set, R, of unlabelled examples to be labelled; *Assign-Labels*(R, S, A) generates labels for the instances selected by *Choose-Label-Set* (U, L, A); *Stopping-Condition*(S, S') dictates when the algorithm should stop iterating.

```
Algorithm Iterative labelling Algorithm
  function ITERATIVELABELLING(U, L, A)
    repeat
      R ← Choose-Label-Set(U, L, A)
      R ← Assign-Labels(R, S, A)
      U ← Replace-Instances(U, R)
    until Stopping-Condition(S, S') = True
```

For our choice of supervised learning algorithm, we selected the MLP classifier after experimenting with different supervised models and finding it to perform best. We used the trained MLP classifier's predictions to label unlabelled instances in the *Assign-Labels* function. We set our stopping condition such that the iteration stops when either all the unlabelled data is exhausted or there begins to be a continued deterioration in performance as more data is labelled. Along with the class of an applied instance, we also compute the model's confidence in its classification. Our algorithm, inspired by Truncated Expectation-Maximization (EM) [44], then grows L based on the confidence of our model's classification. When an instance from R is classified, if the confidence of the classification is greater than some set threshold θ, the instance is labelled. Considering this, our algorithm falls within the *confidence-based* category of iterative labelling or self-training algorithms because it selects instances for which the trained classifier has a high confidence in its predictions.

Confidence-based iterative labelling algorithms can tend toward excessively conservative updates to the hypothesis, since training on high-confidence examples that the current hypothesis already agrees with will have relatively little effect [44]. In addition, it has been proven that in certain situations, many semi-supervised learning algorithms can significantly degrade the performance relative to strictly supervised learning [45, 46].

3.2.2 Co-training model. To address the problems of self-training, we take some ideas from *co-training* [47] to try to improve our algorithm. Co-training requires different views of the data so that multiple classifiers can be maintained for the purpose of labelling new instances. Recall that each tweet can be represented as a feature vector T_i with various features. We now distinguish two representations. The first is a concatenation of our n-*grams*, *Word Classes*, *Denotes Laughter* and *Negative Emojis/Emoticons* features. We represent this feature space as X_1. The second kind of feature vector is a concatenation of our n-*grams*, *Positive and Negative Word Counts*, *Denotes Laughter* and *Negative Emojis/Emoticons* features. We represent this feature space as X_2. We can think of X_1 as the **taxonomical** feature space as is characterised by its inclusion of the *Word Classes* feature while X_2 can be the **sentimental** feature space and this is characterised by its inclusion of the *Positive and Negative Word Counts* feature. As such, X_1 and X_2 offer different, though overlapping, views of the dataset. Each tweet is then represented as a feature vector from each of these spaces.

We now maintain two separate classifiers trained on different views of the data. During the iterative labelling process, we only label instances for which at least one of the classifiers has a high confidence in its prediction and take the result of that classification as the label. Similar to self-training, at the end of each iteration, the newly labelled data is incorporated into each of the classifiers to update their hypotheses. Once the iterative labelling process is completed, the prior training examples for both classifiers as well as the newly labelled examples are joined together and used to train a new classifier using all the features which will then be applied in practice. The benefit of co-training is that the examples labelled by one classifier are also presented to the other classifier to update the hypothesis on the complementary view. Thus, the examples, as represented in each view, receive at least some of their labels from a source other than the classifier that will be updated with them [42].

3.2.3 Correcting the class imbalance. We started with an initial set of manually labelled data contained 3,500 tweets. This consisted of 23.7% tweets that were labelled as relevant and 76.3% labelled as irrelevant. Imbalanced data causes well known problems to classification models [48]. We initially tried both oversampling and undersampling techniques to create a balanced training dataset as well as just using the unbalanced data. We found no major difference between the balancing approaches, but they gave some advantage over the unbalanced data, so we opted for over sampling. The class distribution over the balanced training set had 47% of tweets as relevant and 53% as irrelevant. The test set was not balanced.

3.2.4 Performance metrics. Another important aspect of imbalanced data and of classification in general is having the right performance metric for assessment of classification model [49]. Overall accuracy is a misleading measure [50] as it may only be reflecting the prevalence of the majority class. This is called the accuracy paradox, i.e. we could get high accuracy by classifying all tweets as irrelevant. That would, however, not improve our signal. The aim of our endeavour is to identify tweets which might suggest an increase of cases for a particular syndrome (asthma/difficulty breathing) for the purpose of syndromic surveillance. Our signal for some syndromes is quite weak as not many cases may occur at a national level and even less may be talked about on Twitter. Because of this, we are very concerned with identifying and keeping instances of the positive class (relevant tweets). We would like to reduce the number of irrelevant tweets but not at the expense of losing the relevant tweets. This means that for our classifier, errors are not of equal cost. Relevant tweets that are classified as irrelevant, also

known as False Negative (FN) errors, should have a higher cost and hence be minimised; we can have more tolerance of irrelevant tweets classified as relevant, also known as False Positive (FP) errors. Those subtleties are well captured by alternative measures of model performance such as recall, the probability that a relevant tweet is identified by the model and precision, the probability that a tweet predicted as relevant is actually relevant [51]. Precision and recall are often trading quantities. A measure that combines precision and recall is the *F*-measure or *F*-score [52]. We primarily make use of the F_2 measure which weighs recall higher than precision and may be more suited to our purpose.

3.2.5 Assessment of features and key words. We also assessed the discriminative ability of each of our features by performing feature ablation experiments [53]. We evaluated the performance of a given classifier when using all our features, and then again after removing each one of these features. The difference in the performance is used as a measure of the importance of the feature. We chose to use the difference in F_1 metric over F_2 in this analysis because we wanted to convey how the features performed in the general task of tweet classification.

We also performed some analysis on the word (i.e. *n*-gram) features to learn which words in our vocabulary were the best indicators of relevant tweets. We analysed the *n*-gram component of our compound feature vectors in order to calculate the *informativeness*, or *information gain* of each word unigram. The information gain of each feature pair is based on the prior probability of the feature pair occurring for each class label. A higher information gain (hence, a more informative feature,) is a feature which occurs primarily in one class and not in the other. Similarly, less informative features are features which occur evenly in both classes. The information gain idea is pivotal to the decision tree algorithm but generalises to others and was adapted in the NLTK package for use in a broader sense. In NLTK, informativeness of a word w was calculated as the highest value of $P(w = feature\_value|class)$ for any class, divided by the lowest value of $P(w = feature\_value|class)$ [41]. This informativeness I, is summarised below:

$$I = \frac{\forall c \in C : max(P(feature = feature\_value|c))}{\forall c \in C : min(P(feature = feature\_value|c))}$$

where C is the set of all classes and c is a possible class.

Recall that to collect tweets, we made use of Twitter's streaming API which allowed us to specify keywords to restrict the data collection to tweets containing those specific terms. We measured the usefulness of the keywords we selected. To do this, we assessed their information retrieval performance. Specifically, we used the precision-recall metric. In an information retrieval context, precision and recall are defined in terms of a set of retrieved documents and their relevance. We use our original set of labelled tweets for this assessment (i.e. the set of 3500 tweets). In our scenario, the labelled tweets make up the set of retrieved documents and the tweets labelled as belonging to the "relevant" class make up the set of relevant documents. In this context, recall measures the fraction of relevant tweets that are successfully retrieved while precision measures the fraction of retrieved tweets that are relevant to the query.

4 Results

4.1 Classifiers

The results of our fully-supervised and semi-supervised classification are presented in Table 7. The original data was divided into a 70:30 training and test split through random sampling and the results presented are measures obtained from the test data. Of the fully-supervised classifiers, Logistic Regression, SVM and MLP are very sensitive to hyper-parameters. The values for these hyper-parameters were found using grid-search with a hold-out validation

Table 7. Results of relevance classification on the test data. Naive Bayes (NB), Decision Tree (DT), Logistic Regression (LR), Support Vector Machine (SVM) and Multi-layer Perceptron (MLP) algorithms are reported together with the self-training and co-training iterative labelling algorithms.

Supervised Algorithms	Precision	Recall	Accuracy	F_1 Score	F_2 Score
NB	0.636	0.804	84.2%	0.710	0.764
DT	0.915	0.629	89.7%	0.554	0.671
LR	0.885	0.739	91.5%	0.805	0.764
SVM	0.864	0.722	90.6%	0.787	0.747
MLP	0.928	0.878	95.5%	0.903	0.888
Semi-Supervised Algorithms	**Precision**	**Recall**	**Accuracy**	**F_1 Score**	**F_2 Score**
Self-training	0.897	0.924	95.6%	0.910	0.919
Co-training	0.881	0.942	95.5%	0.910	0.929

https://doi.org/10.1371/journal.pone.0210689.t007

setup. 90% of the training data (70% of the labelled set) was used to build the classifiers and the remaining 10% was used to optimise the hyper-parameters. In the following evaluation, we use the discovered optimal hyper-parameters according to the grid-search. For Logistic regression, we used L2 regularisation with a regularisation strength C of 0.00001. For the SVM, we used a Radial Basis Function kernel and C of 0.01. For the MLP, we used 2 hidden layers, each with 128 neurons, a learning rate of 0.001, a regularisation α of 0.0001, a batch size of 200 and trained for 100 epochs. The Adam optimiser [54] was used in minimising the loss function. For the iterative labelling experiments, we varied and tuned the confidence thresholds until we found the best results and reported those. Below, we also discuss in more detail how the confidence threshold affected the iterative labelling performance as it is a key aspect of the algorithms. The best fully-supervised approach according to a combination of the F_1 and F_2 scores was the MLP, which achieved an F_2 score of **0.888** on the test data. This equated to an overall prediction accuracy of **95.5%**. The best semi-supervised approach, which was the co-training algorithm (using the best fully-supervised classifier—MLP as its base), achieved an F_2 score of **0.929** on the test data, also with a predictive accuracy of **95.5%**. Overall, the semi-supervised approach achieves higher F scores. To confirm what we concluded from the results, we applied a paired t-test to test the difference in F_2 scores between the fully-supervised MLP algorithm and the co-training algorithm. Before carrying out this test, we confirmed that the data satisfied the assumptions necessary for the paired t-test to be relevant—continuous, independent, normally distributed data without outliers. This resulted in a t-statistic of 7.7 and a **p-value of** $\mathbf{1.7 \times 10^{-13}}$ which suggests that the difference between the F_2 scores of the two algorithms was not due to chance.

We also computed the precision, recall and F-score for the best fully-supervised approach and best semi-supervised approach on the minority class, i.e. the relevant tweets. The results of this experiment are presented in Table 8. The semi-supervised approach produces both a stronger $F1$ and $F2$ score on the minority class. To give a better understanding of how the different measures manage to balance the number of FP and FN, we also present the confusion

Table 8. Performance results of the best fully-supervised approach and best semi-supervised approach on the minority class—"*relevant*."

Algorithms	Metric on Relevant Class			
	Precision	Recall	F_1 Score	F_2 Score
Fully-supervised	1.000	0.835	0.910	0.864
Semi-supervised	0.839	1.000	0.912	0.963

https://doi.org/10.1371/journal.pone.0210689.t008

Table 9. Confusion matrix for MLP fully-supervised classification on the test data.

	Actual True	Actual False	Total
Predicted True	TP (256)	FP (20)	276
Predicted False	FN (35)	TN (891)	926
Total	291	911	$N = 1202$

https://doi.org/10.1371/journal.pone.0210689.t009

matrices for both the best performing fully-supervised and semi-supervised methods on the test data. These confusion matrices are shown in Tables 9 and 10 respectively. From the confusion matrices, we see that the semi-supervised approach performs better for the purpose of syndromic surveillance as it yields only 17 false negatives even though it also yields 37 false positives. Considering that our aim is to develop a filtering system to identify the few relevant tweets in order to register a signal for syndromic surveillance, it is critical to have high recall, hopefully accompanied by high precision, and therefore high accuracy. The semi-supervised method is able to identify and retain relevant tweets more often, while also being able to identify irrelevant tweets to a reasonable degree. Hence, even with a shortage of labelled data, the semi-supervised algorithms can be used to filter and retain relevant tweets effectively.

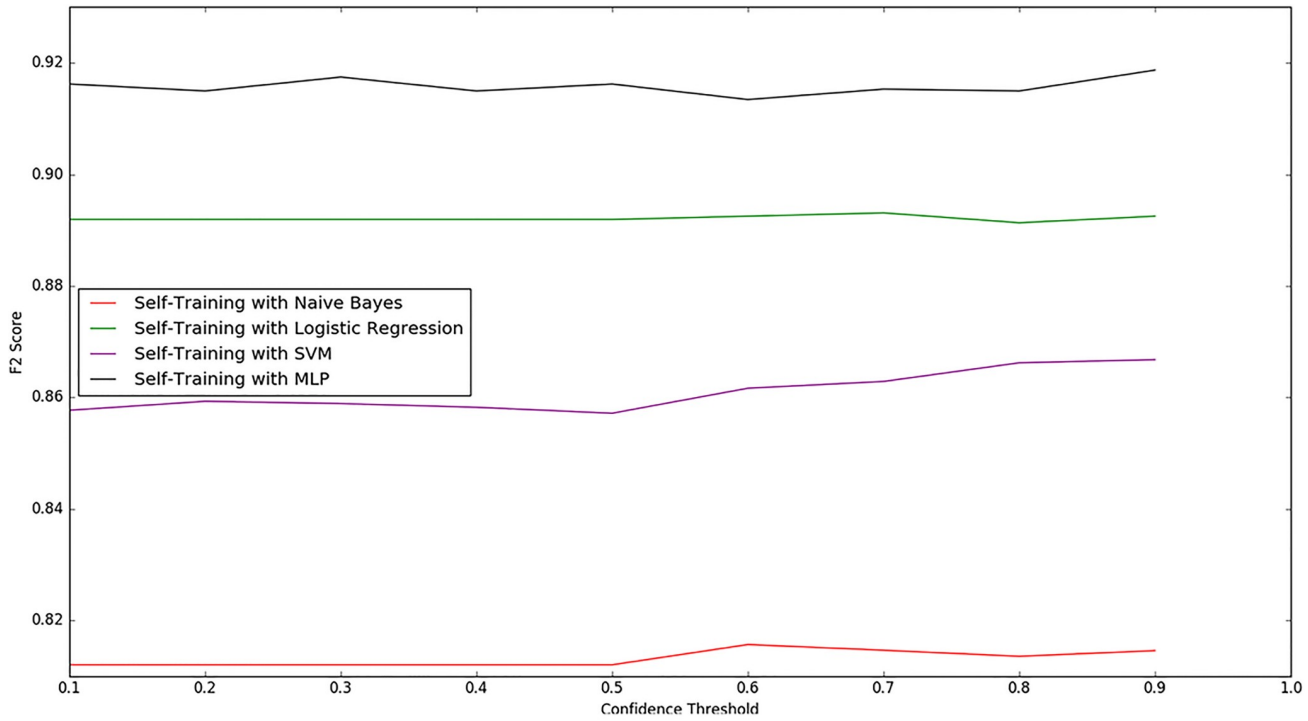
Fig 2 shows how the performances of the semi-supervised systems change as the confidence threshold changes. The confidence threshold controls how conservatively the semi-supervised system assimilates unlabelled instances as it represents how confident the semi-supervised system needs to be in its classification before assimilating the instance to inform future decisions. We observed co-training with MLP to perform best. We also observed that for lower confidence thresholds between 0.1 and 0.5, self-training performance is usually lower and does not change much between thresholds. Co-training on the other hand, appears to be less sensitive to this parameter. Fig 2 also reiterates what we learned from Table 7 that the MLP is our strongest fully-supervised model. In addition, while the logistic regression classifier does not perform as well as the MLP, it appears to be robust to different confidence thresholds when used in an iterative labelling context. We hypothesise that this advantage arises because the logistic regression classifier has considerably less hyper-parameters to optimise. This means that if a set of hyper-parameters, which is impactful on performance, is not optimal for a certain threshold, such a set would be less of a hinderance to the logistic regression model.

The main issue with iterative labelling algorithms is that, because the classifiers are not perfect and do not have 100% accuracy, we cannot be sure that the unlabelled instances that they label for assimilation are always correct. This means that their initial performance before any labelling iterations is vital. Consider a classifier, initially of poor performance (with an accuracy of 0.2 for example). When classifying unlabelled instance with which to train itself, 80% of its classifications will be wrong, so it will assimilate false hypotheses, which will in turn make its performance in the next iteration even worse and so on. Conversely, if the initial accuracy is high, it is more likely to correctly classify unlabelled instance and be less resistant to the drop in performance from assimilating false hypotheses. We conducted an experiment to measure the quality of the automatically labelled instances assimilated by our

Table 10. Confusion matrix for Co-training semi-supervised algorithm on the test data.

	Actual True	Actual False	Total
Predicted True	TP (274)	FP (37)	311
Predicted False	FN (17)	TN (874)	891
Total	291	911	$N = 1202$

https://doi.org/10.1371/journal.pone.0210689.t010



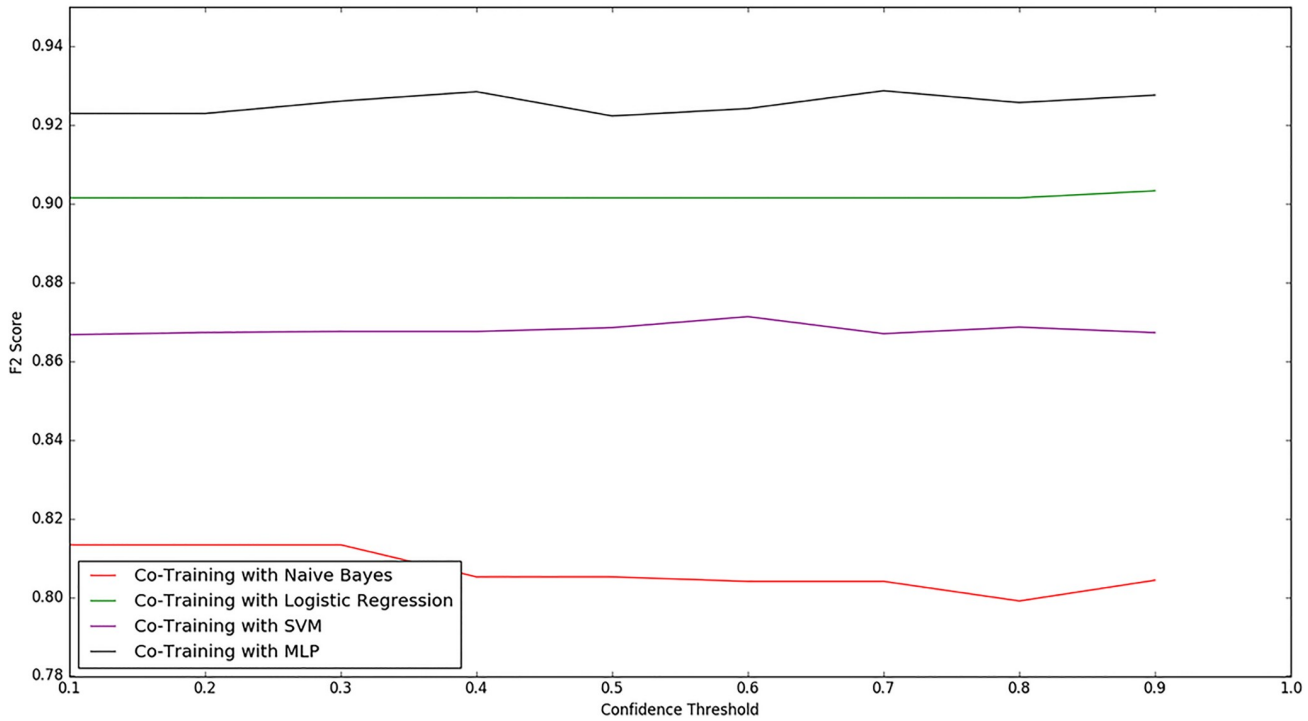


Fig 2. Graph of F2 performance of Iterative Labelling using different confidence thresholds.

https://doi.org/10.1371/journal.pone.0210689.g002

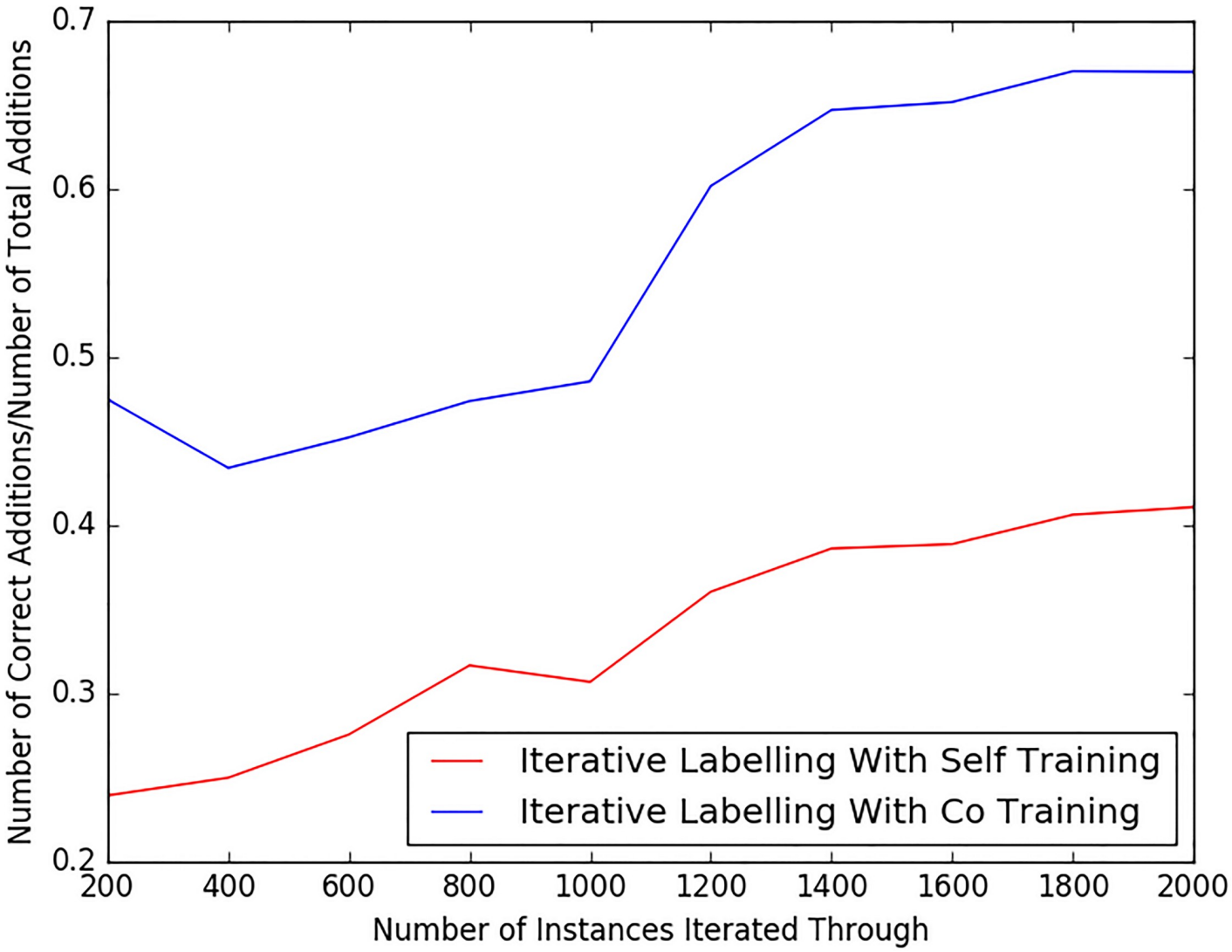


Fig 3. Graph showing how many correct assimilations the iterative labeling algorithms make per iteration using labelled data from a different time period.

https://doi.org/10.1371/journal.pone.0210689.g003

semi-supervised classifiers. For this exercise, we used the second set of labelled tweets from a different time period as the "unlabelled" set with to which the iterative labelling is applied to. The same training set as in our other experiments was used for the initial training stage. The self-training and co-training processes were initiated, applying these classifiers to the alternative set of labelled data (around 2000 instances) in steps of 200. Fig 3 shows a plot of the proportion of correctly classified instances that the iterative labelling process assimilated. The co-training approach had a higher rate of being correct when making new additions. This was in fact the aim of adopting co-training with its multiple different views of the same data. The proportion of correct assimilations of both the self-training and co-training methods rises as more data is assimilated, due to the fact that the systems are getting more intelligent. Although we could not test beyond 2000 instances (because of our limited labelled data), we believe that the proportion of correct assimilations will increase until a certain point, after which it will plateau. We expect this plateau due to the fact that at a certain point, the iterative learning classifiers will have nothing new to learn from new data after having been exposed to so much.

As with the features constructed, we tested how the classifiers would perform for new data collected at a different time period to assess if shifts in language and colloquialisms could have an impact on performance. Our classifiers were built on data from the first collection period (see Table 2). For a simple assessment, we applied our trained model to tweets collected in the most recent collection period, which had a time gap of two years from the original data. Our semi-supervised approach based on co-training achieved a precision of 0.418 and a recall of 0.587 on the 2,000 labelled tweets from the most recent collection period. This means an F_1 score of 0.488 and more importantly, an F_2 score of 0.543. For comparison purposes, we also applied the fully-supervised MLP algorithm to the data from this new time period. This yielded a precision of 0.420 and a recall of 0.410. This meant an F_1 score of 0.415 and an F_2 score of 0.412. In both cases, we observe a deterioration in performance when introduced to tweets from a different time period. This poses an important issue to consider about how language online changes moving forward. Although it changes very gradually, after a period of one or two years, the changes are substantial enough to render the natural language-based models less effective.

4.2 Feature analysis

Table 11 shows the results of the feature ablation experiments. Our central feature was the *n*-gram. Without it, we see the performance of our systems drop by around 0.2. We also found that our supporting features yield some additional improvements in performance on top of the *n*-gram features. For each of these supporting features, their omission results in a drop in performance of around 0.1. Of our additional features, we found that *Negative Emojis/Emoticons* were the most discriminative, followed by the *Denotes Laughter* feature in the supervised approach, which also captures emojis as well as colloquialisms, and *Positive/Negative Word Count* in the semi-supervised approach. All three of these features capture the mood of a tweet.

Table 12 shows the words found to be most informative. For example, the table shows that, of the tweets containing the word *chest*, 96% are relevant and only 4% are irrelevant. The training data is used for this calculation. A surprising negative predictor was the word *health*. When *health* appeared in a tweet, the tweet was irrelevant 94% of the time. The word *pollution* shows a similar trend. This suggests that when Twitter users are expressing health issues, they may not use precise or formal terms, opting for simple symptomatic and emotional words such as *chest*, *cold* or *wow*. The more formal terms may be more often associated with news items or general chat or discussion. Using this information, we could include some of the more relevant but perhaps unexpected keywords as keywords when collecting streaming tweets from Twitter in order to better target and collect relevant tweets.

We also investigated which emojis were most prevalent in our data set as well as how often each emoji appeared in tweets for each class. Fig 4 shows the frequency with which each emoji

Table 11. F1 scores after feature ablation.

Ablated Feature	Supervised F_1 Score	Semi-supervised F_1 Score
None	0.710	0.714
Denotes Laughter	0.628	0.643
Negative Emojis/Emoticons	**0.627**	**0.620**
Word Classes	0.677	0.637
Positive/Negative Word Count	0.648	0.625
n-grams	**0.561**	**0.596**

https://doi.org/10.1371/journal.pone.0210689.t011

Table 12. Most informative words measured by their *Informativeness* and their relevant:irrelevant prior probabilities.

Word	*I* (Relevant:Irrelevant)	Relevant Prior Probability	Irrelevant Prior Probability
chest	22/4	0.96	0.04
throat	17/1	0.95	0.05
wow	17/1	0.95	0.05
health	1/17	0.06	0.94
cold	16/1	0.94	0.06
moment	15/1	0.94	0.06
forecast	1/14	0.07	0.93
awake	13/1	0.93	0.07
awful	13/1	0.93	0.07
sick	13/1	0.93	0.07
cough	12/1	0.92	0.08
pollution	1/12	0.08	0.92
bed	11/1	0.91	0.09
hate	11/1	0.91	0.09
watch	10/1	0.91	0.09

https://doi.org/10.1371/journal.pone.0210689.t012

occurred in the labelled tweets. It shows that only a few emojis appear very frequently in tweets collected in our context. This means that only a few important emojis were needed for determining tweet relevancy as opposed to monitoring for the full emoji dictionary. Fig 5 shows a list of some emojis and the distribution of classes that tweets belonged to whenever they contained said emoji. Overall, it can be seen that each of these emojis tends to lean heavily toward one class. This suggests that they could be quite discriminative and useful indicators of class membership and hence, helpful features.

4.3 Keyword analysis

Table 13 shows the results of the assessment of keywords used in tweet collection from an information retrieval point of view. We found that *asthma*, *pollution* and *air pollution* were the keywords that yielded the most results at 1313, 757 and 509 out of a total of 3500. *Wheezing*, *fumes* and *inhaler* were next with 219, 132, 121 tweets respectively. The remaining keywords return very few results (below 40) or no results. In an information retrieval scenario, precision refers to the fraction of retrieved results that are relevant to the query while recall refers to the fraction of relevant results that are successfully retrieved. *Asthma* had the highest recall but not very high precision so most of its results were irrelevant. *Wheezing*, *inhaler*, *wheeze*, *cannot breathe*, *can't breathe*, *difficulty breathing* and *short of breath* have good precision although their recall is not that high. Some of those keywords express direct symptoms of the syndrome under investigation, hence, we expect good precision. *Tight chest* and *pea souper* have very high precision but only appeared in two tweets each. Of the keywords used, *wheezing* was the most useful in that it brought in a lot of results, most of which were relevant. We included a common misspelling of asthma, the keyword with the highest recall power—*asma*. We found that *asma* only appeared in 4 tweets. We hypothesise that this is due to the fact that most users of Twitter post from devices capable of autocorrect hence it may not be necessary to worry about misspelling of keywords.

The informativeness, *I*, was calculated when the keywords were also features in the classifiers and is presented in Table 14. Most of the keywords were not informative from a feature selection point of view, with an information gain ratio of 1:1 for relevant:irrelevant tweets so

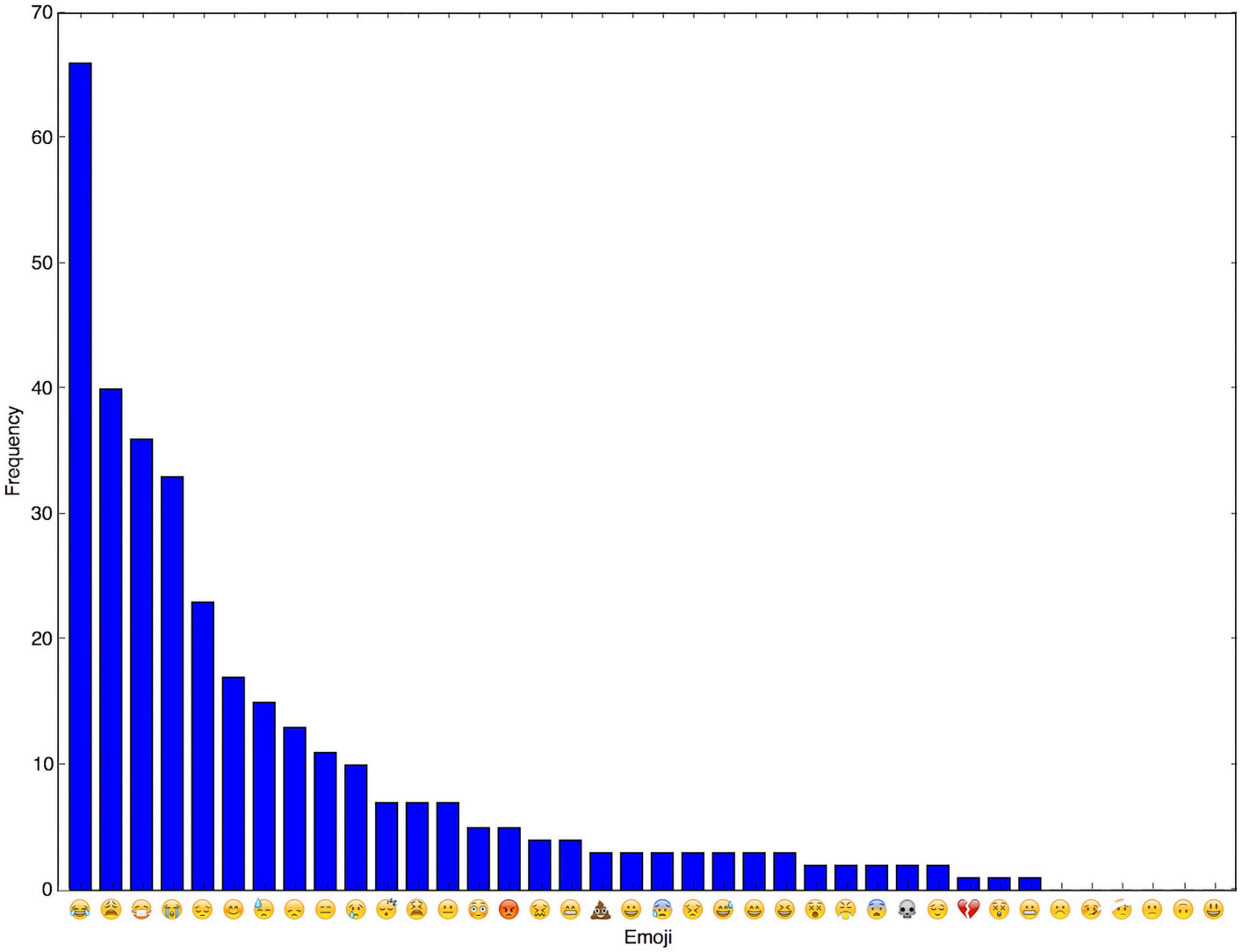


Fig 4. Bar chart showing emoji frequency in labelled data.

https://doi.org/10.1371/journal.pone.0210689.g004

they are not included. We found that some of the keywords used for collecting tweets were also informative in the model, although not always associated with the relevant class. For example, *pollution*, which was a keyword, appeared in the ranking of top 15 most informative words associated with the irrelevant class. Similarly *fumes* also associates with the irrelevant class more. On the other hand *wheezing* had good information retrieval performance and associated with relevant tweets.

4.4 Syndromic surveillance comparisons

As we discussed earlier, the purpose of our relevance filtering is a syndromic surveillance application. After constructing and evaluating our semi-supervised filtering systems, we assessed their utility for syndromic surveillance purposes by retrospectively applying them to Twitter data and comparing the results against data for existing syndromic indicators from the Public Health England (PHE) Real-time Syndromic Surveillance Service. For this experiment,

Emoji	I (Relevant:Irrelevant)	Emoji	I (Relevant:Irrelevant)
😂	17/49	😴	5/2
😩	31/9	😫	6/1
😷	27/9	😐	5/2
😭	21/12	😳	3/2
😔	17/6	😡	4/1
😊	11/6	😖	3/1
😓	12/3	😁	3/1
😞	10/3	💩	3/0
😑	11/0	😀	0/3
😢	8/2	😰	2/1

Fig 5. Most frequent emojis in labelled data and their distributions.

https://doi.org/10.1371/journal.pone.0210689.g005

Table 13. Assessment of the quality of the search keywords from an information retrieval perspective. Precision is the fraction of retrieved tweets that are relevant to the query. Recall is the fraction of the relevant tweets that are successfully retrieved.

Keyword	Precision	Recall	Keyword	Precision	Recall
asthma	0.174	0.475	poor air quality	0.000	0.000
pollution	0.009	0.015	murk	0.000	0.000
air pollution	0.008	0.008	can't breathe	0.556	0.010
wheezing	0.406	0.185	difficulty breathing	0.125	0.002
fumes	0.030	0.008	short of breath	0.333	0.004
inhaler	0.198	0.050	respiratory disease	0.000	0.000
smog	0.023	0.002	asma	0.000	0.000
gasping	0.025	0.002	tight chest	0.500	0.002
puffing	0.033	0.002	pea souper	0.500	0.002
wheeze	0.138	0.008	itchy eyes	0.000	0.000
panting	0.043	0.002	could not breathe	0.000	0.000
cannot breathe	0.412	0.015	coudn't breathe	0.000	0.000
trouble breathing	0.100	0.002	chest tightness	0.000	0.000
sore eyes	0.100	0.002	acid rain	0.000	0.000

https://doi.org/10.1371/journal.pone.0210689.t013

Table 14. Information gain or keyword Informativeness *I* of keywords, conceptualised as the ratio relevant: irrelevant.

Streaming Keyword	*I*	Relevant Prior Probability	Irrelevant Prior Probability
pollution	1/12	0.08	0.92
wheezing	5/1	0.84	0.16
fumes	1/3	0.24	0.76
panting	1.6/1.0	0.62	0.38

https://doi.org/10.1371/journal.pone.0210689.t014

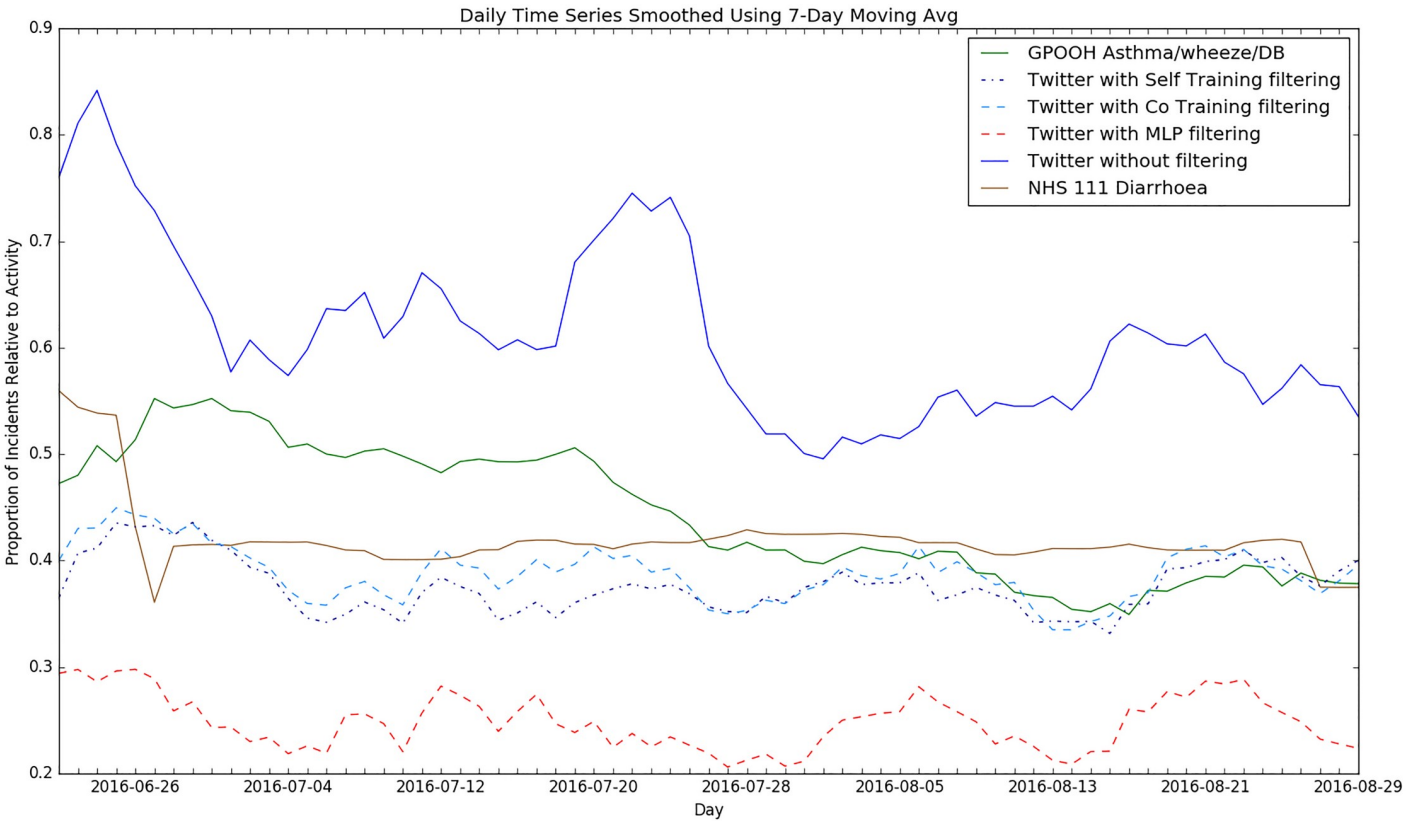


Fig 6. Time series plots comparing GP asthma/wheeze/difficulty breathing data to signals from supervised and semi-supervised Twitter analysis and unrelated diarrhoea.

https://doi.org/10.1371/journal.pone.0210689.g006

we made use of unlabelled tweets from our second collection period, June to August 2016. Syndromic surveillance data on the proportion of daily GP Out-of-Hours (GPOOH) calls for ***asthma, wheezing or difficulty breathing*** and tele-health calls (NHS 111) for ***difficulty breathing*** for the period June to August 2016 were compared to the signal detected from our Twitter data using our semi-supervised and fully-supervised filtering systems. As a sense check, we also compared our detected Twitter signal time series against non-respiratory syndrome data in the form of NHS 111 calls for ***diarrhoea***. This was to provide some form of control which should not correlate with the Twitter signal.

The resulting time series shows the daily proportion of relevant symptomatic tweets and consultations/calls as observed on Twitter and recorded by PHE (Figs 6 and 7). The signals were smoothed using a 7-day moving average to remove the fluctuations in daily activity for GPOOH data as that service receives more usage over the weekends. We also included a time series showing the Twitter signal without any filtering for further perspective. We see that the time series plots of the self-training and co-training filtering follow a similar trend to the GP data time series. Also, the time series for the Twitter data without any filtering has lots of spurious peaks in relation to the ground truth data (i.e. the syndromic surveillance data). Both of these observations together suggest that Twitter data might mirror the health activity of a population and that relevance filtering is useful in reducing noise and obtaining a clearer picture of such activity. Additionally, we see that while the unfiltered Twitter signal does not match well with the *asthma/wheeze/difficulty breathing* or *difficulty breathing* signal, it still seems to match better than that of the diarrhoea signal.

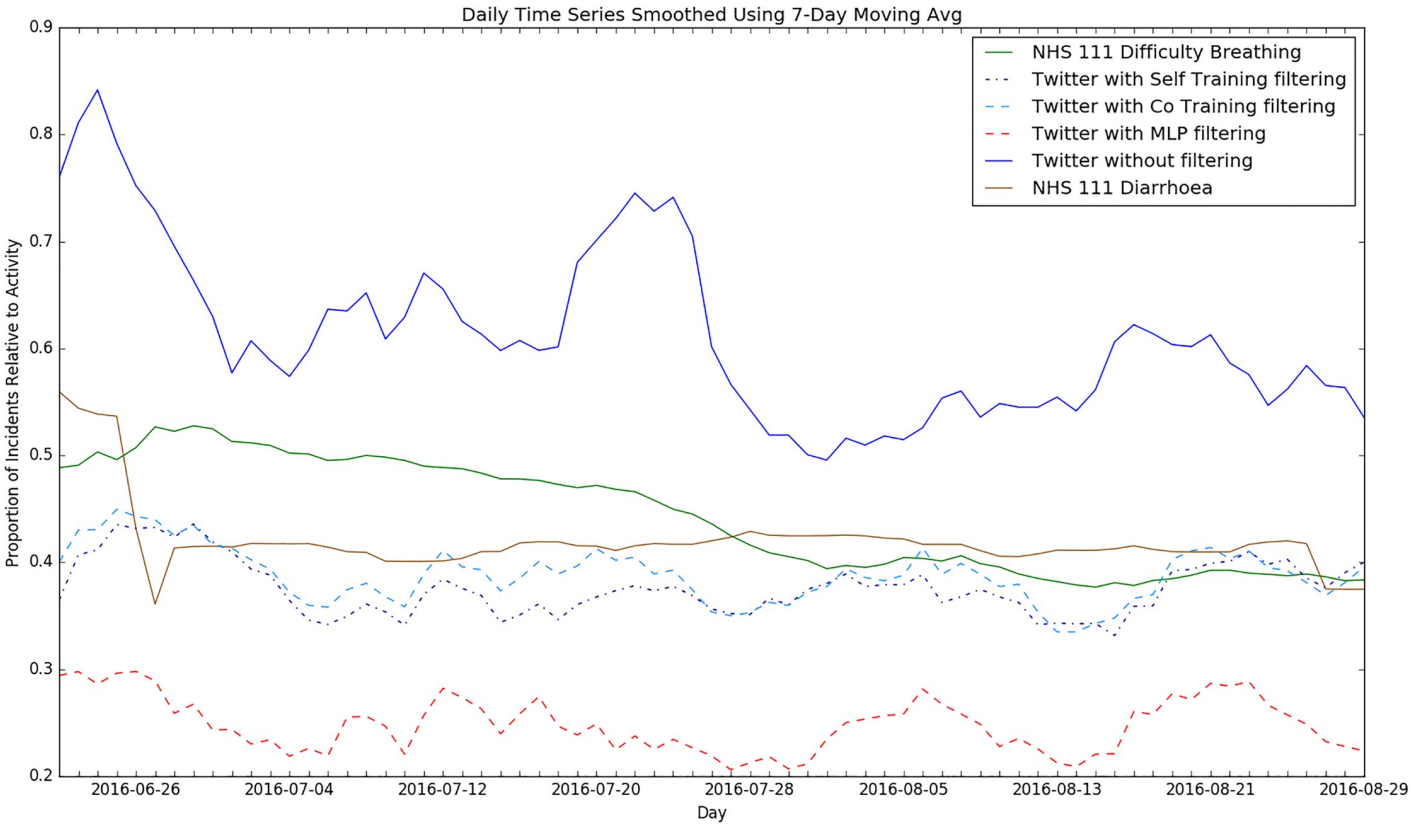


Fig 7. Time series plots comparing NHS 111 difficulty breathing data to signals from supervised and semi-supervised Twitter analysis and unrelated diarrhoea.

https://doi.org/10.1371/journal.pone.0210689.g007

To further evaluate the quality of our detected signal, we calculated the Pearson correlation coefficient to determine the strength and direction of any monotonic relationship between the indicators (Table 15). We observed a weak but statistically significant correlation between the Twitter signals and the asthma and difficulty breathing syndromic surveillance data. The signal produced by the co-training method achieved a stronger correlation than that the signal produced by the best fully-supervised method—the MLP. For the diarrhoea syndrome, there was no statistically significant correlation with the Twitter time series. To corroborate this, it can be seen from the time series that the plots for our detected Twitter signals and GP data follow a similar downward trend but the diarrhoea signal does not. This further suggests that our semi-supervised system can detect activity similar to that detected by traditional syndromic surveillance systems recording GP visits and NHS 111 calls. This suggests that they could potentially be further explored as an additional form of syndromic surveillance, even in an environment with scarce labelled data.

Table 15. Pearson correlations and P-Values for detected signals with syndromic surveillance signals.

Syndrome	Relevance Filtering Algorithm		
	Self-Training	Co-Training	MLP Neural Network
Asthma/Wheezing/DB	0.249($p = 0.04$)	0.414($p = 0.0004$)	0.255($p = 0.03$)
Difficulty Breathing	0.228($p = 0.04$)	0.424($p = 0.0002$)	0.214($p = 0.07$)
Diarrhoea	−0.01($p = 0.9$)	0.04($p = 0.7$)	0.05($p = 0.7$)

https://doi.org/10.1371/journal.pone.0210689.t015

5 Discussion

Twitter is a noisy data source for syndromic surveillance but through data processing and tweet classification, we have been able to identify relevant tweets among the noisy data for a specific syndrome or incident (asthma/difficulty breathing). This in turn allowed us to extract a signal of potential use for syndromic surveillance that correlated positively with real-world public health data. Using a semi-supervised method of classification for filtering tweets, we achieve an accuracy of 95.5% and F_1 and F_2 scores of 0.910 and 0.929 respectively. We argued that recall is very important for us because we want to keep all the relevant tweets so that we can have some signal, even if amplified by some misclassified irrelevant tweets. The best recall, obtained by the semi-supervised algorithm equated to retaining over 90% of the relevant tweets after classification. Also, the semi-supervised approach allowed us to use 8000 previously unlabelled tweets before it started to see a deterioration in performance. This allowed us to make more use of the data collected.

Tweet classification using supervised learning has received a lot of attention [19, 20, 55–57] and gave us good results with F_1 and F_2 scores of 0.910 and 0.919 respectively for the MLP neural network. Tweet labelling, required for supervised classification, is time consuming, however, so often researchers do not use all of the data available to build the model. Semi-supervised methods for tweet classification have been used for sentiment analysis [58, 59]. They can enable more of the collected data to be used for training the classifier bypassing some of the labelling effort. Johnson el al. [60] used a method called label propagation and reported accuracy of 78%. Baugh [30] proposed a hierarchical classification system with self-training and reported accuracy of 61% and an F_1 score of 0.54. We have implemented an iterative labelling semi-supervised approach which seems to have competitive performance and also enables us to use more of the training data without the effort of labelling. Furthermore, we get an improvement on recall over the supervised method, which is important given that the signal we are trying to preserve for syndromic surveillance may be weak. We compare our semi-supervised system to others above but we acknowledge that applications in different domains might weaken the comparison. Baugh [30] also applied semi-supervised systems to tweet classification but not for syndromic surveillance so this comparison might be of more value.

We have also identified strong and novel features in the context of tweet classification: emojis. We have hinted at the growing use of emojis [39] and their importance in establishing the tone of at tweet which in turn is important to relevance classification. Emojis cross language boundaries and are often used by people expressing conditions of interest to syndromic surveillance. Our custom features constructed based on Twitter colloquialisms including emojis proved effective in improving classification performance. Of all our custom features, the one that stood out most was the *Negative Emojis/Emoticons* feature. Emoticons have been used previously [11]. Emojis work even better than emoticons and their uniformity is a real advantage. A smile emoticon could be illustrated in the form ":-D" or ":D". However, because emojis are actually unicode encoded pictographs with a set standard [61], there exist no variants of the same emoji. In a learning scenario, this reduces fragmentation or duplication of features making them more ideal as features than emoticons.

In terms of geolocation of tweets, we have found that most of the obvious location indicators are not well populated, and those that are, may not be accurate. Hence, future work must tackle geolocation as a real part of the problem for establishing a proper signal from Twitter. After comparing our extracted Twitter signal to real world syndromic surveillance data, we found a positive, albeit weak correlation. This suggests that there is a relationship between asthma related Twitter activity and syndromic surveillance data for asthma and breathing-related incidents. While the actual correlation value indicates a weak relationship, it still

suggests that we can detect relevant activity on Twitter which is similar or complementary to that which is collected by traditional means. The strength of the correlation might be affected by the weak location filtering that we have been able to perform. As we discussed, the syndromic surveillance data relates to England but the Twitter data has only been located (not accurately) to the UK. As future work, we plan to assess the full detection capability of Twitter by repeating this analysis prospectively over a longer time period, and for different syndromes, allowing us to determine whether Twitter can detect activity that is of potential benefit to syndromic surveillance.

We also found that "what to collect" is problematic as the data collection of tweets by keywords requires a carefully chosen list of keywords. Furthermore, our experimentation with different type of features like emojis also tell us that the vocabulary used in Twitter is different to expression in other settings (e.g. as part of a medical consultation). Hence we may need to widen our data collection terms to include emojis, emoticons and other types of informal expressions. We may also need to develop adaptive systems in which the set of data collection keywords is dynamically updated to collect truly relevant tweets. So an idea for future research is to begin with a set of keywords, collect tweets, perform relevance analysis and then update the keyword/token list to reflect those that associate with the most relevant tweets, eliminating any keywords/tokens that are not performing adequately.

We saw also that vocabulary and use of tokens change over time. *Negative emojis/emoticons* appeared more often in the second time period, up from 5.5% to 14.4% of labelled tweets containing them. This could suggest that over the past two years, the use of emojis as a form of expression has grown. However their prevalence in each class also changed, which may explain the classification performance showing some marked deterioration in precision. We performed our research on data collected within a two year period, but further data collection and experimentation would be beneficial to understand the temporality of models generated as Twitter conversations change over time.

Supporting information

S1 List. Appendix. Twitter data collection keywords. Pollution, smog, poor air quality, wheeze, wheezing, difficulty breathing, asthma, inhaler, air pollution, itchy eyes, sore eyes, trouble breathing, cannot breathe, could not breathe, can't breathe, coudn't breathe, asma, short of breath, tight chest, chest tightness, respiratory disease, pea souper, murk, fumes, acid rain, gasping, puffing, panting.
(TXT)

S2 List. Appendix. Positive Word Dictionary. Adore, adorable, accomplish, achievement, achieve, action, active, admire, adventure, agree, agreeable, amaze, amazing, angel, approve, attractive, awesome, beautiful, brilliant, bubbly, calm, celebrate, celebrating, charming, cheery, cheer, clean, congratulation, cool, cute, divine, earnest, easy, ecstasy, ecstatic, effective, effective, efficient, effortless, elegant, enchanting, encouraging, energetic, energized, enthusiastic, enthusiasm, excellent, exciting, excited, fabulous, fair, familiar, famous, fantastic, fine, fit, fortunate, free, fresh, friend, fun, generous, genius, glowing, good, great, grin, handsome, happy, hilarious, hilarity, lmao, lol, rofl, haha, healthy, ideal, impressive, independent, intellectual, intelligent, inventive, joy, keen, laugh, legendary, light, lively, lovely, lucky, marvel, nice, okay, paradise, perfect, pleasant, popular, positive, powerful, pretty, progress, proud, quality, refresh, restore, right, smile, success, sunny, super, wealthy, money, cash, well, wonderful, wow, yes, yum.
(TXT)

S3 List. Appendix. Negative Word Dictionary. Abysmal, adverse, alarming, angry, rage, annoy, anxious, anxiety, attack, appalling, atrocious, awful, bad, broken, can't, not, cant, cannot, cold, collapse, crazy, cruel, cry, damage, damaging, depressed, depression, dirty, disease, disgust, distress, don't, dont, dreading, dreadful, dreary, fail, fear, scare, feeble, foul, fright, ghastly, grave, greed, grim, gross, grotesque, gruesome, guilty, hard, harm, hate, hideous, horrible, hostile, hurt, icky, ill, impossible, injure, injury, jealous, lose, lousy, messy, nasty, negative, never, no, nonsense, crap, shit, fuck, fukk, fuxk, nausea, nauseous, pain, reject, repulsive, repulse, revenge, revolting, rotten, rude, ruthless, sad, scary, severe, sick, slimy, smelly, sorry, sticky, stinky, stormy, stress, stuck, stupid, tense, terrible, terrifying, threaten, ugly, unfair, unhappy, unhealthy, unjust, unlucky, unpleasant, upset, unwanted, unwelcome, vile, wary, weary, wicked, worthless, wound, yell, yucky.
(TXT)

Acknowledgments

We acknowledge support from NHS 111 and NHS Digital for their assistance with the NHS 111 system; Out-of-Hours providers submitting data to the GPOOH syndromic surveillance and Advanced Heath & Care.

Author Contributions

Conceptualization: Oduwa Edo-Osagie, Gillian Smith, Iain Lake, Obaghe Edeghere, Beatriz De La Iglesia.

Data curation: Oduwa Edo-Osagie, Gillian Smith, Iain Lake, Obaghe Edeghere, Beatriz De La Iglesia.

Formal analysis: Oduwa Edo-Osagie, Beatriz De La Iglesia.

Funding acquisition: Gillian Smith, Iain Lake, Obaghe Edeghere, Beatriz De La Iglesia.

Investigation: Oduwa Edo-Osagie, Beatriz De La Iglesia.

Methodology: Oduwa Edo-Osagie, Beatriz De La Iglesia.

Project administration: Gillian Smith, Iain Lake, Obaghe Edeghere, Beatriz De La Iglesia.

Software: Oduwa Edo-Osagie.

Supervision: Iain Lake, Obaghe Edeghere, Beatriz De La Iglesia.

Validation: Oduwa Edo-Osagie, Gillian Smith, Iain Lake, Obaghe Edeghere, Beatriz De La Iglesia.

Writing – original draft: Oduwa Edo-Osagie.

Writing – review & editing: Oduwa Edo-Osagie, Gillian Smith, Iain Lake, Obaghe Edeghere, Beatriz De La Iglesia.

References

1. World Health Organisation WHO. The world health report 2007—A safer future: global public health security in the 21st century; 2007. http://www.who.int/whr/2007/en/.
2. Elliot AJ, Smith S, Dobney A, Thornes J, Smith GE, Vardoulakis S. Monitoring the effect of air pollution episodes on health care consultations and ambulance call-outs in England during March/April 2014: A retrospective observational analysis. Environmental pollution. 2016; 214:903–911. https://doi.org/10.1016/j.envpol.2016.04.026 PMID: 27179935

3. Triple S. Assessment of syndromic surveillance in Europe. Lancet (London, England). 2011; 378 (9806):1833.
4. Mooney SJ, Pejaver V. Big data in public health: terminology, machine learning, and privacy. Annual review of public health. 2018; 39:95–112. https://doi.org/10.1146/annurev-publhealth-040617-014208 PMID: 29261408
5. Wong ZS, Zhou J, Zhang Q. Artificial Intelligence for infectious disease Big Data Analytics. Infection, disease & health. 2019; 24(1):44–48.
6. Chretien JP, Burkom HS, Sedyaningsih ER, Larasati RP, Lescano AG, Mundaca CC, et al. Syndromic Surveillance: Adapting Innovations to Developing Settings. PLOS Medicine. 2008; 5(3):1–6. https://doi.org/10.1371/journal.pmed.0050072
7. Khatua A, Khatua A, Cambria E. A tale of two epidemics: Contextual Word2Vec for classifying twitter streams during outbreaks. Information Processing & Management. 2019; 56(1):247–257. https://doi.org/10.1016/j.ipm.2018.10.010
8. Charles-Smith LE, Reynolds TL, Cameron MA, Conway M, Lau EH, Olsen JM, et al. Using social media for actionable disease surveillance and outbreak management: A systematic literature review. PloS one. 2015; 10(10):e0139701. https://doi.org/10.1371/journal.pone.0139701 PMID: 26437454
9. De Quincey E, Kostkova P. Early warning and outbreak detection using social networking websites: The potential of Twitter. In: International Conference on Electronic Healthcare. Springer; 2009. p. 21–24.
10. Serban O, Thapen N, Maginnis B, Hankin C, Foot V. Real-time processing of social media with SENTINEL: A syndromic surveillance system incorporating deep learning for health classification. Information Processing & Management. 2019; 56(3):1166–1184. https://doi.org/10.1016/j.ipm.2018.04.011
11. Lamb A, Paul MJ, Dredze M. Separating Fact from Fear: Tracking Flu Infections on Twitter. In: HLT-NAACL; 2013. p. 789–795.
12. Broniatowski DA, Paul MJ, Dredze M. National and Local Influenza Surveillance through Twitter: An Analysis of the 2012-2013 Influenza Epidemic. PLoS ONE. 2013; 8(12). https://doi.org/10.1371/journal.pone.0083672 PMID: 24349542
13. Copeland P, Romano R, Zhang T, Hecht G, Zigmond D, Stefansen C. Google Disease Trends: an update. In: International Society of Neglected Tropical Diseases 2013; 2013. p. 3.
14. Brownstein JS, Freifeld CC, Reis BY, Mandl KD. Surveillance Sans Frontiéres: Internet-Based Emerging Infectious Disease Intelligence and the HealthMap Project. PLoS Medicine. 2008; 5(7):1–6. https://doi.org/10.1371/journal.pmed.0050151
15. Sadilek A, Kautz H, Silenzio V. Predicting Disease Transmission from Geo-Tagged Micro-Blog Data. In: AAAI Conference on Artificial Intelligence; 2012. http://www.aaai.org/ocs/index.php/AAAI/AAAI12/paper/view/4844.
16. Culotta A. Towards Detecting Influenza Epidemics by Analyzing Twitter Messages. In: Proceedings of the First Workshop on Social Media Analytics. SOMA'10. New York, NY, USA: ACM; 2010. p. 115–122. http://doi.acm.org/10.1145/1964858.1964874.
17. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature. 2009; 457(7232):1012–1014. https://doi.org/10.1038/nature07634 PMID: 19020500
18. Achrekar H, Gandhe A, Lazarus R, Yu SH, Liu B. Twitter Improves Seasonal Influenza Prediction. In: Healthinf; 2012. p. 61–70.
19. Sriram B, Fuhry D, Demir E, Ferhatosmanoglu H, Demirbas M. Short text classification in twitter to improve information filtering. In: Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval. ACM; 2010. p. 841–842.
20. Dilrukshi I, De Zoysa K, Caldera A. Twitter news classification using SVM. In: Computer Science & Education (ICCSE), 2013 8th International Conference on. IEEE; 2013. p. 287–291.
21. Hu H, Moturu P, Dharan K, Geller J, Iorio S, Phan H, et al. Deep learning model for classifying drug abuse risk behavior in tweets. In: 2018 IEEE International Conference on Healthcare Informatics (ICHI). IEEE; 2018. p. 386–387.
22. Lee K, Qadir A, Hasan SA, Datla V, Prakash A, Liu J, et al. Adverse Drug Event Detection in Tweets with Semi-Supervised Convolutional Neural Networks. In: Proceedings of the 26th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee; 2017. p. 705–714.
23. Dai X, Bikdash M, Meyer B. From social media to public health surveillance: Word embedding based clustering method for twitter classification. In: SoutheastCon 2017. IEEE; 2017. p. 1–7.
24. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems; 2013. p. 3111–3119.

25. Lee SH, Levin D, Finley P, Heilig CM. Chief complaint classification with recurrent neural networks. arXiv preprint arXiv:180507574. 2018;.
26. Xi G, Yin L, Li Y, Mei S. A Deep Residual Network Integrating Spatial-temporal Properties to Predict Influenza Trends at an Intra-urban Scale. In: Proceedings of the 2nd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery. ACM; 2018. p. 19–28.
27. Edo-Osagie O, De La Iglesia B, Lake I, Edeghere O. Deep Learning for Relevance Filtering in Syndromic Surveillance: A Case Study in Asthma/Difficulty Breathing. In: Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods—Volume 1: ICPRAM,. INSTICC. SciTePress; 2019. p. 491–500.
28. Zhao J, Lan M, Zhu TT. ECNU: Expression-and message-level sentiment orientation classification in Twitter using multiple effective features. SemEval 2014. 2014; p. 259.
29. Becker L, Erhart G, Skiba D, Matula V. Avaya: Sentiment analysis on twitter with self-training and polarity lexicon expansion. In: Second Joint Conference on Lexical and Computational Semantics (* SEM). vol. 2; 2013. p. 333–340.
30. Baugh W. bwbaugh: Hierarchical sentiment analysis with partial self-training. In: SemEval@NAACL-HLT. Atlanta, Georgia, USA; 2013. p. 539.
31. Liu S, Zhu W, Xu N, Li F, Cheng Xq, Liu Y, et al. Co-training and visualizing sentiment evolvement for tweet events. In: Proceedings of the 22nd International Conference on World Wide Web. ACM; 2013. p. 105–106.
32. Oliver A, Odena A, Raffel CA, Cubuk ED, Goodfellow I. Realistic evaluation of deep semi-supervised learning algorithms. In: Advances in Neural Information Processing Systems; 2018. p. 3239–3250.
33. Lim S, Tucker CS, Kumara S. An unsupervised machine learning model for discovering latent infectious diseases using social media data. Journal of biomedical informatics. 2017; 66:82–94. https://doi.org/10.1016/j.jbi.2016.12.007 PMID: 28034788
34. Thelwall M, Buckley K, Paltoglou G, Cai D, Kappas A. Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology. 2010; 61(12):2544–2558. https://doi.org/10.1002/asi.21416
35. Morstatter F, Pfeffer J, Liu H, Carley KM. Is the sample good enough? comparing data from twitter's streaming API with twitter's firehose; 2013. arXiv preprint arXiv:1306.5204.
36. Tang H, Zhao X, Ren Y. A multilayer recognition model for twitter user geolocation. Wireless Networks. 2019; p. 1–6. https://doi.org/10.1007/s11276-018-01897-1
37. Fleiss JL. Measuring nominal scale agreement among many raters. Psychological bulletin. 1971; 76 (5):378. https://doi.org/10.1037/h0031619
38. Sato K, Young C, Patterson D. An in-depth look at Google's first Tensor Processing Unit (TPU). Google Cloud Big Data and Machine Learning Blog. 2017; 12.
39. Ljubešic N, Fišer D. A Global Analysis of Emoji Usage. ACL 2016. 2016; p. 82.
40. blog btrax com. How Americans and the Japanese Use Emoji Differently; 2015. https://blog.btrax.com/how-americans-and-the-japanese-use-emoji-differently/.
41. Hardeniya N. NLTK essentials. Packt Publishing Ltd; 2015.
42. Hanneke S, Roth D. Iterative Labeling for Semi-Supervised Learning. Urbana, IL, USA: University of Illinois; 2004.
43. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. Journal of the royal statistical society Series B (methodological). 1977; p. 1–38. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x
44. Lee G, Scott C. EM algorithms for multivariate Gaussian mixture models with truncated and censored data. Computational Statistics & Data Analysis. 2012; 56(9):2816–2829. https://doi.org/10.1016/j.csda.2012.03.003
45. Cohen I, Huang TS. Semisupervised learning of classifiers with application to human-computer interaction. University of Illinois at Urbana-Champaign, Champaign, IL. 2003;.
46. Pierce D, Cardie C. Limitations of co-training for natural language learning from large datasets. In: Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing; 2001. p. 1–9.
47. Blum A, Mitchell T. Combining labeled and unlabeled data with co-training. In: Proceedings of the eleventh annual conference on Computational learning theory. ACM; 1998. p. 92–100.
48. Abdel-Hamid NB, ElGhamrawy S, El Desouky A, Arafat H. A Dynamic Spark-based Classification Framework for Imbalanced Big Data. Journal of Grid Computing. 2018; 16(4):607–626. https://doi.org/10.1007/s10723-018-9465-z
49. Powers DMW. Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation. Journal of Machine Learning Technologies. 2011;2(1):37–63.

50. Valverde-Albacete FJ, Peláez-Moreno C. 100% Classification Accuracy Considered Harmful: The Normalized Information Transfer Factor Explains the Accuracy Paradox. PLoS ONE. 2014; 9(1):1–10. https://doi.org/10.1371/journal.pone.0084217

51. Bruckhaus T. The business impact of predictive analytics. Knowledge discovery and data mining: Challenges and realities. 2007; p. 114–138.

52. Hripcsak G, Rothschild AS. Agreement, the F-measure, and reliability in information retrieval. Journal of the American Medical Informatics Association. 2005; 12(3):296–298. https://doi.org/10.1197/jamia.M1733 PMID: 15684123

53. Litkowski K. Feature Ablation for Preposition Disambiguation. Damascus, MD, USA: CL Research; 2016.

54. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980. 2014;.

55. Nishida K, Banno R, Fujimura K, Hoshide T. Tweet classification by data compression. In: Proceedings of the 2011 international workshop on DETecting and Exploiting Cultural diversiTy on the social web. ACM; 2011. p. 29–34.

56. Yerva SR, Miklós Z, Aberer K. What have fruits to do with technology?: the case of orange, blackberry and apple. In: Proceedings of the International Conference on Web Intelligence, Mining and Semantics. ACM; 2011. p. 48.

57. Read J. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In: Proceedings of the ACL student research workshop. Association for Computational Linguistics; 2005. p. 43–48.

58. Rao D, Yarowsky D. Ranking and semi-supervised classification on large scale graphs using map-reduce. In: Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing. Association for Computational Linguistics; 2009. p. 58–65.

59. Yong R, Nobuhiro K, Yoshinaga N, Kitsuregawa M. Sentiment classification in under-resourced languages using graph-based semi-supervised learning methods. IEICE TRANSACTIONS on Information and Systems. 2014; 97(4):790–797.

60. Johnson C, Shukla P, Shukla S. On classifying the political sentiment of tweets; 2012. http://www.cs.utexas.edu/~cjohnson/TwitterSentimentAnalysis.pdf.

61. Consortium TU. Unicode Emoji; 2017. http://unicode.org/emoji/.