Using Twitter To Generate Signals For Syndromic Surveillance Systems - The Use Of Semi-Supervised Classification To Identify Relevant Tweets

Oduwa Edo-Osagie[1], Gillian Smith[2], 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 Field Epidemiology Service, National Infection Service, Public Health England, Birmingham, United Kingdom

Abstract

In this paper we investigate the use of Twitter data to deliver signals that may be used in the context of syndromic surveillance. We exemplify our work by focusing on detecting a signal for a specific syndrome, asthma, or an incident of air pollution that may result in increased respiratory problems. We begin by outlining the method for data collection from Twitter that may work in real time using the Twitter streaming API. We perform analysis and pre-processing of the collected data to eliminate noise. We focus on preprocessing operations to eliminate duplicate or irrelevant tweets (e.g. those relating to news items or authored by spam accounts) or tweets that cannot be located to a specific location (in this case the UK). We note that many of the tweets collected using selected keywords may not be relevant because they represent chatter, or talk of awareness instead of suffering a particular condition, as other authors have found. We therefore 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 the wealth of Twitter data collected without needing to label it all. We propose the use of emojis and other special features 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 bag of words approach. Also, Keyword analysis shows that keyword selection is a difficult problem, as the vernacular used in twitter may involve colloquial terms which may not be immediately obvious. We put forward a methodology for collecting and monitoring tweets for the purpose of syndromic surveillance. Our combination of pre-processing and classification algorithms will enable us to build a system to collect relevant tweets for a specific syndrome.

1 Introduction

Surveillance, described by the World Health Organisation (WHO) as "the cornerstone of public health security" [76], 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, however, 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 epidemics earlier than traditional surveillance and is also aimed at monitoring outbreak distribution and spread and characterising affected populations [15]. The fundamental objective of syndromic surveillance according to Henning [34] is to identify illness clusters early, before diagnoses are confirmed and reported to public health agencies, and to mobilize a rapid response, thereby reducing morbidity and mortality. Syndromic surveillance systems rely on the monitoring of health data, for example, reporting of specific symptoms associated with a given syndrome to sentinel Emergency Departments, although they can also use other data sources (e.g. over-the-counter prescription sales or school absenteeism). Given the expanding access to communications and technology in developing countries, it is becoming more feasible to implement syndromic surveillance systems in such settings too and some early examples in Indonesia and Peru have indicated reasons for optimism [15]. In particular, some advantages of such systems may involve the identification of new or changing diseases or the detection of unusual morbidity trends.

Hence syndromic surveillance refers to the monitoring of clinical syndromes that have significant impact on public health, medical resource allocation or health policy. It involves collecting, analysing and interpreting (often in an automated or partially automated manner) health-related data to provide an early warning to public health threats, which might require public health action. Syndromes such as seasonal influenza, asthma, scarlet fever and others are monitored through various diagnostic indicators. Other incidents that can affect public health such as air pollution, flooding, heatwaves, etc, can also be monitored. Data may be analysed on a daily basis with routine reports being produced by such systems, often weekly. Statistical analysis is applied to the different indicators for each specific syndrome. Statistically significant aberrations or "signals" are investigated to determine their public health importance. If deemed appropriate by human experts, alerts are converted to alarms and appropriate action is taken.

The use of social media such as Twitter data for disease surveillance has been gaining momentum (e.g. [1, 11, 22, 24, 26, 29, 41]) and may be able to capture a population that do not seek medical help via the more established means. Furthermore, in developing countries where access to medical experts may be restricted but where use of mobile phones and social media is becoming more common, it is possible that such data may provide insights into the health of the population that are not otherwise available, alert to outbreaks and also provide a platform to spread information to combat such outbreaks. It is in this light that we develop our work. Hence, our aim is to establish if social media data and specifically Twitter data can be used in the context of syndromic surveillance in order to generate or contribute to alarms for specific syndromes, and what the timeliness and accuracy of a Twitter signal may be in such a context.

Our first objective is to develop a reliable means of extracting a signal from the Twitter stream for different syndromes 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 tweets are considered "relevant" for our purpose of syndromic surveillance. This requires some text classification because other authors [8, 41] 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, and is therefore irrelevant to the syndromic surveillance effort. Such irrelevant content may greatly magnify the signal and lead to incorrect results and over estimation as was observed by Copeland et al. [19] in relation to the Google Flu Trends and Google Dengue Trends, two disease surveillance tools build on internet query data from Google.

The analysis of Twitter data is becoming topical in many different scenarios so our

text mining effort to identify "relevant" tweets in the context of syndromic surveillance has much wider applicability in other scenarios in which Twitter data analysis is important. We investigate features that may be relevant for text mining of tweets, e.g. emojis and emoticons, and also apply novel semi-supervised classification algorithms which can take advantage of unlabelled data to produce good classifications without requiring the substantial effort of manually labelling a large set of tweets to train the classifiers. Our paper therefore makes an important contribution to the text mining analysis of Twitter data.

Our paper is structured as follows: section 2 discusses relevant work in the context of our objectives that we can learn from; section 3 presents the methods we use to analyse twitter data including cleaning and pre-processing operation, the definition of features and the classification algorithms that we will apply; section 4 presents our results which we discuss in section 5. We present our concluding remarks in section 6.

2 Related work

In a survey carried out in 2015, Charles-Smith et al. [11] 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 [9] to the prediction of an individual's health and when they may fall ill [67]. Because syndromic surveillance is quite a broad subject, there have been a number of studies looking at different aspects of the problem from different angles. Those most relevant to us are those concerning *epidemiology* and *the use of social media*.

The most commonly studied disease is influenza or influenza like illness (ILI) [20]. As a communicable disease, one advantage of studying influenza is that many countries including the USA and the UK keep records of infection rates. Such counts can act as "ground truth" when applying machine learning algorithms to construct influenza prediction models.

De Quincey and Kostkova [22] 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. [1] 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. Chen et al. [13, 14] managed to distinguish different biological phases of the influenza from the content of tweets using a temporal topic model. Many of the published work on tracking influenza or ILI is based on the USA where the volume of tweets is greatest. Some work has also been done in the UK [42, 43].

Diaz-Aviles and Stewart [24] also demonstrated the potential of Twitter for syndromic surveillance during an E. Coli outbreak in Germany. They showed that monitoring tweets for outbreak signs generated an alarm a day ahead of well established monitoring systems such as MedISys. The use social media data has also appeared in the context of non-communicable diseases such as asthma: Ram et al. [61] developed a model to predict asthma emergency department visits, using twitter data along with other sources. It has also been used to unearth instances of food born illnesses and to ask those affected to report to the authorities [33]. Cancer has also been studied using

Twitter data [46].

In our work, we begin by focusing on syndromes that are monitored using syndromic surveillance but have received less attention than influenza. In particular, we begin by looking at asthma and air pollution incidents as an exemplar. Our effort is to collect relevant tweets for our given syndrome - asthma. We proceed to the initial data collection by using a set of possibly related keywords. However, 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). For example, in relation to keyword "asthma" we encounter tweets with content 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, walking up 366 stairs. Other than that fine thanks" or "why is my asthma so bad today?" expresses a person currently affected and we would like to consider as relevant. This then becomes a text (or tweet) classification problem.

A number of papers have looked at the tweet classification problem using supervised learning for different applications. Sriram et al. [69] 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 such as the presence of abbreviated words. Dilrukshi et al. [25] applied a Support Vector Machine (SVM) to classify tweets to different news categories. Nishida et al. [54] tried to classify tweets as being related to a given interesting topic. They proposed a method called Compression-based Tweet Classification (CTC) that uses data compressing for classifying a tweet. This method evaluates the compressibility of every tweet (using compression algorithms such as gzip) and works with the assumption that if a tweet is compressible then it is an interesting tweet and uninteresting otherwise. Yerva et al. [79] classified tweets using variations of the Naive Bayes classifier to identify whether they were related to a company or not. An example of this would be determining whether a tweet containing the word *apple* was about the fruit or the technology company. Read [64] used and showed the value of emoticons in sentiment classification.

The most relevant work in the context of tweet classification is that of Dredze and his colleagues [8, 41] as they used Twitter data to investigate influenza surveillance. They argue that for accurate social media surveillance it is essential to distinguish between tweets that report infection and those that express concern or awareness. They first set out to find the type of features that may help with such classification task [8] and they try out many type of features including word classes (e.g. infection, possession, concern, ...), stylometry (e.g. emoticons, hashtags, ...) and part of speech templates (e.g. pairings of first pronoun and last noun, phases that begin with a verb, ...). They found that simple n-grams have the most power as features but they can be enhanced with word class features. In further work [41] they also found that filtering tweets in a multi stage process as relevant/irrelevant to health and then relevant/irrelevant to influenza and indicative of actual infection was an effective approach. A normalised prevalence measure obtained from the filtered tweets compared favourably to the CDC's ILI rate. In comparison, the weekly number of tweets containing influenza keywords provided a weaker correlation. They concluded that filtering tweets according to relevance demonstrates significant improvements and is less sensitive to Twitter chatter. They also noted that their system could be used for real time analysis in addition to retrospective analysis, which is an important consideration for our work.

One problem with the previous approach is that it relies 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. For example Dredze and collegues [8, 41] relied on Amazon Mechanical Turk, a crowdsourcing Internet marketplace, to perform the labelling. They had over 3 billion tweets available but only used 11,990 labelled tweets to build their systems. Since the syndromes/events we wish to study are often rare and may not be mentioned frequently in a Twitter feed, we wish to use as many tweets as possible to build our models. To this effect semi-supervised classification approaches try to produce models using a small set of labelled data but also taking into account the larger set of unlabelled data.

Zhang et al. [83] investigated the semi-supervised classification of tweets for organization name disambiguation, a problem previously tackled with a supervised approach by Yerva et al. [79]. Zhang et al. compared Label Propagation and Transductive Support Vector Machines (TSVMs): both methods utilise unlabelled data in the classifier. Label Propagation (LP) is a graph-based semi- supervised algorithm which uses pair-wise similarities between instances to extend labelling to unlabelled instances and thereby enhance classification accuracy. TSVMs, are an extension to Support Vector Machines which use unlabelled data to find the maximum margin separation hyperplane that separates the positive and negative instances of the labeled data and the unlabelled data. The basic idea of TSVM is to seek a decision surface away from the dense regions of unlabelled data. It belongs to a group of semi-supervised methods based on low-density separation. TSVMs were found to perform better than fully supervised methods for the problem of named entity disambiguation.

Another set of techniques that are categorised as heuristic semi-supervised methods are self-training and co-training [71]. In self-training [78], first a supervised algorithm is trained based on small labelled data only and then it is applied to the unlabelled data to extend the labelling using only the most confident predictions as new labels at each stage. Co-training [6] is an extension of self-training in which two classifiers are trained on different feature sets, ideally independent, with each classifier then providing their most confident labels for the other feature set. A further extension of this that is sometimes used is multi-view training which does not require explicit feature splits or the iterative mutual training procedure.

Some work in semi-supervised tweet classification make use of graph-based methods to propagate labels to unlabelled data. A number of papers have investigated the use of the graph-based label propagation (LP) for text classification [63, 80]. Johnson et al. [37] used label propagation for public political opinion mining on Twitter. They collected tweets containing the (case-insensitive) string "*Obama*" and classified their political sentiments. They compared rule-based, supervised and semi-supervised methods. They found that the semi-supervised label propagation technique worked best yielding an accuracy of 78%. The graph-based approaches require the computation of similarities between instances so the use of a suitable similarity measure is of utmost importance and is key to the success of the algorithm. Unfortunately, similarity measures are usually heavily dependent on the domain and task of interest.

A number of papers have looked at using semi-supervised learning for sentiment analysis, and in particular self-training [4, 84]. Baugh [3] proposed a hierarchical classification system with self-training incorporated into it . His goal was to classify tweets as *positive*, *negative* or *neutral*. The hierarchical classification system consists of two binary classifiers. First, a subjectivity classifier, which classified tweets as *subjective* or *objective* and a polarity classifier which classified tweets as *positive* or *negative*. If the output of the first level is subjective, then the second-level polarity classifier decides if the instance is positive or negative. Otherwise, the instance is classified as neutral. The unlabelled tweets were classified one at a time. If the system classified the tweet as subjective, it was used to train the polarity classifier. This approach achieved an accuracy of 61% and an F_1 score of 0.54. Liu et al. [49] proposed a semi-supervised

framework for sentiment classification in tweets that was based on co-training. They converted tweets into two kinds of distinct features - textual and non-textual. Two Random Forest (RF) classifiers were trained with the same labeled data but one with textual features and the other with non-textual features.

Some work has also recently appeared using semi-supervised deep learning and this may be a promising avenue. Convolutional Neural Networks (CNNs) [44] are very popular neural network models in the context of image analysis because they can discover features as part of the learning process. Convolution layers are used, for example in the context of 2D images, with each layer responding to a small region of input and using convolution operators to produce different features, those operators are known as filters or kernels. A CNN can then learn about the features that matter in the image analysis problem, e.g texture, shape, etc. In text mining, CNNs are also gaining attention as they can also be applied to learn word embeddings from text regions and other features of interest. Johnson et al. [38] proposed a general semi-supervised framework for document classification using Convolutional Neural Networks. Lee et al. [47] applied this framework to the classification of tweets as being related to adverse drug effects or not. Lee et al. use a semi-supervised CNN model which leverages different types of unlabelled data to construct robust tweet classification models. The downside to using neural networks is that in order to build a truly powerful one, a lot of computing power is required. Unfortunately, this may not be available when implementing syndromic surveillance in developing countries.

In this paper, we build classification models for tweets based on the relevance in the context of a specific syndrome/event. As part of our feature investigation which is an important part of text classification, we experiment with different types of features extracted or constructed from our tweet corpus, taking into consideration suggestions from previous work. While we found some literature making use of emoticons in text classification, we did not find any employing emojis. We consider and investigate emojis in tweet classification, and show their worth in a syndromic surveillance context. We compare both supervised and semi-supervised approaches to text classification (particularly self-training and graph-based methods) to understand how to include more of the data that we can collect.

3 Methodology

Here, we describe the methodology for our study. We first collected tweets over multiple periods using the Twitter Streaming API. We want to distinguish tweets that present the view of specific users and not for example news items, and which can be interpreted as a user describing some symptoms or describing the presence of the syndrome under consideration. Consequently, we discuss the data collection, pre-processing and analysis of the collected tweets in order to establish a relevant signal for a given syndrome. We start by narrowing our efforts to asthma and air pollution.

3.1 Data Collection and Pre-processing

Tweets were collected over multiple periods which lasted between 6-8 weeks each and a final longer period. It was important to collect data over more than one period because we are concerned with outbreaks, which in real life may occur to different degrees at different times [51,53,72]. Furthermore, since we are collecting data as it occurs we have no guarantee that the period chosen is going to contain an event so repeated data collection gives us a better chance to observe something happening. Another motivation for the diversity in collection periods is that it could help us establish if there are changes in the use of Twitter as well as in the language used on Twitter over time. We

collected 10 million tweets over three collection periods - September 2015 till November 2015; June 2016 till August 2016; and January 2017 till July 2017. We started with an Autumn period, followed by a summer period and then collected winter through to Spring.

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. The actual percentage of total tweets received from the streaming API varies based on the filters and specific parameters of the stream and the current traffic. 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 [52].

The streaming API has a number of parameters that can be used to restrict the Tweets obtained. We wanted to concern ourselves only with Tweets in the English language since expanding the analysis to more languages would have brought many additional complications, hence we set "language=en" to place the restriction. It is also possible to restrict the data collection to tweets with specific terms. This was important to us since we wanted to concern ourselves with Tweets that may be relevant to a particular syndrome. For this, in conjunction with experts from PHE, we created a set of terms that may be connected to the specific syndrome under scrutiny, in this case asthma and air pollution. We then expanded on this initial list using various synonyms from regular thesauri as well as from the urban dictionary as that may capture some of the more colloquial language used in Twitter. Examples of our keywords are "asthma", "wheezing", "couldn't breathe" etc. A full list of terms used is provided in the appendix. The restriction on specific terms can be implemented in the Twitter API by by using the parameter "track" followed by specific comma-separated list of phrases which will be used to determine the Tweets that will be delivered on 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, text for hashtags, and screen_name for user.

We collected 10 million tweets obtained over three collection periods - September 2015 till November 2015; June 2016 till August 2016; and January 2017 till July 2017. The general characteristics of the collected tweets were analysed and are reported in table 1.

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

Table 1. Information on the data corpus collected before cleaning

The anatomy of a Tweet is well presented in the Status Map in Fig 1 which we reproduce here by courtesy of the author. There are a number of attributes that are associated with a Tweet and would be available to our analysis. We did not consider them all necessary so we collected the attributes that could help us in our task, 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 well suited to the storage of documents.

Fig 1. Map of a Tweet from the Twitter API

3.1.1 Location Filtering

Because our project is concerned with syndromic surveillance systems in the United Kingdom, we wanted to exclude tweets originating from outside the UK. Doing this will give a more realistic signal when the next stage comes. However, inferring the location of Twitter users is notoriously difficult. According to [25], although users of Twitter have the option to disclose their city-level location, fewer than 14% do so and up to 30% may give false or fictitious locations [54]. Also, although Twitter allows for mobile devices to include as meta-data the location coordinates, which would give a very accurate estimate of current location, less than 0.5% of Twitter users turn on the location function [25] owing to concerns over privacy.

A number of approaches have been used to infer location for Twitter data, which include using text mining of the tweet itself, using the friends' network, URLs, Geotags, etc. They are discussed in [25]. For our current purpose, which is relevance analysis in the context of a particular syndrome, accurate location is of relative importance since the type of tweets that we want to analyse to identify relevance could come from different locations and still fit the relevance model. Hence, in this work we do not concern ourselves with accurate location information.

The *time_zone*, *coordinates* and *place* attributes, which we collected, can help in the geolocation of a tweet but are not always present. The time_zone attribute can be optionally set by a Twitter user as part of their Twitter profile and may not be accurate, or may represent the user's home location but not that where a specific tweet originated. The place attribute is also optional for the user to set with the same caveats. The coordinates attribute is available when a user chooses to publish their location at the time of tweeting from a mobile device. The coordinates attribute is the most reliable and the one that is most likely to represent location at the time of tweeting but unfortunately, only a very small percentage of tweets expose their coordinates. As is shown in figure 2, less than 1% of the tweets we collected contained coordinates. The place attribute was also similarly rare with only around 2% of the collected tweets containing the place attribute. In contrast, the timezone was made available by users more than 50% of the time. As a quick check on accuracy, we compared the provided coordinates and time_zone when they were both available. We found that the timezone only matched with the coordinates 11% of the time while the place matched with the coordinates 84% of the time. Therefore, while the timezone attribute is more readily available, it is not reliable.

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%

Table 2. Availability of geolocation attribute in collected Twitter Data

Given our considerations about location, and in the interest of using a sufficiently large dataset for tweet classification, we decided to employ all three geolocation fields in our location filtering. Hence we only filtered out tweets that did not have a UK timezone, a place in the UK or coordinates in the UK.

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

Table 3. Information on the data corpus collected after cleaning

3.1.2 Cleaning the Data

After the data was collected, we focused on the suitability of the data for our purpose. In particular we noticed the following data quality problems that needed to be solved by developing suitable pre-processing algorithms:

- **Retweets:** A Retweet (sometimes abbreviated to RT) is a re-post of a tweet. The retweet feature allows users to quickly share an existing post - which could have been made by them or some other user - while also attributing creation of the post to the original poster [27]. If some user finds a tweet relatable, they may decide to retweet it. This could lead to duplication in our dataset and could result in the raising of false signals. Retweets are always of two forms: either they contain the original post with the username of the original poster in quotes; or they just contain the original post along with the username of the original poster with the word "RT" preceding it or following it. Tweets in our dataset which satisfied either of these criteria were removed. This step removed 34% of the tweets filtered by location as originating from the UK.

- **Duplicate Tweets:** Even when not immediately identifiable as retweets, tweets containing exactly the same text with maybe different URLs appended may be duplicates and may give rise to amplified signals. After removing retweets, we still noticed a significant number of duplicate tweets of this type, many associated with news items or blogs. Those were dealt with by removing tweets that contained exactly the same text, once URLs where discounted. However, it is worth noting that it is possible for different people to express the same sentiment with the same or very similar words, as twitter encourages very short communication (e.g. "asthma bad" or "can't breathe").

  We also reasoned that a user expressing a condition should only be counted once per day for the purposes of syndromic surveillance, so we removed multiple tweets for the same user on a given day. The removal of all these tweets removed 10% of the tweets remaining from the Retweet removal.

- **URLs**: Some tweets contain web links to other pages. Usually these links point to pages which explain the content of the tweet in more detail. However, tweets containing URLs often constitute external reports (such as news articles and blog posts) as opposed to individual reports (ie. an individual expressing concern or opinion). Because of this, we removed URLs and replaced them with the token "<URL>". This not only allowed us to avoid introducing noise to our system, but also helped us recognize tweets that originate from individual user accounts as opposed to news and spam accounts.

- **Spambots and Articles:** A "bot" is the term used for when a computer program interacts with web services that are intended for human use. It is possible to create a Twitter account and then through the use of the Twitter API

automatically post tweets, follow other users and even send direct messages to other users. Tweets made by such accounts are not genuine individual sentiments and so are not of any relevance to our endeavour. News accounts and web blog accounts which usually post links to news and blog articles fall into a similar category and are not relevant in the context of our investigation. Tweets made from such accounts were removed from the dataset.

To recognise those, we looked for tweets containing a URL, and then we check if the user had a very skewed following/follower ratio [75] and a relatively high number of tweets as those characteristics may be associated with spam accounts. We trained and employed a K-Nearest-Neighbour (KNN) [2] classifier to automatically determine which tweets were not posted by individuals. The removal of tweets not posted by individual accounts removed 1% of the tweets remaining from the removal of duplicates.

An overview of the data after cleaning is shown in table 3.

3.1.3 Labeling

10% of the cleaned tweets were manually labeled as "relevant" or "not relevant". A tweet was labeled 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 labeled the 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 the third person who made the decision on which label to use. 23% of the labeled tweets were labeled as "relevant" while 77% were labeled as "irrelevant".

3.1.4 Features

Although it is possible to use any sequence of letters or language tokens to represent text (i.e. n-grams), words are often identified and used in text mining. Word n-grams have been used successfully in language modelling and speech recognition [40, 57, 81]. Words are identified after a process of tokenisation and can then be used to represent a document by their presence or absence without trying to retain any information on the ordering of words or their relationship to one another. That approach is called "bag of words" and despite its relative simplicity can work well in many text mining scenarios [39]. However, some authors [5, 41, 79] have argued that more complex features will dramatically decrease the feature space while leading to better classification performance.

We began by employing a bag of words model with word unigrams. We applied a weighting scheme to the bag-of-words features using the tf-idf (Term Frequency-Inverse Document Frequency) weighting scheme [62]. We chose to do this because a binary representation or a simple word count (or term frequency count) cannot guarantee that the importance of the words is recognised. For example, a word that occurs frequently in many tweets may not have a high discriminative power. Tf-idf overcomes this problem because it incorporates the inverse document frequency which takes account of local and global parameters by considering a word in the context of the set of documents it belongs to [68].

Classification of tweets may be more challenging than classification of standard documents as tweets are very short and in our scenarios the classes may share common vocabularies. That is, both relevant and irrelevant tweets could contain the same words. For example, the tweet "Went out today and I couldn't breathe" is relevant, while the tweet "Justins performance, I couldn't breathe!' is not but they both contain

overlapping keywords. Twitter also has specific language and styles that people use. Hence we investigate additional features that could help in tweet classification.

3.1.4.1 ***Hashtags, emojis and emoticons*** After investigating the peculiarities of Twitter messages, we found that *hashtags*, *emojis and emoticons* are the most promising:

- A hashtag is a token made entirely of alphanumeric text except for the "#" symbol at the beginning of the token. It is used to add additional context to a tweet or categorize and tag tweets. Hashtags serve as a means of creating "groupings" on Twitter, helping users to emphasize or group important information. This also makes it possible for users to easily retrieve information on a topic [26]. We chose not to remove hashtags in our data as we thought they would give our system further insight on the nature or intent of the tweets. In fact, studies have shown that hashtags can be important indicators for improving the accuracy of filtering [21, 30, 65].

- 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 [74]. :-) 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. While emoticons are made up of ASCII characters and emojis are made up of unicode characters they are both generally used for the same purpose and can be used interchangeably. However, emojis have seen a recent surge in popularity, presumably due to the fact that emojis provide colorful graphical representations as well as a richer selection of symbols. In fact, as table 1 shows, a quarter of the tweets collected contained emojis. This was a much higher proportion than other secondary element such as URLs and hashtags. A further advantage is that they 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 a good indication that the tweet is not relevant to our context of syndromic surveillance.

3.1.4.2 Word Classes Following from work by Bergsma et al. [5], we also extended our feature set with further syntactical features in order to make up for the shortcomings n-grams may present when applied to Twitter data. Word classes are labels that Lamb et al. [41] found useful in the context of analysing tweets to categorize 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 Because we applied lemmatisation and stemming, we did not include multiple inflections of the words in our word classes.

3.1.4.3 Positive and Negative Word Counts: We constructed two dictionaries of positive and negative words respectively. These dictionaries are shown in the appendix. This feature computes for every tweet, the number of positive words and

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

Table 4. Our list of word classes

negative words it contains. It uses our hand-crafted dictionary as a reference point for which words are positive or negative. Words that do not appear in either of our dictionaries are not counted. This is because our dictionaries are concerned specifically with positive and negative words which are likely to appear in the context of health. In short, this feature produces two figures for every tweet - a positive count and negative count. Our hypothesis is that tweets which contain more negative words than positive words are likely to be relevant in the sense that they are an individual reporting symptoms or expressing concern over a syndrome. It is then the duty of our learning algorithm to learn a matching between ratios of positive to negative counts to tweet relevance.

3.1.4.4 Indicates Personal Asthma Report: This feature tests for the presence of the word "asthma" in close proximity together with a personal pronoun. In particular, we check for close proximity with "i'm", "im", "my", "i", "am" and "me". The aim of this feature is to determine when a tweet has a user reporting concern over their condition and distinguish this from a tweet where a tweet just happens to mention asthma. For some perspective, only 44% of the tweets in the dataset contain the word "asthma" and 35% of these tweets are relevant. When applied to the dataset, the feature had a value of "True" in 8% of the dataset and of this 8%, 53% were relevant. As such, this feature shows that it is better to check that a tweet contains "asthma" used in a particular way than to just check whether a tweet contains "asthma".

3.1.4.5 Contains "Asthma-Verb" Conjugate: This is a very specific feature to our syndrome. A verb conjugate is a form of a verb derived from its base form according to the rules of grammar, due to a change in person, tense, number or other grammatical categories [56]. *Contains "Asthma-Verb" Conjugate* is a textual and syntactic binary feature which is concerned with whether or not there is a verb form appearing with the word asthma (or its symptoms and keywords) as its object. For example the tweet "I can't believe I'm only just recovering from my asthma attack" contains the word *asthma* used as the object of the verb *recover*, while the tweet "People with asthma shouldn't come to school" sees it being used as part of the subject of the verb. The WordNet [55] interface of NLTK (Natural Language ToolKit) was used to perform Part of Speech (POS) tagging in order to extract these type of features. NLTK is a platform for building Python programs to work with human language data. The advantage of using syntactical analysis is that it may help us to produce more discriminatory features, since the same words may appear in both classes so presence or absence of a word may not be discriminatory whereas its use in the syntax may be. As shown on table 5, 20.9% of the tweets contained such features whereas 79.1% of the

tweets did not. We found that of the instances with a negative value for this feature, 81.1% were found to belong to the "not relevant" class. This showed that absence of this feature seems related to non-relevant tweets.

3.1.4.6 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 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 youth 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.

3.1.4.7 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 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 three. 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.

Feature	Value Distribution		Class Distribution	
			Relevant	Not Relevant
Contains Asthma-Verb Conjugate	TRUE	20.9%	45.6%	54.4%
	FALSE	79.1%	18.9%	81.1%
Indicates Personal Asthma Report	TRUE	8.3%	53.6%	46.4%
	FALSE	91.7%	20.1%	79.9%
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%

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

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 dimensionalities of our custom features C (*Word Classes*, *Positive and Negative Word Counts*, *Contains Asthma-Verb Conjugate*, *Indicates Personal Asthma Report*, *Denotes Laughter* and *Negative Emojis/Emoticons*). This gives us

$$T_i = \{t_{i1}, t_{i2}, ..., t_{in}\} \cup \{C_i^1\} \cup \{C_i^2\} \cup \{C_i^3\} \cup \{C_i^4\} \cup \{C_i^5\}$$

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 (ie. binary, continuous and categorical).

3.2 Text Classification

Once labeled data is available, a classification algorithm for text can be used to produce a model 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 and Support Vector Machines. We used the Python implementations found in the Natural Language ToolKit (NLTK) and Sci-Kit Learn. The set of fully supervised classifiers that were tested included linear and non-linear models as well as parametric and non-parametric models.

The Naive Bayes classifier is a linear probabilistic model based on the popular Bayes probability theorem [66]. Naive Bayes classifiers are often used in text classification scenarios and are known to be well suited to such problems [82]. These models work best when features are mutually independent. Textual data satisfies this constraint as, for the most part, the occurrence of one word does not affect the occurrence of any other word. Decision Trees are a non-parametric model. Their goal is to predict the value of a target variable (or class membership) by learning simple decision rules inferred from the data features. The trees produced are easy to visualize and interpret, so they allow for an easy high-level understanding of problems.

Decision trees very quickly create biased trees if some classes dominate the dataset. For this reason, balancing the dataset becomes a necessity when using Decision Trees. We applied a variation of the ID3 decision tree algorithm found in NLTK.

Logistic Regression is a probabilistic, non-parametric model. Logisitic Regression is sensitive to outliers and is best-suited to uniform data. It assumes that there is no high intercorrelations among the predictors or feature vectors. This can be assessed by using a correllation matrix. Tabachnick and Fidell [70] suggest that as long as correlation coefficients among independent variables are less than 0.90 the assumption is met.

Support Vector Machines (SVMs) are non-parametric models and can be linear or non-linear depending on the kernel used. We made use of a non-linear Support Vector Machine using the Radial Basis Function as our kernel function. SVMs aim to learn the optimum hyperplane separating instances of classes in the dataset into distinct regions. SVMs are efficient in high-dimensional spaces, however, if the number of training instances is less than the dimensionality of the problem space, their performance is very poor. Also, SVMs do not directly provide probability estimates with their classifications but these can be obtained using cross-validation (which could be expensive). SVMs have been tested and proven to be well-suited to text classification [36]. We experimented with each of these algorithms, applying them to our data and investigating their optimum performance.

We then implemented a semi-supervised approach based on Naive Bayes. 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. Naive Bayes models are more suited to small to medium sized datasets [28]. This was important to us because we start with a relatively modest number of labeled tweets. 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

Due to the relatively limited number of labeled instances in our data set, we decided to take a semi-supervised approach to learning. More specifically, we adopted an *Iterative Labelling Algorithm* for semi-supervised learning [31]. Iterative labelling algorithms are closely related to and are essentially extensions of the Expectation-Maximization (EM) algorithm put forward by Dempster et al. [23]. The iterative labelling algorithm is a sort of *meta-algorithm* which uses a data set S of labeled 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 which make up its main steps:

Choose-Label-Set(U, L, A) takes in U, the set of unlabelled instances; L, the set of labelled instances; A, a supervised learning algorithm; and returns a new set, R, of unlabelled examples to be labelled.

Assign-Labels(R, S, A) takes in R, the set of unlabelled instances to be labelled; S, the set of labelled instances; and a supervised learning algorithm, A, and returns the set R, but now with labels. The purpose of this function is to generate labels for the instances selected by ***Choose-Label-Set***(U, L, A).

Stopping-Condition(S, S') takes in the original data set, S, and $S' \leftrightarrow R \cup L$ and maps to $D \in \{True, False\}$. It dictates when the algorithm should stop iterating.

Algorithm Iterative Labeling Algorithm

```
function IterativeLabeling(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 Naive Bayes classifier and chose to use the trained classifier's predictions to label unlabeled instances in the *Assign-Labels* function. Along with the class of an applied instance, we also compute the model's confidence in its classification. Our algorithm, inspired by the Truncated Expectation-Maximization (EM) [45], 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.

It has been found that confidence-based iterative labelling algorithms have a problem of tending 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 [45]. In addition to this, it has been proven that in certain situations, many semi-supervised learning algorithms can significantly degrade the performance relative to strictly supervised learning. Cohen and Huan [16] discuss situations in which it would be better to completely discard the unlabelled data than to use self-training for semi-supervised learning. Pierce and Cardie [59] explain that after a number of iterations, the supervised learning algorithm's past mistakes can create a snowball effect which then has a negative impact on the accuracy of the system.

3.2.2 Co-training model

To address the problems of self-training algorithm, we make some modifications to the self-training algorithm by taking some ideas from *co-training*([7], [17]). In this model,

each tweet can be represented as a feature vector x. We distinguish between two variations of feature vectors. The first kind is a concatenation of our *Bag-of-Words*, *Word Classes*, *Indicates Personal Asthma Report*, *Contains "Asthma-Verb" Conjugate*, *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 *Bag-of-Words*, *Positive and Negative Word Counts*, *Indicates Personal Asthma Report*, *Contains "Asthma-Verb" Conjugate*, *Denotes Laughter* and *Negative Emojis/Emoticons* features. We represent this feature space as X_2. X_1 is the **taxonomical** feature space as is characterised by its inclusion of the *Word Classes* feature while X_2 is 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 views of the dataset. Each tweet is then represented as a feature vector from each of these spaces such that $x = (x_1, x_2)$. Instead of just having the one classifier train itself like in the self-training approach, we maintain two separate classifiers trained on different views of the data. However, co-training makes the strong assumption that each type of the two views alone is sufficient for classification. It is important to note that for co-training to be effective, this assumption must be met. It is less clear to what extent it has to be met for co-training to yield good results. We explore this in more detail in section 5.

During the iterative labeling 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 labeled data is incorporated into each of the classifiers to update their hypotheses. Once the iterative labeling process is completed, the prior training examples for both classifiers as well as the newly labeled 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 labeled 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 [31].

4 Experiments and Results

4.1 Classification

We first employed and assessed a fully supervised approach. We had an initial set of manually labelled data contained 3,500 tweets. This contained 24.7% tweets that were labeled as relevant and 76.3% labeled as irrelevant. We divided the data into a 70:30 training and test split using random sampling. We decided to address the balancing issue when training our models because the classes were very imbalanced and this could lead to poor models that mostly focus on predicting the irrelevant class as this would still lead to high overall accuracy. Imbalanced data causes well known problems to classification models [12]. We initially tried both oversampling and undersampling techniques to create a balanced training dataset. We found no major difference as a result of either technique so opted for undersampling. The class distribution over the balanced training set had 47% of tweets as relevant and 53% as irrelevant.

Another important aspect of imbalanced data and of classification in general is having the right performance metric for assessment of classification model [60, 73]. Overall accuracy is a misleading measure 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 a surge of cases for a particular syndrome (e.g. asthma/air pollution incident) 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 or False Negative (FN) errors should have a higher cost and hence be minimised; we can have more tolerance of irrelevant tweets classified as relevant or False Negative (FP) errors. Those subtleties are well captured by alternative measures of model performance [10] such as Recall, the probability that a relevant tweet is identified by the model, defined as

$$Recall = \frac{TP}{TP + FN}, \tag{1}$$

and Precision, the probability that a tweet predicted as relevant is actually relevant

$$Precision = \frac{TP}{TP + FP}. \tag{2}$$

where TP and TN stand for True Positives and True Negatives respectively.

Precision and recall are often trading quantities. A measure that combines precision and recall is the harmonic mean of precision and recall, the traditional F-measure or balanced F-score [35] is defined as:

$$F = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3}$$

A variation of this, the $F2$ measure, may be more feasible for our purpose as it weighs recall higher than precision. When the formula for F measure has β set to 2 it becomes the $F2$ measure:

$$F_\beta = (1 + \beta^2) \times \frac{Precision \times Recall}{(\beta^2 \times Precision) + Recall}. \tag{4}$$

The best fully supervised approach was the Logistic Regression classifier, which achieved a prediction accuracy of **91.5%** and F_2 score of **0.764** on the test data while the best semi-supervised approach, which was the iterative labeling algorithm, achieved a predictive accuracy of **82.7%** with an F_2 score of **0.814** on the test data. The iterative labeling algorithm (*self-training*) was implemented as described in section 3.2.1 with and without the *co-training* modification also described in section 3.2.1. We tried both variants of the iterative labeling algorithm and report the results of the better variant which we found to be plain self-training. Table 6 shows a breakdown of values obtained for these metrics for different approaches on the test data.

To give a better understanding of how the different measures manage to balance the number of FP and FN. We also present the confusion matrices for both the best performing fully supervised and semi-supervised methods on the test data. These confusion matrices are shown in tables 7 and 8. We see that while the fully supervised approach yields a higher accuracy and precision, the semi-supervised approach yields a higher recall and F_2 score. 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 boasts a lower accuracy but is able to identify relevant tweets much better with a high recall value, while also being able to identify irrelevant tweets to a reasonable degree. While the F_1 score of the fully supervised approach is higher than that of the semi-supervised approach, remember that the F_1 score places equal weighting on precision and recall. The F_2 measure, which

places more weighting on recall than precision, shows the semi-supervised approaches to give better performance. This means that the semi-supervised algorithm outperforms the fully supervised algorithm for our purposes. This is an important finding because it means that when there is a shortage of labeled data, and fully supervised algorithms cannot be used, semi-supervised algorithms which are more robust to this shortage, can be used to gain performance comparable to or better than fully supervised algorithms.

Supervised Algorithms	**Precision**	**Recall**	F_1 **Score**	F_2 **Score**
Naive Bayes	0.636	0.804	0.710	0.764
Decision Tree	0.915	0.629	0.554	0.671
Logistic Regression	0.885	0.739	0.805	0.764
Support Vector Machine	0.864	0.722	0.787	0.747
Semi-supervised Algorithms	**Precision**	**Recall**	F_1 **Score**	F_2 **Score**
Iterative Labeling with Self-training	0.593	0.897	0.714	0.814
Iterative Labeling with Co-training	0.650	0.835	0.731	0.787
Label Propagation	0.602	0.654	0.627	0.643
Label Spreading	0.502	0.665	0.572	0.624

Table 6. Results of relevance classification on the test data

		Actual Response		
		True	False	Total
Predicted Response	True	TP (215)	FP (28)	243
	False	FN (76)	TN (883)	959
	Total	291	911	$N = 1202$

Table 7. Confusion matrix for logistic regression fully supervised classification on the test data

		Actual Response		
		True	False	Total
Predicted Response	True	TP (261)	FP (179)	440
	False	FN (30)	TN (732)	762
	Total	291	911	$N = 1202$

Table 8. Confusion matrix for the iterative labeling semi-supervised classification without co-training on the test data

Fig 2 shows how the performances of the semi-supervised systems change as they assimilate more and more unlabeled data. We plotted the precision and recall of the four semi-supervised methods as they were applied to the unlabeled data to increase the number of labels. The self-training and co-training variants followed opposite trends. For self-training, as more instances were assimilated, precision dropped while recall rose. The opposite happened for co-training: its precision rose while its recall dropped. Label propagation saw an increase in precision as it assimilated more unlabeled data. Unfortunately this was accompanied by a more or less equal decrease in recall. The precision of the Label Spreading approach remained roughly the same with slight fluctuations while its recall rose to be even higher than both the self-training and co-training variants of the iterative labeling algorithm. The precision and recall of the iterative labeling algorithm changes steadily and predictably as they assimilate more data. This is probably due to their conservative labeling approach where they do not

incorporate instances whose predictions they are not sufficiently confident about. As mirrored by table 6, the iterative labeling algorithms yield the best performance, with self-training being the superior option in our case due to its higher recall.

Fig 2. Graph of precision and recall on the test set as the semi-supervised algorithm incorporated the unlabeled data

The main issue with iterative labeling algorithms is that, because the classifiers are not perfect and do not have 100% accuracy, we cannot be sure that the unlabeled instances that they label for assimilation are always correct. This means that how well they initially perform before starting any iterations is vital. Consider a classifier, initially of poor performance (with an accuracy of 0.2 for example). When classifying unlabeled 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. Conversely, if the initial accuracy is high, it is more likely to correctly classify unlabeled 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 labeled instances assimilated by our semi-supervised classifiers. The labeled data set was used in this exercise. The labeled dataset was balanced (again using undersampling) and 500 of these instances were used for the initial labeled training stage. The self-training and co-training processes were initiated, applying these classifiers to the remainder of the labeled data (around 2000 instances) in steps of 500. Fig 3 shows a plot of the proportion of correctly classified instances that the iterative labeling process assimilated.

Fig 3. Graph showing how many correct assimilations the iterative labeling algorithms make per iteration

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 and peaks after the first iteration and then begins to drop. This decline appears to occur steadily. Unfortunately, due to the limited size of our labeled dataset with which we conducted this experiment, we could not observe any further iterations past 2000 instances. This makes it difficult to understand how steady the decline is or whether, either approach may see another rise. However, from our understanding of the workings of the self-training and co-training algorithms discussed in section 3.2, we believe it is highly unlikely that either approach sees another substantial rise in the proportion of correct additions.

4.2 Feature Analysis

We also assessed the discriminative ability of each of our features in order to test if any of our features was hindering our performance. To do this we performed feature ablation experiments [48]. We evaluated the performance of a given classifier (we chose the fully supervised Naive Bayes classifier for its simplicity) 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. Table 9 shows the results of this experiment. We chose to use the F_1 metric over the F_2 because we wanted to convey how the features performed in the general task of tweet classification. We found that negative emojis/emoticons were the most discriminative of our features and improved the system by a reasonable margin. We also found that our additional features proved effective for classifying tweets as relevant or not and yielded an improvement over the bag-of-words baseline.

We performed some analysis on the outcomes of the features using the constructed model. First, we sought to learn which words in our vocabulary were the best indicators of relevant tweets. We analysed the bag-of-words component of our compound feature

Ablated Feature	F1 Score
None	0.710
Contains Asthma-Verb Conjugate	0.693
Denotes Asthma Possession	0.695
Denotes Laughter	0.628
Negative Emojis/Emoticons	**0.627**
Word Classes	0.691
Positive/Negative Word Count	0.689

Table 9. F1 scores after feature ablation

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 generalizes to others and was adapted in the NLTK package for use in a broader sense. In NLTK, informativeness of a bag-of-words word w was calculated as the highest value of $P(w = feature\_value|class)$ for any class, divided by the lowest value of $P(w = feature\_value|class)$ [32]. This informativeness I, is summarised below with N representing the number of instances (more precisely, their feature representations) from the labeled data.

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

where C is the set of all classes and c is a possible class. Table 10 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 performed some experiments investigating the role of emojis in the data. Particularly we investigated which emojis where most prevalent in our context (keyword filtered tweets, specifically labeled data) as well as how often each emoji showed up in a tweet of each class.

Fig 4 shows the frequency with which each emoji occurred in the labeled tweets. It shows that only a few emojis appear very frequently in tweets collected in our context. This is good news because this means that only a few important emojis could be needed for determining tweet relevancy as opposed to monitoring for the full emoji dictionary. Table 11 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 tend to lean heavily towards one class. This shows that they can be quite discriminative and useful indicators of class membership hence helpful features. We believe it would be interesting to perform some further investigations to explore how far their usefulness can generalize even to other domains.

Fig 4. Bar chart showing emoji frequency in labeled data

Word	Relevant Prior Probability	Irrelevant Prior Probability
chest	0.96	0.04
throat	0.95	0.05
wow	0.95	0.05
health	0.06	0.94
cold	0.94	0.06
moment	0.94	0.06
forecast	0.07	0.93
awake	0.93	0.07
awful	0.93	0.07
sick	0.93	0.07
cough	0.92	0.08
pollution	0.08	0.92
bed	0.91	0.09
hate	0.91	0.09
watch	0.91	0.09

Table 10. Most informative words and how they split the data

Emoji	Occurrences for Relevant:Not-relevant Classes	Emoji	Occurrences for Relevant:Not-relevant Classes
😂	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

Table 11. Most frequent emojis in labeled data and their distributions

4.3 Keyword Analysis

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 specific terms. We measured the usefulness of the keywords we used in order to gain a better understanding of what would work best. To do this, we assessed the information retrieval performance of the keywords in data collection. 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 set of labeled tweets for this assessment (i.e. a set of 3500 tweets). In our scenario, the labeled tweets make up the set of retrieved documents and the tweets labeled as belonging to the "relevant" class make up the set of relevant documents. Here, recall measures the amount of results a keyword yielded while precision measures the amount of relevant results yielded. 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* yielded the next most results at 219, 132, 121. *Itchy eyes*, *could not breathe*, *coudn't breathe*, *chest tightness* and *acid rain* did not yield any results. The

remaining keywords did not return very many results (i.e. returned between 2 to 44 results). Table 12 shows the results of the assessment. *Asthma* had the highest recall

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

Table 12. Assessment of the retrieval quality of the search keywords

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. *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 the keyword with the highest recall power - *asma*. We found that *asma* only appeared in 4 tweets. We hypothesize that this is due to the fact that most users of Twitter post from devices capable of autocorrect. We discovered that *asthma* as a keyword does not fetch very many relevant tweets even though it fetches a lot of tweets. We believe this is because asthma is a big topic that is discussed in many different contexts. We still believe that it is still a good keyword because while its precision compared to its recall is not great, it brings in a lot of tweets (1313) and our semi-supervised algorithm is effective enough to filter a lot of the irrelevant tweets out while retaining the relevant tweets. The experiments show that *wheezing* is also a good keyword. It has a relatively high recall value and an even higher precision value so it returns a reasonable amount of tweets (219) and a lot of the tweets it returns are relevant.

Streaming Keyword	Relevant Prior Probability	Irrelevant Prior Probability
pollution	0.08	0.92
wheezing	0.84	0.16
fumes	0.24	0.76
panting	0.62	0.38

Table 13. Keyword Informativeness I of keywords

The informativeness, I, was calculated when the keywords were present as a feature. Most of the keywords were not informative as features and had an informativeness ratio of 1:1 for relevant:irrelevant. Table 13 shows the results for the informative tweets. The keywords not present in table 13 were totally uninformative as characterized as above. We found an overlap between the informativeness of the words in tweets (modeled as

bag-of-words features) and streaming keywords (shown in tables 10 and 13 respectively). *Pollution*, which was a keyword, appeared in the ranking of top 15 most informative words. Also, *wheezing* is relatively informative which mirrors and confirms our results in table 12 in which it is shown to be a good keyword as it had relatively high precision and recall.

5 Discussion

Twitter is a noisy data source for syndromic surveillance but through data processing and tweet classification, we have been able to to identify relevant tweets among the noisy data with an accuracy of 82.7%. We have also identified a strong and novel feature in the context of tweet classification in the form of emojis. Our work has given us very good insights into the user of Twitter for syndromic surveillance. One of the first problems we encountered is that of establishing location effectively, which is an important aspect of the surveillance effort. Although we have not tackled this problem in the paper, it is worth noting that we have found that most of the obvious location indicators are not well populated, and those that are, may not be accurate. Hence, our future work must tackle location as a real part of the problem for establishing a proper signal from Twitter.

We also found that "what to collect" is problematic as the data collection of tweets by keywords requires very carefully chosen keywords. Furthermore, our experimentation with different type of features like emojis and features constructed to capture negative or positive forms of expressions 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 forms of 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 adapt the keyword list to reflect those keywords that associate with the most relevant tweets, eliminating any keywords that are not performing adequately.

From our experiments, we have found that we are able to distinguish between tweets relevant to a given syndrome, in this case asthma/air pollution, to a reasonable degree. Our custom features constructed based on Twitter colloquialisms and norms proved effective in improving our understanding of the tweets and improving classification performance. Of all our custom features, the one that stood out most was the *Negative Emojis/Emoticons* feature. While emoticons have reportedly been used as a feature to success [41], our results prove that emojis could also be used to a similar or even greater effect. Emojis work even better than emoticons not only because they are becoming more widespread and established in popular culture [50], but because of their uniformity including across different languages. A smile emoticon could be illustrated in the form ":-D" or ":D". However, because emojis are actually unicode encoded pictographs with a set standard [18], 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. Negative emojis in particular, worked well for our purposes. To understand why, first recall that we found that a third of tweets used emojis. Upon exploring the tweets, we found that laughter-like and positive emojis were rarely used in relevant tweets (where the tweeter was expressing concern (personal or otherwise)) over the specific syndrome or its symptoms. They were found in relevant tweets where the tweeter was employing sarcasm but these were fairly uncommon. An example of such a tweet is "Can't breathe but thats okay 😊". Sad or negative emojis on the other hand, were only ever found in relevant tweets. In more formal terms, given s as the set of negative emojis, p as the set of positive emojis and Y and N as the sets for relevant and

non-relevant tweets respectively, we find that:

$$s \in \{Y\}$$

$$p \in \{Y, N\}$$

Negative emojis were more discriminative as they were more exclusively used with one of our target classes (ie. the "relevant" class).

We attempted to use semi-supervised classification algorithms because labelling of tweets is an expensive process. We compared semi-supervised algorithms to supervised algorithms to see if including unlabelled data could improve our performance in terms of relevance classification. We found that for the purposes of a system that provides us with a good balance on recall and precision as measured by the F_2 measure, semi-supervised learning using a self-training approach showed some advantage. Semi-supervised learning using co-training was also effective and actually yielded a better accuracy and F_1 score but it had a lower recall (than self-training) and so a lower F_2 score. We argued that recall is more important because the signal is weak and we wish to keep as much of the signal as possible while discarding only irrelevant tweets.

Co-training makes the assumption that each separate view used must be adequate for classification in itself. We were curious as to how this played out for us. For the taxonomical view, precision and recall were 0.57 and 0.88 respectively, while for the sentiment view, precision and recall were 0.86 and 0.77 respectively. We see from the final results of co-training that these values sort of averaged each other out. With this, we are able to affirm that the assumption was met. We are also able to affirm that the purpose of co-training is to ensure that the iterative labeling process assigns labels correctly. Fig 3 illustrates this point as it shows that the plot of the co-training approach is always higher than that of the self-training approach.

Feature	Value Distribution		Class Distribution	
			Relevant	Not Relevant
Contains Asthma-Verb Conjugate	TRUE	9.5%	52.1%	47.9%
	FALSE	90.5%	29.6%	70.4%
Indicates Personal Asthma Report	TRUE	4.3%	55.8%	44.2%
	FALSE	95.7%	30.7%	69.3%
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%

Table 14. Distribution of constructed features and classes across tweets from a different time period 2 years apart from our that of our labelled dataset

We were also curious as to how robust the system was to the changing of times and the shifts in language and colloquialisms that would accompany it. Our trained model (which was reported and assessed in section 4) was built on data from the first collection period. Table 2 details our collection periods. For a simple assessment, we applied our trained model to tweets collected in the most recent collection period. The tweets in this period and the tweets used in building the model have a time gap of two years between them. We believe that this is sufficient time for changes in slang, twitter behaviours and attitudes to occur, if any. Table 14 shows the class and feature distributions for the tweets in the different time period, for comparison to those used to build our models (shown in table 5). We see that *contains ashtma-verb conjugate* and *indicates personal asthma report* is now true less often than it was in our original time

period. *Denotes laughter* and *negative emojis/emoticons* on the other hand, return true more often than they did in our original time period. An interesting point to note is that *contains ashtma-verb conjugate* and *indicates personal asthma report* are syntactical and grammatical based features while *Denotes laughter* and *negative emojis/emoticons* are based on emojis. These results suggest that over the past two years, the use of emojis as a form of expression has grown. Conversely, grammar, at least in this context, may have become a less reliable way of constructing features. This might be due to the fact that slang and vernacular have become more widespread on Twitter and obstructed the meanings of certain words (for example "wheezing" being used as a term for laughter). Also, their convenience and popularity have people less concerned with grammatical and syntactical rules or conventions. Our semi-supervised approach based on self-training achieved a precision of 0.375 and a recall of 0.830 on the tweets from the most recent collection period. This means an F_1 score of 0.516 and more importantly, an F_2 score of 0.668. This suggests that there is some deterioration in performance over long periods of time. The deterioration in precision is minimal (only 0.067) while it is more noticeable in the recall (0.218). This means that the models will need to incorporate new tweets periodically. However, it appears that the deterioration does not occur very fast and so an update period of a year or 9 months should suffice. Alternatively, the system could be modified to allow some degree of autonomy such that it could modify itself as time went on. This provides some food for thought on the subject of future work and next steps.

6 Conclusions and future work

We have performed experiments to extract a relevant signal for a specific syndrome, asthma or an incident of air pollution that may cause an increase in asthma, using Twitter data. We have presented a thorough study looking at the process of data collection, the necessary pre-processing operations to eliminated noise, the extraction of useful features and the application of a number of pertinent algorithms to classify relevant tweets. All those operations, successfully applied, can be the building blocks of a real-time system that can extract a syndromic surveillance signal from Twitter data.

We have found that the signal is initially very noisy but we can improve on it by performing preprocessing and relevance analysis. We have found that some features that may not readily be considered, such as emojis or emoticons, may be quite strong for the purpose of understanding when people are expressing concern (personal or otherwise) over the specific syndrome or its symptoms. We also found that semi-supervised classification algorithms for relevance had two advantages: the first is that they allow us to use unlabelled tweets while building the models, thereby utilising more of the (relatively scarce) data available; second they provide higher recall which is of interest to us. We have analysed the set of keywords used for data collection as well as some of the prominent keywords associated with relevant tweets. We found that the vocabulary associated with relevant keywords may be less obvious than we may expect and include a number of colloquial expressions. Some of the keywords, on the other hand, were not always informative. We also found that the language used on Twitter may change over time. For these reasons, data collection keyword selection may be an area of research for the future.

While we performed our research on data collected within a somewhat wide time period (over the span of about a year and a half), we believe that further data collection and experimentation would be beneficial as it would allow us to further confirm how far it would generalize as time and topics change in Twitter conversations. We would also like to explore methods for automatic keyword selection. In particular, we would like to develop automated keyword selection algorithms which could explore an unfiltered

stream of public tweets to find words that might make good filters for some purpose. We will also work on the problem of location of tweets as an additional requirement for the overall approach to work.

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.

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

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

Acknowledgments

This research received support from the National Institute for Health Research (NIHR), Health Protection Research Unit in Emergency Preparedness and Response at King's College London.

References

1. Achrekar, H., Gandhe, A., Lazarus, R., Yu, S.-H., and Liu, B. Twitter improves seasonal influenza prediction. In *Healthinf* (2012), pp. 61–70.

2. Altman, N. S. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician 46*, 3 (1992), 175–185.

3. Baugh, W. bwbaugh: Hierarchical sentiment analysis with partial self-training. In *SemEval@NAACL-HLT* (Atlanta, Georgia, USA, 2013), p. 539.

4. Becker, L., Erhart, G., Skiba, D., and Matula, V. Avaya: Sentiment analysis on twitter with self-training and polarity lexicon expansion. In *Second Joint Conference on Lexical and Computational Semantics (* SEM)* (2013), vol. 2, pp. 333–340.

5. Bergsma, S., Post, M., and Yarowsky, D. Stylometric analysis of scientific articles. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (2012), Association for Computational Linguistics, pp. 327–337.

6. Blum, A., and Mitchell, T. Combining labeled and unlabeled data with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory* (New York, NY, USA, 1998), COLT' 98, ACM, pp. 92–100.

7. Blum, A., and Mitchell, T. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory* (1998), ACM, pp. 92–100.

8. Broniatowski, D. A., Paul, M. J., and Dredze, M. National and local influenza surveillance through twitter: An analysis of the 2012-2013 influenza epidemic. *PLoS ONE 8*, 12 (12 2013).

9. Brownstein, J. S., Freifeld, C. C., Reis, B. Y., and Mandl, K. D. Surveillance sans frontiéres: Internet-based emerging infectious disease intelligence and the healthmap project. *PLoS Medicine 5*, 7 (07 2008), 1–6.

10. Bruckhaus, T. The business impact of predictive analytics. *Knowledge discovery and data mining: Challenges and realities* (2007), 114–138.

11. Charles-Smith, L. E., Reynolds, T. L., Cameron, M. A., Conway, M., Lau, E. H., Olsen, J. M., Pavlin, J. A., Shigematsu, M., Streichert, L. C., Suda, K. J., et al. Using social media for actionable disease surveillance and outbreak management: A systematic literature review. *PloS one 10*, 10 (2015), e0139701.

12. Chawla, N. V. *Data Mining for Imbalanced Datasets: An Overview.* Springer US, Boston, MA, 2010, pp. 875–886.

13. Chen, L., Hossain, K. S. M. T., Butler, P., Ramakrishnan, N., and Prakash, B. A. Flu gone viral: Syndromic surveillance of flu on twitter using temporal topic models. In *2014 IEEE International Conference on Data Mining* (Dec 2014), pp. 755–760.

14. Chen, L., Tozammel Hossain, K. S. M., Butler, P., Ramakrishnan, N., and Prakash, B. A. Syndromic surveillance of flu on twitter using weakly supervised temporal topic models. *Data Mining and Knowledge Discovery 30*, 3 (2016), 681–710.

15. Chretien, J.-P., Burkom, H. S., Sedyaningsih, E. R., Larasati, R. P., Lescano, A. G., Mundaca, C. C., Blazes, D. L., Munayco, C. V., Coberly, J. S., Ashar, R. J., and Lewis, S. H. Syndromic surveillance: Adapting innovations to developing settings. *PLOS Medicine 5*, 3 (03 2008), 1–6.

16. Cohen, I., and Huang, T. S. Semisupervised learning of classifiers with application to human-computer interaction. *University of Illinois at Urbana-Champaign, Champaign, IL* (2003).

17. Collins, M., and Singer, Y. Unsupervised models for named entity classification. In *Proceedings of the joint SIGDAT conference on empirical methods in natural language processing and very large corpora* (1999), pp. 100–110.

18. Consortium, T. U. Unicode emoji. `http://unicode.org/emoji/`, 2017.

19. Copeland, P., Romano, R., Zhang, T., Hecht, G., Zigmond, D., and Stefansen, C. Google disease trends: an update. In *International Society of Neglected Tropical Diseases 2013* (2013), p. 3.

20. Culotta, A. Towards detecting influenza epidemics by analyzing twitter messages. In *Proceedings of the First Workshop on Social Media Analytics* (New York, NY, USA, 2010), SOMA '10, ACM, pp. 115–122.

21. Davidov, D., Tsur, O., and Rappoport, A. Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of the 23rd international conference on computational linguistics: posters* (2010), Association for Computational Linguistics, pp. 241–249.

22. De Quincey, E., and Kostkova, P. Early warning and outbreak detection using social networking websites: The potential of twitter. In *International Conference on Electronic Healthcare* (2009), Springer, pp. 21–24.

23. Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)* (1977), 1–38.

24. Diaz-Aviles, E., and Stewart, A. Tracking twitter for epidemic intelligence: case study: Ehec/hus outbreak in germany, 2011. In *Proceedings of the 4th annual ACM web science conference* (2012), ACM, pp. 82–85.

25. Dilrukshi, I., De Zoysa, K., and Caldera, A. Twitter news classification using SVM. In *Computer Science & Education (ICCSE), 2013 8th International Conference on* (2013), IEEE, pp. 287–291.

26. Doan, S., Ohno-Machado, L., and Collier, N. Enhancing twitter data analysis with simple semantic filtering: Example in tracking influenza-like illnesses. In *Healthcare Informatics, Imaging and Systems Biology (HISB), 2012 IEEE Second International Conference on* (2012), IEEE, pp. 62–71.

27. Doctor, V. What is a retweet? `https://www.hashtags.org/featured/what-is-a-retweet/`, 2012.

28. Forman, G., and Cohen, I. Learning from little: Comparison of classifiers given little training. In *European Conference on Principles of Data Mining and Knowledge Discovery* (2004), Springer, pp. 161–172.

29. Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., and Brilliant, L. Detecting influenza epidemics using search engine query data. *Nature 457*, 7232 (2009), 1012–1014.

30. Go, A., Bhayani, R., and Huang, L. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford 1*, 12 (2009).

31. Hanneke, S., and Roth, D. Iterative labeling for semi-supervised learning. Tech. rep., University of Illinois, Urbana, IL, USA, 2004.

32. Hardeniya, N. *NLTK essentials*. Packt Publishing Ltd, 2015.

33. Harris, J. K., Mansour, R., Choucair, B., Olson, J., Nissen, C., and Bhatt, J. Health department use of social media to identify foodborne illness - chicago, illinois, 2013-2014. *MMWR Morb Mortal Wkly Rep 63*, 32 (Aug 2014), 681–685.

34. Henning, K. J. What is syndromic surveillance? *Morbidity and Mortality Weekly Report 53* (2004), 7–11.

35. Hripcsak, G., and Rothschild, A. S. Agreement, the F-measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association 12*, 3 (2005), 296–298.

36. Joachims, T. Text categorization with support vector machines: Learning with many relevant features. *Machine learning: ECML-98* (1998), 137–142.

37. Johnson, C., Shukla, P., and Shukla, S. On classifying the political sentiment of tweets. Available at: `http://www.cs.utexas.edu/~cjohnson/TwitterSentimentAnalysis.pdf`, 2012.

38. Johnson, R., and Zhang, T. Semi-supervised convolutional neural networks for text categorization via region embedding. In *Advances in neural information processing systems* (2015), pp. 919–927.

39. Joshi, P. D., Bewoor, M., and Patil, S. In text mining.

40. Jurafsky, D. *Speech & language processing*. Pearson Education India, 2000.

41. Lamb, A., Paul, M. J., and Dredze, M. Separating fact from fear: Tracking flu infections on twitter. In *HLT-NAACL* (2013), pp. 789–795.

42. Lampos, V., and Cristianini, N. Tracking the flu pandemic by monitoring the social web. In *Cognitive Information Processing (CIP), 2010 2nd International Workshop on* (2010), IEEE, pp. 411–416.

43. Lampos, V., De Bie, T., and Cristianini, N. Flu Detector: Tracking Epidemics on Twitter. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases* (2010), ECML PKDD '10, pp. 599–602.

44. Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE 86*, 11 (Nov 1998), 2278–2324.

45. Lee, G., and Scott, C. Em algorithms for multivariate gaussian mixture models with truncated and censored data. *Computational Statistics & Data Analysis 56*, 9 (2012), 2816–2829.

46. Lee, K., Agrawal, A., and Choudhary, A. Real-time disease surveillance using twitter data: Demonstration on flu and cancer. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY, USA, 2013), KDD '13, ACM, pp. 1474–1477.

47. Lee, K., Qadir, A., Hasan, S. A., Datla, V., Prakash, A., Liu, J., and Farri, O. Adverse drug event detection in tweets with semi-supervised convolutional neural networks. In *Proceedings of the 26th International Conference on World Wide Web* (2017), International World Wide Web Conferences Steering Committee, pp. 705–714.

48. Litkowski, K. Feature ablation for preposition disambiguation. Tech. rep., CL Research, Damascus, MD, USA, 2016.

49. Liu, S., Zhu, W., Xu, N., Li, F., Cheng, X.-q., Liu, Y., and Wang, Y. Co-training and visualizing sentiment evolvement for tweet events. In *Proceedings of the 22nd International Conference on World Wide Web* (2013), ACM, pp. 105–106.

50. Ljubešic, N., and Fišer, D. A global analysis of emoji usage. *ACL 2016* (2016), 82.

51. Monaghan, A. J., Morin, C. W., Steinhoff, D. F., Wilhelmi, O., Hayden, M., Quattrochi, D. A., Reiskind, M., Lloyd, A. L., Smith, K., Schmidt, C. A., et al. On the seasonal occurrence and abundance of the zika virus vector mosquito aedes aegypti in the contiguous united states. *PLoS currents 8* (2016).

52. Morstatter, F., Pfeffer, J., Liu, H., and Carley, K. M. Is the sample good enough? comparing data from twitter's streaming API with twitter's firehose. arXiv preprint arXiv:1306.5204, 2013.

53. Nair, H., Brooks, W. A., Katz, M., Roca, A., Berkley, J. A., Madhi, S. A., Simmerman, J. M., Gordon, A., Sato, M., Howie, S., et al. Global burden of respiratory infections due to seasonal influenza in young children: a systematic review and meta-analysis. *The Lancet 378*, 9807 (2011), 1917–1930.

54. Nishida, K., Banno, R., Fujimura, K., and Hoshide, T. Tweet classification by data compression. In *Proceedings of the 2011 international workshop on DETecting and Exploiting Cultural diversiTy on the social web* (2011), ACM, pp. 29–34.

55. Oram, P. *WordNet: An electronic lexical database*. Cambridge Univ Press, Cambridge, MA: MIT Press, 2001.

56. Passer, M. B. Verb classifiers- misfits of nominal classification? In *35th TABU-dag* (June 2014).

57. Peng, F., and Schuurmans, D. Combining naive bayes and n-gram language models for text classification. In *European Conference on Information Retrieval* (2003), Springer, pp. 335–350.

58. Petti, C. A., Polage, C. R., Quinn, T. C., Ronald, A. R., and Sande, M. A. Laboratory medicine in africa: A barrier to effective health care. *Clinical Infectious Diseases 42*, 3 (2006), 377–382.

59. Pierce, D., and 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), pp. 1–9.

60. Powers, D. M. W. Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation. *Journal of Machine Learning Technologies 2*, 1 (2011), 37–63.

61. Ram, S., Zhang, W., Williams, M., and Pengetnze, Y. Predicting asthma-related emergency department visits using big data. *IEEE Journal of Biomedical and Health Informatics 19*, 4 (July 2015), 1216–1223.

62. Ramos, J., et al. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning* (2003).

63. Rao, D., and 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* (2009), Association for Computational Linguistics, pp. 58–65.

64. Read, J. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the ACL student research workshop* (2005), Association for Computational Linguistics, pp. 43–48.

65. Reyes, A., Potthast, M., Rosso, P., and Stein, B. Evaluating humour features on web comments. In *LREC* (2010).

66. Rish, I. An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence* (Sicily, Italy, 2001), pp. 41–46.

67. Sadilek, A., Kautz, H., and Silenzio, V. Predicting disease transmission from geo-tagged micro-blog data. In *AAAI Conference on Artificial Intelligence* (2012).

68. Sparck Jones, K. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation 28*, 1 (1972), 11–21.

69. Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., and 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* (2010), ACM, pp. 841–842.

70. Tabachnick, B. G., Fidell, L. S., and Osterlind, S. J. *Using multivariate statistics*. Allyn and Bacon Boston, 2001.

71. Triguero, I., García, S., and Herrera, F. Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study. *Knowledge and Information Systems 42*, 2 (2015), 245–284.

72. University, U. Warming climate puts europe at risk for seasonal outbreaks of dengue fever. `https://www.sciencedaily.com/releases/2016/04/160406124746.htm`, 2017.

73. Valverde-Albacete, F. J., and Peláez-Moreno, C. 100% classification accuracy considered harmful: The normalized information transfer factor explains the accuracy paradox. *PLoS ONE 9*, 1 (01 2014), 1–10.

74. Wikipedia. Emoticon. `https://en.wikipedia.org/wiki/Emoticon`, 2003.

75. Wikipedia. How to spot a twitter spambot. `http://mashable.com/2013/11/08/twitter-spambots/#x0RY3kS2ssqP`, 2003.

76. World Health Organisation WHO. The world health report 2007 - a safer future: global public health security in the 21st century. Available at: `http://www.who.int/whr/2007/en/`, 2007.

77. world Health Organisation (WHO), T. Seasonal Influenza fact sheet. http://www.who.int/mediacentre/factsheets/fs211/en/, 2014. [Online; accessed 22-Sep-2016].

78. Yarowsky, D. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics* (Stroudsburg, PA, USA, 1995), ACL '95, Association for Computational Linguistics, pp. 189–196.

79. Yerva, S. R., Miklós, Z., and 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* (2011), ACM, p. 48.

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

81. Zhang, D., and Lee, W. S. Question classification using support vector machines. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval* (2003), ACM, pp. 26–32.

82. Zhang, H., and Li, D. Naïve bayes text classifier. In *Granular Computing, 2007. GRC 2007. IEEE International Conference on* (2007), IEEE, pp. 708–708.

83. Zhang, S., Wu, J., Zheng, D., Meng, Y., and Yu, H. Semi-supervised classification of twitter messages for organization name disambiguation. In *IJCNLP* (2013), pp. 869–873.

84. Zhao, J., Lan, M., and Zhu, T. T. Ecnu: Expression-and message-level sentiment orientation classification in twitter using multiple effective features. *SemEval 2014* (2014), 259.