TOWARDS EARLY DISCOVERY OF SALIENT HEALTH THREATS: A SOCIAL MEDIA EMOTION CLASSIFICATION TECHNIQUE

BAHADORREZA OFOGHI[1] and MEGHAN MANN[1] and KARIN VERSPOOR[1,2]

[1]*Department of Computing and Information Systems*
[2]*Health and Biomedical Informatics Centre*
The University of Melbourne
Parkville, Victoria 3010, Australia

Online social media microblogs may be a valuable resource for timely identification of critical ad hoc health-related incidents or serious epidemic outbreaks. In this paper, we explore emotion classification of Twitter microblogs related to localized public health threats, and study whether the public mood can be effectively utilized in early discovery or alarming of such events. We analyse user tweets around recent incidents of Ebola, finding differences in the expression of emotions in tweets posted prior to and after the incidents have emerged. We also analyse differences in the nature of the tweets in the immediately affected area as compared to areas remote to the events. The results of this analysis suggest that emotions in social media microblogging data (from Twitter in particular) may be utilized effectively as a source of evidence for disease outbreak detection and monitoring.

Keywords: Twitter, Ebola, Emotion classification, Shift detection

1. Introduction

Syndromic surveillance involves monitoring of public health information resources, to facilitate early detection of disease outbreaks, and to monitor the size, spread, and tempo of epidemic outbreaks.[1] Many jurisdictions have regulations for reporting on infectious diseases to public health officials, for instance requiring that laboratory-confirmed cases of influenza be notified to the government (see, e.g., the Australian National Notifiable Diseases Surveillance System[a]). However, it is important to have surveillance mechanisms in place that identify weaker signals of disease activity, in particular for diseases with potentially severe public health consequences, such as Botulism or Ebola, that public health officials want to be able to respond to quickly. Social media posts are a major source of uncurated user-generated feedback, that may have a positive impact on critical applications related to public health and safety.[2]

There have been a number of efforts to develop computational approaches that enable automated monitoring and *early warning* systems making use of online resources. In recent work,[3–6] prediction of near future Influenza events as well as the spread of N1H1 and Ebola cases were studied using descriptive statistics extracted from Twitter messages as well as utilizing data from Google Flu Trends.[7] This work supports the usefulness of Twitter data for pandemic event surveillance. However, it mostly focuses on descriptive statistics at the level of single tweets (or single sentiments) over time and does not consider the *combination* or *distribution* of sentiments across a *collection* of tweets as an early warning signal.

RSS feeds have also been classified as relating to certain pathogens without necessarily having explicit evidence or mention of the pathogen (i.e., from reported symptoms).[8] Such

[a]www.health.gov.au/internet/main/publishing.nsf/Content/cda-surveil-nndss-nndssintro.htm

systems have been demonstrated to produce similar predictions to that of government health organizations.[7,8] However, these methods may not be appropriate for identifying salient outbreaks where only a small number of people are infected, such as isolated Ebola cases, while an influx of web and social media messages regarding the disease is encountered.

We approach disease outbreak detection from the perspective of the emotional stance of a user towards a disease. The underlying hypothesis of our work is that a proximal disease incident will trigger the expression of concerns about the incident, and that these expressions will differ qualitatively (emotionally and linguistically) from the typical chatter around a distant or less immediate threat. We propose a model, building on this hypothesis, to detect a shift in the nature of the conversations around a specific disease on the basis of changes in the distribution of emotions expressed in tweets containing some response to a public health incident. Public mood has been demonstrated to relate to major socio-economic events,[9] and identifying shifts in emotions may also provide a useful early indicator that a new public health incident has occurred. This strategy removes the need for classification of textual documents into pre-defined syndromes or explicit prediction of future events; instead, it has a focus on the distribution of emotional expressions in the texts of microblogs in specific periods of time.

We therefore explore the relationship between public mood and salient public health threats in this paper. We believe that users express different emotions, thoughts and speculations and may post different types of informational links and resources at times prior to and following major epidemic incidents. This may be particularly true when a user feels directly impacted by an incident, e.g., due to geographical proximity to an event. We do not pre-suppose that there are specific emotions that will be consistently identifiable across distinct public health issues, but rather focus on whether there is a change in the distribution of emotions.

We examine the distribution of emotion classes in tweets to estimate the differences between emotional features before and after likely outbreaks with two component strategies:

- Emotion classification of tweets, using a trained classification model to assign each tweet to one of several emotion classes.
- Emotion shift detection through statistical analysis of tweet corpora, comparing the distribution of emotions expressed in tweets immediately prior to and after relevant incidents.

To explore our hypothesis, a case study of two recent events in London, United Kingdom where a health worker was found to have been exposed to the Ebola virus is provided. The emotions of all tweets in London around these events explicitly mentioning Ebola were analysed. We demonstrate that by monitoring Twitter microblogs, it is possible to capture likely outbreaks through detection of emotional shifts in user tweets.

2. Emotion Classification of Tweets

We begin by developing an emotion classifier for outbreak-related tweets, using a new annotated data set and an emotion inventory that adopts Ekman's six basic emotions[10] ("anger", "disgust", "happiness", "sadness", "surprise", and "fear"), and extends it with three additional "attitudinal" classes, "sarcasm", "news-related", and "criticism".

Emotion detection from textual data has been previously tackled using various unsupervised[11] and supervised approaches.[12] Aman and Szpakowicz[13] utilized corpus-based unigrams, emotion-related words extracted from Roget's thesaurus, and features derived from WordNet-Affect to train a supervised emotion classifier. They employed Ekman's inventory plus a no-emotion class in their work. Wang et al.[14] utilized an overlapping emotion inventory and similar features. They found improvements in tweet emotion classification through consideration of the sentiment of words (positive or negative) as features.

We approach emotion detection using similar textual features to these previous studies, testing both lexicon-based (unsupervised) and supervised methods. We developed a binary classifier for each emotion class, experimenting with several representations of tweet texts, as will be discussed in the following sections.

2.1. *Lexicon-Based Classification*

A simple unsupervised baseline emotion classifier was implemented for each emotion class, using a lexicon-based vector model. We constructed a *reference vocabulary* consisting of terms corresponding to each emotion class.[b] These terms include emotion-related terms from Emotion Vocabulary,[15] lexical units derived from the FrameNet[16] frame *Judgement-communication* (for class "criticism"), emotion terms from the Profile of Mood States,[9,17] and emoticons. We also include terms specifically for the "news-related" category, corresponding to popular international news agency names. The resultant reference vocabulary contains 499 terms. Each emotion class is represented as a binary vector with respect to this reference vocabulary; any term from the vocabulary relevant to the emotion was marked 1 and irrelevant terms 0.

For classification, tweets were also mapped into this lexical vector representation, with a 1 indicating that the tweet contains a given term from the vocabulary. The cosine similarity score between this tweet vector and each of the nine emotion class vectors was computed; the class with the highest similarity was returned as the classification of the tweet.

2.2. *Machine Learning-Based Classification*

A Naïve Bayes classifier, implemented in MALLET toolkit,[18] was used for our machine learning-based classification. The basic features used to represent tweets were bag-of-words. This set of features was augmented in a feature engineering step. The extra features included: i) the lexicon-based similarity score for each of the nine classes obtained from the baseline lexicon-based classifier, ii) emotion vocabulary from the same reference vocabulary that the lexicon-based classifier made use of, iii) emoticons, iv) punctuations including question and exclamation marks, and v) sentiment classification of the tweet text from the Stanford Sentiment Analyzer (i.e., negative, neutral, or positive).[19]

2.3. *Data*

To train the emotion classifier on tweets relevant to an active public health threat, we collected recent tweets regarding Ebola using the Twitter API. A total of 12,101 tweets that contained

[b]The vocabulary is available at: `https://bitbucket.org/readbiomed/socialsurveillance`

Table 1. The distribution of the nine emotion classes over the 4,405 *Ebola* tweet set labelled by Mechanical Turk workers.

Class	Sarcasm	News-rel.	Criticism	Fear	Surprise	Anger	Happiness	Disgust	Sadness
#Tweets	1,322	2,572	166	81	67	62	61	51	23

the word "ebola" were collected from all over the world in the second half of March 2015. Non-English tweets were filtered out, leaving 7,039 tweets. After initial pre-processing of the tweets and removal of redundant (identical) tweets, 4,405 tweets remained.

The tweet texts were normalized in the pre-processing step. All URLs, email addresses, mentions (i.e., @replies and @usernames), and hash tags were replaced by "url", "emailAddress", "atSign", and "hashTag", respectively. Only "#Ebola" tags were retained and converted to "ebola" in order to preserve mentions of the disease. The "RT" tags at the beginning of the re-tweets were also removed and any redundant tweets (e.g., re-tweets of the same text) were then filtered out. This resulted in preserving only those re-tweets for which the original tweets were missed in the time frame when data capture was in process.

Amazon's Mechanical Turk[20] was used to acquire human judgements of the emotion labels for each tweet in the set of 4,405 *Ebola* tweets. The qualification criteria for Mechanical Turk workers who labelled the data included: i) they were "categorization masters", ii) located in the US (as a proxy to ensure their English was of reasonable standard), and iii) achieved at least 90% accuracy on a test that involved labelling of 10 tweets in to one of the nine emotion classes. Table 1 summarizes the distribution of classes over the resultant tweet set.

A second round of pre-processing was carried out on the labelled dataset before training the binary Naïve Bayes classifiers for the nine classes. This included tokenization, lowercasing of tokens, removal of stop-words, and lemmatization.

2.4. *Experiments and Discussion*

The two classifiers were applied to the Ebola emotion dataset. The baseline classifier, as it is unsupervised, was tested on the full dataset. The ML classifier was trained and tested in a 10-fold cross-validation scenario. The macro average of precision, recall, and F1 measures were calculated over the nine classes for each classifier, with each feature set (see section 2.2). Table 2 summarizes the results (some results not shown for clarity).

The results in Table 2 demonstrate that the baseline lexicon-based classifier is strongly outperformed by the ML-based Naïve Bayes classifier, even with the basic bag-of-words features. Adding features beyond the bag-of-words features to the ML classifier had an incremental effect on the performance of the classifier. In general, bag-of-words features may result in higher classification performances as the number of input texts grow.[19] Since user tweets are mostly short pieces of text, the incremental effect of additional features is expected. On the other hand, in most cases, lemmatization of tweet tokens had only marginal impact on the results, so we have elided results without lemmatization from Table 2, except for the scenario with the overall best performance. When lemmatization was off, the highest classification performance was achieved by the ML-based classifier that utilizes all the possible feature sets. Among the different feature sets used in combination with bag-of-words features, the

Table 2. Binary emotion classification results on the set of 4,405 tweets with different feature sets. The Lexicon-Based (LB) measures were obtained on the entire data set as the test set while the Naïve Bayes (NB) measures were calculated using 10-fold cross validation. Note: bow=bag-of-words, LBsim=lexicon-based similarity measure (see section 2.1), eVoc=emotion vocabulary, emt=emoticon, punc=punctuation, sent=sentiment, rest=LBsim+sent+punc, p=precision, r=recall, f=F1 score, M.avg.=macro average, +[*] means NB/bow+[*], and lem=lemmatization. Except for +sent, all results are with +lem only.

Class	Metric	LB	NB/bow	+LBsim	+eVoc	+emt	+punc	+sent(iment)		+rest
								-lem	+lem	
Sarcasm	p	.517	.782	.793	.786	.787	.801	.798	.784	.794
	r	.910	.781	.791	.785	.787	.801	.798	.783	.791
	f	.659	.781	.790	.784	.786	.800	.797	.783	.791
News-rel.	p	.0	.827	.821	.824	.823	.828	.835	.829	.830
	r	.0	.821	.820	.822	.816	.825	.830	.825	.830
	f	.0	.823	.820	.823	.819	.826	.832	.827	.830
Anger	p	.466	.799	.726	.815	.807	.760	.776	.789	.769
	r	.774	.780	.726	.802	.805	.742	.768	.790	.755
	f	.582	.777	.720	.800	.798	.731	.747	.771	.733
Criticism	p	.0	.679	.688	.636	.636	.662	.675	.675	.661
	r	.0	.680	.683	.637	.636	.665	.670	.667	.660
	f	.0	.670	.679	.633	.633	.660	.663	.665	.652
Surprise	p	.473	.489	.566	.617	.609	.681	.658	.624	.707
	r	.791	.512	.546	.590	.603	.646	.618	.619	.677
	f	.592	.472	.486	.574	.560	.626	.594	.591	.659
Fear	p	.513	.711	.688	.697	.665	.695	.653	.717	.666
	r	.963	.701	.702	.681	.654	.687	.652	.709	.653
	f	.669	.678	.671	.664	.611	.676	.629	.673	.642
Happiness	p	.487	.717	.714	.760	.740	.668	.745	.842	.758
	r	.918	.702	.735	.750	.726	.658	.758	.820	.747
	f	.636	.681	.669	.735	.710	.646	.722	.824	.741
Disgust	p	.471	.742	.681	.695	.648	.686	.667	.743	.692
	r	.784	.716	.666	.681	.635	.684	.674	.716	.690
	f	.588	.661	.614	.654	.603	.661	.626	.696	.679
Sadness	p	.537	.829	.675	.785	.767	.729	.717	.821	.646
	r	.956	.829	.717	.771	.771	.729	.712	.754	.708
	f	.687	.799	.646	.720	.724	.686	.656	.739	.622
M.avg.p		.385	.731	.706	.735	.720	.723	.725	**.758**	.725
M.avg.r		.677	.725	.710	.724	.715	.715	.720	**.743**	.723
M.avg.f		.491	.705	.677	.710	.694	.701	.696	**.730**	.706

sentiment features (with lemmatization) resulted in the highest classification macro average values (precision=0.758, recall=0.743, and F1=0.730). This classification setting was therefore selected for the next experiments to detect shifts in emotions expressed in user tweets.

3. Unsupervised Emotion Shift Detection

We then explored the measurement of emotional shifts in user tweets around public health incidents using the best-performing emotion classification model. We propose a method for shift detection, and test it on a focused dataset collected for the study.

3.1. *Data*

3.1.1. *Ebola Incident Tweet Corpora*

We collected Twitter microblogs (tweets) around the time of two reported cases of possible Ebola infection in London. The first event involved a health worker named Pauline Cafferkey, who was diagnosed with Ebola in Glasgow on December 29, 2014 and transferred to London the following day. News reporting of the event began on December 30, 2014. The second event involved another healthcare worker who suffered a needle-stick injury in Sierra Leone and was flown to London for treatment. News reporting of the event began on January 31, 2015.

Specifically, we collected sets of tweets from London containing a mention of the word "ebola", for a period of 7 days prior and 7 days after each event. In this way, the time windows for tweets analysed for each of these events are disjoint. We refer to these datasets as the *ebola-event* datasets, and the subsets corresponding to the two time periods under study as the *pre-event corpus* (7 days prior) and the *post-event corpus* (7 days after), maintained separately for each event. We consider the day the event was reported as the split point. Tweets on that day are included in the *post-event corpus*.

To establish a reference dataset, we then downloaded a set of tweets in a similar way to the *ebola-event* datasets, but at a time period distinct from the events in London, and from a region remote to those events. We selected tweets mentioning Ebola from Australia in the time period December 09-22, 2014. As Australia has had no known cases of Ebola infection, and it is an issue that likely does not directly impact the Twitter users writing the collected tweets, this dataset should capture "normal", background dialogue about Ebola. We refer to this as the *ebola-background* dataset. We divide this dataset into two subsets, arbitrarily at the mid-point of the time period to obtain two subsets of Ebola-related tweets representing a comparable time frame to each of the *ebola-event* datasets. We refer to these subsets as the *pre-ebola-background* corpus and the *post-ebola-background* corpus. Figure 1 summarizes the tweet datasets that were collected and analyzed in this work.

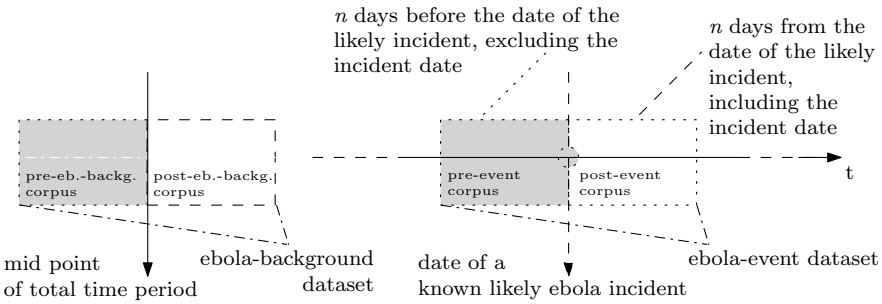

Fig. 1. The tweet collection schema used for analyzing and monitoring changes in user expressed emotions.

3.1.2. *Tweet Retrieval and Processing*

The tweets for the *ebola-event* datasets and the *ebola-background* dataset were collected using the Twitter API. The geo-codes of the tweets were used to retrieve only tweets posted by the

Table 3. Tweet and vocabulary statistics of the *ebola-event* datasets and *ebola-background* dataset with a window of 7 days before and after set dates, tied to likely Ebola incident in the region under study for the *event* data. The size of the vocabulary (|vocab.|) is equal to the number of distinct tokens.

Dataset	Date (±7)	pre-corpus		post-corpus	
		#tweets	|vocab.|	#tweets	|vocab.|
ebola-event-1	Dec/29/14	73	204	337	906
ebola-event-2	Jan/31/15	165	700	90	417
ebola-backg.	Dec/16/14	429	1453	340	1208

users from the specific regions under study. A radius of 200 kilometers was used around a specific geo-code, which for a city roughly corresponds to the geographic center of the city. Re-tweets and non-English tweets were excluded from the retrieved set of tweets. Only tweets containing a mention of the specific keyword "ebola" (either as hashtag or an individual word) were retained for both the *ebola-event* datasets and *ebola-background* dataset. No further analysis was performed for either finding the location of tweets that did not have explicit geo-code tags, or identifying any other tweets that may have been related to Ebola with no explicit mention of the disease.

The *ebola-event* datasets and *ebola-background* dataset were organized by retrieving tweets using the above query parameters, restricted to 7-day windows. This is the maximum number of days that one can move back in history of tweets and retrieve microblogs at any given time when using the current Twitter API. All of the tweets retrieved for each time window were put together in one text corpus, i.e., two tweet corpora were created for each dataset; one containing the tweets related to up to 7 days prior to the likely incident and another containing 7 days of tweets starting from the date of the likely incident in the region. Table 3 shows the statistics of the two datasets.

Textual modeling of retrieved tweets required some pre-processing of the tweet corpora, including tokenization, surface normalization, and removal of stop-words from the dictionary of terms for each tweet corpus. This processing was performed using MALLET.[18]

3.2. *Experiments and Discussion*

Our experiments assess the emotion class distributions in tweet corpora, in order to determine whether there are discernible differences in the emotions expressed in user tweets on the topic of an infectious disease that arise when the threat shifts from being abstract to being more immediate. To examine this, we considered the differences *within* the various datasets that we have collected — comparing the pre-event and post-event corpora.

3.2.1. *Corpus-Level Emotion Distribution Analysis*

Each tweet in each of the tweet corpora was first classified into one of the nine emotion and non-emotion classes introduced in section 2. For this, the highest-performing emotion classifier model trained on the distinct set of labelled Ebola tweets (see Section 2.4) was utilized to predict an emotion class label for each tweet. Table 4 summarizes the results of this step. Then, the differences in the distributions of classes between pairs of corpora were measured using

Table 4. Distribution of nine classes over the different tweet corpora obtained using the best-performing emotion classifier. For each data set we report two numbers, (X,Y): the number of instances classified as positive=X and the number of instances classified as negative=Y, Critic.=Criticism, Happ.=Happiness, pre/post-e-x=pre/post-ebola-event-x, pre/post-bkg.=pre/post-ebola-background.

Dataset	Sarcasm	News-rel.	Anger	Critic.	Surprise	Fear	Happ.	Disgust	Sadness
pre-e-1	71,2	73,0	73,0	73,0	73,0	73,0	73,0	73,0	63,10
post-e-1	298,39	322,15	300,37	325,12	265,72	288,49	308,29	298,39	277,60
pre-e-2	133,32	162,3	114,51	165,0	113,52	115,50	105,60	138,27	112,53
post-e-2	77,13	88,2	79,11	89,1	75,15	85,5	76,14	79,11	77,13
pre-bkg.	428,1	428,1	373,56	395,34	338,91	405,24	362,67	407,22	395,34
post-bkg.	255,85	332,8	255,85	295,45	283,57	230,110	308,32	309,31	294,46

Table 5. Statistical paired t-test analysis of class distributions in the different datasets in terms of positive and negative classified instances. A † shows a statistically significant p-value at the 5% level.

Classes	Dataset	p-value
6 emotions	ebola-event-1	0.004†
	ebola-event-2	0.002†
	ebola-backg.	0.259
6 emotions + 3 non-emotions	ebola-event-1	0.009†
	ebola-event-2	0.007†
	ebola-backg.	0.079

statistical paired t-tests. The statistical significance analysis of the differences between class distributions was performed for two groups of instances per class per pair of tweet corpora: group 1) all of the instances that were classified as positive (e.g., happy), and group 2) all of the instances that were classified as negative (e.g., not-happy). This analysis was done in terms of the pure-emotion classes (i.e., the six basic emotions) as well as all of the nine emotion and non-emotion classes.

The distribution differences *within* each tweet corpus (i.e., the two *ebola-event* datasets and the *ebola-background* dataset) were then calculated. Each time-delimited corpus of each dataset was compared against its neighboring counterpart; that is, the *pre-ebola-event-1* and the *post-ebola-event-1* corpora were compared with each other, the *pre-ebola-event-2* and the *post-ebola-event-2* corpora were compared with each other; and finally the two subsets of the *ebola-background* dataset were compared. Table 5 summarizes the results of this experiment.

In Table 5, all of the p-values obtained for comparing the tweet corpora before and after the incidents in London indicate statistically significant differences (at the 5% level) between class distributions. On the other hand, none of the comparisons between the tweet corpora in the *ebola-background* dataset shows a statistically significant difference. This suggests that the distribution of six basic emotions and/or the nine emotion and non-emotion classes in user tweets shift significantly as a result of salient health incidents such as Ebola.

Table 6. KL-divergence analysis of emotion classification distributions for the three datasets. Since KL-divergence is non-symmetric, X,Y values mean X=KL-divergence of tweet corpus before vs. after, and Y=KL-divergence of tweet corpus after vs. before.

Class	ebola-event-1	ebola-event-2	ebola-backg.
Sarcasm	0.077,0.119	0.013, 0.012	0.395, 1.377
News-rel.	0.066, ∞	0.001, 0.001	0.023, 0.048
Anger	0.168, ∞	0.175, 0.139	0.063, 0.074
Criticism	0.052, ∞	0.016, ∞	0.020, 0.024
Surprise	0.347, ∞	0.096, 0.083	0.010, 0.009
Fear	0.227, ∞	0.436, 0.278	0.312, 0.494
Happiness	0.130, ∞	0.186, 0.154	0.028, 0.024
Disgust	0.177, ∞	0.011, 0.010	0.016, 0.019
Sadness	0.009, 0.009	0.144, 0.119	0.022, 0.026

3.2.2. *Emotion-Level Distribution Analysis*

To understand how the emotions expressed in user tweets shift as a result of likely Ebola incidents, further analysis was carried out. Here, we measure the distribution of the tweets that were classified positive vs. negative with respect to each of the nine emotion classes before and after the likely incidents.

For this analysis, Kullback-Leibler divergence[21] was utilized. KL-divergence, also known as cross-entropy or information divergence, is a non-symmetric measure for the difference in two probability distributions P and Q over the same event space. On a finite set χ, the KL-divergence between the two probability distributions P and Q is calculated using Equation 1. In this case, P and Q represent probability distributions of positive and negative instances of a specific emotion class for tweet corpora prior to and after the likely incidents. The measure has been shown to be useful for comparing linguistic corpora in prior work.[22,23]

$$D_{kl}(P\|Q) = \sum_{x\in\chi} P(x)\log_n \frac{P(x)}{Q(x)} \quad (1)$$

Table 6 shows the results of KL-divergence analysis of the distribution of the results of emotion classifications (per emotion) between the two tweet corpora in each dataset. In addition to revealing the distributions of positive and negative classified instances per emotion class in each dataset, the results in Table 6 demonstrate how the differences in classifications distributions across the datasets vary. For instance, it can be observed that the KL-divergence values for class "surprise" are larger in the two event datasets (0.347,∞ and 0.096,0.083 for positive and negative classified instances) compared with those between the two corpora of the *ebola-background* dataset (0.010,0.009). In this particular case, the KL-divergence measures re-confirm our findings in previous sections; differences in class distributions are only significant between the tweet corpora in the two *ebola-event* datasets (note that smaller KL-divergence values indicate more similar probability distributions).

4. Limitations

This work has tackled the problem of understanding emotional shift as a result of likely disease outbreaks in particular regions of the world. However, it is important to note several limitations

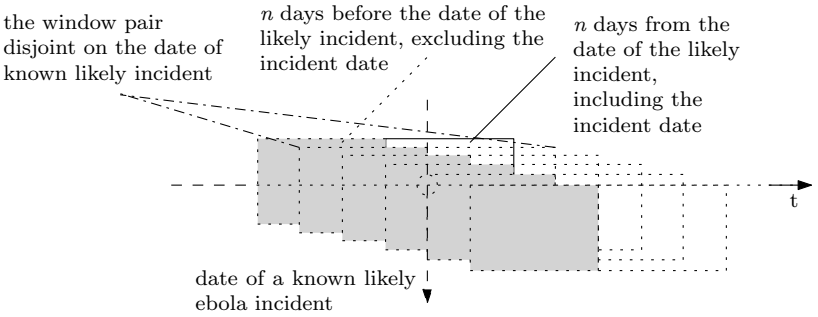

Fig. 2. The detection architecture for capturing microblogs and monitoring emotional changes in user posts.

of the study, which we will address in future work.

First, the three tweet datasets that were collected may have been more directly comparable if they were collected from exactly the same region. While the two *ebola-event* datasets were from the same urban center in Europe (i.e., London, UK), *ebola-background* dataset was collected from Australia. While this allowed us to compare the event-related tweets with a neutral background set, a dataset from the same geographical region but separate in time from the events would provide a better assessment of the methods. An additional dataset from Australia collected over the same time period as the active events would also have been preferable as a background set. Together, these datasets would allow us to contrast geographical separation and temporal separation in terms of vocabulary. Due to the history restrictions on the Twitter API, it is not possible to re-create such datasets after the fact.

Second, due to the sparseness of geo-location meta-data in tweets, a number of related microblogs from the specific region were missed and not included in the three datasets we collected. Other researchers have investigated this problem in other contexts.[24] An extension of this study would be to utilize other tweet features for locating microblogs to improve the data collection procedure.

5. Towards a Detection Model

The long-term objective of our work is to provide practical evidence for early discovery and timely alarming of localized pandemic outbreaks and salient health threats. We propose that this aim can be achieved through continuous monitoring of user microblogs, specifically through identification of sudden emotion shifts. Figure 2 depicts the architecture of the proposal; in which microblog emotions are analyzed and monitored for changes in the distribution of emotions. The size and significance of the changes in emotion distributions can subsequently be utilized, either individually or in combination with other sources of evidence, to detect likely incidents or outbreaks that are of concern to the public. It is expected that the proportion of emotional shift reaches its highest value for the two time windows that are disjoint on the specific date when a putative disease-related incident or outbreak occurs. Although our focus has been on early identification of localized incidents, more generally, the proposed methodology can be utilised to detect any wave of panic in public related to other phenomena.

We have taken the initial steps towards reaching this goal by validating the underlying

assumption that it is possible to observe emotion changes in neighboring sets of Twitter microblogs across a given time-point corresponding to the start of a reported health threat. We have established the viability of the approach, although further experiments are required to explore its application to real-time streaming data from Twitter, and to determine its effectiveness for early detection. Of particular interest is whether Twitter provides a meaningful information source for detecting concern about major diseases ahead of news reporting.

We also intend to capture a larger number of tweets and tweet corpora over time to further our understanding of the nature of *vocabulary or lexical shifts* around health threats, in addition to emotion shifts already studied in this work. We would like to implement an active monitoring and detection procedure over specific regions of the world for any outbreaks of Ebola or similar pandemic threats that may be both emotionally and lexically monitored and detected.

6. Conclusions

We have analyzed the variation in emotion in Twitter microblogs that are posted by users prior to and after an identified health threat, building on a text-based emotion classifier to produce a statistical assessment of emotion distributions. The combined classification and corpus analysis approach has promising application in online monitoring and detection of outbreaks in streaming textual data.

Different strategies for emotion classification in the context of serious public health events were studied in this work, including an unsupervised lexicon-based approach and a supervised machine learning-based classifier. Our experiments on a large set of Ebola tweets demonstrated that the ML-based classifier achieved the highest emotion classification performance when the tweets were represented using sentiments derived from the Stanford Sentiment Analyzer, combined with lemmatized bag-of-words features.

We considered differences in the distributions of emotion class labels assigned to microblogs across tweet corpora collected from two recent salient Ebola threats, examining variations in both corpus-level emotion and emotion-level changes. In our experiments, we found that there were statistically significant differences in the distribution of emotions in the tweet corpora that belong to the time periods before and after likely incidents of Ebola. There were no such differences for the two tweet corpora in a background dataset that was not aligned to any Ebola incident. This suggests that the distribution of predicted emotion class labels for tweets, based on Ekman's six basic emotions plus the three non-emotion classes "sarcasm", "news-related", and "criticism", can be used as an indication of the occurrence of pandemic health threats.

We will explore the broader capacity of this work to detect emergent health threats of concern to a localized community, prior to formal reporting, in future work. Currently, we are developing this study to consider lexical shifts in the tweet corpora alongside the emotion class distributions discussed in this work. We are working to identify the lexical items that distinguish tweets from before and after likely health threats.

7. Acknowledgments

We thank the Victoria State Government's Department of Health and Human Services in Melbourne, Victoria, Australia for funding support.

References

1. K. J. Henning, *Morbidity and Mortality Weekly Report (MMWR)* **53(suppl)**, 5 (2004).
2. N. Elhadad, L. Gravano, D. Hsu, S. Balter, V. Reddy and H. Waechter, Information extraction from social media for public health, in *KDD at Bloomberg: The Data Frameworks Track*, (New York City, US, 2014).
3. M. J. Paul, M. Dredze and D. Broniatowski, *PLOS Currents: Outbreaks* (2014).
4. C. Chew and G. Eysenbach, *PloS one* **5**, p. e14118 (2010).
5. J. Gomide, A. Veloso, W. Meira, V. Almeida, F. Benevenuto, F. Ferraz and M. Teixeira, *Proceedings of the ACM WebSci'11, June 14-17 2011, Koblenz, Germany.* , 1 (2011).
6. M. Odlum and S. Yoon, *American Journal of Infection Control* **43**, 563 (2015).
7. J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski and L. Brilliant, *Nature* **457**, 1012 (2009), 10.1038/nature07634.
8. N. Collier, S. Doan, A. Kawazoeand, R. M. Goodwin, M. Conway, Y. Tateno, Q. Ngo, D. Dien, A. Kawtrakul, K. Takeuchiand, M. Shigematsu and K. Taniguchi, *Bioinformatics* **24**, 2940 (2008).
9. J. Bollen, A. Pepe and H. Mao, *CoRR* **abs/0911.1583** (2009).
10. P. Ekman, Universals and cultural differences in facial expression of emotion, in *Nebraska Symposium on Motivation*, (Lincoln, Nebraska, 1972).
11. A. Agrawal and A. An, Unsupervised emotion detection from text using semantic and syntactic relations, in *Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology - Volume 01*, 2012.
12. S. Aman and S. Szpakowicz, Using roget's thesaurus for fine-grained emotion recognition, in *Proceedings of the Third International Joint Conference on Natural Language Processing*, 2008.
13. S. Aman and S. Szpakowicz, Using Rogets thesaurus for fine-grained emotion recognition, in *Proceedings of the Third International Joint Conference on Natural Language Processing*, (Hyderabad, India, 2008).
14. W. Wang, L. Chen, K. Thirunarayan and A. P. Sheth, Harnessing Twitter "big data" for automatic emotion identification, in *Proceedings of the International Conference on Privacy, Security, Risk and Trust and the 2012 International Confernece on Social Computing (SocialCom)*, (Amsterdam, 2012).
15. T. Drummond, Emotion vocabulary http://www.sba.pdx.edu/ faculty/ mblake/ 448/ FeelingsList.pdf.
16. C. F. Baker, C. J. Fillmore and J. B. Lowe, The Berkeley FrameNet project, in *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98 (Association for Computational Linguistics, Stroudsburg, PA, USA, 1998).
17. D. McNair, M. Loor and L. Droppleman, *Profile of Mood States* 1971.
18. A. K. McCallum, MALLET: A Machine Learning for Language Toolkit http://mallet.cs.umass.edu, (2002).
19. R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng and C. Potts, Recursive deep models for semantic compositionality over a sentiment Treebank, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*, (Seattle, USA, 2013).
20. M. Buhrmester, T. Kwang and S. D. Gosling, *Perspectives on Psychological Science* **6**, 3 (2011).
21. S. Kullback and R. Leibler, *The Annals of Mathematical Statistics* **22**, 79 (1951).
22. P. Rayson and R. Garside, Comparing corpora using frequency profiling, in *Proceedings of the Workshop on Comparing Corpora, held in conjunction with ACL 2000*, 2000.
23. K. Verspoor, K. B. Cohen and L. Hunter, *BMC Bioinformatics* , p. 10:183 (2009).
24. J. Mahmud, J. Nichols and Drews, *ACM Trans. Intell. Syst. Technol.* **5**, 47:1 (July 2014).