A NOVEL APPROACH TO TRACK PUBLIC EMOTIONS RELATED TO EPIDEMICS IN MULTILINGUAL DATA

Vinay Kumar Jain [1], Shishir Kumar [1], Neha Jain [1], Payal Verma [2]

[1] Jaypee University of Engineering and Technology, Guna, India
[2] Samara State Aerospace University, Samara, Russia

Emergence of new epidemic and re-appearance of older diseases causes great impact towards public health. Surveys based techniques which are costly and time-consuming are the most popular methods to measure information related to public health and used in decision making. Early monitoring of these epidemics helps in rapid decision making. Social media platforms provide rich source of information related to public health in forms of blogs, tweets, public posts etc., but these data is in unstructured form contains multiple languages words. This research focused on developing an automatic system for detecting public emotions related to epidemics in multilingual unstructured data to gain deeper understanding of public emotions and health related information. This approach gives timely information related to epidemics, corresponding symptoms, prevention techniques and awareness, which can help government and health agencies for rapid decision making. Experimental analysis of data set provides results that significantly beat the baseline term counting methods used for sentiment analysis.

Keywords: Social Media, Swine flu, Influenza, Naïve Bayes, H1N1

Introduction

Social media platforms produce a massive amount of unstructured data, which create lots of opportunities in the field of information retrieval and text mining to uncover hidden patterns [1]. India will account for the third-largest user base on micro-blogging site Twitter at 18.1 million by the end of this year [2].

The Internet provides one of the important resources in the field of surveillance systems for tracking disease outbreaks and helps in rapid decision making[3]. It provides an opportunity for low-cost and fast computation of data in comparison to existing traditional surveillance systems [4].

The medical diagnosis system worked in various levels to deal with different aspects of diagnosis such as the importance of symptoms, symptom patterns and the relationship between themselves. Delay in identifying the beginning of infectious epidemic results a big damage towards a society [5].An example is shown in Fig.1, which represent the year wise case of Influenza-A H1N1 [6].Consequently, there is strong interest in reducing these delays. This paper deals with developing an automatic real-time data analysis system which will detect and characterize unusual activity related to public health in the form of emotions.

Computational approaches to emotion analysis have focused on various emotion modalities, resulting in a large number of multi-modal emotion-annotated data. However, only limited work has been done in the direction of automatic recognition of emotion in text.

India social media user's share their opinions in multiple native languages and it is hard to detect emotions contains in their opinions. This paper present a framework which deals in identifying public emotions related to Swine Flu (H1N1) from multilingual data using two well known emotion theories given by Ekman [7] and Muni [8].

Literature Survey

Social media data provide a rich source of information for detecting and forecasting disease outbreak in all around the world. Earlier works carried out by authors are based on volume of keywords and sentiment analysis. Chew & Eysenbach [9] used specific keywords related for outbreak detection in 2009 H1N1 pandemic.Hu et al. [10] used Google web search queries related to influenza epidemic using specific keywords. Lampos & Cristianin [11] used content based methods with regression models to monitor and measure public concern and different levels of disease H1N1 pandemic. Chunara et al. [12] detected cholera outbreak using Twitter. Machine learning techniques like SVM for predicting influenza rates in Japan is used by Aramaki et al. [13]. Stewart et al. [14] developed a real time data analysis of disease using social media with an early warning system.Bodnar et al. [15] applied various classification techniques for detecting influenza. A low cost framework for tracking public health condition trends via Twitter is developed by Parket et al. [16].

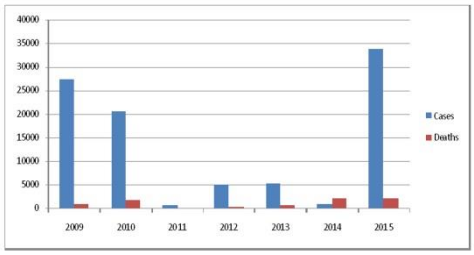

Fig. 1. Year wise cases of Influenza-A H1N1[6]

Human emotion can be sensed from such cues as facial expression, speech, gestures and writings. Research in emotion has focused on all these aspects [17]. Read et al. [18] carried out appraisal annotation of a corpus of book reviews, a genre that provide ample instances of the various kinds of appraisal classes. Mihalcea and Strapparava[19] present results in favor of automatic recognition of humor in texts. Neviarouskaya et al. [20] proposed a system for augmenting online conversations with a graphical representation (avatar) of the users. Ghazi et al. [21] used hierarchical classification to classify the six Ekman[7] emotions. Neviarouskaya et al. [22] developed EmoHeart, a lexical rule-based system in which emotion is detected from text and emotion expressions are visualized in a virtual environment. Strapparava et al. [23] developed a linguistic resource for lexical representation of affective knowledge named WordNet – Affect [24]. Dung et al. [25] exploited the thought that emotions are related to human mental states that are caused by some emotional events. Dey et al. [26] developed a system for extracting emotions from real time chat messenger Shaheen et al. [27] proposed a framework for classification of emotions in English sentences where the emotions are considered as generalized concepts extracted from the sentences.

Emotion Theories

Emotion recognition in text is one of the fundamental tasks of making computers to sense and respond towards the emotions. Emotion provides a complex, subjective experience which is accompanied by biological and behavioral changes. It involves activation of the nervous system which reflects in the forms of feeling, excitement, thinking, and physiological changes and behavioral changes such as facial expressions.Emotion theories such as Tomkins [28], Plutchik [29], Izard [30], Ortony et al. [31] and Ekman [7] are based on English language(roman) and are widely used in the field of physiology.

The most popular emotion theory in English language is given by Ekman [7] who substantiated Darwin's theory that human emotions are an evolved, biological response shared throughout cultures worldwide. The author gives six basic emotions, happiness, anger, sadness, disgust, surprise, and fear.

Emotion theory given by Muni [8] is based on Hindi language. According to Muni[8], emotions are the basic gastric juices which represent every part of the world into sentiments. Muni[8] divide emotion in eight class given in Table 5.

Data Collection

In this phase, identification of relevant tweets indicates the presence of flu or flu symptoms are fetched from the Twitter using Twitter API [32] with the help of relevant keywords and medical science terms. Relevant keywords collection methodology is based Jain & Kumar [3] which gives dynamic keywords which are popular during a particular time period and related to public sentiments. Some of the trending keywords are:

Keywords:{#SwineFlu,#flu,#H1N1,#Swine,#swinevirus,##h1n1,#influenza,#swinefluindia, #infuenzavirus, #delhiSwineflu ,#Fluvirus}

Data collected during different intervals of time is presented in Table 1.

Normalization of the tweets are carried out by applying text preprocessing steps like tokenization, stop words removal, duplicate tweets removal, stemming, lemmatization, feature weighting, dimensionality reduction and frequency based methods. Thus, the purpose of this step is to decrease the amount of noise from the tweets and filter out irrelevant tweets.

Table 1. No. of tweets collected weekly (Feb 2015 to Mar 2015)

Weeks	No. of tweets
01 Feb-07 Feb 2015	**7020**
08 Feb-14 Feb 2015	**10504**
15 Feb-21 Feb 2015	**22089**
22 Feb-28 Feb 2015	**17653**
1 Mar-07 Mar 2015	**10674**
08 Mar-14 Mar 2015	**7542**
15 Mar-21 Mar 2015	**8530**
21 Mar-28 Mar 2015	**7483**
Total	91495

Corpus Based Features

The corpus-based features exploit the statistical characteristics of the Twitter data set on the basis of the distribution of n-grams. In the experiments, unigrams (n=1) are used as features. Following most occurring unigrams related to swine flu symptoms, prevention methods and medicine are filtered out using count based technique and presented in Table 2.

Table 2. Relevant Occurring unigrams

Category	**Unigrams**
Symptoms	cold symptom, respiratory failure, cough, asthma, runny nose, blocked nose, problem in of breath, breathing difficulties, breathing trouble, pneumonia, sore throat, bronchitis, pain in the chest, tonsillitis, vomiting, abdominal pain, dizziness
Prevention	tulsi, kapoor, mask, face mask, wash your hands, avoid touching your eyes, nose, and mouth
Medicine	tamiflu, flu vaccine, garlic, ayurvedic medicine, homeopathy medicine, turmeric, tulsi, neem

Proposed Method

Emotion detection in text is referring to as a classification problem where various nominal labels are assigned to a sentence from a group of target emotion labels.Proposed framework for emotion detection in tweets given as follow:

Let t is a tweet and k is an emotion label. Let e be a set of n possible emotion categories (excluding neutral) where e = {e1, e2, e3… en}. The objective is to label t with the best possible emotion label k, where k ∈ {e1 , e2,e3.. en, neutral}. For the classification experiments, Naïve Bayes (NB) algorithm is used for detection of emotions.

Table 3. Category distribution in the Twitter dataset used in emotion/non-emotion classification

Category	**Number of Tweets**	**Percentage**
Emotion	17592	34.4%
Non-emotion	33600	65.6%
Total	51192	100%

The dataset collected from Twitter is divided into two high-level categories, namely, emotion and non-emotion. The distribution of this dataset is shown in Table 3. Feature words are taken from two most popular dictionaries WordNet-Affect [24] and Hindi WordNet-Affect [33]. The Naïve Bayes classifier is used as baseline for the classification and gives an accuracy of 65.6 %, in case of Non-emotion category given in Table 3.After classification of tweets into emotion and non-emotion categories, identification of fine-grained emotion categories according to Ekman [7] and Muni [8] models has been performed.

Emotion in a tweet can be discerned on the basis of its lexical content. For detection of lexical content naïve based is used. The presence of one or more seed words of a particular emotion category in a tweet provides a good premise for interpreting the overall emotion of the tweet. This kind of approach relies on a list of words with prior knowledge about their emotion type, and uses it for tweet-level classification.

Experimental Results

For evaluation purposes, a baseline system has been developed that counts the number of emotion words of each category in a tweet, and then assigns the category with the largest number of words to the corresponding tweet. For obtaining prior knowledge about emotion-bearing words, seeds words related to emotions words have been extracted from WordNet-Affect [24] and

Hindi WordNet-Affect [33] into basic emotion categories. Two different sets of experiments were performed to test the effectiveness and contribution of the different feature groups:

1. Using only features words from WordNet-Affect [24] (WNA) in Ekman Model presented in Fig.2.

2. Using features from WordNet-Affect [24] (WNA) + Hindi WordNet-Affect [33] (HWNA) in Muni [8] Model presented in Fig.3.

A comparative performance evaluation of Naïve bayes algorithm in terms of correctly predicted emotions containing in tweets has been examined using F-Measure given calculated using Equation (3).

$$\Pr ecision(P) = \frac{TE}{TE + FE} \quad (1)$$

$$\mathrm{Re}\, call(R) = \frac{TE}{TE + FN} \quad (2)$$

$$F - Measure = \frac{2P * R}{P + R} \quad (3)$$

Where,

TE (True Emotional)- No. of Sentences that are correctly classified as emotional.

FE (False Emotional)- No. of sentences incorrectly classified as emotional.

TN (True Neutral) – No.of sentences correctly classified as Non-emotional or neutral.

FN (False Neutral) – No. of sentences incorrectly classified as neutral

The results are explained in terms of precision, recall and F-measure using two models and represented in Table 4 and Table 5. The results presented in *Fig. 2* and *Fig. 3* shows that combination of features words from WordNet-Affect [24] (WNA) with Hindi WordNet-Affect [33] (HWNA) in Muni [8] Model detects maximum emotional words in tweets as compared to Ekman model. This approach showed that combination of different emotion theories can help in better emotion detection in multilingual unstructured data.

During epidemics or any event these emotion classes present in Fig. 2 and Fig. 3 help in determining the public emotions. For example, public feel joy and fear corresponds to H1N1 in Fig. 2 and Soka (Sorrow) and Rati(Love) from Fig. 3. Levels of emotion classes represent public moods and hence help in decision making.

Conclusion

Social media data offers unique challenges and opportunities for monitoring and surveillance public health. This paper presented an effective approach for tracking public emotions during epidemic of H1N1 2015 in India using two popular emotion theories given by Ekman and Muni. This approach could give rapid information of public feeling related to epidemic. Corpus based analysis gives various meaningful inference such as symptoms corresponding swine flu, prevention techniques etc. Results showed that combination of Ekman and Muni models performed better in classify the emotions in multilingual tweets. Presented results also support that

the use of social media for tracking public emotions will be utilized for knowing the public health in the society.

Possibilities of improvement in results are also there, when proposed approach will be applied for the events of those countries where social media is used by most of the citizens for expressing their opinions.

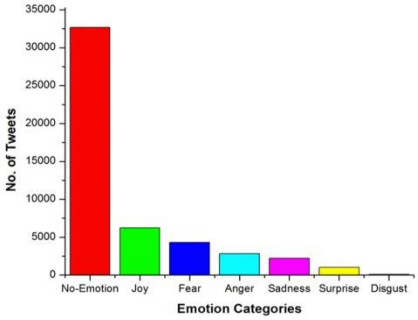

Fig. 2. Emotion detection using Ekman Model

Table 4. Performance metrics of the baseline system (Ekman model)

Class	Precision	Recall	F-Measure
Happiness	0.519	0.320	0.396
Sadness	0.527	0.283	0.368
Anger	0.612	0.212	0.312
Disgust	0.914	0.097	0.176
Surprise	0.320	0.216	0.258
Fear	0.714	0.325	0.447
No-emotion	0.621	0.467	0.534

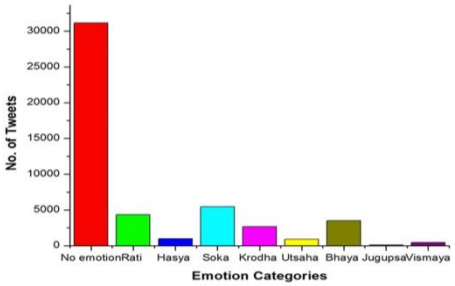

Fig. 3. Emotion detection using Muni Model

Table 5. Performance metrics of the baseline system (Bharata muni Model)

Class	Precision	Recall	F-Measure
Rati (Love)	0.512	0.310	0.389
Hasya (Mirth)	0.501	0.213	0.299
Soka (Sorrow)	0.681	0.262	0.379
Krodha (Anger)	0.868	0.119	0.209
Utsaha (Energy)	0.318	0.296	0.306
Bhaya (Terror)	0.804	0.315	0.453
Jugupsa (Disgust)	0.934	0.067	0.125
Vismaya (Astonishment)	0.534	0.667	0.593
No emotion	0.627	0.383	0.475

References

1. Pang B and Lee L. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval.2008; 2:1–135.
2. Emarketer. http://www.emarketer. .com. Date accessed: 01/10/2015.
3. Jain V K, Kumar S. An Effective Approach to Track Levels of Influenza-A (H1N1) Pandemic in India Using Twitter. Procedia Computer Science.2015;70(1):801–807.
4. Ceron AC. Using Sentiment Analysis to Monitor Electoral Campaigns: Method Matters—Evidence From the United States and Italy. Social Science Computer Review.2014;33(1):3-20.
5. Glik D. Risk communication for public health emergencies. Annual Reviews of Public Health;2007;p.33-54.
6. Preventive Measures for Swine Flu http:// pib.nic.in/newsite/ PrintRelease.aspx ?relid=115710. Date accessed: 01/05/2015.
7. Ekman P. An argument for basic emotions. Cognition and Emotion.1992; 6(1):169–200.
8. Raghavan V. The Number of Rasa-S. Theosophical Publishing House;1967.
9. Chew C M. Pandemics in the age of twitter: A content analysis of the 2009 h1n1 outbreak. Master's thesis. University of Toronto;2010.
10. Hu X., Tang L, and Liu H. Enhancing accessibility of microblogging messages using semantic knowledge. In Proceedings of the 20th ACM International Conference on Information and Knowledge Management, New York, USA.2011;p.2465-2468. 2011
11. Lampos V and Cristianini N. Tracking the flu pandemic by monitoring the social web. In 2nd IAPR Workshop on Cognitive Information Processing (CIP 2010),IEEE Press.2010;p.411–416.
12. Chunara R, Andrews JR, Brownstein JS. Social and news media enable estimation of epidemiological patterns early in the 2010 haitian cholera outbreak. Am J Trop Med Hyg.2012;86(1):39-45.
13. Aramaki E , Maskawa S , Morita M. Twitter catches the u: detecting inuenza epidemics using Twitter. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics.2011;p.1568-1576.
14. Stewart A and Diaz E. Epidemic intelligence: for the crowd, by the crowd .In Proceedings of the 12th international conference on Web Engineering, ICWE'12,Berlin, Heidelberg,Springer-Verlag.2012;p.504–505.
15. Bodnar T, Barclay VC, Ram N, Tucker CS and Salathé M. On the ground validation of online diagnosis with twitter and medical records. WWW Companion.2014;p.651–656.
16. Parker J, Wei Y, Yates A, Frieder O, Goharian N.A framework for detecting public health trends with Twitter, Proceeding ASONAM '13 Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining.2013;p.556-563.
17. Cowie R, Douglas-Cowie E ,Tsapatsoulis N, Votsis G, Kollias S, Fellenz W & Taylor J. Emotion recognition in human-computer interaction. IEEE Signal Processing Magazine.2011;18(1),32–80.
18. Read J, Hope D and Carroll J. Annotating expressions of Appraisal in English. Proceedings of the ACL-2007 Linguistic Annotation Workshop. Prague Czech Republic. June 2007;p.93-100.
19. Mihalcea R and Strapparava C. Making Computers Laugh: Investigations in Automatic Humor Recognition, In Proceedings of the Joint Conference on Human Language Technology/Empirical Methods in Natural Language Processing (HLT/EMNLP) Vancouver, Canada.2005;p.531-538.
20. Neviarouskaya A, Prendinger H and Ishizuka M. Analysis of affect expressed through the evolving language of online communication. In Proceedings of the 12th International Conference on Intelligent User Interfaces (IUI-07), Honolulu, Hawaii, USA.2007;p.278-281.
21. Ghazi D, Inkpen D, and Szpakowicz S. Hierarchical versus flat classification of emotions in text. Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text. 2010;p.140-146.
22. Neviarouskaya A, Prendinger H and Ishizuka M. Narrowing the Social Gap among People involved in Global Dialog: Automatic Emotion Detection in Blog Posts, In Proceedings of the International Conference on Weblogs and Social Media (ICWSM 2007), Boulder, Colorado, USA.2007; p.293-294.
23. Strapparava C, Valitutti A and Stock O. The affective weight of lexicon. In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006), Genoa, Italy.2006; p.423–426.
24. Miller GA.WordNet: a lexical database for English. Communications of the ACM 38.11.1995; p.39-41.
25. Dung and Cao TH.A high-order hidden Markov model for emotion detection from textual data. Knowledge Managementand Acquisition for Intelligent Systems. Springer Berlin Heidelberg.2012;p.94-105.
26. Dey L, Afroz N and RPD Nath. Emotion extraction from real time chat messenger.In proceedings of 3rd International Conference On Informatics, Electronics & Vision,Dhaka. May 2014;p.1-5.

27. Shaheen S,El-Hajj W, Hajj H, Elbassuoni S. Emotion Recognition from Text Based on Automatically Generated Rules, 2014 IEEE International Conference on Data Mining Workshop., Shenzhen. Dec 2014;p.383-382.
28. Tomkins SS. Affect, imagery, consciousness. The positive affects. New York, Springer; 1962.
29. Plutchik R. A General Psycho evolutionary Theory of Emotion. In Plutchik, R. and Kellerman, H. (eds.), Emotion: Theory, Research and Experience: Vol. 1. Theories of Emotions. New York: Academic; 1980.p.3-33.
30. Izard CE. Human emotions. New York: Plenum Press; 1977.
31. Ortony A, Clore GL and Collins A. The Cognitive Structure of Emotions. Cambridge University Press; 1988.
32. Twitter Developer Page. https://dev.twitter.com/docs/.Date accessed: 01/01/2015.
33. Hindi Word Net [Internet]. IIT Mumbai; 2015 [Cited 2015 Nov 15] Available from: http://www.cfilt.iitb.ac.in/~wordnet/wn.old/