

Analysis of Public Health Concerns using Two-step Sentiment Classification

P. Naresh Behera
Department of Computer Science
JNT University Kakinada
Kakinada, India -533003

Mrs. Suneetha Eluri
Assistant Professor,
Department of Computer Science,
University Kakinada
Kakinada, India -533003

Abstract— Our aim is to develop a sentiment analysis tool for public health officials to monitor the spreading epidemics in a certain region and time period. Analyzing the public concerns and emotions about health related matters is an important issue to know the spreading of a disease. In this work, sentiment classification of Twitter messages is focused to measure the Degree of Concern (DOC) of the people about a disease spreading. In order to achieve this goal, the disease related tweets are extracted based on time and geographical location. Then, a novel two-step sentiment classification is applied to identify the personal negative tweets. First, the clue-based algorithm is used to classify the personal tweets from non personal tweets by using subjectivity clues. Next, lexicon-based algorithm and Naïve Bayes classifiers are applied to classify negative and non-negative personal tweets. The personal negative tweets are used to measure Degree of Concern. The Public Health Surveillance System (PHSS) is also developed by using visualization techniques such as maps, graphs and charts to visualize the Degree of Concern (DOC) of the epidemic related twitter data. The visual concern graphs and charts can help health specialists to monitor the progression and peaks of health concerns of people for a disease in particular space and time, so that necessary preventive actions can be taken by public health officials. Negation Handling and Laplacian Smoothing techniques are used with Lexicon Based classifier and Naïve Bayes classifier to improve performance.

Keywords— Degree of concern; Disease Spreading; Public Health; Polarity; Sentiment Analysis; Social Network; Twitter

I. INTRODUCTION

A. Sentiment Analysis

Whenever we need to make a decision we need to know other's views, opinions and advice. It is necessary for both individual and organizations. With exponential growth of the social network content in the internet, the views and opinions of people can be easily extracted. These users not only use the available resources in the web, but also give their feedback, so that additional useful information is generated. To evaluate and analyze this huge amount of information, Sentiment Analysis is originated.

Sentiment Analysis or Opinion Mining [5] is a task that extracts information from social networks and it identifies the user's opinions, views and emotional feelings in the form of positive, negative and neutral, by using Natural Language Processing technique. Sentiment Analysis on social media is widely used in different areas [6], such as marketing, business, election prediction, education, medical and

communication. But the recent challenge task for sentiment analysis is on health related data such as public health surveillance, disease ontology, health maps, spreading epidemics, disease detection etc.

B. Monitoring Public Health Concern

Monitoring the Public Health, disease spreading and controlling it, are the important responsibilities for Public Health Officials. They analyze the public opinions, emotions and concerns about health related matters when there is an indication of a sudden disease outbreak. Different monitoring strategies have been developed to analyze the public health. These strategies include household surveys, laboratory based surveillance, sentinel surveillance systems, and the most-recent IDSR (Integrated Disease Surveillance and Response) [7]. Among these strategies, people's emotional changes, due to sudden disease outbreak, have caught increasing attention of health officials. X.Zhu et al. [16] analyzed the mental state of people of china during the outbreak of SARS (2003).Based on their analysis, during the outbreak 94.6% people are surveyed, and reported the emotional changes. Among them 54.8% are panic, 34.0% are nervousness, 7.6% are fear and 23.3% are admitted to irrational behaviors such as seeking shelters, going on a shopping spree etc. Thus, it is critical to monitor health issues for public health officials and Government decision makers. However, it is hard to monitor public health and their emotional changes using traditional surveillance system. The existing methods such as, questionnaires and clinical tests are very slow and can only cover limited number of people. A novel system must be developed to supplement the existing system. This tool must track the real time statistics of public emotions related to different health issues, to provide early warning and to help public health officials and government decision makers to prevent necessary actions.

Social networks, such as Google news, blogs, search engines, twitter, facebook etc. has abundant resources for monitoring threats of public health. Twitter, a micro-blog service provider has many advantages than others for disease surveillance. Twitter has more than 500 million users posted more than 400 millions tweets per day. It is up-to-date and most tweets are public related. It is fixed length message i.e., 140 characters. Twitter API [15] enables to extract the tweets along with related information, such as, geographic location, time and hyperlinks.

In this work, Twitter is used as ultimate resource for extract the opinions of public related to health matters. It helps the government decision makers and public health officials to gauge the degree of concern (DOC) calculated in the tweets of Twitter users who are under impact of disease. The early detection of public health concerns can assist health officials to take timely decision to counter rumors, thus prevent potential social crises. In order to calculate DOC of user tweets, a classification technique is developed to analyze the sentiments of disease related tweets. This technique involves two steps. First, using subjectivity clues it separates personal tweets from news (non-personal) tweets. Personal tweets are posted by individual users and non-personal tweets or news tweets are released by online media and possibly re-tweeted by twitter users. In second stage, sentiment classification is used to classify the personal negative tweets from personal non-negative (neutral) tweets by using Lexicon based classifier and Naïve bayes classifier. Finally, Public Health Surveillance System (PHSS) is a visualization system with graphs and charts used to visualize DOC to the public health officials.

II. RELATED WORK

A. Sentiment Classification using Twitter

In sentiment analysis, B.Pang et al. trained an algorithm to classify the sentiments of online movie reviews. Pandey and Iyer [12] proved that instead of using common text features used in traditional information retrieval tasks, the domain specific features has more significance. Barbosa and Feng [13] focused on the process in which, the training data is automatically generated. They used three sources: Tweet Feel, Twitter Sentiment, and Twendz to label the sentiments of tweets. The Naïve Bayes classifier is reported by Yu et al. [14] as the best in terms of precision and recall, when applied to sentiment classification of news articles.

B. Monitoring of Disease Spreading

Ginsberg, [1] used search engines to analyze the sentiments of users based on their queries. Aramaki, [4] used different Machine Learning methods to classify epidemic-related tweets into two classes (positive or negative). Collier et al. developed a model that classifies the Twitter messages automatically into six fixed syndromic categories, such as Respiratory and Gastrointestinal. Signorini et al. analysed H1N1-related tweets using a SVM-based estimator, and estimated the ILI rate before the official announcement by one to two weeks. Using online news, Brownstein et al. [20] developed the system, Health- Map, which collects reports from Google News and classifies the news into disease related and unrelated reports and filter the disease related new into “warnings” “Breaking News” and “old news”. Similarly, Culotta [2] correlated user tweets with CDC statistics using number of regression models and using a large number of Twitter messages they provided a relatively simple method to track the ILI rate. Lamos [18] et al. used a method which helps to compute flu scores using a set of markers, and get a high association with HPA flu score, which is equal to the CDC score in UK. Salathé and handelwal [19] analysed the reaction of the Twitter users towards the H1N1 vaccine using sentiment analysis. They categorised user tweets into four

categories: positive, negative, neutral, and irrelevant. They used the relative difference of positive and negative messages and then calculated the H1N1 vaccine sentiment score.

III. IMPLEMENTATION ISSUES

A. System Architecture

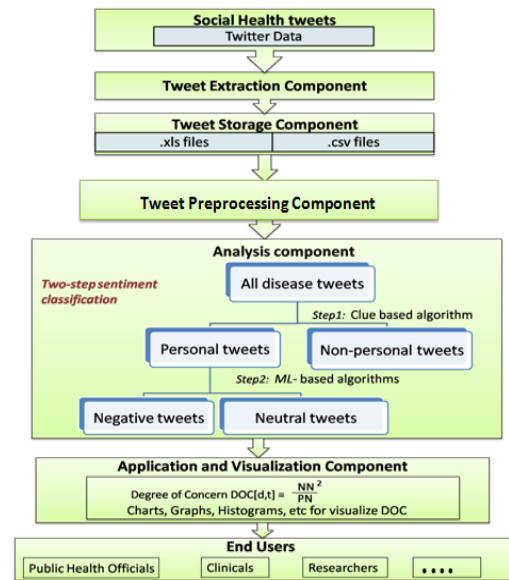


Fig: System Architecture

B. Tweet Extraction

Using Twitter OAuth Authentication[15] the disease related tweets are extracted based on time and geographical location(latitude and longitude) using the keywords related to diseases. The disease related keyword, time and geo-code are used as parameters to extract the disease related tweets. The geo-code involves latitude, longitude and radius (miles or kilometres). For Example; if we want to extract the tweets form India then the geo-code will be, 78,21,3000mi or 78,21, 3700km. The latitude and longitude may be either positive or negative.

C. Tweet Preprocessing

In pre-processing techniques we removed the following:

- Remove Special symbols and digits:** The digits, punctuation marks and symbols like !, @, #, \$, %, ^, +, -, *, / etc. are removed.
- Remove URLs:** The URLs in tweets are removed.
- Remove Duplicate Tweets:** The tweets which start with ‘RT’ and repeated tweets are removed.
- Remove Stop words:** The stop words i.e. a, an, the, if, on, of etc. are removed.
- Normalize Elongated words:** Some characters are repeated multiple times in a word, for example instead of good, the word typed as gooooood or for super, suuuupppppeeerrr should be normalized.
- Replace emoticons:** The emoticons like, (:,): D: D etc. are replaced by their correspondent emotional words based on the logical meaning of the emoticons.

TABLE-I: LIST OF EMOTICONS

negative	neutral	positive
:(:	:)
:(:)
:-(-		:-)
(:D
:(:P

D. Tweet Classification

A Novel Two-Step Sentiment Analysis Technique [2] is used:

Step-1: A clue based algorithm is used to classify the personal and non-personal tweets using subjectivity clues [10].

Step-2: Lexicon Based algorithm and Naive Baye's algorithm is applied to personal tweets for classifying positive, negative and neutral tweets.

Negation Handling and Laplacian Smoothing is also used for improving accuracy of classification.

E. Calculating Degree of Concern (DOC)

The personal negative tweets are used to measure the Degree of Concern, DOC [d, t], for a particular disease 'd' and a particular time 't'.

$$DOC [d,t] = \frac{NN^2}{PN} \quad \text{-----(1)}$$

d- a particular disease

t- a particular time

NN- number of negative personal tweets

PN-number of personal tweets

F. Visualization

The experiment is done with three diseases, Malaria, Cancer and Swine-Flu. The visual concern graphs and charts are used to visualize the Degree of Concern of these three diseases.

IV. ALGORITHMS

A. Clue based classifier for Personal Tweets

Clue-based classifier divides each tweet into a set of words

and matches them with a corpus of personal clues. For personal versus non-personal classification, subjective corpus [10] are used, if there are enough subjective clues in the tweet, it can be regarded as personal tweet, otherwise it is a news tweet. The corpus from the literature [9] contains 8,221 words, 5569 clues are strongly subjective clues and 2652 clues are weakly subjective clues.

We counted the number of strongly subjective [9] terms, the number of weakly subjective terms, in each tweet and experimented with different thresholds. A tweet is classified as personal if its count of subjective words exceeds the chosen threshold; otherwise it is classified as a non-personal tweet.

B. Twitter Sentiment Classifiers

Lexicon based algorithm and Naïve Bayes algorithms are used to classify the polarity[14] of tweets, such as, positive, negative and neutral. Negation Handling and Laplacian Smoothing techniques are used to improve the accuracy of the classifiers

C. Lexicon-Based Classifier

- Step: 1 Divide a message M into words
 $M_i = \{w_1, w_2, w_3, \dots\}, i=1, 2, n$
- Step 2: for each w_i , compare with data dictionary of +ve and -ve words and Return +ve polarity and -ve polarity.
- Step 3: Calculate overall polarity of a word = sum(+ve polarity) - sum(-ve polarity)
- Step 4: Repeat step 2 until end of words
- Step 5: add the polarities of all words of a message i.e. total polarity of a message.
- Step 6: Based on that polarity, message can be positive or negative or neutral.
- Step 7: repeat step 1 until M is NULL

D. Naïve Bayes Classifier

Probability of a word belonging to a particular class is given by the expression:

$$P(x_i|c) = \frac{\text{Count of } x_i \text{ in message of class } c}{\text{Total no. of words in messages of class } c} \quad \text{-----(2)}$$

According to the Bayes Rule, the probability of a particular tweet 'd' belonging to a class C_i is given by,

$$P(c_i | d) = \frac{P(d | c_i) * P(c_i)}{P(d)} \quad \text{-----(3)}$$

$$P(c_i | d) = \frac{(\prod P(x_j | c_j)) * P(c_j)}{P(d)} \quad \text{-----(4)}$$

- $P(C_i | d)$ = probability of instance d being in class C_i
- $P(d | C_i)$ = probability of generating instance d in given class C_i
- $P(C_i)$ = probability of occurrence of class
- $P(d)$ = probability of instance d occurring

E. Laplacian Smoothing

If the classifier encounters a word that has not been seen in the training set, the probability of both the classes would become zero and there won't be anything to compare between. This problem can be solved by Laplacian smoothing,

$$P(x_i | c_j) = \frac{\text{Count}(x_i) + k}{(k + 1) * (\text{No. of words in class } c_j)} \quad \text{-----(5)}$$

Usually, k is chosen is 1

F. Negation Handling

Algorithm:-

Negated: = False

For each word in document:

If negated = True:

Transform word to “not ” + word.

If word is “not” or “n’t”:

If a punctuation mark is encountered

Negated: = False

V. RESULTS

A. Classification of Disease related Tweets

The tweets are extracted based on keywords of the major diseases malaria, cancer and swine flu and preprocessed the tweets. Then, sentiment Analysis is used to find the sentiments of each tweets of every disease.

date	text	emotion	polarity
1	12-07-2015 after decades of denial national cancer institute finally admits that "cannabis kills cancer"	unknown	negative
2	12-07-2015 weed does not slow the growth of cancer cells it actually speeds up growth by up to	unknown	negative
3	12-07-2015 cancer have a hard time saying goodbye	unknown	negative
4	12-07-2015 cancer have a hard time saying goodbye	unknown	negative
5	11-07-2015 studies show omega in flaxseedcan reduce growth of cancer cells	unknown	negative
6	11-07-2015 augustat pm detect throat cancer early	unknown	negative
7	11-07-2015 ppl that are pro life are always like it could have cured cancer mhm it could also be the next hitler hoe shut up	joy	neutral
8	11-07-2015 weed does not slow the growth of cancer cells it actually speeds up growth by up to	unknown	negative
9	11-07-2015 lies virgo buzas a pisces when you have a bad day you call a cancer	sadness	negative
10	11-07-2015 hithanks for the post i would like to discuss more about breastcancer medicalresearch	joy	positive
11	11-07-2015 rt i told her her eye was looking abnormal and her previous breast cancer dr said theres a chance its cancer because	unknown	negative
12	11-07-2015 hellothanks for posting can we connect about breastcancerresearch medicalresearch	unknown	positive
13	10-07-2015 normal is nice msu softball player winning cancer fight	unknown	neutral
14	10-07-2015 rt i love when archers ears are ringing from explosives and gunshots breast cancer episodes onnow	joy	negative
15	10-07-2015 "what if an aborted fetus would've found the cure to cancer"	unknown	positive
16	10-07-2015 rough rider leg knife cancer kb stainless clip blade folding knife	unknown	negative
17	10-07-2015 thanks for posting would love to talk more about breastcancerresearch medicalresearch	joy	positive
18	09-07-2015 president jimmy carter talks about his cancer	unknown	negative
19	09-07-2015 rt register free methodismadhatther celebrating breast cancer fighters ampamp survivors addfw	unknown	neutral
20	09-07-2015 lets break the internet spread the wordopen some eyes its true kids do get cancer lets do somethi...	sadness	negative
21	09-07-2015 omg so im doing a mud run tomorrow to raise money and awareness for a cancer organization and its for a video sc	unknown	negative
22	09-07-2015 fuck you cancer	unknown	negative
23	09-07-2015 prez jimmycarter has more dignity in tip of his finger thanamp his birther dad combined	unknown	positive
24	09-07-2015 weed does not slow the growth of cancer cells it actually speeds up growth by up to	unknown	negative

Fig-1: polarity and emotion classification of cancer related tweets

Total tweets extracted are, for malaria 1500, for cancer 1500 and for swine flu 1498. Among them, polarities for malaria (positive=800, negative=470 and neutral= 230), for cancer (positive=300, negative=850 and neutral =350) and for swine flu (positive=630, negative=770 and neutral =100).

The Degree of Concern and count of disease (i.e. malaria, cancer and swine flu) related tweets and their polarities are shown below:

	parameters	cancer	swineflu	malaria
1	Total Tweets	1499	1493	1496
2	Total negative tweets	852	752	466
3	Total positive tweets	293	639	795
4	Total neutral tweets	354	102	235
5	DOC	484.2588	378.7703	145.1578

Fig -2:=Degree of concern of different diseases

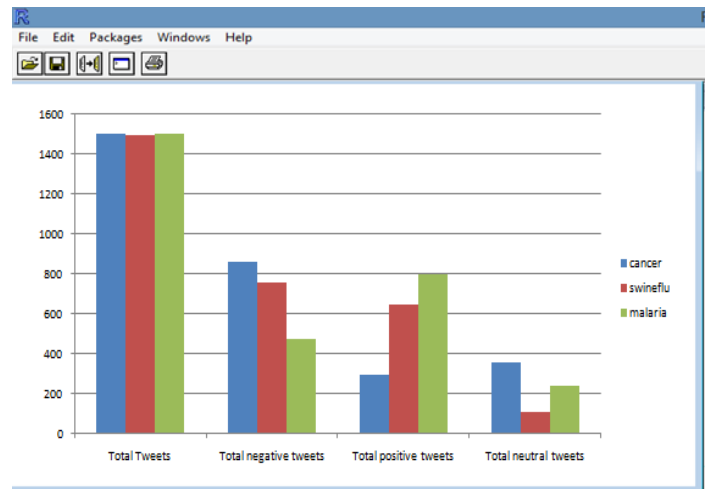


Fig -3:-No. Of Total tweets, negative, positive and neutral tweets for the diseases cancer, swine flu and malaria

B. Month-wise Degree of Concern

The disease related tweets are extracted and analyzed from the march 2015 to august 2015 and analyzed the tweets by calculating DOC for every month.

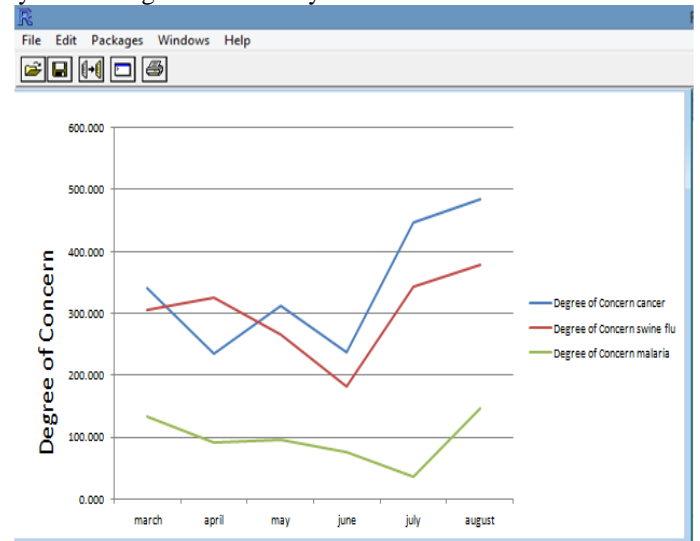


Fig-4:- Month-wise Degree of concerns for diseases

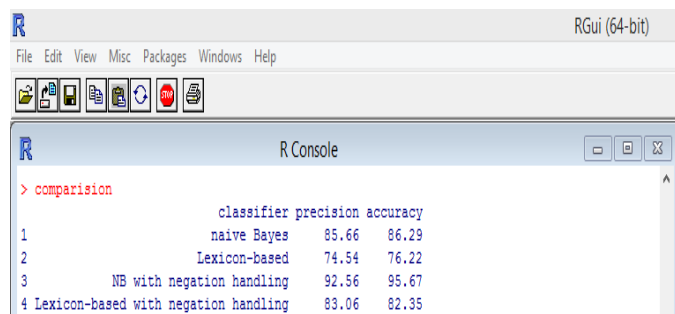
C. Comparison of Various techniques used

Precision and Accuracy is computed to compare performance of various algorithms.

$$\text{Precision} = \frac{\text{number of true positives}}{\text{number of true positives} + \text{false positives}} \quad \text{---- (6)}$$

$$\text{Accuracy} = \frac{\text{no. of true positives}}{\text{no. of true positives} + \text{false positives} + \text{false negatives} + \text{true negatives}} \quad \text{---- (7)}$$

The tweets which are classified by sentiment classifiers are compared with manually classified tweets and calculated the Precision and Accuracy. For simple Naïve Bayes classifier, precision and accuracy are 85.66% and 86.29%, for Lexicon based classifier precision=74.54% and accuracy=76.22%. Then, the negation handling technique is combined to both the classifiers and performance is improved, which are shown in below figure.



	classifier	precision	accuracy
1	naive Bayes	85.66	86.29
2	Lexicon-based	74.54	76.22
3	NB with negation handling	92.56	95.67
4	Lexicon-based with negation handling	83.06	82.35

Fig-5: Performance measure of sentiment classifiers

VI. CONCLUSION AND FUTURE WORK

This work presents the tweet classification approach to identify the negative sentiment of personal health tweets to measure the degree of concern (DOC) for monitoring the public sentiments for a disease. The charts, table and graphs are developed to visualize the DOC. If the DOC is more for a disease in a particular location and period means spreading of that disease is more in that region. So that, public health officials will take preventive actions.

We can extend the number of disease events to be monitored by implementing disease ontology. We can also use the symptoms of various diseases to detect and predict the disease. We can analyze the death toll rate due to spreading of diseases. In Addition with twitter, extract input from facebook, personal blogs, news forums etc.

REFERENCES

- [1] Ginsberg, M. H. Mohebbi, M. S. Smolinski, and L. Brilliant, "Detecting influenza epidemics using search engine query data," *Nature* 457, 2009, pp. 1012-1014.
- [2] Ji, X., Chun, S. A. and Geller, J. Monitoring Public Health Concerns Using Twitter Sentiment Classifications. In *Proceedings of International Conference on Health Informatics*, Philadelphia, PA, 2013.
- [3] A. Culotta, "Towards detecting inuenza epidemics by analyzing Twitter messages," 1st Workshop on Social Media Analytics (SOMA '10), Washington, DC, USA, 2010..
- [4] E. Aramaki, S. Maskawa, and M. Morita, "Twitter Catches The Flu: Detecting Influenza Epidemics using Twitter," *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2011. R. Nicole, "Title of paper with only first word capitalized," *J. Name Stand. Abbrev.*, in press.
- [5] Jayashri Khairnar, Mayura Kinikar, "Machine Learning Algorithms for Opinion Mining and Sentiment Classification" *International Journal of Scientific and Research Publications*, Volume 3, Issue 6, June 2013, ISSN 2250-3153.
- [6] Yongin, South Krea; Khattak, A.M.; Sungyoung Lee; Maqbool, J, "Precise Tweet Classification and Sentiment Analysis" Published in:
- [7] Disease Control Priorities Project, <http://www.dcp2.org/file/153/dcpp-surveillance.pdf>, accessed on 02/15/2013..
- [8] Stopwords, http://web.njit.edu/~xj25/eosds_beta/files/newsstopword.xlsx
- [9] E. Riloff and J. Wiebe, "Learning Extraction Patterns for Subjective Expressions," In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP-03)*, 2003.
- [10] Subjectivity lexicon, <http://www.cs.pitt.edu/mpqa/>, accessed on 7/15/2012.
- [11] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, Vol. 2, No 1-2 pp. 1-135, 2008.
- [12] V. Pandey and C.V.K. Iyer, "Sentiment Analysis of Microblogs," *Technical Report*, Stanford University, 2009
- [13] L. Barbosa, and J. Feng, "Robust Sentiment Detection on Twitter from Biased and Noisy Data," In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, 2010.
- [14] H. Yu and V. Hatzivassiloglou, "Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences," In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, 2003.
- [15] Twitter developers documentation, <https://apps.twitter.com/docs>, accessed on 2/15/2015.
- [16] X. Zhu, S. Wu, D. Miao, and Y. Li, "Changes in Emotion of The Chinese Public In Regard to The SARS Period," *Social Behavior & Personality*, Vol. 36, Issue 4, pp. 447, 2008.
- [17] B. Pang, L. Lee, S. Vaithyanathan, "Thumbs up? sentiment classification using machine learning techniques," In *proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 79-86. July 2002.
- [18] Vasileios Lampos and Nello Cristianini, "Tracking the flu pandemic by monitoring the Social Web," *2nd International Workshop on Cognitive Information Processing (CIP)*, 2010.
- [19] M. Salathé and S. Khandelwal, "Assessing Vaccination Sentiments with Online Social Media: Implications for Infectious Disease Dynamics and Control," *PLoS Comput Biol* 7(10): e1002199. doi:10.1371/journal.pcbi.1002199, 2011.
- [20] J. S. Brownstein, C. C. Freifeld, B. Y. Reis, and K. D. Mandl, "Surveillance Sans Frontières: Internet-Based Emerging Infectious Disease Intelligence and the HealthMap Project," *PLoS Med* 5(7): e151. doi:10.1371/journal.pmed.0050151, 2008.