Sentiment Analysis and Information Diffusion on Social Media: The Case of the Zika Virus

Chuan-Jun Su and Jorge A. Quan Yon

***Abstract*—First identified 50 years ago, the Zika virus has recently made global headlines due to a high profile outbreak in Brazil coinciding with the Olympics. Mentions of Zika on social media platforms exploded following initial reports of the outbreak, and this unprecedented surge of heterogeneous data can be processed using Big Data analysis techniques to acquire further insights and knowledge into general public opinion. Twitter data streams have previously been used to predict outcomes of real world events. Twitter data filtered for the keyword "Zika" was subjected to analysis using a sentiment analysis lexicon-based framework to establish the polarity of the messages. The World Health Organization (WHO) recommends avoiding exposure to Zika-infected mosquitoes as the most effective approach to prevention. The diffusion of Twitter messages citing the WHO recommendations is analyzed to help public health professionals and health agencies formulate an effective response. Our results show that the WHO recommendations were largely ignored in Twitter-based discussions related to the Zika virus.**

***Index Terms*— Zika virus, sentiment analysis, social media, twitter, vector control, big data.**

I. Introduction

First identified in 1947 in Uganda, the Zika virus is currently experiencing a major outbreak in South and Central America. Zika has been found to cause microcephaly in human fetuses, and no vaccine or treatment is currently available. Diagnosis is also difficult and, though the disease exhibits symptoms similar to those of Dengue and Chikungunya, no test is commercially available. The World Health Organization (WHO) has declared Zika to be a 'Global Emergency'.

Public opinion can be a useful resource to public health professionals seeking to better understand the diffusion of the health information. This can be especially advantageous in formulating communication and educational strategies to counter emergency threats. To this end, public health institutions and organizations typically conduct face-to-face, telephone or online surveys to understand and measure public attitudes and behavioral responses to health-related products and services [1]. Face-to-face surveys suffer from several disadvantages, such as high cost per respondent, geographical limitations, interviewer bias, and time pressure on respondents [2], [3]. Although telephone surveys offer better geographical coverage and lower cost while still maintaining a degree of personal interaction, they suffer from reduced response rates, interviewer bias, and the inability to use visual aids [4]. Online surveys have cost and efficiency advantages, while also allowing the use of interactive visual prompts and offering increased flexibility which can increase response rates. Online surveys can take advantage of new tools like Skype videoconferencing, social networking site-based surveys, and surveys optimized for mobile devices [5].

Both individuals and organization rely on the opinions of trusted sources in making decisions [6]. "Social media" has emerged as an important and efficient mass communications tool, and is increasingly used by individuals to express feelings towards products, services, issues or events that impact their daily lives. Social media typically refers to social network sites (SNSs) such as Facebook, video or image sharing sites such as YouTube, and venues for sharing opinions and comments such as blogs and Twitter. The creation and sharing of information and ideas between people is a central feature of social media [7]. Traditional surveys have been used to research public opinion responses to H1N1 [8], obesity [9], and bioterrorist attacks, but such methods require significant time and resources to execute [10].

Social media platforms like Twitter allow users to easily connect and share information. This constant interaction between billions of users makes social media an ideal place for mining trends and patterns of interest through the use of Big Data analysis techniques. Tweets have been used to provide real-time insight into dynamic issues such as the outcomes of HIV Young, Rivers and Lewis [11], to track and predict flu dissemination patterns [12], [13], suicide risk factors [14], and even the prevalence of healthy/unhealthy foods, with results indicating that such social media data can accurately reflect real world health issues. This study analyzes Twitter data to assess public awareness of WHO health recommendations issued in response to the recent Zika virus outbreak, measures which focus on preventing people from being bitten by Zika-infected mosquitoes and by eliminating mosquito breeding sites [15].

II. Literature Review

Twitter has emerged as the most popular and influential micro-blog service, and Twitter user data has attracted considerable attention among researchers. Unlike many other social networking services, Twitter makes most user data publicly accessible.

A. Zika Virus

The Zika virus is named after the Zika region in Uganda, where the disease was first identified in a rhesus monkey [16].

Manuscript received March 13, 2018; revised April 26, 2018.

The authors are with the Department of Industrial Engineering and Management, Yuan Ze University, Taiwan, ROC (e-mail:iecjsu@saturn.yzu.edu.tw, s1038908@saturn.yzu.edu.tw).

doi: 10.18178/ijiet.2018.8.9.1123

Zika virus is a single stranded RNA arbovirus member of the genus Flavivirus which includes viruses such as yellow fever, dengue, West Nile, and Japanese encephalitis [17], [18]. Transmission to humans occurs mainly through the bite of an infected female *Aedes* mosquito, but some cases have been reported in which the virus was transmitted through sexual contact [19]. Common symptoms of the Zika virus include fever, fatigue, rash, joint pain, or conjunctivitis (red eyes), however patients can frequently remain asymptomatic for days after transmission, thus delaying medical treatment [20].

B. Social Media

The ideological and technological foundations of Web 2.0 have given rise to a group of Internet-based applications that allow users to create and exchange content; these applications have been broadly characterized as types of "Social Media" [21], and the five most popular social media platforms today include Facebook, Twitter, LinkedIn, Pinterest, and Google Plus [22], each of which focuses on different forms of communication and serve different purposes. Google Plus combines the functionalities of Facebook and Twitter but is relatively new. Pinterest focuses on sharing static and moving images. LinkedIn is a business-oriented social networking service, mainly used for professional networking. As for the top two social networking platforms, Facebook is seen as a social network while Twitter is an information network.

Of these platforms Twitter offers the greatest availability of data and associated programming interfaces, along with a huge and widely diffuse user base.

C. Twitter

Twitter's user data have attracted considerable research interest in a wide range of domains. Twitter is also a recognized source of breaking news and other high-value information. For example, Keith Urbahn, the chief of staff for the former US defense secretary Donald Rumsfeld, tweeted, "*So I'm told by a reputable person they have killed Osama Bin Laden. Hot damn.*" A great many highly informed and influential individuals are regular Twitter users, and frequently use the platform as a venue for sharing potentially important information.

Communication on Twitter is shaped by several important features of the service. "Hash tags" (#) are used to identify individual messages with trending topics or specific events. The @ symbol is included with the screen ID of other Twitter users to automatically notify the target user of the new tweet, in effect inviting the target user to respond and create a conversation. "RT" is used as an abbreviation for "re-Tweet", indicating that the message content was found elsewhere and was not created by the poster.

Twitter users can "follow" each other on the platform, with new messages from each user automatically being distributed to his/her "followers". The system standard default settings are set so the user can follow any other user unless his/her profile is set to 'private'. In this case, an initial request for approval is required. The concept of followers is an important aspect in the Twitter community in that it determines the degree to which information is distributed through the network. As such, one's number of followers has become a critical form of measurement for popularity and influence amongst Twitter users.

Twitter data has been analyzed for use in predicting future events with promising results in domains including stock market movements [23], flu outbreaks[24] and even election outcomes [25]. As of February 2016, Twitter boasted 320 million monthly active users, 80% of whom accessed the service via a mobile device [26].

One of Twitter's defining characteristics is that it provides real-time communication. A research project in Japan [27] successfully built an earthquake reporting system based on tweets. The resulting system detected 96% of earthquakes of earthquakes registering an intensity of 3 or above on the Japan Meteorological Agency (JMA) seismic intensity scale, and was able to distribute emergency alerts faster than the JMA's broadcast. In the area of product sales, Twitter data was used to predict film box office revenues during the first two weeks of general release [28], finding that the tweet-rate time series has a strong correlation with movie earnings and producing predictions better than those from the Hollywood Stock Exchange prediction market.

Research in HIV [11] suggests the feasibility of using geolocated Twitter data to identify HIV detection outcomes based on risk-related communications. Tweets with HIV prevalence data were linked to aidsvu.org, which provides data from the US CDC (Centers for Disease Control and Prevention) national HIV surveillance database. Similar studies were made for flu surveillance [29], finding a high correlation between the geolocated tweets extracted and the CDC database. Similar successes have been obtained in Spain [13]. Research on 2010 Haitian cholera outbreak was performed by [30] obtaining a good correlation between the analyzed the data from Twitter, health maps, and the Haitian Ministry of Public Health. These examples suggest that Twitter data can be of great value for predicting real world outcomes in many fields.

D. Twitter Data Extraction Techniques – Twitter API

Twitter message content and social network relations can be retrieved in real-time through Twitter's API. We used Python to process and analyze the data, the python library "Tweepy" to gather the desired information, and the library "Pymongo" to store the information as a MongoDB database, which is a document-based non-SQL database. MongoDB is a robust and well-documented database that works well for small and large datasets. It provides powerful query operators and indexing capability to significantly reduce the amount of analysis needed for custom Python code. Pymongo is a Python distribution containing tools for working with MongoDB, and is recommended for use with MongoDB from Python[31]. The Tweepy driver connects directly to the Twitter Streaming API. With the streaming API continuously delivering a real-time stream of tweets, we gathered the public available tweets containing the keyword "zika". The tweets were saved in JSON format, which is a lightweight data-interchange format[32]. Only English-language tweets were collected without location constraints.

E. Sentiment Analysis

Textual information can be broadly classified into two

main categories, facts and opinions [33]. Sentiment analysis, also known as "opinion mining", is an opinion-oriented Natural Language Processing (NLP) method, which is typically used to define the polarity (positive or negative) of a piece of text. Various approaches are used for sentiment analysis, including lexicographical analysis and machine learning techniques.

According to Liu [34], opinions can be classified as regular or comparative. Regular opinions are often referred to simply as "opinions". In a comparative opinion, two or more entities are compared in terms of their similarities or differences; usually using comparative or superlative adjectives or adverbs [35]. Go et al. provided one of the first studies of sentiment analysis for Twitter data [36], using machine-learning algorithms to classify message sentiment using distant supervision and training data consisting of Twitter messages with emoticons, which are used as noisy labels.

Sentiment analysis has been widely used in the medical domain, primarily to process web-based medical documents [37]. Other applications include mining and retrieving personal health information and opinions such as drug reviews [38], messages regarding hearing loss [39], and personal health information [40]. Other studies have applied sentiment analysis to analyze the emotional affects of messages related to suicide intention [37], infertility treatments [41], and cancer [42].

F. Sentiment Analysis Web Services

Several sentiment analysis assessment tools are now available as web services and have been validated in various studies [43]-[45]. Despite the potential convenience of using such services, the present study relied on a homemade sentiment analysis framework to ensure the results present a deep understanding of the data manipulation and classification processes.

III. Research Methodology

A. Classifying the Tweet

There are two basic approaches to performing sentiment analysis: Machine Learning and Lexicon-based. Lexicon-based approaches require a set of predefined terms that represent polarity, called a sentiment lexicon. The Machine Learning approach requires domain specific labeled data to extract features to train the classifier.

We decided to perform lexicon-based analysis since it doesn't require manual labeling to train the classifier. Moreover, the labeled data must be domain-specific to achieve acceptable results.

B. Pre-processing Tweets

The language found in the Twitter corpus is distinct from conventional texts, and the special features and attributes characteristic of Twitter messages must first be processed using natural language processing tools. The core idea is to pre-process the raw data and perform different transformations to remove the noise and then feed it to the classifier. All tokens are converted to lowercase using a command built in python. Special characters, such as "RT" for retweet, can be removed to reduce noise without lexically significant impact. The Hashtag symbol is removed using regular expressions. The content of the hashtag is retained because it is usually used in the context of the tweet. To remove punctuation, a special algorithm had to be implemented since emoticons are mostly composed of punctuation marks. Usernames were removed as they don't hold any useful information. URLs in different formats were removed. Words were tokenized and texts segmented by splitting them by spaces using NLTK, to produce a set of words for subsequent analysis. The dataset was also searched for an NLTK list of common stop words and high frequency words like "a", "the", "of", "and", "an" were also removed to reduce the dimensionality of the data sets. The filtering process normalized the tweet content for processing with the sentiment classifier.

C. Sentiment Classifier

The sentiment analyzer component applies a three-way classification algorithm to classify the tweets as positive, negative or neutral. Unsupervised classification approaches do not require any training data, and this study used the three-way classification algorithm approach proposed by Khan et al. [44] including Polarity Based Emoticon Classifier (PBEC), Improved Polarity Classifier (IPC), and SentiWordNet Classifier (SWNC) (Fig. 1). The result of the PBEC returns a neutral tweet, this output is then reprocessed in the IPC and, if it is still classified as neutral, it is then processed by the SWNC. If the three processes classify the tweet as neutral, we declare it as neutral.

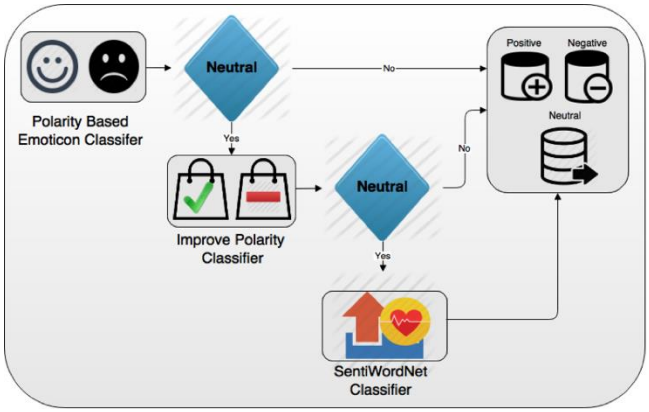

Fig. 1. Sentiment analyzer structure.

A set of emoticons, manually tagged as positive or negative, are used to classify the polarity of the tweet. With the use of regular expressions in python, we detect the presence of the established emoticons, which are then used to classify the tweet as positive or negative. We used the set of 70 positive and 75 negative emoticons proposed by [44]. If the emoticon is found in the positive set, then it is declared positive. The emoticon is declared negative if it is found in the negative set. If the emoticon is not found in either set, we continue the process with the IPC. We take into account the polarity of the words expressed in the tweet found with an emoticon and help resolve inconsistencies resulting from differing polarities between the text and the emoticon.

The IPC is based on the "bag of words" approach which uses a predefined list of words classified as positive or negative. The first set was created by Liu [46] and is

composed of 2006 positive words and 4784 negative words, for a total of 6790 words. The second set was created by Loughran & McDonald [47] and is composed of 354 positive words and 2349 negative words, for a total of 2703 words. The IPC includes a total of 9493 trained words. If the word is found in the positive set, then it is declared positive. It is declared negative if found in the negative set. If the word is not found in either set, we declare it neutral and process it using SWNC for the final polarity assessment. SentiWordNet is an English lexical dictionary based on WordNet for sentiment classification and opinion mining. Each WordNet synset (sets of synonyms) is assigned a triple polarity score (positivity, negativity, and objectivity), where the sum of these scores is always 1 [48]. After processing we will have three databases for positive, negative, and neutral tweets.

IV. Implementation

A. Collected Tweets

We extracted the raw data from Twitter based on the keyword "Zika". With 65 days of data extraction, we managed to acquire 1,072,648 tweets from 331,284 unique users, with an average of 3 tweets per user, taking into account that some users post more frequently than others. A summary of the data is presented in Table I.

TABLE I: Summary of Extracted Tweets

Metric	Description
Total Dataset Size	532 MB
Total Number of Tweets	1,072,648
Total Unique Tweets	346,189
Number of Retweets	471,685
Number of Tweets with URL	937,194
Total Number of Unique Users	331,284
Data Extraction Start	28-Mar-16
Data Extraction End	31-May-16

Fig. 2 shows a general overview of the relationship of the top users number of tweets and their related number of followers accounted for the last day in the period of extraction. The number of followers is considered one type of influence referred as the "indegree influence" [49]. The influence is a very important aspect to analyze, as the information that has been distributed among the influential network can give an insight of the general diffusion of information.

A user named @AdamNeira posted with a total of 11,799 related to the Zika virus, but the impact of these tweets is relatively low because he has very few followers (only 76). In contrast, the user @Zika_News posted 4,486 Zika-related tweets to its 5,223 followers. This user is a "bot" designed to automatically retweet Zika-related posts in Twitter. As shown in figure 2, among the most active users, the most influential account is from the Canadian news agency @Reuters_Health. Of the user accounts analyzed the accounts with the greatest number of followers tended to be news agencies and health organizations. News tweets can contain positive or negative information, but this study focuses on the impact of public opinion, and thus tweets from these accounts with the word "news" or similar terms were excluded from analysis.

The Twitter accounts of large news agencies (e.g., CNN, BBC, New York Times) are easily identifiable by their large number of followers (see Fig. 3). Several of these organizations have multiple twitter accounts. Although these users are highly influential, their posting frequency is considerably lower than the average user. Our data includes 6,721 users with the word "news" in the account name, responsible for a total of 70,324 Zika-related tweets.

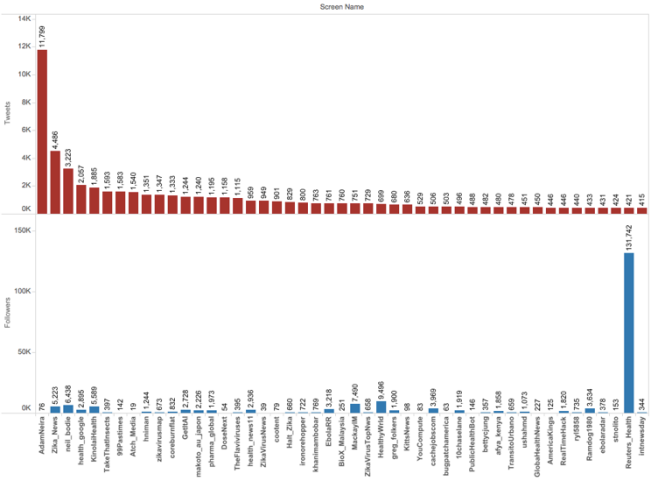

Fig. 2. General overview of frequent users.

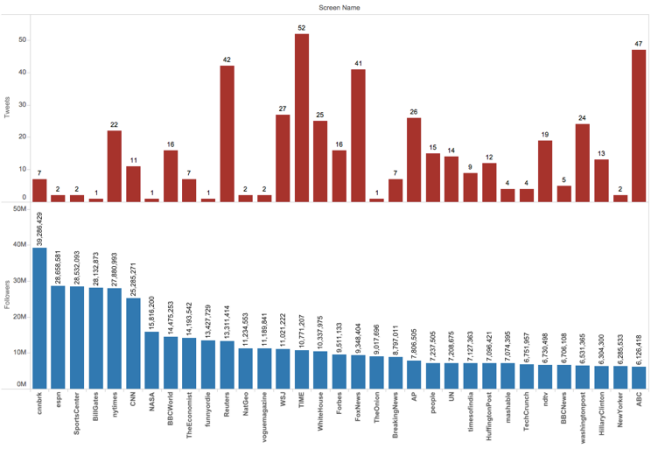

Fig. 3. Top accounts by number of followers.

B. Word Frequency

Figure 4 presents a word cloud for the 995,902 unique words in the data set (from a total of 1,072,648 tweets), with the size of the word representing its relative frequency. Words which appeared fewer than 4,000 were excluded.

Fig. 4. Word cloud.

The association of various words with "zika" can be used to identify distinct topics. As expected, the word most commonly associated with "zika" is "virus", followed by "mosquitoes" with nearly 37 thousand mentions. Surprisingly, the word "WHO", referring to the World Health Organization, is only mentioned 528 times. We found 86,669 tweets with the word "mosquito" or "mosquitoes", accounting for only 8% of the data extracted.

C. Net Sentiment Rate

The net sentiment rate (NSR) is a metric used to estimate the overall sentiment expressed towards particular topics in social media [50], [51]. NSR is defined by the following equation:

$$NSR = (Positive\ Tweets - Negative\ Tweets) / Total\ Tweets \quad (1)$$

In the general context of our extracted tweets, we obtained a -0.36 NSR, with 61% negative, 25% positive, 9% neutral tweets, and news related tweets accounting for 5%.

D. Retweets Analysis

Retweeting accounted for 471,685 tweets from 79,946 unique users, or 44% of the total number of tweets. Retweeting can have a powerful reinforcing effect for particular messages [49], but many retweets could be due to bots, and does not reflect the value judgment of individual users.

Figure 5 shows that many of the most prolific users in our sample tend to retweet frequently. For example, the user "TransistoUrbano" only retweets and does not post original messages, suggesting that this account may be a bot. Although this research doesn't extend to bot analysis, reviewing this user account we found that all his posting activity consists of retweeting, mainly traffic activity in Brazil.

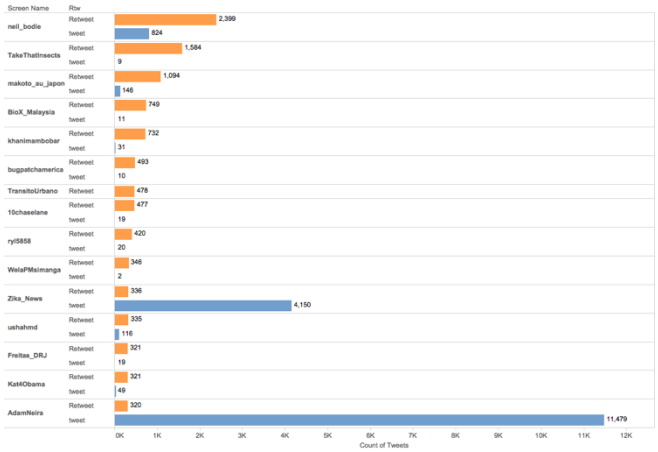

Fig. 5. Top retweet users.

The general NSR for the retweets analysis is -0.4. We now focus on the polarity of heavily retweeted posts and their corresponding impact measure based on the number of followers that received the message. The Table 2 shows the top five most retweeted post in our negative dataset, taking into account that these tweets all contain URLs.

The retweet "zika causes microcephaly and other birth defects cdc concludes" was retweeted by 2,439 unique users, though some retweeted the same message multiple times. Although the original tweet was produced from nine different sources, the main content was identical and produced the same post-processing output. Some links refer to a news article (from Forbes) and others link to the Science Daily report from CDC scientists.

TABLE II: MOST NEGATIVE RETWEET

Tweet	Frequency
zika causes microcephaly and other birth defects cdc concludes	2683
zika virus may now be tied to another brain disease	2,596
zika virus tested in brain precursor cells	2,522
link between zika virus and fetal brain damage confirmed	2,414
the genetic evolution of zika virus	2,379

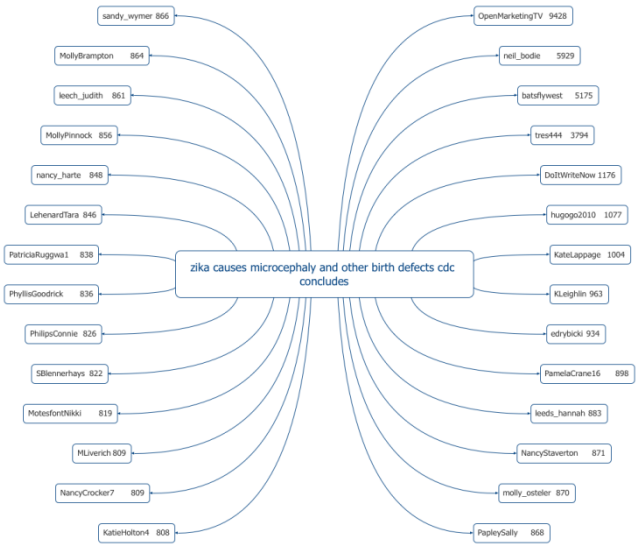

Fig. 6. Tweet followers networks.

Fig. 6 shows the top users and their number of followers. However, we lack a deeper insight into the relationships between followers and users for the accounts that received this message as the results do not account for multiple users of a single account, or single users having multiple accounts. Retweeted messages are received by an average of 151 users, with an average of 4,905 followers. The use of URLs in tweets suggests the content includes news-related information, as it provides the user with access to follow up information beyond that included in the tweet content. As shown in Table II, Twitter-based discussion of Zika is focused on the virus' impact on brain disease and birth defects.

The top retweets presented in Table II are all news related, and all contain links to different news agencies, but none mention the WHO recommendations.

This analysis attempts to filter noise and unrelated information from the data to focus on public opinion as reflected in the tweets. Hashtags can help exclude tweets on unrelated topics. The hashtag #news clearly states that the information shared is news-related from a news agency and not a reflection of public opinion, and such tweets are thus excluded from analysis. Topics can be differentiated by their

associated hashtags for sentiment analysis. Table 4 shows hashtags containing the word "zika", including #zikavirus and #zikasummit, which refers to the CDC summit where doctors shared the latest updates, including implications for pregnant woman and strategies for mosquito control. The hashtag #Mosquito is the fourth most negative hashtag in the NSR.

TABLE III: TOP FREQUENT HASHTAG

Hashtag	Frequency	NSR
#health	60,543	-0.54
#news	49,980	-0.39
#rio2016	29,412	-0.27
#brazil	20,844	-0.1
#virus	15,618	-0.82
#olympics	12,816	-0.57
#microcephaly	12,282	-0.14
#ebola	10,881	-0.15
#mosquito	9,696	-0.53
#science	9,132	-0.42

Fig. 7. Mosquito hashtag word cloud.

TABLE IV: TOP HASHTAG WITH THE WORD "ZIKA"

Hashtag	Frequency	NSR
#zikavirus	118,302	-0.48
#zikasummit	9,717	0.18
#reuterszika	5,847	-0.1
#zikaolympics	4,683	-0.96
#openzika	2,340	0.36
#combatzika	2,337	-0.06
#zika-related	1,308	-0.26
#knowzika	1,284	0.34
#zikahack	1,077	0.34
#zika-linked	1,059	-0.6

Fig. 7 shows a word cloud for all hashtags containing the word "mosquito" in the negative dataset, including alternative spellings "mosquitoes" and "mosquitos". Hashtags such as #mosquitocontrol, #stopthemosquito, and #mosquitorepellent can be used to promote awareness. The high frequency of these hashtags in our data set represents that the related topics are currently under discussion and important. Hashtags can help us classify the dataset into specific events of discussion for further analysis. They can also help us identify anomalous or irrelevant information (spam) for exclusion from the final analysis. After reviewing the top hashtags, we can separate our analysis in different topics: the 2016 Brazil Olympics Games, Ebola, and mosquitoes. These three topics are important, but this research focuses on "mosquitoes" because of their relevance to the WHO recommendations.

E. Hashtag Frequency Analysis

Hashtag frequency analysis can show topics related to the main hashtag. In this case, "#Zika" is hashtagged in 520,656 tweets which also prominently feature the hashtags #health, #virus, #science. Other prominent hashtags include those related to the 2016 Rio Summer Olympics Games, including #rio2016, #brazil, #olympics. This association is due to Brazil being the epicenter of the recent Zika outbreak, which coincided with the run-up to the Olympics. The hashtag #ebola was also found to be associated with #Zika because the US had shifted funds originally slated to combat Ebola to fight the Zika threat. Finally, the hashtags #mosquito and #microcephaly are associated with the virus' cause and effect. Table 3 displays the frequency and NSR of the top ten hashtags found in the tweets data set.

F. Timeline Analysis

Distinct spikes in tweeting activity can indicate reaction to important events. Overall the majority of Twitter traffic is generated within the United States, but traffic spikes can occur in other countries due to special circumstance. While Twitter hosts a global average 5,700 tweets per second (TPS), on August 3, 2013 Japanese Twitter users responded to the broadcast of a popular animated movie with a TPS spike of 143,199. Such spikes can be driven by other events including natural disasters, military actions, political events, or sporting events [52].

Our data set includes 3 notable spikes for Zika-related posts. We analyzed the top retweets and most-frequent hashtags during the spikes. The first spike was on April 12 with 33,712 tweets. Reviewing the data from this day we found 2,522 retweets focused on testing for Zika in brain precursor cells but, none of the following tweets were related to mosquito control, but rather included hashtags such as #congress, #money, #fight, and the expected #zika and #health.

The second spike was on April 14 with 39,460 tweets generally focused on CDC confirmation that the zika virus is highly related to birth defects such as microcephaly. This focus on this particular aspect of Zika led to some confusion and misapprehension that birth defects were the only serious impact of the virus. For example, one tweet read: "Wait. A 70 year old man died from the Zika virus in Puerto Rico?? But.... I thought it only affected pregnant women."

The biggest spike, with 52,963 tweets, occurred on May 20

in response to a CDC announcement that 157 women in the US had tested positive for Zika. Amongst the top 5 tweets we can see political implications. In one, US Senator Bernie Sanders discusses funding for Zika prevention, while in the other Hillary Clinton says that 1 of 4 people in Puerto Rico were at risk for developing Zika. Thus far, all the popular tweets found in the timeline are related on the effects of the virus rather than prevention.

G. Evaluation

The confusion matrix was used to evaluate the performance of the classifier used in this research, showing the number per class of well classified and mislabeled instances. It is a simple and understandable way to show the classifier performance. Based on the survey performed by Serrano-Guerrero et al. 2014 where they evaluated the performance of various web sentiment analysis services, they found that the best service is Alchemy API with an accuracy of 62.5% performed on tweets classification [45]. Although Alchemy API provides its services for free, they have a limitation for 1000 transactions (tweets) daily. Performing evaluation on our proposed system SM-PIF we found an accuracy of 67.97% using the training data provided by the International Workshop on Semantic Evaluation (SemEval) 2014 [53].

V. Conclusions

In our period of analysis, the top Zika-related discussion trends on Twitter focus on the virus' causes and news of recent outbreaks, while relatively few discussions are related the WHO prevention recommendations. While providing updated information on virus outbreaks is important, but a consistent focus on prevention measures is also of vital importance. The development of social media platforms has changed the way people obtain and share information, with individuals and organizations now able to broadcast news and opinion from any location through mobile devices and wireless Internet.

Crowd-sourced opinion, in the form of individual tweets, is expected to present a more accurate representation of events than traditional surveys. The gathering and integration of such social media information can contribute to better-informed decision making. The outcomes derived furnish valuable information for health organization to target the required improvement to help reduce the Zika pandemics and educate the population. This research applies big data analysis techniques to Twitter feeds to examine the diffusion of important information related to the Zika virus. WHO recommendations emphasize the need for mosquito control, diagnostics, and public education. Mosquito are the cause of many other diseases, distributing the information on mosquito control should not be limited to the novelty of the outbreak but should be a constant flow of information to create a stable awareness and prevent major calamities.

References

[1] B. Liu, *Opinion Mining Encyclopedia of Database Systems*, 2008, Springer.

[2] A. L. Holbrook, M. C. Green, and J. A. Krosnick, "Telephone versus face-to-face interviewing of national probability samples with long questionnaires: Comparisons of respondent satisficing and social desirability response bias," *Public Opinion Quarterly*, 2003, pp. 79-125.

[3] P. L. Alreck, and R. B. Settle, *The Survey Research Handbook*, 1995, Irwin.

[4] K. M. Goldstein,and M. K. Jennings, "The effect of advance letters on cooperation in a list sample telephone survey," *The Public Opinion Quarterly*, 2002, pp. 608-617.

[5] G. Szolnoki and D. Hoffmann, "Online, face-to-face and telephone surveys—Comparing different sampling methods in wine consumer research," *Wine Economics and Policy*, 2013, pp. 57-66.

[6] B. Liu, "Web usage mining," *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*, 2007, pp. 449-483.

[7] M. Lövheim, A. Jansson, S. Paasonen, and J. Sumiala, "Social media: implications for everyday life, politics and human agency," *Approaching Religion*, 2013, pp. 26-37.

[8] G. J. Rubin, R. Amlôt, L. Page, and S. Wessely, "Public perceptions, anxiety, and behaviour change in relation to the swine flu outbreak: Cross sectional telephone survey," *Bmj*, 2009, p. b2651.

[9] J. E. Oliver and T. Lee, "Public opinion and the politics of obesity in America," *Journal of Health Politics, Policy and Law*, 2005, pp. 923-954.

[10] R. J. Blendon, J. M. Benson, C. M. Desroches, and K. J. Weldon, "Using opinion surveys to track the public's response to a bioterrorist attack," *Journal of Health Communication*, 2003, pp. 83-92.

[11] S. D. Young, C. Rivers, and B. Lewis, "Methods of using real-time social media technologies for detection and remote monitoring of HIV outcomes," *Preventive Medicine*, 2014, pp. 112-115.

[12] H. Achrekar, A. Gandhe, R. Lazarus, S.-H. Yu, and B. Liu, "Predicting flu trends using twitter data," presented at 2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), 2011, IEEE.

[13] A. Moreno-Sandoval and E. Moro, "Big data versus small data: The case of gripe (Flu) in Spanish," *Procedia - Social and Behavioral Sciences*, 2015, pp. 339-343.

[14] J. Jashinsky *et al*., "Tracking suicide risk factors through Twitter in the US," *Crisis*, 2014.

[15] Organization. (2016). W.H. Zika virus and complications: Questions and answers. [Online]. Available: http://www.who.int/features/qa/zika/en/

[16] G. Dick, S. Kitchen, and A. Haddow, "Zika virus (I). Isolations and serological specificity," *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 1952, pp. 509-520.

[17] J. A. Tetro, "Zika and microcephaly: Causation, correlation, or coincidence?" *Microbes and Infection*, 2016.

[18] C. Chang, K. Ortiz, A. Ansari, and M. E. Gershwin, "The Zika outbreak of the 21st century," *Journal of Autoimmunity*, 2016, pp. 1-13.

[19] *Zika Virus - Fact Sheet*. (2016). [Online]. Available: http://www.who.int/mediacentre/factsheets/zika/en/

[20] *Symptoms, Diagnosis, & Treatment*. (2016). [Online]. Available: http://www.cdc.gov/zika/symptoms/

[21] A. M. Kaplan and M. Haenlein, "Users of the world, unite! The challenges and opportunities of social media," *Business Horizons*, 2010, pp. 59-68.

[22] eBizMBA. *The 15 Most Popular Social Networking Sites*. (2016). [Online]. Available: http://www.ebizmba.com/articles/social-networking-websites

[23] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science*, 2011, pp. 1-8.

[24] V. Lampos and N. Cristianini, "Tracking the flu pandemic by monitoring the social web," presented at 2010 2nd International Workshop on Cognitive Information Processing (CIP), 2010.

[25] S. Harald *et al*., "The power of prediction with social media," *Internet Research*, 2013, pp. 528-543.

[26] Twitter. *Fourth quarter 2015*. (2016). [Online]. Available: https://investor.twitterinc.com/results.cfm

[27] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: Real-time event detection by social sensors," in *Proc. the 19th International Conference on World Wide Web*, 2010, ACM: Raleigh, North Carolina, USA, pp. 851-860.

[28] S. Asur and B. A. Huberman, "Predicting the future with social media," in *Proc. the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 2010, IEEE Computer Society, pp. 492-499.

[29] J. Chon, R. Raymond, H. Wang, and F. Wang, *Modeling Flu Trends with Real-Time Geo-tagged Twitter Data Streams, in Wireless Algorithms, Systems, and Applications*, 2015, Springer, pp. 60-69.
[30] R. Chunara, J. R. Andrews, and J. S. Brownstein, "Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak," *The American Journal of Tropical Medicine and Hygiene*, 2012, pp. 39-45.
[31] MongoDB. *PyMongo 3.2.2 Documentation*. (2015). [Online]. Available: https://api.mongodb.org/python/current/
[32] JSON. *Introducing JSON*. (2016). [Online]. Available: http://www.json.org
[33] B. Liu, *Opinion Mining and Sentiment Analysis, in Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*, 2011, Springer Berlin Heidelberg: Berlin, Heidelberg, pp. 459-526.
[34] N. Indurkhya and F. J. Damerau, *Handbook of Natural Language Processing*, 2010, Chapman \& Hall/CRC. 704.
[35] N. Jindal and B. Liu, "Mining comparative sentences and relations," in *Proc. the 21st National Conference on Artificial Intelligence*, 2006, AAAI Press: Boston, Massachusetts, pp. 1331-1336.
[36] A. Go, L. Huang, and R. Bhayani, "Twitter sentiment analysis," *Entropy*, 2009.
[37] J. P. Pestian *et al*., "Sentiment analysis of suicide notes: A shared task," *Biomedical Informatics Insights*, 2012, pp. 3.
[38] J.-C. Na *et al*., "Sentiment classification of drug reviews using a rule-based linguistic approach," 2012, Springer Berlin Heidelberg: Berlin, Heidelberg, pp. 189-198.
[39] T. Ali, D. Schramm, M. Sokolova, and D. Inkpen, "Can I hear you? Sentiment analysis on medical forums," *IJCNLP*, 2013.
[40] M. Sokolova, S. Matwin, Y. Jafer, and D. Schramm, "How Joe and Jane tweet about their health: Mining for personal health information on twitter," *RANLP*, 2013.
[41] M. Sokolova and V. Bobicev, "What sentiments can be found in medical forums?" *RANLP*, 2013.
[42] P. Biyani *et al*., "Co-training over domain-independent and domain-dependent features for sentiment analysis of an online cancer support community," *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2013.
[43] K. Meehan, T. Lunney, K. Curran, and A. McCaughey, "Context-aware intelligent recommendation system for tourism," 2013 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops), 2013.
[44] F. H. Khan, S. Bashir, and U. Qamar, "TOM: Twitter opinion mining framework using hybrid classification scheme," *Decision Support Systems*, 2014, pp. 245-257.
[45] J. Serrano-Guerrero, J. A. Olivas, F. P. Romero, and E. Herrera-Viedma, "Sentiment analysis: A review and comparative analysis of web services," *Information Sciences*, 2015, pp. 18-38.
[46] B. Liu, M. Hu, and J. Cheng, "Opinion observer: analyzing and comparing opinions on the web," in *Proc. the 14th International Conference on World Wide Web*, 2005, ACM.
[47] T. Loughran and B. McDonald, "When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks," *The Journal of Finance*, 2011, pp. 35-65.
[48] S. Baccianella, A. Esuli, and F. Sebastiani, "SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining," *LREC*, 2010.
[49] M. Cha, H. Haddadi, F. Benevenuto, and P. K. Gummadi, "Measuring user influence in twitter: The million follower fallacy," *ICWSM*, 2010, pp. 30.
[50] N. K. Rana and N. Kapoor, *Sentiment Analysis Using Integrated Approach of Na ïve Bayes and Principal Component Analysis*.
[51] S. Afyouni, A. E. Fetit, and T. N. Arvanitis, "Perspectives through social media analysis," *Enabling Health Informatics Applications*, 2015, pp. 243.
[52] R. Krikorian, *New Tweets per Second Record, and How!* (2013). (2016). [Online]. Available: https://blog.twitter.com/2013/new-tweets-per-second-record-and-how
[53] S. Rosenthal, N. Farra, and P. Nakov, "SemEval-2017 task 4: Sentiment analysis in Twitter," in *Proc. the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 2017.

Chuan-Jun Su is professor of the Industrial Engineering and Management Department at the Yuan Ze University, Taiwan. Prior to joining the University, he was an assistant professor with Hong Kong University of Science and Technology. His research interests include information systems, mobile agent technology, virtual reality, and intelligent design and manufacturing systems.

Jorge A. Quan Yon was born in Guatemala on August 8, 1984. He received his master's degree in industrial engineering and management at Yuan Ze University, Taiwan, in 2014. He is currently a Ph.D student at the intelligent information systems laboratory at the Industrial Engineering and Management Department at Yuan Ze University, Taiwan. His research interests include internet of things, big data analysis, social media prediction systems, and smart sustainable hydroponic systems.