Predicting Community Engagement on Twitter on Environmental Health Hazards

Adel Alshehri[1,2]
Dept. of Computer Science and Engineering
[1] University of South Florida
Tampa, Florida 33620
[2] King Abdulaziz City for Science and Technology
Riyadh, Saudi Arabia 11442
Email: Adelalshehri@mail.usf.edu

Wainella Isaacs
Dept. of Civil and
Environmental Engineering
University of South Florida
Tampa, Florida 33620
Email: wainellai@mail.usf.edu

Aseel Addawood
Illinois Informatics Institute
University of Illinois at
Urbana Champaign
Champaign, Illinois 61820
Email: Aaddaw2@illinois.edu

Maya Trotz
Dept. of Civil and
Environmental Engineering
University of South Florida
Tampa, Florida 33620
Email: matrotz@usf.edu

Sriram Chellappan
Dept. of Computer Science and
Engineering
University of South Florida
Tampa, Florida 33620
Email: sriramc@usf.edu

***Abstract*—In this empirical study, a framework was developed for binary and multi-class classification of Twitter data. We first introduce a manually built gold standard dataset of 4000 tweets related to the environmental health hazards in Barbados for the period** 2014 - 2018**. Then, the binary classification was used to categorize each tweet as relevant or irrelevant. Next, the multi-class classification was then used to further classify relevant tweets into four types of community engagement: reporting information, expressing negative engagement, expressing positive engagement, and asking for information. Results indicate that (combination of TF-IDF, psychometric, linguistic, sentiment and Twitter-specific features) using a Random Forest algorithm is the best feature for detecting and predicting binary classification with (87% F1 score). For multi-class classification, TF-IDF using Decision Tree algorithm was the best with (74% F1 score).**

***Index Terms*—Community Engagement, Barbados, Sewage, Mosquito, Environmental, Social media, Crisis, NLP**

I. Introduction

Environmental hazards like unsafe water, poor sanitation, urban air pollution and rising temperatures cause significant disease burden globally [1]. During infectious disease outbreaks, early epidemiological assessment is hindered when data, which may not be available for weeks, is only collected through official reporting structures like hospitals. To get timely estimates of disease burden and dynamics, near real-time data from informal sources (e.g. online social media) can be used. For example, during the 2010 cholera outbreak in Haiti, HealthMap news media reports and Twitter posts were positively correlated with official government cholera cases reported [2]. This unofficial data for a water-related disease was available up to two weeks earlier than official reported cases.

Twitter has been used as a formal source of information. It has been used in the United Kingdom to share and exchange information between the public, emergency responders, and water service providers [3]. Researchers showed that within social media, residents could report, request and obtain crisis-related information, while engaging in disaster response and rescue efforts [4].

Barbados (The Case Study Site) is a country in Caribbean has experienced significant and consistent water and sewage crises, which impact incidence of many diseases, including mosquito borne diseases. In this paper we use data mining and machine learning algorithms to detect and predict community engagement on twitter about water and sewage environmental health hazards in Barbados.

To the best of our knowledge, no studies have been conducted to classify and predict the types of community engagement (reporting information, expressing negative engagement, expressing positive engagement, and asking for information) on Twitter related to water, sewage and mosquito borne disease health risks.

II. LITERATURE REVIEW

This interdisciplinary research intends to merge environmental health hazards, machine learning, and social media analysis to detect and predict community engagement. We outline the relevant recent work concerning:

Health-related topics on social media: Multiple studies have used social media platforms for the exploration of public health issues [5],[6]. In [5], the team collected 300 million tweets for approximately one year and seven months focusing on the expansion of influenza. They implemented a support

978-1-5386-6783-5/19/$31.00 ©2019 IEEE
DOI 10.1109/ICSC.2019.00087

vector machine (SVM) classifier to distinguish between related and unrelated tweets. Their approach performed well in detecting influenza epidemics with an 0.89 correlation. Further, Antonio A. Ginart et al. [6], formed and validated a machine learning classifier to separate relevant from irrelevant tweets for understanding health behavior about marijuana.

The role of Twitter in community engagement: One study [7] investigated factors connected with an engagement of U.S. Federal Health Agencies via Twitter. They studied numbers of retweets in addition to the time between the agencys initial tweet and both the first and last retweets. They noticed that a third of the tweets had zero retweets. The analysis shows that hashtags, URLs, and user-mentions are positively associated with retweets. A text analysis of 1,583 tweets, where the numbers of retweets and favorites were included as engagement signs found that the American Heart Association, American Cancer Society, and American Diabetes Association varied in the degree of using the retweet, hashtags and hyperlink features for broadcasting health information [8].

Communication during Environmental disasters: The primary aim of [9] is to examine health-related warning messages sent by public safety agencies over Twitter during the 2013 flooding in Boulder, Colorado. They found that tweets focused on drinking water 41%, floodwater exposure 18%, general crisis information 16%, sanitizing 14%, and sewage 8%. Pascal Beaudeau et al., [10] proposed a framework to find out how climate change could affect health risks concerning drinking water.

In this paper, we provide a Twitter annotated dataset and propose a framework to track tweets and predict the different types of peoples engagement during crisis. Our study may contribute important considerations for decision-makers and other individuals in efforts to prepare for crisis situations and save time in choosing where to focus their limited resources.

III. Method

In this section, we propose a binary and multi-class model to classify online social community engagement during environmental health risks that have caused disruptions to communities. With a 280-character limit, twitter [1] has been widely used for sharing news, beliefs, and activities during crises and natural emergencies, and to gather support for social and public health monitoring [11].

Framework:
Figure 1 presents a high-level picture of the framework used to collect a series of data over a given time-frame for a given location, Barbados to be specific. Its components are split into four stages: (1) Data (tweets) collection; (2) Data Preprocessing; (3) Binary classification; and (4) multi-class classification.

Data Collection and Filtering: We used Crimson Hexagon [12] to crawl any tweet, including the keyword Barbados from January 1, 2014, to May 31, 2018. 5,532,419 tweets were captured (all languages), of which 3,897,789 were in English. Then, a filtering method with a list of keywords containing the following was adopted: crisis, wastewater, sewage, Water, Climate Change, waterborne, Zika, mosquito, dengue fever, yellow fever, chikungunya, Aedes aegypti, West Nile, malaria, and infectious disease. After applying the aforementioned filters, the total sample size was reduced to 30,358 tweets. This much smaller sample size was likely due to 1) Barbados is a small island compared to other populations, and people might use other social media platforms like Facebook or Instagram, 2) Although English is the official language of Barbados it is not the only language used (e.g. large international tourist population speak multiple languages), and this study excluded potential content concerning community engagement on this topic written in other languages.

Data Preprocessing: The goal behind preprocessing is to tokenize sentences into words in order to represent each tweet as a feature vector. Twitter users tend to use idioms, abbreviations, and grammatical errors in their posts. Therefore, text processing methods like stop-word removal, punctuation, stemming (converting a word to its root), and removing unnecessary white spaces using the Natural Language Toolkit (NLTK) library available in Python were applied. Also, all characters were lowercased, and after initially saving these features all URLs and mentions were eliminated.

Binary and Multi-class Classifications : In this stage, A two step-procedure was used for twitter community engagement classification. Step 1 used binary classification to classify the tweets as relevant or irrelevant. In our case, Irrelevant tweets include personal messages, holiday greetings, chatter, ambiguous tweets, and spam. Step 2 used multi-class classification to further describe relevant tweets as four types: (1) asking for information, (2) reporting information, (3) expressing negative engagement, and (4) expressing positive engagement.

IV. Experimental Setup

In this section, we will focus on (1) the process of crowdsourcing Twitter annotations, (2) extracting and selecting features to improve algorithm performance and (3) classification of algorithms and WEKA (data mining software tool).

A. Crowdsourced Annotation

A small random sample (4000 tweets) was selected and uploaded to Amazon Mechanical Turk (MTurk)[2]. To facilitate the labeling process, some guidelines and samples were provided to the 3 annotators. The relevant tweets are divided into four types: (1) Asking for information, verification or instructions for handling specific situations; e.g. *"what is going on in both sewage plants on the south coast?"* (2) Reporting facts, activities, events, and observations; such as *"our beaches on the south coast are full of sewage"* (3) Expressing negative engagement such as complaints, frustration, or sarcasm; e.g., *"Barbados in a crisis and all Kellman thinking about is an airport,"* and (4) Expressing positive engagement like

[1] https://twitter.com/

[2] https://www.mturk.com/

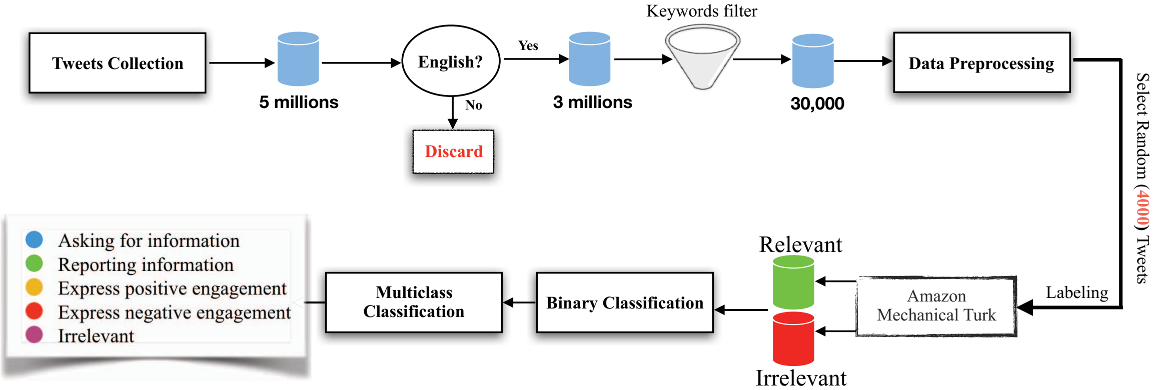

Fig. 1: Twitter Community Engagement Detection Framework

proposing a solution or showing satisfaction, or counter the spread of misinformation. E.g., *"This is bullshit. I live in Barbados, and NOTHING happened. Take this misinformation down "* If the tweet contained no content related to public health issues or the sewage leak issue in Barbados, they were labeled as irrelevant. E.g., the tweet *"good morning Barbados, a day spent on the water enjoying snorkeling "*. The *"No Agreement"* tweets that belongs to Multi-class labeling were excluded from the model to avoid bias. From 4000 tweets, in binary labeling, 40% of final tweets were relevant, and 60% determined as irrelevant. The results of the annotation process are shown in Table I. See section V for more details.

B. Feature Sets

1) TF-IDF Features refer to N-gram features, which rely on the word count for each given unigram that appears in the tweet. The three main components that affect the importance of a word in a dataset are (1) Term Frequency (TF) which presents the frequency of the word in the dataset. (2) Inverse Data Frequency (IDF): applied to calculate the weight of rare words overall tweets in the dataset. (3) TF/IDF is a technique which uses the product of TF and IDF to determine the weight of each word. In formulas (1, 2 and 3), t is a term, d is the tweet in which t occurs, and D is the dataset.

$$TF(t) = tf(t, d) \tag{1}$$

$$IDF(t) = log\left(\frac{|D|}{1 + |\{d : t \in d\}|}\right) \tag{2}$$

$$TF - IDF(t) = TF * IDF \tag{3}$$

2) Psychometric Features are connected more with mental abilities and behavioral characteristic. We adopted the linguistic inquiry and word count (LIWC version 2015) tool to extract these features [13]. Psychometric features involve: emotional, social words, and personal concerns. It also includes drives and needs which are represented by words related to personal power, accomplishment, reward, and risk.

3) Linguistic Features include the following two types: (1) Grammatical features, which produce a rate of words that are verbs, adverbs, pronouns, and other punctuation. (2) Summary variables, which include (analytical thinking, and emotional tone). [13] Analytical thinking is the percentage of terms in which people use words that suggest formal and hierarchical thinking patterns. While clout refers to the social situation or leadership that people demonstrate within their writing. Lastly, with emotional tone, LIWC merges both positive emotion and negative emotion scores into a single summary variable.

4) Twitter-specific features refer to characteristics unique to the Twitter platform. There are various forms that users on Twitter engage: (a) Retweet ratio is a metric of fame for a tweet since it implies both endorsement and distribution [14]. (b) Mention ratio is a technique in Twitter to ask other users to engage or follow a discussion in the form of (@username). (c) Hashtag ratio is an essential characteristic of Twitter which can be injected anyplace in a message. Some Hashtags are devoted mainly to events such as (Barbados) which can be used as search key on Twitter [15]. (d) Url ratio is the number of inserted links in a tweet which used to share extra information about the situation. (e) The number of Followers and Followings.

5) Sentiment Features is the computational study of opinions, emotions, and disturbances shown in the text [16]. In our experiment, we decided to use the sentiment labels provided by the Crimson Hexagon tool because we found it produces more accurate results than we would have had otherwise.

C. Feature Selection

Typically, any machine learning algorithm represents a model as as a function f that predicts the output Y given the input X { x_1, x_2, ..., x_R} where x_i is selected input features and R is a real number. It is commonly right that not all input feature x affords the same value of information about the output Y, rather just a small subset of them { x_1, x_2,..., x_s} where (S < R), that addresses important information about Y.

The total number of initial features that we extracted was 6540. This is a very large number of features. To optimize the features, we used the an Information Gain [17] approach for extracted relevant features only.

Information Gain is the variation of the volume of information that can be carried to the classification model when a feature is included or not. Therefore, to compute information gain, we need first to determine the information entropy. The information entropy H(T_r) and the information Gain (IG) for a feature F_i calculated is as follows:

$$IG(T_r, F_i) = H(T_r) - \sum_{c \in F_i} P(c)H(c), where \tag{4}$$

$$H(T_r) = -\sum_{x \in S} p(x) \log_2 p(x) \tag{5}$$

D. Classifiers

We split the dataset into two parts. The larger part we use for training (80%) and the smaller part we use for evaluation (20%). An experiment was conducted to evaluate the performance of the model under the selected 4 supervised learning classifiers using a machine learning tools named WEKA [18]. We compared classifiers that have frequently been used in related work: Support Vector Machine (SVM) [19]; Decision Trees (DT) [20]; Naive Bayes (NB) [21]; and Random Forest (RF) [22].

E. Performance Measurements

More generally, to evaluate the effectiveness of our model, we used the standard classification metrics: (1) Accuracy (total number of correct predictions); (2) F score (harmonic mean of precision (P) and (R) recall) and (3) Area Under Curve (AUC) which describe by false positive rates on the horizontal axis and true positive rates on the vertical axis. Based on classification of True Positives (TP), True negatives (TN), False Positives (FP) and False Negatives (FN), we have

$$F1 = 2 * \frac{R * P}{R + P} \tag{6}$$

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{7}$$

V. Experiment

A supervised classifier is trained to predict which tweets contain environmental health risks and which do not. Then a multi-class classification is performed to identify the types of engagements during the crisis.

Experiment 1 Binary classification: In the first part of this experiment, to classify which tweet is relevant and which is not relevant to Barbados crisis, we extracted (TF-IDF, Psychometric, Linguistic, Twitter-specific, and Sentiment) features. Then, we combined all features into one file (all-features), then we ranked all features before selecting the top features that are assumed to improve the system's performance. Finally, we ran the WEKA used four supervised classification algorithms (mentioned earlier) to evaluate the performance of automatic detection. The F score, AUC and Accuracy values obtained from the binary classification are presented in Table I and Figure 2

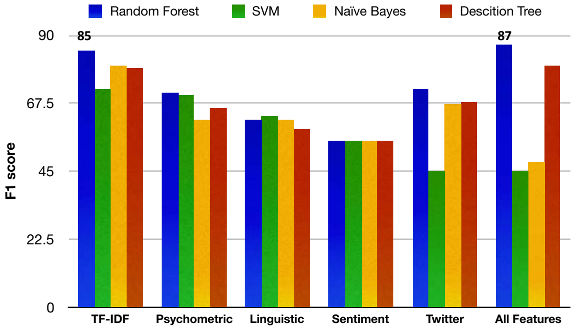

Fig. 2: Best F-score of models for the Binary classification among the four types of features and the combination of all features

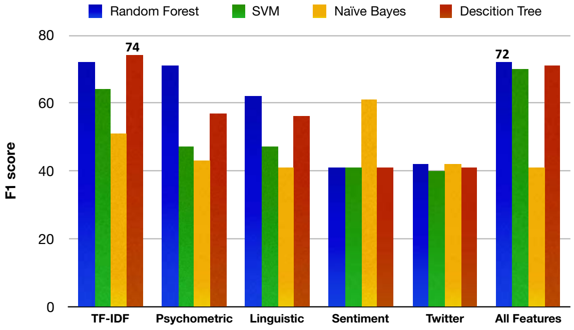

Fig. 3: Best F-score of models for the Multi-class classification among the four types of features and the combination of all features

Experiment 2 Multi-class classification: In the second part of this experiment, five classes (irrelevant, expressing negative engagement, reporting information, expressing positive engagement, and asking for information) were trained with the four classifiers as we mentioned previously using the WEKA platform. The agreement between the three MTurk annotators, measured using Cohen's kappa coefficient, was substantial (kappa = 0.64). The F score, AUC and Accuracy values obtained from the Multi-class classification are presented in Table II and Fig. 3 .

VI. Results and Discussion

Annotation Results: Here we present the results of Amazon MTurks labeling of community engagement type of the 4000 tweets according to the binary and multi-class classifications. More than half of tweets were classified as irrelevant (2245). Of the 1755 relevant tweets, reporting information (984) was the most common type of engagement. Expressing

TABLE I: Binary Results of RF, SVM, NB, and DT using (80%) Training data and (20%) evaluation

Feature set	**RF**			**SVM**			**NB**			**DT**		
	Acc.	AUC	F 1	Acc.	AUC	F 1	Acc.	AUC	F1	Acc.	AUC	F 1
TF-IDF	**85.13**	**91**	**85**	74.75	69	72	80.5	85	80	79.75	84	79
Psychometric	71.25	76	71	70	68	70	62	68	62	66.25	67	66
Linguistic	64.13	68	62	64.88	61	63	62.88	63	62	61.38	61	59
Sentiment	61.16	59	55	61.16	54	55	61.17	59	55	61.17	59	55
Twitter	72.96	81	72	59.66	50	45	66.19	75	67	67.91	70	68
All-Features	**87.31**	**96**	**87**	59.67	50	45	49.46	59	48	79.76	87	80

TABLE II: Multi-class Results of RF, SVM, NB, and DT using (80%) Training data and (20%) evaluation

Feature set	**RF**			**SVM**			**NB**			**DT**		
	Acc.	AUC	F 1	Acc.	AUC	F 1	Acc.	AUC	F1	Acc.	AUC	F 1
TF-IDF	**76.16**	86	72	70.82	63	64	40.13	83	51	**77.12**	80	**74**
Psychometric	71.25	76	71	61.64	50	47	35.75	64	43	60.27	62	57
Linguistic	64.13	68	62	61.64	50	47	36.98	61	41	57.81	60	56
Sentiment	56.67	49	41	56.67	50	41	72.37	48	61	56.67	50	41
Twitter	50.66	52	42	55.74	50	40	50	55	42	46.34	50	41
All-Features	**75.93**	**87**	72	70.27	62	70	33.5	80	41	72.57	77	71

negative engagement was the second most common form of engagement (296). These tweets shared resulting disgust and inconvenience caused by the environmental health hazards (sewage). 95 tweets were categorized as expressing positive engagement. Asking for information was the least common type of tweet engagement (26) . The inability of users to directly communicate with Barbados water utility on twitter may have contributed to this low type of engagement. When users asked for information they tended to direct their tweet towards a specific individual or organization.

Experiment Results and Interpretations: In our results, the modeling and classification were attempted on a machine with Intel Core i5 CPU @2.7 GHz with 8GB RAM configuration. In the first experiment, the best results were reached by using Random Forests classifier with (all-features) with 87% F1 score. It is interesting to note TF-IDF feature became the second highest features which recorded an 85% F1 score. In the second experiment using the multi-class model to detect the types of community engagement, results across the training techniques were comparable; Decision Tree with TF-IDF features achieved the highest F1 score with 74%. Whereas Random Forest with (TF-IDF and All-Features) reported the second highest F1 score with 72% scores.

To determine and rank the relevant attributes of each feature, we applied information gain equation 4. The top 10 features with the most significant weight for each class are listed in Table III. To distinguish between relevant and irrelevant tweets, as displayed in the table, the top affected feature in binary classification was Twitter-features with the following attributes: Following, Followers, Mention, and Hashtag. Whereas, in Multi-class classification, the (TF-IDF) got the most informative features among all other features where: zika, water, virus, case, and sewage recorded the majority of the weight.

VII. CONCLUSION, LIMITATIONS AND FUTURE WORK

The effectiveness of social media platforms in the field of environmental health can support decision makers and health organizations by measuring the feedback of people's responses about crises. In this paper, a framework for binary and multi-class community engagement classifications was proposed. The framework included choosing essential features of tweets, applying feature selection algorithm and training the dataset using machine learning algorithms.

While we consider that our outcomes are encouraging, we do see that there is still some uncertainty in the efficacy of our technique for classification. The first reason is the quality of labeling, knowing there are many advantages of using MTurk, such as easy access to a vast topic pool, the low rate of performing experiments and faster with producing results. The quality of annotation of the data can be much better if we could hire domain experts. The second reason is some types of community engagement might overlap in the meaning or might have very similar features which make the process too difficult to distinguish between tweets. For instance, we excluded 354 out of 4000 tweets because the three annotators could not agree with them (No Agreement). Based on our experiences, if the annotators are authorized to view the entire discussion preceding the tweet and are familiar with the content of the URLs this could mitigate these misunderstandings. Moreover, specifying five labels possibly caused confusion for the annotators; It may be possible to narrow the variety of the choices and make them only three for the future work. In addition, it is essential to note that our model results may not be adaptable to other fields without further examination. Knowing public opinions and beliefs towards environmental health risks topics may help academics, health agencies, and policymakers generate better strategies and guidelines for maintaining public health.

TABLE III: Top 10 ranked attributes for extracting relevant tweets of Binary and Multi-class classification, where T, L, P, S, and TF are shortened form of Twitter, Linguistic, Psychometric, Sentiment and TF-IDF

Binary classification										
Ranking	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Feature	Following (T)	Followers (T)	Exclam (L)	WPS (L)	Mention (T)	Hashtag (T)	problems (TF)	see (P)	test (TF)	hit (TF)
Weight	0.0592	0.0279	0.0271	0.0185	0.011	0.009	0.0035	0.003	0.0027	0.002
Multi-class classification										
Feature	zika (TF)	water (TF)	virus (TF)	case (TF)	ingest (P)	sewage (TF)	Tone (S)	positive (S)	Negative (S)	leisure (P)
Weight	0.157	0.092	0.079	0.076	0.065	0.051	0.05	0.04	0.037	0.03

In future research, we plan to employ graph theory such as community detection technique to explore other types of community engagement that may not be presented in our data. For more validation, we will work on another social media platform to predict community engagement.

Acknowledgment

Mr. Adel Alshehri and Ms. Aseel Addawood received academic scholarship from king Abdulaziz City for Science and Technology and from Al Imam Mohammad Ibn Saud Islamic University (Saudi Arabia), respectively.

References

[1] Health and E. L. Initiative. (2005) Priority environment and health risks. [Online]. Available: http://www.who.int/heli/risks/en/t

[2] 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*, vol. 86, no. 1, pp. 39–45, 2012.

[3] S. Bunney, S. Ward, and D. Butler, "Inter-organisational resilience for flood focussed emergency planning: examining multi-agency connectedness through twitter," *Water Practice and Technology*, vol. 13, no. 2, pp. 321–327, 2018.

[4] S. Vieweg, A. L. Hughes, K. Starbird, and L. Palen, "Microblogging during two natural hazards events: What twitter may contribute to situational awareness," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '10. New York, NY, USA: ACM, 2010, pp. 1079–1088. [Online]. Available: http://doi.acm.org/10.1145/1753326.1753486

[5] E. Aramaki, S. Maskawa, and M. Morita, "Twitter catches the flu: Detecting influenza epidemics using twitter," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP '11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 1568–1576. [Online]. Available: http://dl.acm.org/citation.cfm?id=2145432.2145600

[6] A. A. Ginart, S. Das, J. K. Harris, R. Wong, H. Yan, M. Krauss, and P. A. Cavazos-Rehg, "Drugs or dancing? using real-time machine learning to classify streamed dabbing homograph tweets," in *2016 IEEE International Conference on Healthcare Informatics (ICHI)*, Oct 2016, pp. 10–13.

[7] S. Bhattacharya, P. Srinivasan, and P. Polgreen, "Engagement with health agencies on twitter," *PLoS One*, vol. 9, no. 11, p. e112235, 2014.

[8] H. Park, B. H. Reber, and M.-G. Chon, "Tweeting as health communication: Health organizations use of twitter for health promotion and public engagement," *Journal of Health Communication*, vol. 21, no. 2, pp. 188–198, 2016, pMID: 26716546. [Online]. Available: https://doi.org/10.1080/10810730.2015.1058435

[9] J. Sutton, C. League, T. L. Sellnow, and D. D. Sellnow, "Terse messaging and public health in the midst of natural disasters: The case of the boulder floods," *Health Communication*, vol. 30, no. 2, pp. 135–143, 2015, pMID: 25470438. [Online]. Available: https://doi.org/10.1080/10410236.2014.974124

[10] P. Beaudeau, M. Pascal, D. Mouly, C. Galey, and O. Thomas, "Health risks associated with drinking water in a context of climate change in france: a review of surveillance requirements," *Journal of Water and Climate Change*, vol. 2, no. 4, pp. 230–246, 2011.

[11] G. Lotan, E. Graeff, M. Ananny, D. Gaffney, I. Pearce, and danah boyd, "The arab spring— the revolutions were tweeted: Information flows during the 2011 tunisian and egyptian revolutions," *International Journal of Communication*, vol. 5, no. 0, 2011. [Online]. Available: http://ijoc.org/index.php/ijoc/article/view/1246

[12] S. Etlinger and W. Amand, "Crimson hexagon [program documentation]," *Retrieved September*, vol. 15, p. 2016, 2012.

[13] J. W. Pennebaker and M. E. Francis, "Cognitive, emotional, and language processes in disclosure," *Cognition and Emotion*, vol. 10, no. 6, pp. 601–626, 1996. [Online]. Available: https://doi.org/10.1080/026999396380079

[14] T. Takahashi and N. Igata, "Rumor detection on twitter," in *The 6th International Conference on Soft Computing and Intelligent Systems, and The 13th International Symposium on Advanced Intelligence Systems*, Nov 2012, pp. 452–457.

[15] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling, "Twitterstand: News in tweets," in *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, ser. GIS '09. New York, NY, USA: ACM, 2009, pp. 42–51. [Online]. Available: http://doi.acm.org/10.1145/1653771.1653781

[16] N. Indurkhya and F. J. Damerau, *Handbook of natural language processing*. CRC Press, 2010, vol. 2.

[17] C. Lee and G. G. Lee, "Information gain and divergence-based feature selection for machine learning-based text categorization," *Information Processing Management*, vol. 42, no. 1, pp. 155 – 165, 2006, formal Methods for Information Retrieval. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0306457304000962

[18] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, Nov. 2009. [Online]. Available: http://doi.acm.org/10.1145/1656274.1656278

[19] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their Applications*, vol. 13, no. 4, pp. 18–28, July 1998.

[20] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, Mar 1986. [Online]. Available: https://doi.org/10.1007/BF00116251

[21] D. D. Lewis, "Naive (bayes) at forty: The independence assumption in information retrieval," in *Machine Learning: ECML-98*, C. Nédellec and C. Rouveirol, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 1998, pp. 4–15.

[22] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct 2001. [Online]. Available: https://doi.org/10.1023/A:1010933404324