Discovering Explanatory Models to Identify Relevant Tweets on Zika

Roopteja Muppalla[1], Michele Miller[2], Tanvi Banerjee[1] and William Romine[2]

***Abstract*— Zika virus has caught the worlds attention, and has led people to share their opinions and concerns on social media like Twitter. Using text-based features, extracted with the help of Parts of Speech (POS) taggers and N-gram, a classifier was built to detect Zika related tweets from Twitter. With a simple logistic classifier, the system was successful in detecting Zika related tweets from Twitter with a 92% accuracy. Moreover, key features were identified that provide deeper insights on the content of tweets relevant to Zika. This system can be leveraged by domain experts to perform sentiment analysis, and understand the temporal and spatial spread of Zika.**

I. INTRODUCTION

Zika has been around for decades but the current outbreak that started in 2015 has sparked significant concern. This is the first outbreak of Zika associated with microcephaly and Guillain-Barre syndrome, so management is still an important challenge [1]. The three ways to become infected with the Zika virus are: (i) an infected mother passing the virus to her fetus, (ii) being bitten by an infected *Aedes* mosquito, and (iii) through sexual contact[1]. There are currently no medications or vaccines to prevent or treat the Zika virus. As of January 18, 2017, 4,900 Zika related cases had been reported in the United States[2]. With the growing number of cases, it is important for health officials to recognize Zika virus hot spots and spread the necessary information to the public in real time.

Public health organizations often depend on traditional survey based methods to gather information about a disease outbreak. Though these methods are useful, they take a long time to recognize an outbreak. This is a major roadblock when trying to detect the rapid spread of a disease. However, social media can reduce this time lag while also allowing for studying public opinions on health issues. People often have health related conversations on social media and openly discuss diseases. Platforms like Twitter make it easy to share personal experiences so that people can empathize with each other. In particular, public opinion mining has been studied in the past for exploration of public views on important social issues such as gender-based violence [2], as well as to mine health related beliefs [3], [4]. Studies like the one about a cholera outbreak after the earthquake in Haiti [5] have demonstrated that Twitter data may represent a potential way to track diseases faster in future events.

[1]Department of Computer Science and Engineering,Wright State University Dayton, OH 45435, USA. roopteja, tanvi@knoesis.org

[2]Department of Biological Sciences, Wright State University, Dayton, OH 45435, USA. miller.1232, william.romine@wright.edu

[1]http://www.webcitation.org/6mhnTZk4b

[2]https://www.cdc.gov/zika/geo/united-states.html

Since Twitter has become a common platform for discussions about disease, researchers can have a greater understanding of the disease and can *communicate* and *address* any issues in real time. This study finds the best features which can be used to build a classifier to detect Zika related tweets. A quicker detection of a disease through social media can give more time to prepare the response team and contain the disease spread. Such a system can provide health officials a collective view of the public's health and also detect any future Zika outbreaks.

In this study, techniques from natural language processing (NLP) were used in combination with machine learning techniques to build models to classify Zika related tweets. In particular, we used data science techniques to not only build a strong classifier to identify relevant tweets on Zika, but also extract features that best discriminate the two categories: relevant and not relevant. The main focus of this study is a shift from black box methods that achieve high classification rates but cannot explain the results, to simpler and more explanatory models that provide deeper insights into the model performance. Through exploration of multiple models, we gain a deeper understanding of the content of tweets pertaining to Zika.

II. RELATED WORK

Multiple studies have used Twitter for exploration of public health issues [6], [7], [8]. One study focused on the spread of influenza from November 2008 to June 2010 and collected 300 million tweets [6]. Tweets were identified as relevant to influenza based on their influenza corpus using a support vector machine (SVM) based classifier. Pearson correlation was used to compare estimated values and annotations. Their method performed well in detecting influenza epidemics with high correlation (0.89 correlation).

Alvaro et al. [7] obtained a random sample of tweets over a 12 month period to analyse first-hand experience with selective serotonin reuptake inhibitors or cognitive enhancers. The ground truth consisted of 100 annotated tweets for 15 categories which were then compared to crowd sourced annotators by calculating Kappa values for each of the categories. Using URLs, hashtags, and N-grams from the tweets, Bayesian Generalized Linear Modeling was found to be the best technique for interpretation. In this study, we followed a similar approach in collecting and preparing the data for the classifiers.

Recently, we performed an exploratory study using Zika related tweets to determine what people were tweeting about Zika [8]. Tweets were collected over a period of 2 months. A two-stage classifier was used to find the Zika related tweets and to further classify the tweets into subcategories. The

classifiers used the whole tweet to generate a set of features whereas in this study, we extracted features based on natural language techniques to build a simple model with a fewer number of features.

While these different studies highlight the utility of using social media to monitor peoples thoughts regarding a specific disease outbreak, they did not discuss the role of features and their significance in classification. This study focuses on extracting features using Part of Speech tagging and N-gram techniques and identifying the set of features through model selection which will improve the classification results. Models which are simple and interpretable will help researchers to quickly classify Zika related tweets to address public concerns and misconceptions, similar to the research done on other diseases [9], [10], [11].

III. DATA COLLECTION

Tweets were collected over a period of two months (a total of 1,234,605 tweets) from Twitter based on the keywords *'zika'* and *'zika virus'* using a Twitter streaming application program interface (API). Though the tweets contain the word *'zika'*, not all tweets were relevant to our study. For example, tweet such as *'if you need me I'll b contracting the zika virus to avoid my ap test'* contains the word *'zika'* but it was used in the context of humor. But this study was focused on bridging the gap between public and health organizations in order to tackle the disease, so in our context such tweets are considered irrelevant. This results in the need to perform classification to remove such irrelevant tweets. For this, we took a random sample of 1,467 tweets from the dataset for analysis. This dataset was then annotated by three microbiology and immunology experts as to whether the tweets were relevant to Zika or not (1,137 tweets were labeled relevant). Inter-rater reliability among the annotators was found using Fleiss Kappa [12]. We calculated a Kappa value of 0.71 which indicates substantial agreement among the raters [13].

IV. FEATURE EXTRACTION

Once the data were collected, we needed features to help the learning algorithms or classifiers predict whether or not a tweet was relevant. A simple approach to extract features from the text is using a bags-of-words model where each word is considered a feature. But this results in a large number of features, which makes the learning algorithm difficult to process. Therefore, we made use of the following two ways to extract features from tweets:

A. Parts of Speech (POS) Features

Features were extracted from the tweets with the help of Stanford NLP POS tagger [14]. First, a feature vector with all the 25 POS tags was created. Then the tool annotator identified the features in the tweet and the count of each feature was recorded. For example, some of the features generated by the POS tagger for the tweet, *'RT @nationalpost: Canada confirms its first case of sexually transmitted Zika virus, in Ontario'*, are shown in Table 1. From these 25 POS features, two features were excluded namely *'existential verbal'* and *'proper noun verbal'*, as none of the tweets contained those two features.

TABLE I
POS TAG FEATURES FOR A TWEET

Tag (feature)	**Count (feature value)**	**Sample from the tweet**
discourse marker	2	RT, :
at-mention (@)	1	@nationalpost
proper noun	2	Canada, Zika, Ontario
verb	2	confirms, transmitted
nominal and verb	1	its
punctuation	1	,
adjective	1	first
pre- or post-position	2	of, in
common noun	2	case, viru
adverb	1	sexually

B. N-gram features

Features were extracted with the help of n-grams. N gram [15] is a sequence of n words from a given text which is treated as a single unit. As part of pre-processing, URLs, hashtags, and stopwords were removed from the tweets as these terms appear commonly in tweets and will not help the classifier to learn and distinguish Zika related tweets. For the tweet *'zika makes americans rethink travel'*, the features generated by n-grams are shown in Table 2.

TABLE II
N-GRAM FEATURES FOR A TWEET

N-gram	**Feature**
Unigrams	zika, makes, americans, rethink, travel
Bi-gram	zika makes, makes americans, americans rethink, rethink travel

Several studies use the entire unigram corpus to investigate text content in datasets such as tweets [8]. However, this study was performed by taking the top 10 occurring unigrams and bigrams in the dataset as the features for our analysis after pre-processing. For every n-gram, the count was increased if that selected unigram or bigram existed in the tweet. For example, if a selected unigram like *'zika'* is a feature and it occured 2 times in a tweet then the count of the occurrences was recorded as (2). Higher n-grams were not considered since the frequency of these was far less due to the tweet length constraint of 140 characters[3].

V. ANALYSIS

Using POS, selected unigram, and selected bigram features, there were a total of 43 features. Examples of features include: *'at mention', 'birth defects', 'cdc', 'emoticon', 'fight', 'funding', 'hashtag', 'microcephaly', 'pregnant*

[3]https://dev.twitter.com/basics/counting-characters

women', 'pronoun', 'public health', 'symptoms', 'treatment', 'URL'.

Using R[4] programming language, we created a simple logistic model considering all 43 features. To estimate the relative quality of the model containing all of these features in relation to simpler models which contain subsets of these features, we used Akaikes information criterion (AIC) [16].This gave a value of 957.78 for the full model (*All Features* in Table 4). Then we performed forward/backward stepwise model selection where features were added one at a time and tests whether the AIC will be improved by removing a previously added feature at each step. This process yielded a model with 27 features (*Stepwise* in Table 4) with an AIC value of 934.99.

Principal Component Analysis (PCA) was used to further reduce the number of features and to test whether the model with the reduced features gave us better results. After performing PCA, a scree plot (Figure 1) was used to select 2 components [17]. Since there is a low correlation between the individual features, it is highly unlikely for the features to have high correlation with the principal components. Therefore, a low cut off (0.2) was used to determine which features were associated with the principal components as shown in Table 3.

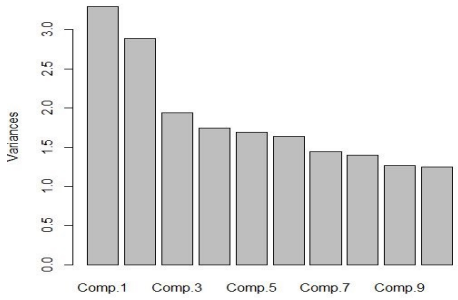

Fig. 1. Scree plot of factor eigenvalues.

Component 1 was comprised of **topical** features generated by n-grams such as *'birth defects', 'cdc', 'microcephaly'*, whereas, component 2 was comprised of **lexical** features generated by POS tagger such as *'adverb', 'pronoun', 'verb'*. These two components were able to explain a total of 16 features based on the cut off value of 0.2.

TABLE III
STRUCTURE MATRIX OF FEATURES ONTO COMPONENTS 1 AND 2.

Feature	Component 1: Topical	Component 2: Lexical
adjective	0.01	-0.13
adverb	-0.01	**-0.33**
birth defects	**-0.34**	-0.10
causes microcephaly	**-0.43**	-0.11
cdc	**-0.34**	-0.08
fight	0.03	0.05
verb	0.11	**-0.36**

[4]https://www.r-project.org/

The model built using just these two principal components had an AIC value of 1412.59. Therefore, a model (*All-2-PC*) was built using the 2 components and the remaining features that did not load onto these components. Similarly another model (*19-Stepwise-2-PC*) was built using these 2 principal components and the remaining features present in the *Stepwise* model that the two principal components did not explain. Finally, the *Stepwise* model was chosen as the best model based on the Akaike weights (w), which are used to give the relative likelihood of existence among the models within a probabilistic framework [18]. Based on this, we observed that PCA did not help in improving the model.

TABLE IV
AIC VALUES FOR DIFFERENT MODELS

Model	AIC	w
All	957.78	1.1E-5
Stepwise	*934.99*	**0.99**
PC	1412.59	0
19-Stepwise-2-PC	1035.50	0
All-2-PC	1009.03	0

We used the *Stepwise* logistic model to classify Zika related tweets from the dataset. Table 5 shows the confusion matrix, which gives the performance of the model in classifying the data. The results for this analysis (F measure of 0.92) were considerably better than the results generated by multiple classifiers for the 1000 unigram features extracted from the dataset through Weka [8], with F measures ranging from 0.82 to 0.89. These results show that the *Stepwise* model has more distinguishing features and is able to achieve high accuracy even with a simple logistic model, as opposed to more complex models such as SVMs, despite using relatively few features as compared to our earlier study [8].

TABLE V
CONFUSION MATRIX FOR STEPWISE MODEL

		Predicted	
		Relevant	**Not relevant**
Actual	**Relevant**	1071(94%)	66(6%)
	Not relevant	108(33%)	222(67%)

The *Stepwise* model contains topical features (as shown in Table 6) such as *'microcephaly', 'funding', 'fight', 'treatment', 'symptoms', 'health'* (part of the top 12 n-grams), which were able to classify Zika related tweets well. This sheds light on topics people tweet the most regarding Zika. We also observed that the *Stepwise* model contains 15 POS tag features, indicating that lexical components were useful in discriminating between the relevant and non-relevant tweets.

Along with the n-gram features and lexical features such as *'hashtag', 'at mention', 'URL'*, we observe that most of the tweets could potentially originate from a news source or retweets of this information. For example, 'Health Tech Forum' retweeted the following message, *'CDCgov: The best way to prevent #Zika is to prevent mosquito bites.*

TABLE VI

FEATURES IN STEPWISE MODEL

Topical features	*treatment, symptoms, microcephaly, first, fight, health, puerto rico, cdc, new, funding, health officials, white house*
Lexical features	*URL, hashtag, discourse marker, coordinating conjunction, interjection, at mention, punctuation, common noun, determiner, emoticon, numeral, verb, verb particle, existential, nominal possessive*

URL', which is indeed tweeted by 'CDC' through their official Twitter handle. These lexical features could also help researchers in analyzing public sentiment [19], [20] regarding Zika.

VI. CONCLUSION AND FUTURE WORK

From this study, we were successfully able to not only improve the performance of the relevance classifier as compared to the state-of-the-art classifiers [8], but were also able to extract meaningful and explanatory features for classification, as compared to the complete set of unigram features (1000 total) used for classification in our earlier study [8]. This not only allows us to better analyze system performance, but also improve the computation time and resources to build a high accuracy, real-time classification system for Zika-related tweets. For our next steps, we want to explore explanatory features for the next stage of classification to categorize the relevant tweets into sub-groups of treatment, symptoms, transmission, and prevention. Future work will also involve using sentiment-based features that can classify public sentiment regarding a specific topic within the tweets. Such a system will enable public health organizations to employ real-time decision making for epidemics like Zika and help address public concerns in a faster and more efficient manner.

VII. ACKNOWLEDGEMENTS

Banerjee and Muppalla would like to acknowledge funding from NIH project 1K01LM012439-01. Romine and Miller were partially supported by Department of Education I3 project U411C140081 and Institute of Educational Sciences (IES) award R305A150364.

REFERENCES

[1] S. S.-Y. Wong, R. W.-S. Poon, and S. C.-Y. Wong, "Zika virus infectionthe next wave after dengue?," *Journal of the Formosan Medical Association*, vol. 115, no. 4, pp. 226–242, 2016.

[2] H. Purohit, T. Banerjee, A. Hampton, V. L. Shalin, N. Bhandutia, and A. P. Sheth, "Gender-based violence in 140 characters or fewer: A# bigdata case study of twitter," *arXiv preprint arXiv:1503.02086*, 2015.

[3] M. J. Paul and M. Dredze, "You are what you tweet: Analyzing twitter for public health.," *Icwsm*, vol. 20, pp. 265–272, 2011.

[4] S. Bhattacharya, H. Tran, and P. Srinivasan, "Discovering health beliefs in twitter.," in *AAAI Fall Symposium: Information Retrieval and Knowledge Discovery in Biomedical Text*, 2012.

[5] C. W. Schmidt, "Using social media to predict and track disease outbreaks," *Environmental health perspectives*, vol. 120, no. 1, p. A31, 2012.

[6] 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*, pp. 1568–1576, Association for Computational Linguistics, 2011.

[7] N. Alvaro, M. Conway, S. Doan, C. Lofi, J. Overington, and N. Collier, "Crowdsourcing twitter annotations to identify first-hand experiences of prescription drug use," *Journal of biomedical informatics*, vol. 58, pp. 280–287, 2015.

[8] M. Miller, T. Banerjee, R. Muppalla, W. Romine, and A. Sheth, "What are people tweeting about zika? an exploratory study concerning symptoms, treatment, transmission, and prevention," *accepted, Journal of Medical Internet Research*, 2017.

[9] A. Signorini, A. M. Segre, and P. M. Polgreen, "The use of twitter to track levels of disease activity and public concern in the us during the influenza a h1n1 pandemic," *PloS one*, vol. 6, no. 5, p. e19467, 2011.

[10] A. Robertson, "Harnessing twitter, partnerships and the power of influence to stop stigma and spread awareness," in *2013 National Conference on Health Communication, Marketing, and Media*, CDC, 2013.

[11] X. Ji, S. A. Chun, and J. Geller, "Monitoring public health concerns using twitter sentiment classifications," in *Healthcare Informatics (ICHI), 2013 IEEE International Conference on*, pp. 335–344, IEEE, 2013.

[12] M. L. McHugh, "Interrater reliability: the kappa statistic," *Biochemia medica*, vol. 22, no. 3, pp. 276–282, 2012.

[13] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *biometrics*, pp. 159–174, 1977.

[14] K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N. A. Smith, "Part-of-speech tagging for twitter: Annotation, features, and experiments," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pp. 42–47, Association for Computational Linguistics, 2011.

[15] R. J. Solomonoff, "An inductive inference machine," in *IRE Convention Record, Section on Information Theory*, vol. 2, pp. 56–62, 1957.

[16] D. Anderson and K. Burnham, "Model selection and multi-model inference," *Second. NY: Springer-Verlag*, 2004.

[17] R. B. Cattell, "The scree test for the number of factors," *Multivariate behavioral research*, vol. 1, no. 2, pp. 245–276, 1966.

[18] E.-J. Wagenmakers and S. Farrell, "Aic model selection using akaike weights," *Psychonomic bulletin & review*, vol. 11, no. 1, pp. 192–196, 2004.

[19] 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*, pp. 36–44, Association for Computational Linguistics, 2010.

[20] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, "Sentiment analysis of twitter data," in *Proceedings of the workshop on languages in social media*, pp. 30–38, Association for Computational Linguistics, 2011.