A Study to Identify Key Topics Bearing Negative Sentiment on Twitter Concerning the 2015/2016 Zika Epidemic

Ravali Mamidi MS
Wright State University
Computer Science and Engineering
3640 Colonel Glenn Hwy, Dayton, OH 45435

Corresponding Author
Michele Miller MS
Wright State University
Biological Sciences
3640 Colonel Glenn Hwy, Dayton, OH 45435
Phone (574)261-3969
Fax (937)775-3320
miller.1232@wright.edu

Dr. Tanvi Banerjee PhD
Wright State University
Computer Science and Engineering
3640 Colonel Glenn Hwy, Dayton, OH 45435

Dr. William Romine PhD
Wright State University
Department of Biological Sciences
3640 Colonel Glenn Hwy, Dayton, OH 45435

Dr. Amit Sheth PhD
Wright State University
Kno.e.sis Computer Science and Engineering
3640 Colonel Glenn Hwy, Dayton, OH 45431

Abstract

Background In order to understand public sentiment regarding the Zika virus, social media can be leveraged to understand how positive, negative, and neutral sentiments are expressed in society. Specifically, understanding the characteristics of negative sentiment could help inform federal disease control agencies’ efforts to disseminate relevant information to the public about Zika related issues.

Objective The purpose of this study was to analyze public sentiment concerning Zika using posts on Twitter and determine the qualitative characteristics of positive, negative and neutral sentiments expressed.

Methods Machine learning techniques and algorithms were used to analyze the sentiment of tweets concerning Zika. A supervised machine learning classifier was built to classify tweets into
three sentiment categories: positive, neutral, and negative. Tweets in each category were then examined using a topic modeling approach in order to determine the main topics for each category, with focus on the negative category.

**Results** A total of 5,303 tweets were manually annotated and used to train multiple classifiers. These performed moderately well (F1 score = 0.69, 0.68) with text-based feature extraction. All 48,734 tweets were then categorized into the sentiment categories. Ten topics for each sentiment category were identified using topic modeling with a focus on the negative sentiment category.

**Conclusions** Our study demonstrates how sentiment expressed within discussions of epidemics on Twitter can be discovered. This allows public health officials to understand public sentiment regarding an epidemic and enables them to address specific elements of negative sentiment in real-time. Our negative sentiment classifier was able to identify tweets concerning Zika with three broad themes: *neural defects*, *Zika abnormalities*, and *reports and findings*. These broad themes were based on domain expertise and from topics discussed in journals such as MMWR and Vaccine. Since the majority of topics in the negative sentiment category concerned symptoms, officials should focus on spreading information about prevention and treatment research.

**Keywords:** Social Media; Machine Learning; Natural language processing; Epidemiology

**Introduction**

**Background**

Zika was discovered in 1947 in Uganda [1]. From the 1960s to 1980s, only 14 cases were diagnosed across Asia and Africa and typically caused mild symptoms [2]. The first large outbreak occurred in 2007 with the virus spreading from Yap across the Pacific with cases reporting mild symptoms. Cases were likely underreported from 1947 to 2008 due to the fact that the symptoms were similar to chikungunya and dengue however. It was not until this most recent outbreak that Zika was linked to Guillain-Barré syndrome and microcephaly [1]. Due to the new-found association of Zika and neurological disorders, people started expressing concern with the Zika virus, especially after an article in the BBC stated that the United States declared the Zika virus scarier than first thought [3].

In our previous exploratory study [4], we collected 1.2 million tweets over a period of two months and developed a two-stage classifier to categorize relevant tweets as concerning 4 disease categories: symptoms, treatment, transmission, and prevention. Tweets in each disease category were then examined using topic modeling to ascertain the top five themes for each category. We demonstrated how discussions on Twitter can be discovered to aid public health officials understanding of societal concerns. Our previous work focused on identifying relevant tweets with little emphasis on public sentiment. Much of the fear around Zika concerns the symptoms it causes [3]. Therefore, in this study, we turn our focus toward an in-depth analysis of symptoms of Zika, and undertake an analysis of specific positive, negative, and neutral sentiments expressed about the Zika virus.

**Related Works**

Identifying sentiment on a specific topic was pioneered by Chen et al. [5-6]. Since, several studies have looked at sentiment analysis on a variety of topics. Two studies focused on personal communication tweets only [7-8]. The study by Daniulaityte et al. [7] collected 15,623,869 tweets from May to November 2015 using keywords related to synthetic
cannabinoids, marijuana concentrates, marijuana edibles, and cannabis. They found that using personal communication tweets only compared to all tweets improved binary sentiment classification (negative and positive) but not multiclass classification (positive, negative, and neutral). A study by Ji, Chun, and Geller [8] collected tweets concerning listeria from September 26-28 and October 9-10 in 2011. They also focused on personal communication tweets only for sentiment classification (negative and not negative) and also found that the classifiers performed well after excluding non-personal communication (with a classification of F Measure=0.82-0.88). Instead of focusing on personal communication tweets alone, we included all relevant tweets after the BBC article about scientists declaring Zika scarier than initially thought [3] in our previous study [4]. A study by Househ collected approximately 26 million tweets and Google News Trends concerning the Ebola virus from September 30th to October 29th 2014 [9]. This study also influenced the decision to use all tweets and not just personal communication when they found that news feeds were the largest Twitter influencers during the Ebola outbreak.

Ghenai and Mejova [10] collected 13,728,215 tweets concerning Zika from January to August 2016. Tweets were annotated as dubunking a rumor, supporting a rumor, or neither. They concluded that mainstream news websites may help spread misinformation and fear. A study by Seltzer et al. [11] collected 500 images from Instagram from May to August 2016 using the keyword ‘Zika’. Of those 500 images, only 342 were related to Zika. Of those 342 images, 193 were coded as ‘health’ and 299 were coded as ‘public interest’. Of the ‘health’ images, the majority related to transmission and prevention, which is similar to what we found in our previous study on Twitter [4]. This shows results can be corroborated across different social media platforms. Seltzer et al. also found that many of the images portrayed negative sentiment and fear. Their study was limited to using images and was only concerned with negative sentiment. Our study will use tweets and will include positive, neutral, and negative sentiment.

In many of these studies, the main topical content within each sentiment category was not explored. We take this additional step in our study to determine the topics of public concern regarding the Zika virus. We also use all tweets including personal communication as well as news articles, because news articles can go viral and include negative sentiment as seen with the BBC article briefly described in the background section [3]. The phenomenon of news articles going viral and including negative sentiment is also discussed in our previous study [4].

**Purpose of the Research**

In this study, public sentiment concerning Zika virus symptoms was explored to determine important topical subcategories for positive, neutral, and negative tweets. Using the framework shown in Figure 1, two main research questions were addressed:

*R1a:* Data Annotation Analysis: What was the distribution of positive, neutral, and negative tweets in the gold standard dataset? What was the agreement between the two annotators’ labels used as the gold standard for the sentiment classification?

*R1b:* Classification Performance: How well can we categorize tweets as positive, neutral, and negative in an automated fashion?

*R2:* Topical Analysis: What were the main topics discussed in the three sentiment categories with a focus on the negative sentiment category?

**Methods**

To address the research questions, we built the following methodological framework in Figure 1. Tweets obtained from the previous study [4] using Twitris 2.0 [12] were preprocessed
and labeled as positive, negative and neutral by domain experts. Features were then extracted using word embeddings and n-grams. A two-staged classifier was built using the extracted features to identify the relevant tweets and then categorize them into the three sentiment categories. Preprocessed unlabeled tweets in each sentiment category were then analyzed using topic modeling techniques to find the top ten topics for each of the three sentiment categories. This allows public health officials to discover public sentiment regarding disease outbreaks, and address apprehensions in real-time.

![Figure 1. Block diagram of content retrieval using two-stage supervised classification followed by unsupervised analysis for characteristics of sentiment content.](image)

**Data Annotation Analysis (Addressing R1a):**

A total of 5,303 random tweets selected from a total of 48,734 tweets were annotated as positive, neutral, or negative by two annotators with domain knowledge related to Zika epidemics. A tweet was considered positive if it mentioned research discoveries related to Zika as seen in this tweet: “Zika structure discovered, raising hopes for new ways to combat virus” or reflected a positive attitude towards treatments, preventions, or funding for Zika (Table 1). A tweet was considered negative if it discussed the defects/disorders caused by Zika like “Zika may be linked to autoimmune brain disorder”. Tweets were considered neutral if they gave information with no emotionally charged wording such as “hope”, “combat”, “severe” etc. An example of a neutral tweet is “cdc offers new zika guidance for family planning”. Agreement was found using Cohen’s kappa which is a robust statistic useful for either inter-rater or intra-rater reliability testing and accounts for the possibility of guessing [13]. These tweets became known as the gold standard dataset once significant agreement was reached (Kappa>0.81) [13].

**Table 1. Example tweets classified as positive, negative, and neutral.**

<table>
<thead>
<tr>
<th></th>
<th>Positive</th>
<th>Negative</th>
<th>Neutral</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>#Bayer scientists aiding in fight against #Zika virus:</td>
<td>1. #news Zika virus may spread to Europe in coming</td>
<td>1. Zika symptoms, diagnosis and treatment, from the</td>
</tr>
</tbody>
</table>


Preprocessing

Before data analysis could begin, tweets had to be preprocessed by removing screen handles (@username), URLs, non-ascii characters, and retweet indicators (RT). Tweets were then further processed by removing single letters such as “a”, “e”, “i”, etc, extra spaces, and stop words. Stop words are the most commonly used words in the English language such as “and”, “in”, “for”, etc. This preprocessed tweet corpus was used for extracting features using the word embeddings (described in section 2.1.2) and N-grams (described in section 2.1.5). These features were extracted similarly to our earlier studies [4, 14].

Word Embedding: Machine learning algorithms are incapable of handling raw text or strings and require numeric data to extract knowledge from textual data and build applications. Word embedding is one technique which maps individual words to a predefined vector space in such a way that the semantic relation between words is preserved [15]. Additionally, words or phrases from the tweets were embedded into the n-dimensional space where n is the number of words in the corpus. After word embedding, a sentence can be considered as a sequence of points which are grouped according to a semantic criterion so that two similar words are close to each other. It captures the context of words, while at the same time reducing the number of features in the data. In order to provide a better understanding of word embedding, we provide an example from a sample of our dataset. For visualizing the high dimensional data, we used a technique called t-Distributed Stochastic Neighbor Embedding (t-SNE) which maps each data point to a lower dimension (of size two) [16]. From Figure 2, we see the spatial distribution of a random sample of 100-word embeddings generated from the word2vec model [17]. We see that the words that are similar eventually form clusters (come spatially closer) in the vector space. For example, words like ‘outbreak’, ‘fears’, ‘postponed’, ‘summer’ (cluster 1) formed a cluster because they are used in the same context; in the case of the Rio Olympics and words like ‘republicans’ and ‘congress’ (cluster 2) are grouped together as they are used in the context of Zika funding.
Figure 2. Visualization of Zika word embedding using t-SNE which shows clusters of related word groups within the context of Zika tweets.

**Models:**
We used two different main models for classification. One was Word2vec [18] and the other was an N-gram model [14].

**N-gram model:** In this model, features were extracted from tweets using the Stanford Natural Language Processing Part of Speech tagger [19] and N-grams [20] where an n-gram represents a sequence of words treated as a single entity or feature. Initially, features were identified from the tweets and the count for each feature was determined. In total, there were 61 features. Examples include: ‘AT_Mention’, ‘Zika’, ‘Discourse_marker’, ‘microcephaly’, ‘fetal’, ‘ProNoun’, ‘health’, ‘birth defects’, ‘Zika infection’, ‘Hashtag’, and ‘brain damage’.

**Word2vec model:** Word2Vec composes two different methods: continuous bag of words (CBOW) and Skip-gram [21]. In the CBOW method, the goal is to predict a word given the surrounding words i.e. the words before and after it [21]. Skip-gram is the opposite: we want to predict surrounding words given a single word [21]. The skip-gram method with negative
sampling works best with the medium or large sized datasets [15]. As our dataset was considered medium sized [15], we used the Skip-gram model with a negative sampling rate of ten.

For the word embeddings, we used the Gensim library version 2.2.0 of Python version 3.5.4 [22] for converting all the words to an n-dimensional space before training the classifiers. The tokenized words were then fed to the word2vec tool and trained with the Skip gram model. We considered a window size of four because the average length of the tweets was less than ten words which means four tokens apart from the target words are considered as adjacent words.

With these collective parameters, we generated the word vectors of size 300 and tested the learned vectors using the similarity functionality of the word2vec. To evaluate the vectors generated using the tool, we selected two words ‘dengue’ and ‘Zika’ which are mosquito borne diseases to assess similarity. Similarity is used to find the distance between two vectors. The closer the similarity is to 1, the more closely related the words [23]. The similarity was 0.92 which indicates the words are closely related or used in a similar context. When words like microcephaly and pregnant were used, it gave related words like woman, women, and infected, among others.

Vector operations such as sum and mean were used to build the final feature vector. The following are the operations performed on the word vectors.

1. **Sum of Word Embeddings:** This is the sum of all word vectors in the tweet.
   \[ FV_{\text{Sum}} = \sum W \]

2. **Mean of word Embeddings:** Average of all the word vectors in the tweet.
   \[ FV_{\text{Mean}} = \frac{1}{n} \sum W \]

‘W’ represents word in a tweet and ‘FV_{\text{Sum}}, FV_{\text{Mean}}’ represents the feature vector of the tweets.

**Classification Performance (Addressing R1b):**

Supervised classification algorithms including logistic regression, support vector machines with rbf kernel, and random forest were used for classifying the tweets into the three sentiment categories. These methods rely on labeled data, in this case the 5,303 randomly selected tweets that were annotated as positive, neutral, or negative by the two annotators from a total of 48,734 tweets. These classifiers were trained to categorize tweets into the specified categories based on the gold standard derived by the annotators.

The performance of each classifier was assessed using the stratified k-fold cross validation as we had an unbalanced dataset. We report k=7 because there was no improvement in the result with increase of k and also saves computation time. The stratified k-fold maintains equal number of samples for each annotator-labeled class [24]. In this method, one subsample (fold) of tweets was used for a testing set and the remaining six for training. This was repeated seven times, with each subsample being used as the testing subsample once [24]. This study reports average recall (indication of category tweets not missed by the classifier), precision (correctly categorized tweets) and F1-scores (weighted average of precision and recall) as measures of classification performance for each classifier.

**Topical Analysis (Addressing R2):**

Previous studies such as Lau, Collier, and Baldwin have shown the usefulness of Latent Dirichlet Allocation (LDA) for grouping tweets into themes in short text documents such as tweets [25]. In this study, we used LDA topic modeling to identify the underlying topics discussed within each of the sentiment categories. This aids better qualitative exploration of the subtopics in each of the three categories.
To determine the number of topics required for topic modeling, we used perplexity, a measure used to evaluate topic models generated by LDA where the smaller the perplexity score, the better the generalization performance [26, 22]. We used this measure to evaluate the topic modeling results by testing a range of 2 to 100 topic models for the three sentiment categories. For calculating the perplexity measure, preprocessed tweets were used. Words that occurred only once or twice in the corpus were removed as they increase the number of topics which will not give generalizable information [26].

Results

In this section the distribution of tweets in the gold standard dataset is discussed. The performance of three different classifiers using the word2vec and Ngram models is also explained. Third, the topic modeling results for the positive, neutral, and negative categories is explored with a focus on the themes that emerged in the negative sentiment category.

Data Annotation Analysis (Addressing R1a):

In order to train the classifiers, the gold standard dataset had to be created as described in the methods section above. The kappa value for the level of agreement between the two annotators was 0.95, indicating near perfect agreement [13]. The distribution of the tweets in the gold standard dataset are shown in Figure 3. The majority of tweets displayed negative sentiment (2,423; 46% of the total tweets) and the fewest displayed positive sentiment (1,010; 19%). As can be seen in Figure 3, there is high class imbalance in the three sentiment categories.

![Figure 3. Distribution of tweets in three sentiment categories.](image)

Classification Performance (Addressing R1b):

Table 2 provides the performance of the two text processing models and the corresponding classifiers. The N-gram model performed slightly better than the word embedding model. For this dataset, classifiers performed reasonably well with F1-scores ranging from 0.48 to 0.68. However, the logistic regression classifier used with the N-gram model performed the best with an F1-score of 0.68. This performance is comparable to similar studies [7, 18].
Table 2. Classifier performance for Sentiment Analysis using seven-fold cross validation. The classifiers used are: Logistic, Support Vector Machine, Random Forest.

<table>
<thead>
<tr>
<th>Model</th>
<th>Classifier</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word2vec</td>
<td><strong>FV</strong> Sum</td>
<td>Logistic Regression</td>
<td>0.68</td>
<td>0.66</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Support Vector Machines</td>
<td>0.67</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Random Forest</td>
<td>0.55</td>
<td>0.53</td>
</tr>
<tr>
<td></td>
<td><strong>FV</strong> Mean</td>
<td>Logistic Regression</td>
<td>0.63</td>
<td>0.63</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Support Vector Machines</td>
<td>0.66</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Random Forest</td>
<td>0.50</td>
<td>0.50</td>
</tr>
<tr>
<td>Ngrams</td>
<td>Logistic Regression</td>
<td>0.69</td>
<td>0.68</td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td>Support Vector Machines</td>
<td>0.65</td>
<td>0.65</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td>Random Forest</td>
<td>0.68</td>
<td>0.67</td>
<td>0.67</td>
</tr>
</tbody>
</table>

Using the n-gram based logistic regression sentiment classifier, we categorized all 48,734 tweets obtained from our previous study (Figure 4) [4]. The total number of negative tweets was almost four times larger than the positive and neutral categories combined. We can clearly see from Figure 4 that this is a highly unbalanced dataset with the majority of tweets belonging to the negative sentiment category.
Topical Analysis (Addressing R2):

Within the tweets with negative sentiment, the perplexity decreased rapidly until about 10 topics and then leveled off (Figure 5). The perplexity graph for the positive and neutral category are available online [28]. This indicates that increasing the number of topics after 10 will not significantly improve generalizability of the results of the LDA models [29]. Therefore, ten topics per sentiment were extracted.

![Perplexity plot measures for the 7-fold cross-validation of topic modeling for the negative sentiment category.](image)

The results of the LDA are discussed below for the positive, neutral, and negative categories. First, the topics for the positive and neutral categories will be briefly discussed. The tables including the theme names, topic words, and example tweets for the positive and neutral topic models are available online [28]. Then, a more detailed explanation of the negative sentiment topics will be presented.

**Topics from Positive and Neutral Sentiment**

Within the positive sentiment themes, there were 4 broad qualitative topics within the 10 topics chosen using the perplexity measure with LDA: mosquito killing methods, models to help understand the Zika virus, detection of the Zika virus in cells, and treatment and prevention discoveries (Table 3). These broader themes were labelled based on domain expertise and from journals such as Vaccine and MMWR, allowing further categorization of the ten topics. For the
broader theme of models that help understand the Zika virus, topic #1 contained tweets concerning a new model researchers were developing to study Zika pathogenesis and topic #2 described 3d printed mini brains used for understanding the Zika virus. For the mosquito killing methods theme, topic #4 contained tweets concerning sweat-emitting Brazilian billboards killing the Zika carrying mosquitoes and topic #10 addressed other ways of killing Zika carrying mosquitoes. In the treatment and prevention discoveries theme, topic #3 comprised of tweets regarding the discovery of how Zika stunts the development of a fetus, topic #5 characterized the development of vaccines to treat Zika, and topic #8 reported about the IBM magic bullet to destroy all killer viruses. This “magic bullet” is actually a macromolecule that will attach to the surface of any virus and prevent it from attaching to a human cell [30]. If the virus cannot attach and enter a cell, infection is prevented. The macromolecule is also basic, neutralizing the acidity of an infected cell in case the virus is already infecting human cells by the time the “magic bullet” is used [30]. In the broader theme of detection of the Zika virus in cells, topic #6 regarded different types of tests for identifying Zika infection, topic #7 outlined the detection of Zika using fetal tissue, and topic #9 detailed the detection of Zika accumulations in the brain.

Overall, the broader themes in Table 3 (model, mosquito, virus discovery, and detection) were present in the Positive sentiment category because they all have to do with helping prevent transmission, or research that could lead to treatments. Both of these topics are viewed positively by the public due to the fact that they help prevent the defects that have become associated with Zika. For example, tweets in the mosquito theme discussed ways to kill mosquitoes, which would help prevent the spread of Zika [31]. Tweets in the model and viral discovery theme addressed discoveries that could help lead to treatments, like the IBM magic bullet [30]. Virus discovery tweets were positive because they pointed to faster ways to detect Zika. Knowing where Zika accumulates would help with developing treatments [32]. Tweets in the positive category also used words with positive connotations such as “understand”, “develop”, “hope”, “discover”, “benefit”, and “reveal”, among others. Themes in the positive sentiment category mainly addressed research to treat Zika and prevention methods. Themes in the neutral category mostly comprised posts from news agencies stating facts.

Table 3. Positive sentiment topic modeling results grouped together based on the broader themes. The numbers reflect the size of the theme. For example, the topic mouse model had more tweets than 3d-printed mini brains.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Words</th>
<th>Tweet</th>
<th>Broader Theme</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1 Mouse Model</td>
<td>Researcher, mouse, model, develop, health, research</td>
<td>new #zika mouse model researchers develop another mouse model of zika infection that mimics the disease in humans</td>
<td>Model</td>
</tr>
<tr>
<td>#2 3d-Printed Mini Brains</td>
<td>Scientist, test, brain, mystery, help</td>
<td>mini 3d printed brains help scientists understand zika virus</td>
<td>Mosquito</td>
</tr>
<tr>
<td>#4 Brazilian Billboards</td>
<td>Rapid, billboard, emit, Brazilian, structure</td>
<td>sweat-emitting Brazilian billboards lure zika-carrying mosquitoes to their death</td>
<td>Mosquito</td>
</tr>
<tr>
<td>#10 Killing Mosquitoes</td>
<td>Mosquito, infect, kill, insight, biomolecular</td>
<td>researchers develop #algae to kill #mosquitoes carrying viruses like #zika</td>
<td></td>
</tr>
<tr>
<td>-----------------------</td>
<td>---------------------------------------------</td>
<td>-------------------------------------------------------------------</td>
<td>Virus Discovery</td>
</tr>
<tr>
<td>#3 Fetal Brain Development</td>
<td>Fetus, human, discover, help</td>
<td>how zika virus stunts foetal brain development researchers have discovered how hijacking a human immune mole…</td>
<td></td>
</tr>
<tr>
<td>#5 Vaccines</td>
<td>Model, infect, vaccine, provide, develop</td>
<td>mouse models of zika virus infection in pregnancy provide basis to develop vaccines, treatments</td>
<td></td>
</tr>
<tr>
<td>#8 IBM Magic Bullet</td>
<td>Kill, develop, understand</td>
<td>IBM research IBM announces magic bullet to zap all kinds of killer viruses, like #zika by seancaptain</td>
<td></td>
</tr>
<tr>
<td>#6 Zika Tests</td>
<td>Urine, discover, pattern, Jamaica, programmable, molecular</td>
<td>#saling follow interim cdc guidance finds urine specimen better than serum for rapid and specific zika testing - cdc</td>
<td>Detection</td>
</tr>
<tr>
<td>#7 Fetal Tissue Research</td>
<td>Fetal, tissue, infect, detect, equip, test</td>
<td>last month, fetal tissue research helped doctors’ understand how the zika virus infects fetus &amp; how to detect its presence much</td>
<td></td>
</tr>
<tr>
<td>#9 Zika Accumulation</td>
<td>Reveal, accumulate, Zika, virus, examine, pregnancy, report</td>
<td>one of the first mouse models of #zika reveals the virus accumulates in the brain</td>
<td></td>
</tr>
</tbody>
</table>

Within the neutral sentiment topics, there were 3 broader qualitative themes: public health messages, knowledge gaps, and Zika characteristics (Table 4). In the public health messages topic #1 explained how scientists were trying to unravel the Zika mystery, topic #2 cautioned about the dangers of Zika infection to pregnant mothers, topic #3 declared that Zika is a mosquito borne disease, topic #4 specified the laws regarding birth control and abortion, topic #5 discussed fighting the mosquitoes, and topic #6 regarded the officials warning the public to be careful not to be bitten at work. Knowledge gaps consisted of topic #7 which discussed knowledge gaps concerning the Zika virus. In the Zika characteristics theme topic #8 affirmed Zika symptoms, topic #9 included comparisons between dengue and Zika, and topic #10 described fetal brain damage from Zika infection.

In this case, the broader themes in Table 4 (public health messages, knowledge gaps, and Zika characteristics) highlight the neutral sentiments because the tweets in these themes were from public health experts and news agencies informing the public and thus are more likely to state facts than opinions. For example, the tweet “Officials: Zika-Infected Couples Should Postpone Pregnancy” is a statement from officials about postponing pregnancy during a Zika outbreak to help prevent babies born with birth defects. Some tweets were neutral even though
they contained words with both positive and negative connotations because the sentiment of the tweet overall is neutral, such as this tweet “#voanews brazil scientists seek to unravel mystery of zika twins scientists struggling to unravel t…” Topics 1 through 6 all contained messages from public health agencies and were therefore labeled as public health messages. Topics 8 through 10 concerned characteristics of the Zika virus and thus were grouped together. Topic 7 did not belong in either category and was therefore made a separate theme. As you have seen, the neutral topics contained tweets from news agencies and public health officials. The negative sentiment topics also contained some tweets from news agencies and public health officials but additionally contained opinion tweets from the public.

Table 4. Neutral sentiment topic modeling results grouped together based on the broader themes. The numbers reflect the size of the theme.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Words</th>
<th>Tweet</th>
<th>Broader Theme</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1 Zika Mystery</td>
<td>Brazil, common, unravel, question, important, disease, issue</td>
<td>#voanews brazil scientists seek to unravel mystery of zika twins scientists struggling to unravel t…</td>
<td>Public Health Messages</td>
</tr>
<tr>
<td>#2 Aedes Mosquito</td>
<td>Mosquito, infect, pregnancy, outbreak, women, child</td>
<td>the zika virus and the dengue mosquito have a common nature. very resistant ones, and very dangerous too. infects mothers with pregnancy!</td>
<td></td>
</tr>
<tr>
<td>#3 Mosquito Born Illness</td>
<td>Symptom, today, health, born, mosquito, effect</td>
<td>zika is a mosquito borne illness that does not present symptoms in many people. that is a very dangerous thing.</td>
<td></td>
</tr>
<tr>
<td>#4 Abortion</td>
<td>Abortion, learn, worse, survive, guideline, paper</td>
<td>zika virus, birth control and abortion our anti-woman laws will make this worse.</td>
<td></td>
</tr>
<tr>
<td>#5 Fight the Bite</td>
<td>Infect, fight, bite, affect, death</td>
<td>only 1 in 4 pple infected w/ #zika will show symptoms. fight the bite, destroy mosquito breeding sites #nobitenozika</td>
<td></td>
</tr>
<tr>
<td>#6 Officials’ Warning</td>
<td>Officials, control, disease, center, researcher</td>
<td>health officials warn against exposure to zika at work the centers for disease control and prevention #atlanta</td>
<td></td>
</tr>
<tr>
<td>#7 Knowledge Gap</td>
<td>People, expert, relate, ebola, cure</td>
<td>various ‘experts’ need to get up to speed on the zika+ front now. time is of an essence. many people are ‘behind the curve’.</td>
<td>Knowledge Gap</td>
</tr>
<tr>
<td>#8 Symptoms</td>
<td>Fever, scarier, infect, eye, first</td>
<td>zika symptoms – fever, rash, joint pain, and/or red eyes. most people infected</td>
<td>Zika Characteristics</td>
</tr>
</tbody>
</table>
typically don’t have symptoms though.

<table>
<thead>
<tr>
<th>#9 Dengue</th>
<th>Dengue, flu, rash, compare, cause, malaria</th>
<th>dengue &amp; zika have a rash, fever etc. 4 dengue strains increasing in ja. docs need to be careful #testedorsuspected</th>
</tr>
</thead>
<tbody>
<tr>
<td>#10 Fetal Brain Damage</td>
<td>Fetus, information, prevent, symptom, damage, fetal</td>
<td>“why fetal tissue research is crucial to saving babies from zika new study uncovers ‘alarming’ information…”</td>
</tr>
</tbody>
</table>

**Topics from the Negative Sentiment:**

We had chosen to focus on the topics from the negative sentiment category because negative sentiment is the sentiment health officials will be most concerned with as there is greater need for intervention and information dissemination in these topics. Additionally, this was the category with the majority of the tweets (Figure 4). Health officials are especially concerned with negative sentiment because it reflects public concerns that must be addressed. By knowing what is and is not of concern, health officials can tailor their messages accordingly. The topic model results for negative sentiment were shown in Table 3. In the negative sentiment topics, there were 3 broader topics: neural defects caused by Zika infection, abnormalities due to Zika infection, and reports and findings concerning the Zika virus. Topics #1 brain defects, #2 neurological effects, #5 fetal effects, and #8 Guillain-Barré Syndrome all concerned the nervous system. Topics #6 Zika abnormalities and #9 Zika effects were both related to abnormalities due to Zika infection. Topics #3 initial reports, #4 Zika impact, #7 ultrasounds, and #10 dengue association all concerned reports and findings concerning the Zika virus. There was significant overlap between topics #3 and #4 because they both addressed reports and findings concerning the Zika virus. However, topic #3 initial reports includes tweets stating the locations where Zika is spreading while topic #4 Zika impact includes tweets concerning the BBC article that describes Zika as scarier than initially thought [3].

The broader themes in Table 5 (neural defects, Zika abnormalities, and reports and findings) were all negative because they addressed topics of concern for the general public. Before this outbreak, Zika was considered a mild illness with only 14 reported cases [2]. It was not until this most recent outbreak that Zika became associated with microcephaly, Guillain-Barré Syndrome, and Congenital Zika Syndrome, all of which caused fear and concern across the globe [1, 4, 33, 34].

Table 5. Negative sentiment topic modeling results grouped together based on the broader themes. The numbers reflect the size of the theme.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Words</th>
<th>Tweet</th>
<th>Broader Theme</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1 Brain Defects</td>
<td>Brain, microcephaly, baby, disorder, confirm, cause</td>
<td>#zikavirus confirmed zika causes brain damage in babies born with microcephaly brain abnormalities in babies</td>
<td>Neural defects</td>
</tr>
<tr>
<td>#2 Neurological Effects</td>
<td>Severe, problem, immune, neural, death, birth</td>
<td>human neural stem cells infected by zika subsequently trigger an innate immune response that leads to cell death</td>
<td></td>
</tr>
<tr>
<td>-------------------------</td>
<td>-----------------------------------------------</td>
<td>--------------------------------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>#5 Fetal Effects</td>
<td>Brazil, fetus, shrink, development, disrupt, outbreak, pregnancy</td>
<td>in brazil zika eats away at fetal brain, shrinks or destroys lobes controlling thought &amp; prevents development.</td>
<td></td>
</tr>
<tr>
<td>#8 Guillain-Barré Syndrome</td>
<td>Syndrome, rare, case, associate, cause, microcephaly</td>
<td>cases of rare nervous disorder guillain-barre syndrome may increase if zika spreads via</td>
<td></td>
</tr>
<tr>
<td>#6 Zika Abnormalities</td>
<td>Brain, eye, abnormality, scientific, consensus, confirm, relate</td>
<td>a9 zika associated complications for pregnancy include miscarriage, stillbirth, brain abnormalities and eye abnormalities. #reuterszika</td>
<td></td>
</tr>
<tr>
<td>#9 Zika Repercussion</td>
<td>Zikavirus, infect, child, adult, fetal</td>
<td>researchers says that zika virus infection can stunt growth of children</td>
<td></td>
</tr>
<tr>
<td>#3 Initial Reports</td>
<td>Report, puertorico, infect, link, defect</td>
<td>puerto rico reports first zika-linked birth defect - puerto rico reports first zika-linked birth defect</td>
<td></td>
</tr>
<tr>
<td>#4 Zika Impact</td>
<td>impact, spread, reuters, mosquito, scarier</td>
<td>#reuters zika spread, impact 'scarier than we initially thought' u.s. health official</td>
<td></td>
</tr>
<tr>
<td>#7 Ultrasounds</td>
<td>ultrasound, doctor, baby, unborn, infect</td>
<td>#chevycar ultrasounds missed zika infection until the one showing serious harm to her baby</td>
<td></td>
</tr>
<tr>
<td>#10 Dengue Association</td>
<td>Expert, warn, sound, dengue, causal, fetus, spread, microcephaly</td>
<td>lab findings hint that dengue antibodies intensify zika infection=&gt;leading to microcephaly &amp; gbs²? evidence</td>
<td></td>
</tr>
</tbody>
</table>

*GBS: Guillain-Barré syndrome*

**Discussion**

In the discussion section, we will address one cause of tweets being misclassified with some examples. The three negative sentiment broader themes, neural defects, Zika abnormalities, and reports and findings, will then be explored and discussed in more detail.

**Classification Analysis**
As seen in Table 2, classification is not 100% accurate, implying that some tweets were misclassified. We will focus on the negative tweets since those were the focus of our discussion. Some tweets were misclassified due to words like ‘active’, ‘saliva’, ‘feds’, ‘busted’, ‘beast’, ‘prenatal’ which were not seen by the model because the count of these words is less than the minimum count (set to 5) parameter given in the Word2vec model and hence were discarded. The minimum count was set to 5 since words used fewer than 5 times do not add any information to the analysis and the default setting in Gensim is set to 5 as the minimum count [35]. Adding more training data could improve these results, however, a study by Nakov et al. annotated 6,000 tweets and had similar F1-scores to our study [36]. Since these words occurred fewer than 5 times, the algorithm was not able to identify these tweets as negative as it was not able to determine the words closer to these words. A couple of examples of tweets that were correctly and incorrectly identified are shown in Table 4.

Table 6. Sample tweets of correctly and incorrectly identified negative tweets by word embedding model.

<table>
<thead>
<tr>
<th>Correctly identified negative tweets</th>
<th>Incorrectly identified negative tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. #3tking Zika virus makes Rio Olympics a threat in #Brazil and abroad, #health expert says <a href="https://t.co/9giNkKNM3B">https://t.co/9giNkKNM3B</a></td>
<td>1. #seattle major zika fail! feds busted for lazy response …</td>
</tr>
<tr>
<td>2. #ap breaking cdc no longer any doubt that zika virus causes birth defects.</td>
<td>2. @DrFriedenCDC Scary how you could substitute prenatal alcohol in place of Zika!Same symptoms,hidden- YetCDC quiet <a href="https://t.co/3JBeHILqeW">https://t.co/3JBeHILqeW</a></td>
</tr>
</tbody>
</table>

**Topic Model**

In this section, we focus on the negative sentiment topics of neural defects, Zika abnormalities, ultrasounds, and dengue association. These themes and topics were chosen because they were topics of public concern, have been addressed by the CDC or WHO [37, 33, 38-40, 41] and can be addressed by officials to help mitigate the concern. Section 4.2.1 describes the first broad theme, neural defects. Section 4.2.2 describes the second broad theme, Zika abnormalities. Section 4.2.3 focuses on ultrasounds, a topic in the reports and findings broader theme. Section 4.2.4 details dengue association, another topic under the broader theme of reports and findings. Zika impact was not addressed because it is the focus of our previous paper [4]. Initial reports were not addressed since it is specific to this outbreak, and officials and the public cannot wholly prevent the spread of the Zika virus.

**Neural Defects**

Neural defects is a broader theme of concern for the public that needs to be addressed to mitigate fear and concern due to the defects to the nervous system caused by Zika virus infection. For Table 5, topics #1 (brain defects), #2 (neurological effects), #5 (fetal effects), and #8 (Guillain-Barré Syndrome) all concern the neural system. For example, topic #1, brain defects, points to brain damage in babies due to microcephaly as seen in this tweet “scans show extent of brain damage in babies with microcephaly associated with zika...”. Microcephaly has been a topic of concern for the CDC since babies born with microcephaly will require assistance
throughout their lifetime [37, 42]. The topic neurological effects (#2) includes tweets discussing the death of neural stem cells which leads to neurological disorders in humans, as seen in this tweet, “zika virus targets human cortical neural progenitors causing cell death & attenuated neural cell growth” [27]. The topic fetal effects (#5) also address brain shrinking or brain damage but additionally the tweets discuss the destruction of the brain lobes that control thought, vision and other functions in fetuses as seen in this tweet, “scans & autopsies show that zika eats away at the fetal brain. it shrinks or destroys lobes that control thought, vision & other functions”. Guillain-Barré Syndrome (topic #8) is a sickness caused by damage of nerve cells. The tweet "human neural stem cells infected by #zika subsequently trigger an innate immune response that leads to cell death" includes information on how Zika can lead to damage of neural stem cells and causes a disease like Guillain-Barré syndrome [44]. The reader can see how topics #1, #2, #5, and #8 all include information on neural issues following Zika infection, but all focus on different issues and are therefore, three separate topics. By looking at these tweets public health officials can see the public is concerned about the neurological defects caused by Zika. Therefore, the next steps officials need to take is to focus on how to prevent mosquito bites, especially when pregnant in order to prevent these neurological defects. The ‘Fight the Bite’ campaign is an example of such an effort [41].

Zika Abnormalities

Zika abnormalities is also an important broader theme to address due to the fear and concern of abnormalities and defects in infants due to Zika virus infection during pregnancy. In Table 3, the topics #6 (Zika abnormalities) and #9 (Zika effects) are both related to abnormalities due to Zika infection but include diverse problems. The topic Zika abnormalities (#6) describes various anomalies associated with the fetus and babies born with Zika infection as seen in this tweet, “birth defects linked to #zika now also incl hearing loss, vision problems, impaired growth, abnormalities in limbs”. These types of abnormalities are termed as Congenital Zika Syndrome by the CDC and includes a collapsed skull, eye scarring, severe muscle tension, and brain calcification [33, 34]. The topic Zika effects (#9) focuses on the stunt in growth and development of children. Again, both of these topics concern abnormalities due to Zika infection but focus on two different abnormalities and are therefore kept as two distinct topics. By pushing prevention like the ‘Fight the Bite’ campaign, officials can help ease fears concerning these abnormalities.

Ultrasounds

Ultrasounds is another important topic to address due to the fact that initial ultrasounds fail to reveal microcephaly and other birth defects, leading to a false sense of security for the couple [38, 45, 39, 46]. As previously stated, Zika is linked to microcephaly; however, ultrasounds were found to have high false negative predictions regarding the presence of microcephaly during the first and second trimesters of a woman’s gestational period [45]. Therefore, the topic of ultrasounds is important to discuss because pregnant women may have a false sense of security after getting an ultrasound and Zika not being detected in their fetus in the early stages of pregnancy. The CDC states on their website that microcephaly is more readily detected late in the second trimester to early in the third trimester [38]. Researchers are also recommending parents have an MRI of their newborn’s head performed because some abnormalities are not apparent at birth but may be detected in an MRI [39]. To address the concern of detecting microcephaly before a baby is born, officials need to keep providing up to
date information on ways to detect microcephaly and to keep striving to improve detection methods to help the public make informed decisions regarding their fetus.

**Dengue Association**

Dengue association may explain why this Zika outbreak is associated with abnormalities and defects and previous infections were not, which is why it is an important topic to address [40, 47-49]. Dengue is in the same family of viruses as Zika and is also spread by the same two mosquitoes as Zika [40]. If a person has been previously infected with one strain of dengue, and then later gets infected with a different strain, they are at risk of developing severe dengue symptoms due to antibody-dependent enhancement (ADE) [47]. In the topic dengue associations (#10), scientists suspected and are starting to confirm that earlier illness of dengue enhances the chances of Zika infection also due to ADE [48, 49]. The fact that this is in the negative sentiment category shows that the public is concerned with dengue interacting with Zika which informs public health officials that their messages concerning this topic are being heard and causing adequate concern. Now that there is evidence that previous dengue infection enhances the chances of more severe Zika infection, public health officials need to proliferate this message across social media sites and encourage those with past dengue infection to continue to take precautions against mosquito bites.

**Limitations**

The tweets in our analysis were limited to the English-language, which limits the generalizability of the study. This is critical since South American countries were the first and hardest hit countries. Future studies can address this limitation by analyzing tweets in Portuguese and Spanish. The keywords used in data collection were Zika, Zika virus, Zika virus treatment, and Zika treatment. Therefore, tweets that refer to this disease in another language would be overlooked. Tweets that refer to the disease without mentioning it by name would also be overlooked.

**Conclusion**

Overall, the negative sentiment topics focused on neural defects and abnormalities caused by the Zika virus. Since these tweets were categorized as negative sentiment, officials could see that the public was concerned with the symptoms caused by the Zika virus. Since the public was concerned, officials could focus on spreading information encouraging prevention. Officials could also see that the top themes all concerned actual symptoms and defects and did not focus on misconceptions or misinformation that they needed to address. Moving forward, officials can also start informing the public that studies are providing evidence for the Zika dengue interaction hypothesis. They should focus these messages in areas where dengue is endemic since they are the ones most at risk of the interaction causing more severe Zika infection.

This is one of the first studies to address Zika sentiment classification using Twitter. Using such a system allows public health officials to ascertain public sentiment concerning disease outbreaks, and address concerns in real-time.

**Future Work**

Future studies could analyze the change in sentiment over time to see if the number of negative tweets decreases as the outbreak subsides and more advances in treatments are discovered. Studies could also look at sentiment by gender or geographic location. Both are
prudent due to Zika’s effect on fetuses, and its comparative prevalence in equatorial regions, respectively. We would also suggest future studies to leverage other sources of information such as other social media sites, newspapers, and blogs. Similar methodologies could also be applied to future pandemics and epidemics to ascertain public sentiment.

Conflicts of Interest
None Declared

Works Cited

**Abbreviations**

CBO: continuous bag of words
LDA: Latent dirichlet allocation
CDC: Center for Disease Control
WHO: World Health Organization
GBS: Guillain-Barre syndrome
ADE: Antibody-dependent enhancement