Zika Exploration: tracking zika tweets and understandin it's topics, sentiment and time-series

Hayeong Song*
Computer Science
Department
University of Colorado
Boulder

Linzi Xing†
Computer Science
Department
University of Colorado
Boulder

ABSTRACT

In this paper, we designed and implemented a new visualization to explore the patterns of zika related tweets' sentiment, topic distribution, time series trend, and connections between sentiment, topic and time series, topic. We collected a zika dataset with appropriate size and preprocessed it from two aspects – sentiment extraction and topic modeling. Then we selected four different forms of visualizations for four views of the data, which were sentiment, topic, time series and tweet content and combined them together into a panel which could work systematically. With the exploration of the system, we observed results that are reasonable and prove the performance of the tool as well. With this tool, user can understand the twitter data better like what are topics mostly dealt with and how people feel about it.

Index Terms: Text analytics, text visualization, Twitter, natural language processing, sentiment analysis

1 INTRODUCTION

Nowadays, the use of social media such as Twitter is increasing at a substantial rate. As social media is up to date most of the time, it gives us a more efficient and accurate way to track and analyze social events in various fields. Compared with news reported by journalists, information from social media can reflect more personal sentiments from ordinary people instead of objective opinions from journalism, and sometimes, even journalists tend to filter out some information to affect public opinions. Also, for news, it takes time to be reported by journalists. For instance, an event happens at 3 pm on Friday, the news of this event may probably come out the next morning. However, for social media like Twitter, it can post the event almost right after it happened. What's more, the trends are easy to access and informative.

In this paper, we generated a dataset by collecting tweets from Twitter using Twitter Streaming APIs[1]. The primary goal of our work is to design an efficient tool to explore the sentiment pattern on different topics and different timelines. With the patterns we found, it can let researchers be aware of the problematic aspects and concentrate more on certain fields. The specific details of our dataset are discussed in the following sections. With this social media data, we tried to resolve the following questions. If we have already get the tweets about one topic in a particular time interval? Which aspects we can analyze and explore? What's the most appropriate way to explore? Since the tweet is a short text, it can be effective to analyze this from the view of semantic features and word usages. We designed a visualization system from these aspects: tweets' sentiment, topic distribution and time series of tweets. For tweets' sentiment, heatmap visualization was selected, and for topic distribution and time series, row chart and line chart was adopted because of their simplicity and capability to keep information easy to understand. All of these visualizations are user interactive and connect with each other. When there is a change of selection in one view, all the others will change as well. This way, patterns of tweets in different time intervals and relationship between topics and sentiments can be analyzed in a convenient way.

Our primary contributions can be summarized as follows:

- We collected a new Twitter dataset which contains tweets related to keyword 'zika' in a particular time interval. Also, to fit into visualization system, we also pre-processed our data to extract sentiment and topic of each tweet in our dataset.
- We designed and built a new visualization system. It's fully connected, interactive and goal-oriented. With this system, it's much easier to explore the pattern, such as the public's sentiment in a period, and the relation between sentiment and different topics of an event.

The rest of this paper is organized in the following frame: in Section 2, we introduce some previous works related to the task we discuss in this paper. In Section 3, we describe our dataset and design process of our visualization system in details. In Section 4, we talk about the results we got from the visualization system we implemented. In Section 5 and 6, we give the summary of our work and future works.

2 PREVIOUS WORK

Before our work, some works were focusing on the exploration of some natural language processing topics like sentiment analysis and evaluation of topic modeling. For the aspect of topic model, Jason Chuang et al. designed a visualization tool named Termite, which could be used to explore the relation of topics and specific terms likely to be utilized in these topics. [9] The visualization format was a term-topic matrix. X axis represented topics and Y axis represented the terms with the highest frequencies according to the dataset. This tool was dense and bright, but since the corpus they used was not in real time and their primary goal was just finding patterns for topics, their design didn't include interactions with other types of visualizations. Allison J. B. Chaney et al. [1] also developed a similar topic model exploration tool to dig the hidden structures and relationships between topics and documents. They designed the tool as a navigation chain. It could guide users to find the most informative and appropriate materials they were looking for. Different from Jason Chuang, they also visualized the relationship between topics, not just topics and words. However as the corpus they used was Wikipedia, almost all the documents were static, it was hard to explore the trend of topic changes. Eric Alexander et al. designed a new visualization tool named Serendip, which can be used to explore the text corpus on the topics generated by topic model. [6] Similar to our work, they focused more on the

*e-mail: Hayeong.Song@colorado.edu
†e-mail:Linzi.Xing@colorado.edu

[1] https://dev.twitter.com/streaming/overview

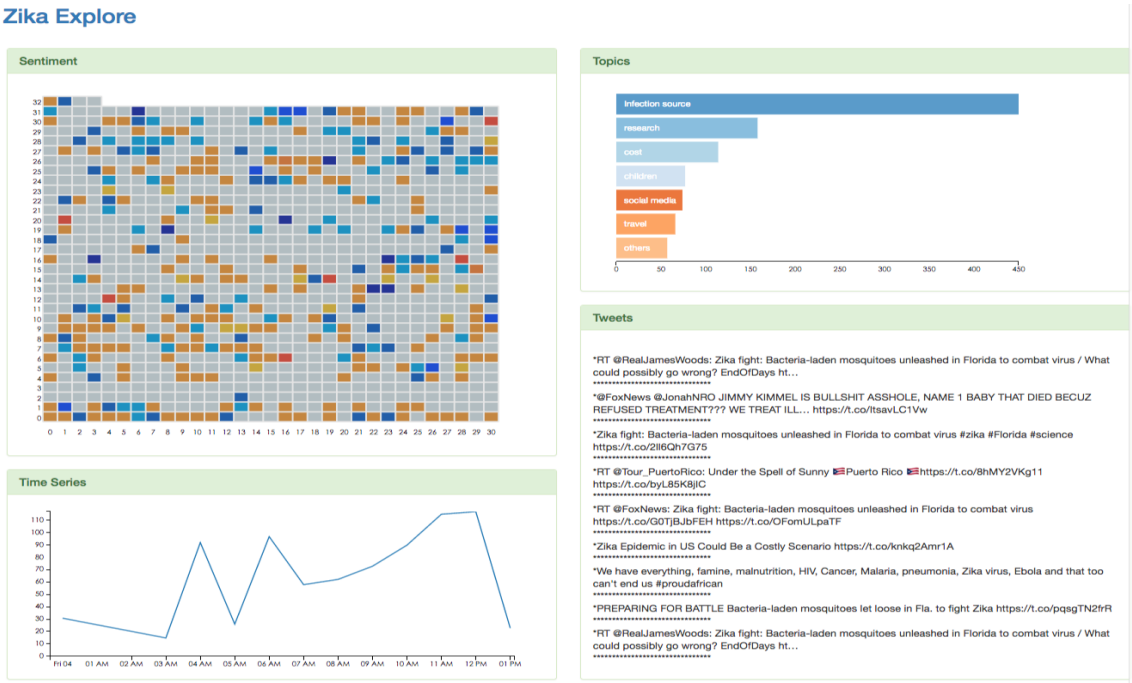

Figure 1: This panel has four components. For Sentiment Heatmap, each cell represents a tweet in dataset and color represents the sentiment of each tweet. The color scale is from red(negative) to dark blue(positive) corresponds to sentiment score range from -1 to 1. The middle value of sentiment score is 0 and is represented by grey(neutral). For Topic Row Chart, each row represents a topic, length corresponds to the number of tweets related to this topic. For Time Series, X and Y axis are detailed time and number of tweets caught at time. When we change the selection of any of these views, all the others will change to filter out other data and highlight the selected ones.

relationship between topic and metadata of documents like genre, publisher and token count. In this way, the trend of documents can be caught and analyzed easier. When users use it, they could even predict which kind of text would be more related to topics based on the trend revealed in the previous period. Carson Sievert et al. also designed a topic model exploration tool named LDAvis. [5] It can show the document-topic relation and topic-term relation.

Also there were works that visualized social media content specifically for tweets. According to Mengdie Hu et al, SentenTree was introduced, which displays sentence pattern that appears occasionally that is extracted from social media content. [7] It allows people to understand the fundamental concept and opinion of the tweets by showing a pattern of sentences using SentenTree. Another work done by Mengdie Hu et al is OpinionBlocks that helps the user understand the opinion text. [8] This system supports the visual summary of the opinions and allows users or crowd to rectify text analytics so that it can overcome limitations of NLP technologies and improve overall quality. [8]

Analyzing tweets for diseases detection has been done in previous work. Alex Lamb et al. analyzed Twitter data to track infection and discriminated tweet that indicated fear and expressing concern about flu. [10] Luciano Barbosa et al proposed an approach making use of noisy label as training data to detect sentiment on Twitter data. [2]

3 Design Process

This section is separated into three subsections: Twitter Dataset, Data Pre-processing, System Visualization.

3.1 Twitter Dataset

The data set size is about 4.362 megabytes which consists of 996 tweets which is about 1000. All tweets in this dataset were collected from "Wed 03 23:39:55 2017" to "May 4 13:08:10 2017", Which is about 13 hours of collection. We used Twitter Streaming APIs to collect data and the collecting procedure are using filter keyword 'zika' and English. It means tweets were returned when it contains keyword 'zika' and if the tweet was written it English. The collected tweet is returned in json objects and includes the subfields of 'created_at', 'text' and 'lang' and more. But these three are the subfields that were mainly used in our system.

3.2 Data Pre-processing

- **tokenize**
  Before feeding the data on the algorithm twitter data is pre-processed. With this application @, emoticons, links or URLs and hashtags() are tokenized into a single token.

- **eliminate stopwords**
  English stopwords such as "a", "the" that appear often but does not contain meaning were eliminated before inputting the Twitter text to algorithm.

- **sentiment analysis**

  For sentiment analysis, we use Twitter text as an input. 'TextBlob'[2], a implemented Python library that is used for natural language processing task, mostly used for processing textual data. It supports sentiment analysis on text data, and we used this library to compute sentiment score. With this

[2]https://pypi.python.org/pypi/textblob

application, it will return the sentence with polarity or sentiment score. The score ranges from -1.0 to 1.0. When the score is between [-1.0,0) it is classified as negative. When 0.0 neutral and (0,1.0] as positive. Also we vectorized corpus of tweets to vector using 'word2vec'[3] model from 'Gensim'[4], and transform sum of the vector of words to 2 dimensional space using PCA from 'Sklearn'[5] Python package. This way we can represent tweet as a 2-dimensional coordinate. And use as two coordinates for visualization.

- **topic modeling:** The topic model we adopted here is latent Dirichlet allocation (LDA) [4]. This is one of the simplest unsupervised topic models. This model assumes for a text corpus, there are a certain number of topics. These topics have various distributions under each document and words have wide-range of distributions under each topic. The training process is a document reconstructing process. Each re-selected word is chosen by the word distribution of a certain topic which selected by the topic distribution of the certain document. All the distributions are initialized by Multinomial Distribution based on Dirichlet Distribution with two variables: α and β. After the Gibbs Sampling procedure [3], variables and distributions we used on the initial stage will be revised till it gets convergence. In our procedure, according to our data size, we set the topic number as 7, and for each topic we picked 20 most frequent words. Since the LDA model can't output the most appropriate words to represent each topic, we summarized these topics manually and got these 7 categories: Infection Source, Research, Cost, Children, Social Media, Travel and Others. Then we fitted topic information into the csv file under the attribute 'topic' for future use. The LDA we implemented was based on the python package 'Gensim'[6].

3.3 Visualization System

The whole visualization system we implemented has four components, which are Sentiment Heatmap, Topic Row Chart, Time Series Line Chart and Tweets Table. All these visualization views were implemented based on the dc.js, which is a dimensional charting javascript library[7] and the browser page frame was based on bootstrap[8]

- **Sentiment Heatmap**: The goal of sentiment analysis in this paper is mainly about exploring how sentiment changes according to time and what is the relationship between topics and sentiment. For example, for some negative topics, tweets related to these are also supposed to be negative. Heatmap has some features which are suitable to our goal. First, it can express data in a systematic way. The organized format of heatmap can help better to see the potential patterns of sentiment. Second, since heatmap is a visualization that is compact, it can save space in the certain level, which makes the whole system look neater. Users can observe the pattern more intuitively and instantly. When we filter by topic and time, certain tweet cells will stay and others will turn into grey. This way sentiment pattern of this topic will be distinct and easy to understand.

- **Topic Row Chart**: The views we should show are just topic name and distribution of tweets that belong to each topic. To keep the system straightforward and easy to understand, we chose row chart to visualize topics. It can also be treated as a filter for sentiment and time series. When we select a topic, cells related to this topic in sentiment heatmap will be highlighted. Also, trends in time series will also change. One of the reasons why we chose row chart is because it is simple and it can be easy for the user to understand the visualization. As people are used to row chart it can be effective. This way, the higher the distribution, the longer the rowchart and user can perceive it quickly.

- **Time Series Line Chart**: The view we want to show about time series is the trend of the number of tweets by time. So there is no need to use relatively complex visualization. Based on the requirement of simplicity and clearness, we thought line chart was the best choice and advantage of line chart is that it is widely adopted for visualization to indicate the particular period. As it is simple of there is a big bump one can see that there was a significant increase or decrease in a number of tweets. Or if the line chart reaches its peak user can perceive that on certain time the tweet was generated the most. Overall, users can understand the pattern with ease. As it is simple and intuitive user can understand the trend quickly Like Topic Row Chart we talked above, this chart can also be used as a filter to explore the relationship between sentiment change and time. For example, when we select a time interval, the correspond tweets will be highlighted in heatmap matrix and it's clear to see in this period, positive or negative posts take the most proportion. Same for topic, when we use together with these two filter, it will be more convenient to extract topic-time pattern. Like for topic 'Infection Source', when is the most popular time.

- **Tweets Table**: This component was designed for users to check the detailed content of each tweet. When we hit one cell in sentiment heatmap, the content of the tweet will show. If more than one tweet is selected, maximum number 9 of tweets will display based on the post time because of the space limitation. Implementing this component with a scrollbar can overcome space limitation.

4 Results

Before the design process, we thought about some questions and treated them as guidelines of design to answer them. Next, we cover the results mainly focusing on answering these questions.

(1) What is the relationship between sentiments and topics and how well this tool can detect? We explore the most related sentiment for each topic and found some interesting patterns which make sense. For example, when we filtered the heatmap by topic 'Infection source', almost all the tweets under this topic were negative, and when we filtered the heatmap by topic 'research', most of tweets related to this were positive. These patterns are reasonable because, in real life, people will talk infection source of zika in negative aspect as it is mostly dealt with the outbreak of diseases which is negative. When it comes to the researches about zika, it is about how to combat zika virus such as vaccination which is a positive aspect of the topic. What's more, when we filtered by topic 'News', the proportion of positive and negative tweets were in the balance. It also makes sense because news usually will cover both good and bad news. This way, it shows that our tool works and supports users to detect sentiment pattern based on topics.

(2) What is the topic distribution on our Twitter dataset? According to the row chart, it's clear that 'Infection source' is the most popular topic people will talk about and it's more than two times to the second popular topic 'research'. Topic 'cost' is the third popular and the rest topics like 'children', 'social media' and

[3]https://radimrehurek.com/gensim/models/word2vec.html
[4]https://radimrehurek.com/gensim/models/ldamodel.html
[5]http://scikit-learn.org/stable/
[6]https://radimrehurek.com/gensim/models/ldamodel.html
[7]https://dc-js.github.io/dc.js/
[8]http://getbootstrap.com/

'others' have similar probabilities to be mentioned. And order of topics change when filtered by certain time-interval

(3) What can we get from post volume of topics in given period? We used topic row chart as a filter and explore the trend of post volume for each topic. We found that two topics show the similar pattern, for example, the trends for topic 'research' and 'social media' show similar pattern in time-series visualization. This is due to the fact researched in the advancement in combating of zika virus are likely to be reported by news and social media. We also found for topic 'children', pattern was different from other topics. For instance, the spike will appear at 4 am, which is before dawn for North American countries mostly a time that is assumed that people are asleep. So based on that, it is likely that these tweets were generated from other nations or it may indicate that this topic is popular in other nations compared to North America.

(4) What can time series visualization tell? It shows a trend of on which time tweet was mostly generated or less. For example number of tweets reached its peak around 12 pm. And it mostly dealt with infection of the diseases such as outbreak of zika. Or actions taken to combat the released of diseases.

(5) Which tweet indicate which sentiment? To see if which tweet is determined as tweet sentiment, as one clicks on a certain cell the tweet linked to it will show. And the tweet identified as positive covers the text that tries to combat the release of zika which indicated that our system works. And the opposite works as well when click on a negative sentiment cell the twitter linked to will cover negative context such as death occurrence due to zika.

(6) Which topic was generated on which time? When exploring the system one can know which topic was generated on certain time, using the filter of the time-series visualization. For example, on time between 4 am to 6 am mostly generated topics were zika related to infection or children. But overall, through all 13 hours of collection infection source and research were the top two topics.

(7) Can system answer when certain tweet with certain sentiment was generated? By exploring the system, one can see that tweets created later on or late night were more positive than the previous ones. As in heatmap visualization, the upper part of tweets has blue cells(positive) more than the lower part. Also when we filter time series visualization to show afternoon cells shown on heatmap will have a higher number of blue cells. On tweet table number of tweets with positive text will outweigh negative tweets.

5 Summary

To summarize, we collected a new Twitter dataset which contains tweets related to keyword 'zika' in a specific time interval(13 hours and about 1000 tweets). Also, to fit into visualization system, we also pre-processed our data to extract sentiment and topic of each tweet in our dataset. In addition, We designed and built a new visualization system that user can explore about Twitter data. With this system, it's easier for users to understand the sentiment of certain tweets or topic and can also know the relationship between sentiment visualizations and different topics of an event or news.

6 Future work

To improve this system, we can add more functions to it. First, for the data pre-processing part, the LDA model we used may not efficient enough for short text like tweets. To better extract and categorize topics depend on the corpus we have, we can use other topic models specially designed for Twitter data like Twitter-LDA [11] to improve our system's accuracy. Second, like the previous work we mentioned in Section 2, we can explore more about topic-term level. Hierarchical Bar Chart can be a good choice. With that, we can select the specific terms under topics and see the sentiment distributions for different terms instead of just topics. Third, we can collect another data set using different keyword such as 'health care' or 'birth'. And build a dashboard both for data set collected from 'zika' and another keyword. By adding new data set and showing it in a dashboard according to different key word, the user can understand the relationship between two words. This way the twitter can better support user to explore and understand what the twitter data contains. Lastly, in order to make the system more interactive to each other adding function to highlight the positive words of negative words in the twitter table will be helpful. To apply it in a system, if the user clicked one cell of heat map visualization the twitter that matches will show on twitter table and when sentiment words or context is highlighted the system will be more interactive and be more informative. As user can see which words indicated that the tweets were positive or negative.

References

[1] B. C. Allison, J. and B. David, M. Visualizing topic models. Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media, Dublin, Ireland, 2012.

[2] 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.

[3] Bishop and M. Christopher. Pattern recognition and machine learning. *Springer*, 2006.

[4] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, pp. 993–1002, Mar. 2003.

[5] S. Carson and S. Kenneth, E. Ldavis: A method for visualizing and interpreting topics. Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, Baltimore, Maryland, USA, 2014.

[6] A. Eric, K. Joe, V. Robin, W. Michael, and G. Michael. Serendip: Topic model-driven visual exploration of text corpora. Visual Analytics Science and Technology, Paris, France, 2014.

[7] M. Hu, K. Wongsuphasawat, and J. Stasko. Visualizing social media content with sententree. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):621–630, 2017.

[8] M. Hu, H. Yang, M. X. Zhou, L. Gou, Y. Li, and E. M. Haber. Opinionblocks: A crowd-powered, self-improving interactive visual analytic system for understanding opinion text. In *INTERACT (2)*, pp. 116–134, 2013.

[9] C. Jason, M. Christopher, and H. Jeffrey. Termite: Visualization techniques for assessing textual topic models. pp. 74–77. AVI 12, Capri Island, Italy, 2012.

[10] A. Lamb, M. J. Paul, and M. Dredze. Separating fact from fear: Tracking flu infections on twitter. In *HLT-NAACL*, pp. 789–795, 2013.

[11] X. Zhao, W., J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li. Comparing twitter and traditional media using topic models. Proceedings of the 33rd European conference on Advances in information retrieval, ECIR11, Berlin, Heidelberg, 2011.