PanTop: Pandemic Topic Detection and Monitoring System (Student Abstract)

Yangxiao Bai and Kaiqun Fu

Department of Electrical Engineering and Computer Science, South Dakota State University
{bai.yangxiao, kaiqun.fu}@sdstate.edu

Abstract

Diverse efforts to combat the *COVID-19* pandemic have continued throughout the past two years. Governments have announced plans for unprecedentedly rapid vaccine development, quarantine measures, and economic revitalization. They contribute to a more effective pandemic response by determining the precise opinions of individuals regarding these mitigation measures. In this paper, we propose a deep learning-based topic monitoring and storyline extraction system for *COVID-19* that is capable of analyzing public sentiment and pandemic trends. The proposed method is able to retrieve Twitter data related to *COVID-19* and conduct spatiotemporal analysis. Furthermore, a deep learning component of the system provides monitoring and modeling capabilities for topics based on advanced natural language processing models. A variety of visualization methods are applied to the project to show the distribution of each topic. We believe that our proposed system accurately reflects how public reactions change over time along with pandemic topics.

Introduction

The rampaging *COVID-19* pandemic has greatly affected emotion and everyone's daily life in the past two years. With the abundance of the generated social media data during the pandemic, more users intend to express their emotions and opinions about social events, such as the mitigation policies or the invention of vaccines. Such dramatic social media data increase provides great research opportunities in social media mining and natural language processing. To further understand the various public behavior, some researchers built a real-time tweets analyzer to get high-frequency words and polarity over time in the United States (Kabir, Madria et al. 2020). Previous work by Qazi (Qazi, Imran, and Ofli 2020) employs a gazetteer-based approach to infer the geolocation of tweets. Accordingly, we deploy an *ElasticSearch* server to manage all the geo-tagged tweets by time and spatial coordinates, which makes it possible to select the appropriate study scope as needed. To mine the information and topics in these tweets, we looked at some of the existing data mining models (Ordun, Purushotham, and Raff 2020). However,due to the suboptimal performance of*LDA* model in the face of large volumes of text. We apply *LDA* to capture hashtag information while using *BERT* (Grootendorst 2022) to generate embedding for the body of tweets. Ultimately, we generate a storyline to show the difference and connections of each topic (Sun et al. 2019).

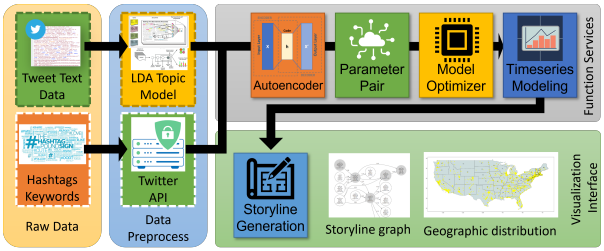

Figure 1: *PanTop* System Diagram

Proposed Method

From a holistic view of the research method Figure 1, our proposed method consists of several relational parts of dataset processing, model training for topic clustering, and visualization. We first import data into the *ElasticSearch* server in batches with a time-space filter. For the processing of the body part, we use the pre-trained *BERT* to generate the corresponding sentence vector. To synthesize the information of a complete tweet and reduce the dimension as much as possible, we use the autoencoder to combine the two weighted vectors.

The number of potential topics differs for each selected domain, and the weight scale of vectors generated from *LDA* and *BERT* also affects performance. Hence, we propose an exploratory analysis to extract potential topics. We obtain the ideal parameter pair by repeated experiments within a specific range. Then we evaluate the performance of a single experiment in two ways. One is calculating the silhouette coefficient to show performance under a different number of clusters. The other is projecting the distribution of clusters into 2D space for subjective judgment. Besides, due to the timeliness of the topic of the epidemic, we use Time Series KMeans to do the clustering and add release time as the z-axis based on 2D projection, based on which we find the optimal weight scale for vectors.

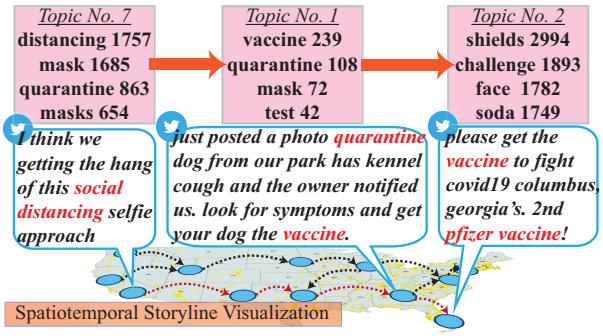

Figure 2: Storyline for Topic clusters

Visualization

A topic often contains thousands of tweets. We hypothesize that the tweet closest to the cluster's center represents the topic the most. Therefore, we calculate the cosine similarity between every tweet and cluster centroid point and pick the nearest tweet as the most representative tweet. Wordcloud and a geographic distribution map are other ways to reflect the topic's content. Through these methods, we can study this topic's content in multiple dimensions.

To generate a graph showing all topic relationships, we calculate the similarity between all the topic pairs. Then sort the calculation results from high to low. Based on it, we set a threshold of edge. An edge will be generated between all topic pairs whose similarity is greater than the threshold. Finally, we import the point and edge information into *Graphviz* and get a directed graph, which can reveal the development trend of each topic to a certain extent.

Results and Discussion

Dataset: The dataset we used is *CORONAVIRUS (COVID-19) GEO-TAGGED TWEETS DATASET* from *IEEE DataPort* (Lamsal 2021). This dataset includes English tweets related to the epidemic from 2020 to 2022 around the world. Due to the tweets spreading policy, Twitter content cannot be accessed directly. So we use tweets ID to request Twitter API in batch by a python package *Twarc* and get completed information, including content, UTC time, and geographic location. In total, 469,414 tweets have been imported to our *Elastic* server. To provide a basis for filtering data, we refer to a *COVID-19* Government Response Database (Cheng et al. 2020). After collecting all the keywords within the policy announcements, we make a comprehensive filtering function to create suitable study samples.

Result: We pulish our project on github: [1] By using the *Pantop* process flow, we first test the most suitable number of nodes and vector weights, then use them to generate visual results. We can get several storylines for the directed graph we generate by traversing the graph nodes. Figure 2 is one example. To assess the storyline's quality, we compare each node's topical content with government policies published in the immediate vicinity of that time in reality. Take topic number 7, for example, released on August 8th. The hot keywords of this topic are very relevant to the epidemic prevention policy. By referring to the *CoronaNet Dataset*, we can find that the governments of some states, such as Kentucky, Nevada, Oklahoma, Alabama, Vermont, Delaware, Ohio, New Jersey, New Hampshire, and Mississippi, have published policies related to the closure and regulation of schools. And most of these states are concentrated in the northwest. For another, the Maryland, Alabama, New Mexico, Mississippi, Delaware, North Carolina, Louisiana, and Nebraska governments published a series of anti-epidemic isolation measures. The regional relevance of policy and the topics can be detected from geographical distribution of clusters. Although topics and policies are not always closely related and the tweeting habits of the people in each state are also different, the study of hot topics can reflect the trend of the epidemic and people's reactions, which is of great significance to public opinion polls.

[1] https://github.com/Baiyangx/PanTop

Conclusion

Our proposed approach successfully extracted the focus topics related to the epidemic. It provided a very intuitive way to show each topic's potential connections and its geographic time information. However, the accuracy of the results is often curbed due to the large amount of noise present in Twitter text, as well as the limitation of text length. In future work, we optimize the model in two ways. 1. Apply text summarization methods to conclude the topics and enhance the interpretability of the results. 2. Use the attention mechanism to reduce the impact of noise on text analytics.

References

Cheng, C.; Barceló, J.; Hartnett, A. S.; Kubinec, R.; and Messerschmidt, L. 2020. COVID-19 government response event dataset (CoronaNet v. 1.0). *Nature human behaviour*, 4(7): 756–768.

Grootendorst, M. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Kabir, M.; Madria, S.; et al. 2020. Coronavis: A real-time covid-19 tweets analyzer. *arXiv preprint arXiv:2004.13932*.

Lamsal, R. 2021. Design and analysis of a large-scale COVID-19 tweets dataset. *Applied Intelligence*, 51(5): 2790–2804.

Ordun, C.; Purushotham, S.; and Raff, E. 2020. Exploratory analysis of covid-19 tweets using topic modeling, umap, and digraphs. *arXiv preprint arXiv:2005.03082*.

Qazi, U.; Imran, M.; and Ofli, F. 2020. GeoCoV19: a dataset of hundreds of millions of multilingual COVID-19 tweets with location information. *SIGSPATIAL Special*, 12(1): 6–15.

Sun, W.; Wang, Y.; Gao, Y.; Li, Z.; Sang, J.; and Yu, J. 2019. Comprehensive event storyline generation from microblogs. In *Proceedings of the ACM Multimedia Asia*, 1–7.