PanTop: Pandemic Topic Detection and Monitoring system

Yangxiao Bai, Abir Mohammad Hadi, Youngsun Jang, Kwanghee Won, Kaiqun Fu
Department of Electrical Engineering and Computer Science
South Dakota State University
Brookings SD, USA
{bai.yangxiao,abirmohammad.hadi,youngsun.jang,kwanghee.won,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. In addition to spatiotemporal data mining, the system implements a sentiment analysis on Twitter users. Furthermore, a deep learning component of the system provides monitoring and modeling capabilities for topics based on the most advanced natural language processing models. The proposed system includes a user-interactive visualization component that provides audience monitoring and analytical toolboxes. Case studies found in abundance in our proposed system can justify empirical analysis. Ultimately, we believe that our proposed system accurately reflects how public sentiments change over time and locations, along with specific pandemic topics.**

***Index Terms*—social media analysis, topic modeling, storyline generation**

I. Introduction

The rampaging COVID-19 pandemic has greatly affected emotion and everyone's daily life in the past two years. The general public's emotions went through a roller coaster-like up-and-downs: shocked by the deadly known disease, overwhelmed by misinformation, upset by the mitigation policies, and regaining confidence with the development of vaccines. Ubiquitous user-input content on social media and online services have generated a tremendous amount of information that can reflect the users' emotion and opinion towards social events. With the abundance of the generated social media data during the pandemic, more users intend to express their emotions and opinions about COVID-19-related social events, such as the mitigation policies or the invention of the vaccines. Such dramatic social media data increase provides great research opportunities in social media mining and natural language processing.

Spatiotemporal sentiment analysis is one of the target research areas in social media analysis for COVID-19. In the case of pandemics, spatiotemporal sentiment analysis provides a reference for the development of epidemic mitigation measures and evaluation techniques. In the context of sentiment analysis, some previous studies accomplished success by using the machine learning technique. For instance, previous work [1] used a deep learning-based sentiment analyzer to collect and curate the COVID-19 dataset. Attention mechanisms were applied to characterize the word-level interactions within a local and global context to capture the semantic meaning of words. Among other logical approaches, Chakraborty et al. [2] propose the implementation of fuzzy logic for taming the fuzziness of sentiments. However, the existing works in social media sentiment analysis lack the consideration of spatial and temporal factors. Our proposed system provides both spatial and temporal aspects of sentiment analysis under the topic of COVID-19.

Topic detection and modeling are the second focus of our proposed system. Topic detection is an important part of addressing the said problem to separate points of interest among many candidates. Several noble studies have been conducted in the last few decades to mine the most appropriate topic associated with a text phrase. A topic graph-based approach was proposed by Batool et al. [3] where a topic graph from vectorized Twitter data was proposed with term frequency as a heuristic and social relationship between the virtual users. Other work [4] proposed a time-dependent burst detection technique that focuses on two and three-word data acceleration as a spread tendency to make early detection based on the keywords. Abd-Alrazaq et al. [5] used Latent Dirichlet Allocation (LDA) for topic modeling on English twitter data. Gencoglu et al. [6] used a large Twitter dataset to analyze semantic topic clusters using Language Agnostic BERT Sentence Embeddings (LaBSE). We combine the advantages of LDA and neural network models to develop a transformative model for topic discovery and monitoring components in the system.

The storyline generation problem was first studied by Kumar et al. [7] as a generic redescription mining technique, by which a series of redescription between the given disjoint and dissimilar object sets and corresponding subsets are discovered. Storytelling is an efficient way to solve the issue of information overload. By extracting critical and connected entities, the original document is structurally summarized. Current works contain two categories: Textual Storytelling [7]–[12] and Visual Storytelling [13]–[15]. A storyline generation component is deployed in the proposed system. The storyline generation is capable of summarizing the highlighted COVID-19-related events with social media data, and a visualization component is also developed to demonstrate the comparisons between the official releases of the mitigation events and the identified topic events from social media.

In this paper, we present the Pandemic Topic Detection and Monitoring system (PanTop), which is capable of 1) collecting and retrieving social media data on COVID-19-related topics, 2) detecting hidden trends of topics from social media posts, and 3) generating and visualizing storylines for the extracted COVID-19-related topics. This paper is structured as follows: Section II discusses the architecture of the PanTop system; Section III illustrates the experiment and data of the system; Section IV demonstrates the case studies revealed by the PanTop system; in section V, we conclude our discoveries from the proposed PanTop system.

II. Method and Data

From a holistic view of research method (Figure 1), it consists several relational parts of the dataset import,dataset Processing, and machine model training for topic clustering

In dataset import process, since the dataset is un-hydrated, we needed to “hydrate” the data by building a process that will fetch the tweet content by querying with the Tweet ID. After hydrating, the data are saved in a local storage, and by using Tweets API the data are completed to be analyzed and stored in the Elasticsearch server. In dataset processing step, after importing data into Elasticsearch server in batches, we choose a research scope and we get the target dataset from Elasticsearch. By doing spatial query and filtering task, the target dataset can be visualized into a map. Furthermore, after processing the data by using techniques such as the regex filtering, tokenization and tagging, and removing stop words, it is possible to apply a topic clustering model to get topics.

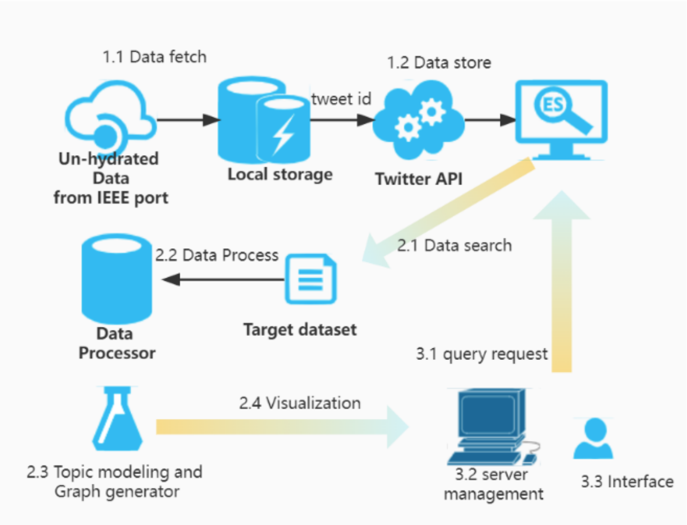


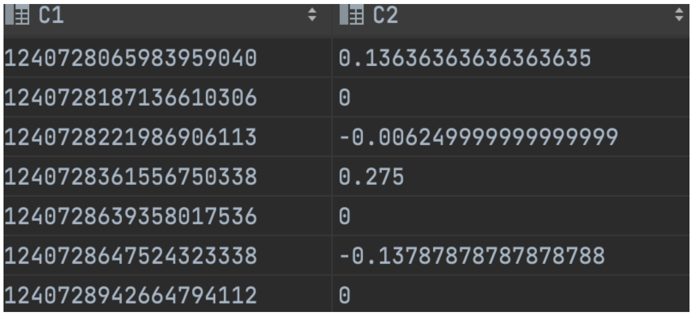
Fig. 1. System structure

A. Tweet Data Collection and Processing

1) Tweet Data Collection process: The dataset we used is CORONAVIRUS (COVID-19) GEO-TAGGED TWEETS DATASET form IEEE DataPort. Due to the tweets spreading policy. We can’t access the completed tweets content directly. Thus. We use a transfer program named Twarc to batch fetch tweets in the dataset. Twarc is a python package that used to export tweets automatically. The main principle is that Twarc will make use of the registration information from twitter developer platform and get the permission from Twitter. Then Twitter can monitor the whole process of getting tweets. This way is completely legal and doesn’t violate the privacy protocols set by Twitter. But it also cause some issues. If the tweets account is banned or the permission has been changed, we can’t access the content anymore. Normally Twitter doesn’t send any warning message but change the content of tweets to a prompt message. So some filtering process is necessary to ensure that the data set is not contaminated with irrelevant information.

We collect data from March 2020 to March 2022. These data comes form hundreds of single CSV files, Since they are stored by date. So a batch process is necessary to import from these files. After that we store all of these data to Elasticsearch, in order to facilitate subsequent steps to call. After delete all the irrelevant information or the missing information data. The whole memory size up to 142Mi. The number of useful pieces of information reaches 388,719. However, the geographical distribution of this tweets is very uneven. So we try to select states with large sample sizes as research subjects.

2) Tweet Data structure: The original data contains tweets ID and sentiment score calculated by database publisher Figure 2. We use tweets ID to request twitter API in batch and get completed information, including content, UTC time, geographic location. All of these information are fully imported into the server after splicing with sentiment score. In this process, we will transfer the longitude and latitude to the geometry, then we can apply some spatial operator to handle them. Time format is another point in this way, we transfer all the time appearing in our project to UTC format. Then all filtering of time will be consistent.



C1	C2
1240728065983959040	0.13636363636363635
1240728187136610306	0
1240728221986906113	-0.006249999999999999
1240728361556750338	0.275
1240728639358017536	0
1240728647524323338	-0.13787878787878788
1240728942664794112	0

Fig. 2. Tweet dataset structure

3) Tweet Data Pre-processing Method: The pre-processing process includes regex filtering, tokenization and tagging, removal of stop words (Figure 3). Regex filtering process used to delete some proper noun or URL. We can also delete all the hashtag or the emojis. Normal phrase or locations can also be a optional. Basically, we will set different filtering rules according to the task goals and model requirements.

Tokenization and tag are important to our project. We use nltk package to process the sentence in the tweets and decompose the content into single words. Then we can filter for parts of speech and get more reasonable results. The purpose of removal of stopwords is to avoid distractions from common words on the topic. The nltk package provide a library of stopwords and we can set our own list of common words based on the theme. It's an important step to get good performance of machine learning models and we can adjust it to fit the current strategy we used. It's a part of fine turning.

```
[('i', 'NN'), ("'m", 'VBP'), ('at', 'IN'), ('superior', 'JJ'), ('court', 'NN'
), ('corona', 'NN'), ('branch', 'NN'), ('in', 'IN'), ('corona', 'NN'), (',',
','), ('ca', 'MD')]
--------
[('joyeux', 'NN'), ('anniversary', 'JJ'), ('-', ':'), ('happy', 'JJ'), ('birt
hday', 'NN'), (',', ','), ('@', 'NNP'), ('titaylea', 'NN'), ('have', 'VBP'),
('fun', 'VBN'), ('and', 'CC'), ('enjoy', 'VB'), ('your', 'PRP$'), ('day', 'NN
'), ('!', '.'), ('we', 'PRP'), ('love', 'VBP'), ('you', 'PRP'), ('.', '.'), (
'@', 'VB'), ('sebsalcedo', 'JJ'), ('#', '#'), ('birthday', 'JJ'), ('#', '#'),
('bday', 'JJ'), ('#', '#'), ('joyeuxanniversaire', 'NN'), ('#', '#'), ('happy
birthday', 'JJ'), ('#', '#'), ('friends', 'NNS'), ('#', '#'), ('bffs', 'NNS')
, ('#', '#'), ('friendship', 'NN'), ('@', 'NNP'), ('corona', 'NN'), (',', ','
), ('california', 'NN')]
--------
```

Fig. 3. Tokenization and tag

B. Topic Clustering

For topic clustering, first we collected the data from Elasticsearch server by filtering using spatial queries. The data had text, geo-tag and tweet ID as data columns. Since the text contained lots of unwanted strings or characters, we needed to clean the data.

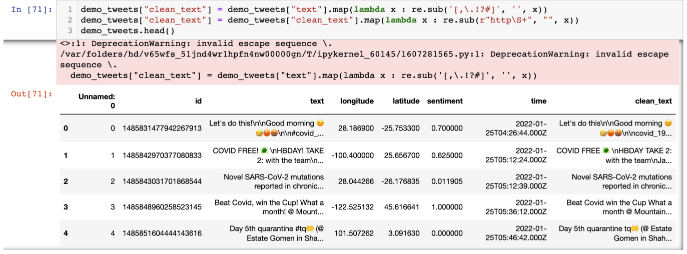
Some tweet contains hypertext URL such as:

```
https://t.co/{tweet_id}
```

and some other URLs as shown in (Figure 4):

We needed to filter these unnecessary texts from the actual twitter data using some regular expression operator and cleaning the part that corresponds to those specific terms. We also needed to remove all the punctuation marks so that it drops the redundant and recurring feature that don't contribute to the clustering.

Example of these regular expression technique is shown in the (Figure **??**):



```
In [71]: demo_tweets["clean_text"] = demo_tweets["text"].map(lambda x : re.sub('[,\.!?#]', '', x))
         demo_tweets["clean_text"] = demo_tweets["clean_text"].map(lambda x : re.sub(r"http\S+", "", x))
         demo_tweets.head()
```

```
<>:1: DeprecationWarning: invalid escape sequence \.
/var/folders/hd/v65wfs_51jnd4wr1hpfn4nw00000gn/T/ipykernel_60145/1607281565.py:1: DeprecationWarning: invalid escape
sequence \.
  demo_tweets["clean_text"] = demo_tweets["text"].map(lambda x : re.sub('[,\.!?#]', '', x))
```

Out[71]:

	Unnamed: 0	id	text	longitude	latitude	sentiment	time	clean_text
0	0	1485831477942267913	Let's do this!\n\nGood morning 😊 🤣😡🤬\n\n#covid_...	28.186900	-25.753300	0.700000	2022-01-25T04:26:44.000Z	Let's do this\n\nGood morning 😊 🤣😡🤬\n\ncovid_19...
1	1	1485842970377080833	COVID FREE! 🦠 \nHBDAY! TAKE 2: with the team\n...	-100.400000	25.656700	0.625000	2022-01-25T05:12:24.000Z	COVID FREE 🦠 \nHBDAY TAKE 2: with the team\nJa...
2	2	1485843031701868544	Novel SARS-CoV-2 mutations reported in chronic...	28.044266	-26.176835	0.011905	2022-01-25T05:12:39.000Z	Novel SARS-CoV-2 mutations reported in chronic...
3	3	1485848960258523145	Beat Covid, win the Cup! What a month! @ Mount...	-122.525132	45.616641	1.000000	2022-01-25T05:36:12.000Z	Beat Covid win the Cup What a month @ Mountain...
4	4	1485851604444143616	Day 5th quarantine #tq👐 (@ Estate Gomen in Sha...	101.507262	3.091630	0.000000	2022-01-25T05:46:42.000Z	Day 5th quarantine tq👐 (@ Estate Gomen in Shah...

Fig. 4. Right-most column showing cleaned text with Regex

In the (Figure **??**), the circle size of the term clusters visualize the frequency of the terms appearing in the documents.

C. Spatial Data Mining and Visualization

1) Data Mining Method: First, we import all data into Elasticsearch server. Then we get to use different type of query instructions to request data on demands. In order to narrow down the scope of our research range, we build two methods and well-package them, which are query by time and query by location. Regionally-targeted requests are often broken down by state. So we use two key steps to accomplish the filtering to the data. First, get the minimum circumscribed rectangle of the state to search. Next, filter out all points outside the boundary by spatial operations.

2) Visualization Method: We use matplotlib to generate the map automatically. At the same time, we distinguish different types of points and areas by marking the map with different colors. Depending on the focus, we'll display a national map or a state map.

III. Experiments

We apply our research method experimentally on a large spatial-temporal dataset collected from IEEE dataport. By comparing the result with real events in Coronanet which records government responses to coronavirus, we evaluate the plausibility of the method. The result shows several points of interest in a specific time period and the details within each topic.

A. Dataset Description and Experiment Setup

Dataset:

The dataset we used is CORONAVIRUS (COVID-19) GEO-TAGGED TWEETS DATASET form IEEE DataPort. Due to the tweets spreading policy. We can't access the completed tweets directly. Thus. We use a transfer program named Twarc to batch fetch every tweet in the dataset. Twarc is a python package that is used to export tweets automatically. The main principle is that Twarc will make use of the registration information from twitter developer platform and get the permission from Twitter. Then Twitter can monitor the whole process of getting tweets. This way is completely legal and doesn't violate the privacy protocols set by Twitter. But it also causes some issues. If the tweets account is banned or the permission has been changed, we can't access the content anymore. Normally Twitter doesn't send any warning message but changes the content of tweets to a prompt message. So some filtering process is necessary to ensure that the data set is not contaminated with irrelevant information.

Metrics:

We use silhouette factor to evaluate the performance of clustering. For storyline generation, we based on rationality of the time series and topic keywords.

Topic mining:

To evaluate the quality of generated topics, we focus on both the content of the topic itself and the distribution of the topic.

conditions	gamma = 7	gamma = 8	gamma = 9
k = 16	0.028	0.025	-0.005
k = 18	-0.007	0.017	0.040
k = 20	-0.006	0.008	0.032

In the experiment, each topic consists of a series of related sentence vectors. We use high-frequency words in WordCloud to show the content of the topic. Besides, We also output one of the most representative sentences in each topic as supporting evidence.

The distribution of clusters in the vector space is another indicator of the quality of the results. The projection of vectors in two-dimensional space can intuitively see the distribution of the cluster. In order to verify the temporal relevance of the topic we also use the time label of the topic vector as the third axis to generate a three-dimensional vector distribution map.

Comparison Models:

We propose a combination of lda and bert methods to produce high-quality sentence vectors. To verify the advantages of this approach, we output the clustering results using bert and LDA separately. We make a comprehensive assessment based on the quality of the topics and timelines generated by each method

B. Results and Discussion

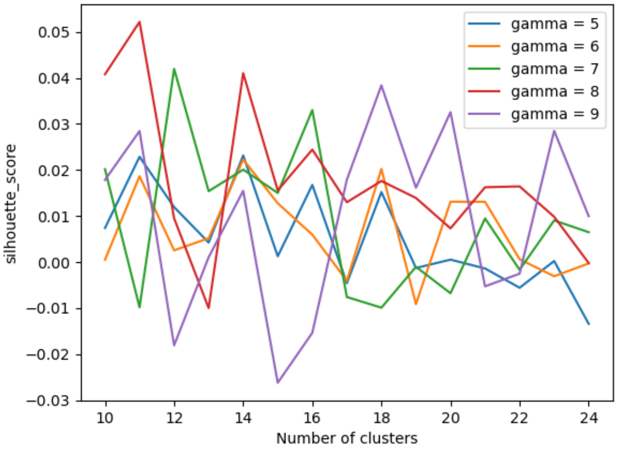


Fig. 5. silhouette score

1) Evaluation of clustering: In Fig 5, we adjust the number of clusters and weight ratio of the vector structure to get the best silhouette score. The quality of the generated storyline will also be a reference. For tweets dataset from May 2020 to August 2021, we found that silhouette score is maximized when the number of clusters is set to 18 and gamma is set to 9. We use this parameter pair for clustering in subsequent experiments.

2) Topic content: For each cluster we get, we focus on keywords and representative tweet within it. Fig 6 shows an example. The wordcloud shows high frequencywords in current topic. As a supplement to the topic, we also obtained the central sentence of the cluster in each topic as the most representative tweet in Fig 7. By these methods above, we show what people really care about in this topic.

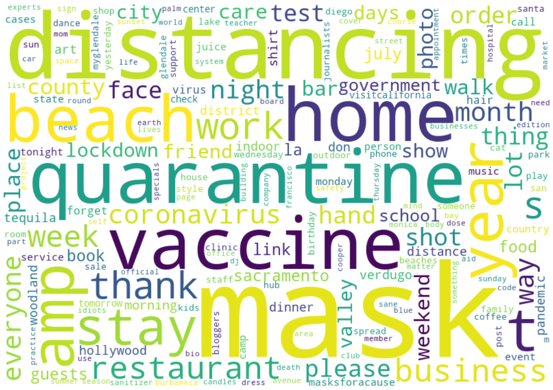


Fig. 6. wordcloud for one topic

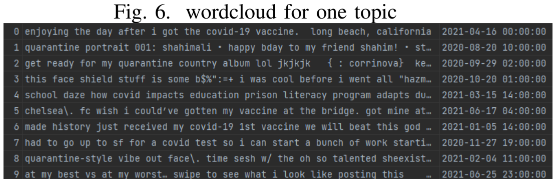

```
0 enjoying the day after i got the covid-19 vaccine.  long beach, california  2021-04-16 00:00:00
1 quarantine portrait 001: shahimali • happy bday to my friend shahim! • st…  2020-08-20 10:00:00
2 get ready for my quarantine country album lol jkjkjk   { : corrinova}  ke…  2020-09-29 02:00:00
3 this face shield stuff is some b$%":=+ i was cool before i went all "hazm…  2020-10-20 01:00:00
4 school daze how covid impacts education prison literacy program adapts du…  2021-03-15 14:00:00
5 chelsea\. fc wish i could've gotten my vaccine at the bridge. got mine at…  2021-06-17 04:00:00
6 made history just received my covid-19 1st vaccine we will beat this god …  2021-01-05 14:00:00
7 had to go up to sf for a covid test so i can start a bunch of work starti…  2020-11-27 19:00:00
8 quarantine-style vibe out face\. time sesh w/ the oh so talented sheexist…  2021-02-04 11:00:00
9 at my best vs at my worst… swipe to see what i look like posting this    …  2021-06-25 23:00:00
```

Fig. 7. centroid tweet set

3) Storyline: Fig 12 shows a result of storyline from LDA-BERT.

Each node in the graph represents a topic. From the label of nodes, we can get a list of high frequency words and published time. The number attached on the edges stand for a measure for correlation. Nodes with dark backgrounds are connected in series to form a storyline. The connection relationship between nodes reflects the connection between topics.

4) Model comparision: For each method, we produce a 2D distribution graph to show the relevance of each cluster in terms of time and content. Fig 9 shows a relatively scattered clustering results. However,in Fig 12 we can see that using lda alone result in too many topics mixed with noise.

Fig 10 shows poor clustering results that lots of clusters mixed together. It is difficult to reflect the inherent characteristics of each topic.

Fig 11 shows a compromised clustering effect. Some distinct clusters can be found at the edges of the graph. Meanwhile the topic generated from this method shows stronger noise immunity, refer to Fig 8

5) Derivation of topic: Our framework also produce a 3D distribution graph for the topic to show the time correlation, refer to Fig 13, where vertical z-axis stand for time. We can find that some topics are time-sensitive, which may reflect a shift in people's attention.

IV. Case Studies

In this project we propose a model mapping some spatio-temporal data into a map once getting such dataset from Twitter. This model successfully take pre-processed twitter

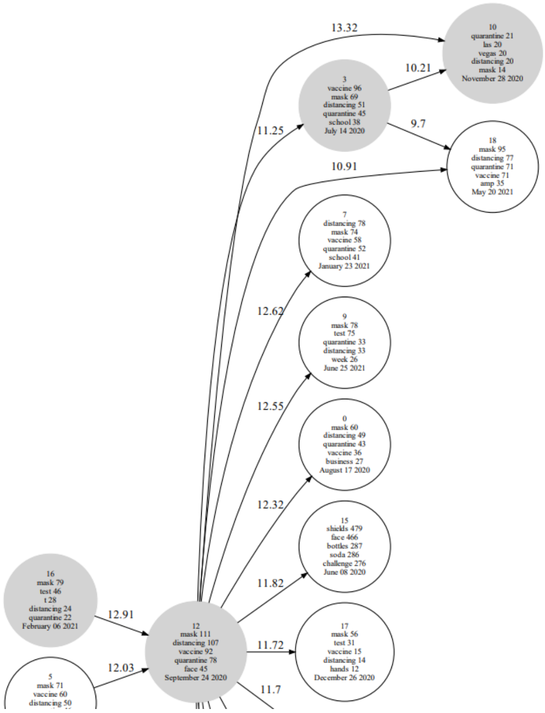


Fig. 8. slice of storyline for LDA-BERT

data and return visualized result of distribution of such data. With machine learning-based binary classification, the spatio-temporal data now can have sentiment information as well, therefore it is possible to visualize the the trend of people's opinion by regions and time. This model has potential to be developed as an API automatically return visualized map of distribution of sentiments on some issues. This may contribute the government decision maker to refer to public opinion on a specific issue they are interested in.

Visualization of the spatial feature flow as the temporal data progresses is an integral part of data warehousing and data mining techniques. Spatial data mining requires specific trend recognition in order to make successful conclusive argument which is visualized using the visualization technique. In this project, we have used Geo Pandas to visualize our spatial data using designated spatial operator. In our future studies, we would like to make the model propose the trends in spatial data flow in maps which can be visualized using the similar libraries.

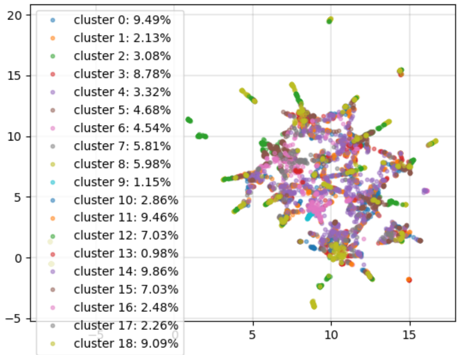


Fig. 9. 2d derivation for LDA

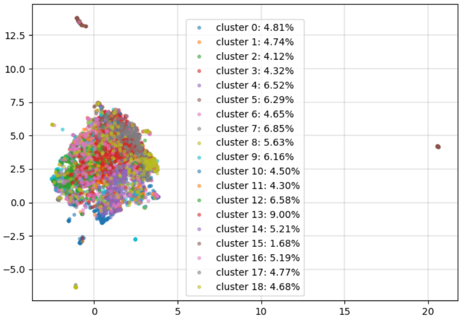


Fig. 10. 2d derivation for BERT

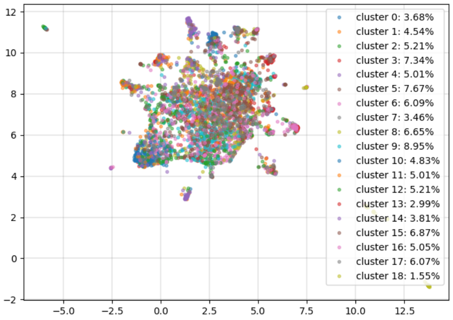


Fig. 11. 2d derivation for LDA-BERT

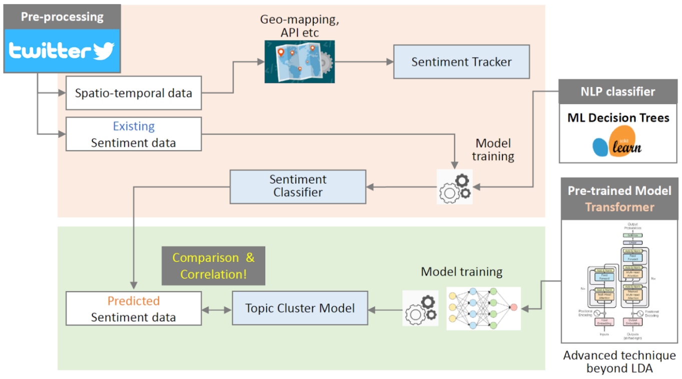


Fig. 14. Future plan

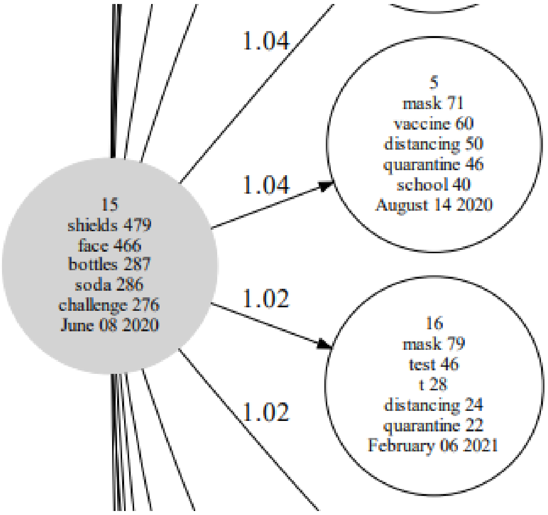


Fig. 12. slice of storyline for lda

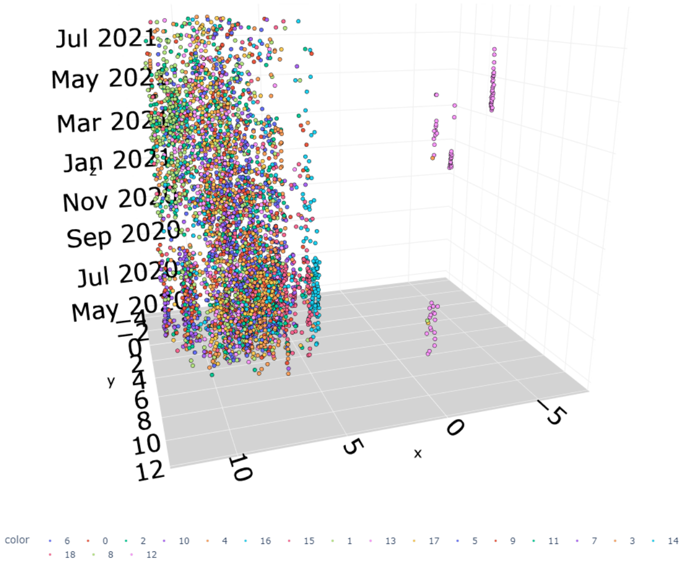


Fig. 13. 3D distribution graph

However it still has a long way to go in that it is too simple for disclosing some deeply meaningful implications. At present, since our model just provides the spatio-temporal view of the public opinion about an issue, it lacks further analysis about what specific topics are to be drawn from this. To dig into this issue, we will do detailed topic clustering task to classify how many different sub topics can be drawn from the overall topic of Covid-19 outbreak. Our topic for twitter dataset are filtered by the topic pandemic. However this is too simple. Therefore in the longer run, we should proceed to sub-topic clustering to get detailed areas under the single main topic. For example, they might be either government's specific policy about the pandemic or public opinion about some vaccine company's business guideline, and etc.

In this project, we have chosen a static method for topic clustering which is Latent Dirichlet Allocation (LDA). Although this is static and an old method compared to the modern deep learning approaches, our result shows its relevancy in this domain is still present. The current research that are going on, a lot of the authors are using LDA to get a primary insight in the dataset that they compare with their model as a baseline. In the coming days, our research will focus on the modern approaches using deep learning to make compact and deployable topic clustering model.

One promising candidate for the model is the Transformer. This is a greatly applied model architecture in Natural Language Processing now, which is based on the core technology called the *attention mechanism*. Our team belive this should be an appropriate alternative current LDA to do the detailed topic clustering task (Figure **??**).

V. Conclusion

In this project, our team carried out a machine learning-based spatial modeling of sentiment towards the pandemic. A rationale for this research comes from the increased interest on analyzing a lot of issues related to pandemic by using machine learning in academia. We have been pursued to combine the ML and Deep Neural Network (DNN) with our spatial-data mining technology so that this study contributes to the field by producing deeper implications.

As a dataset for sentiment analysis the importance of Twitter dataset can not be overstated. In recent days of DNN era, the public opinion expressed in twitter became a valuable resource for numerous tasks including such sentiment prediction. We paid attention on the applicability of DNN on the Twitter dataset for the pandemic issue. By using some applications such as Twarc and Elasticsearch we collected and stored such related dataset into our own repository, and spatial datamining task such as querying is followed. By doing such spatial query and filtering task, the target dataset could be visualized into a map. Furthermore, we develop a sentiment classifier by adopting the decision tree in Machine Learning. This module analyzes people's opinion in tweeter data and classifies it into positive or negative, by getting trained with pre-built sentiment score.

However, there are still another important task remained to achieve our team's final goal. Since our current model simply mine the spatio-temporal dataset and show how people's opinion is distributed differently by regions and time, it requires to do further research to draw deeper implications for contributing academia. Therefore, we plan to build a sub-topic clustering module using Transformer and compare its result with our predicted sentiment score. With adopting such a new NLP model architecture, it is expected to track a distribution of public opinion on detailed issues, e.g., government policy on social distancing, vaccine supply, etc., so that the study can have practical implication to decision makers.

References

[1] Z. Kastrati, L. Ahmedi, A. Kurti, F. Kadriu, D. Murtezaj, and F. Gashi, "A deep learning sentiment analyser for social media comments in low-resource languages," *Electronics*, vol. 10, no. 10, p. 1133, 2021.

[2] K. Chakraborty, S. Bhatia, S. Bhattacharyya, J. Platos, R. Bag, and A. E. Hassanien, "Sentiment analysis of covid-19 tweets by deep learning classifiers—a study to show how popularity is affecting accuracy in social media," *Applied Soft Computing*, vol. 97, p. 106754, 2020.

[3] R. Batool, A. M. Khattak, J. Maqbool, and S. Lee, "Precise tweet classification and sentiment analysis," in *2013 IEEE/ACIS 12th International Conference on Computer and Information Science (ICIS)*. IEEE, 2013, pp. 461–466.

[4] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in *International conference on machine learning*. PMLR, 2016, pp. 478–487.

[5] A. Abd-Alrazaq, D. Alhuwail, M. Househ, M. Hamdi, Z. Shah *et al.*, "Top concerns of tweeters during the covid-19 pandemic: infoveillance study," *Journal of medical Internet research*, vol. 22, no. 4, p. e19016, 2020.

[6] O. Gencoglu, "Large-scale, language-agnostic discourse classification of tweets during covid-19," *Machine Learning and Knowledge Extraction*, vol. 2, no. 4, pp. 603–616, 2020.

[7] D. Kumar, N. Ramakrishnan, R. F. Helm, and M. Potts, "Algorithms for storytelling," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 6, pp. 736–751, 2008.

[8] M. S. Hossain, P. Butler, A. P. Boedihardjo, and N. Ramakrishnan, "Storytelling in entity networks to support intelligence analysts," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2012, pp. 1375–1383.

[9] L. Fang, A. D. Sarma, C. Yu, and P. Bohannon, "Rex: explaining relationships between entity pairs," *arXiv preprint arXiv:1111.7170*, 2011.

[10] N. Voskarides, E. Meij, M. Tsagkias, M. De Rijke, and W. Weerkamp, "Learning to explain entity relationships in knowledge graphs," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2015, pp. 564–574.

[11] H. Lee, M. Recasens, A. Chang, M. Surdeanu, and D. Jurafsky, "Joint entity and event coreference resolution across documents," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2012, pp. 489–500.

[12] D. Shahaf, C. Guestrin, and E. Horvitz, "Metro maps of science," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2012, pp. 1122–1130.

[13] G. Kim, L. Sigal, and E. P. Xing, "Joint summarization of large-scale collections of web images and videos for storyline reconstruction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 4225–4232.

[14] C. C. Park and G. Kim, "Expressing an image stream with a sequence of natural sentences," *Advances in neural information processing systems*, vol. 28, 2015.

[15] D. Wang, T. Li, and M. Ogihara, "Generating pictorial storylines via minimum-weight connected dominating set approximation in multi-view graphs," in *Twenty-sixth AAAI conference on artificial intelligence*, 2012.