Social Media Mining with Dynamic Clustering: A Case Study by COVID-19 Tweets

Hidetoshi Ito
Graduate School of Software and Information Science
Iwate Prefectural University
152-52 Sugo, Takizawa, Iwate 020-0693, JAPAN
Email: g231r004@s.iwate-pu.ac.jp

Basabi Chakraborty
Faculty of Software and Information Science
Iwate Prefectural University
152-52 Sugo, Takizawa, Iwate 020-0693, JAPAN
Email: basabi@iwate-pu.ac.jp

2020 11th International Conference on Awareness Science and Technology (iCAST) | 978-1-7281-9119-5/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICAST51195.2020.9319496

***Abstract*—Recently Social Networking Service (SNS) is used extensively due to proliferation of the Internet and cheaper, compact, easy to use computing devices. Texting, especially via Twitter, is very popular among people of all ages all over the world, and enormous text data is generated regularly which contains various types of information, rumors, sentimental expressions etc. The variety of topics related to the contents of the social media data are prone to changes with the passing of time and sometimes fade out completely after a certain time. Such time varying topics may include beneficial information that could be used for various decision making by general public as well as governmental organization. Especially for the recent pandemic of COVID-19, extraction and visualization of the changing needs of people might help them making some better countermeasures. In this study, COVID-19 related tweets have been collected and analyzed in units of time (hour, day and month) by means of various clustering models to visualize the dynamic changes of topics with time. It is found that Sentence-Bert is the most effective tool among the techniques used here though it is not yet enough for clear understanding of the topics semantically.**

***Index Terms*—topic model, document clustering, social media mining**

I. Introduction

Nowadays Social Networking Services (SNS) have been widened largely for sharing opinion, impression, sentiment,events, or information and contents among the users world-wide. Twitter, a microblogging platform, is one of the most popular SNS that has enormous effect as a communication channel between people. People tweet their opinion and information on various events and happenings via Twitter and check others tweets whenever they can connect to the internet, even if it is during a disaster. Therefore it is an useful media also to share variety of important information via announcements from public organizations during emergency. It also includes the reactions of users to the information, and there is a possibility to know how current topics emerge, spread and change astime goes on. Especially these days, many people tweet about novel coronavirus, COVID-19. The COVID-19 pandemic has started from the end of 2019, and various countermeasures and economic policies have been taken by each government. People post reactions and exchange opinions in Twitter and spread information about this. Then analysis of COVID-19 related tweets have been conducted by several researchers from various perspectives, such as topic extraction, sentiment analyis [1] and transmission of information [2], so far. The tweets also include shifting demands and criticisms. Hence the dynamic changes of topics have also received attention [3]–[5] from the users. Analysis of COVID-19 tweets colleted over a period of time can reveal the change of interests, needs of people through time which in turn may be useful for governments for taking various economic measures.

In this study, COVID-19 related tweets have been collected over a period of time and then the tweets are clustered. At first topic modelling for all tweets have been done by Latent Dirichlet Allocation (LDA), a popular topic modelling tool. Secondly, tweets are embedded in vector space by three pretrained language models; Word2vec; BERT; Sentence-BERT, and are clustered by k-Means clustering in order to compare each vector representation for finding out the best method for clustering tweets with regard to semantics. Thirdly, tweets are grouped according to three units of time; hour; day and month, and dynamic clustering with single-linkage method have been done to extract the change of information regarding COVID-19 with time. The next section represents a brief summary of works on analysis of COVID-19 tweets. The following section describes the methodology of analysis adopted in this paper. Section 4 presents the data set and analysis followed by summary and conclusion in the final section.

II. Analysis of COVID-19 Tweets

Several research papers on analysis of COVID-19 tweets have been published recently. Most of the papers conducted topic modelling by LDA, or used other dynamic topic model for analysis. [1] performed a sentiment analysis to identify emotional aspects of tweeter users and the predominant emotion expressed in the tweets. As a result, it was found that nearly half of all the tweets expressed fear and nearly 30% expressed surprise. [3] analysed topic-level sentiment dynamics with dynamic topic models. The results show that there are various discussions about COVID-19, and the topic sentiment is polarized. The research works in [2] and [6] analysed the diffusion of information about the COVID-19 from twitter data. In contrast to calculation of retweeting times in [6], the authors of [2] fit information spreading with epidemic models characterizing the basic reproduction numbers R_0 for each social media platform.

TABLE I
PAPERS ON COVID-19 TWITTER ANALYSIS

Author	Number Tweets	Time Period	Topic	Sentiment	Transmisison	Mobility	Network	Dynamic	Clustering
Sha, et al. [5]	7881	Jan. 1 - Apr. 7, 2020	x				x	x	
Medford, et al. [1]	126049	Jan. 14 - Jan. 28, 2020	x	x					
Ours	30679	Jan. 21 - Jul. 23, 2020	x					x	x
Cinelli, et al. [2]	1187482	Jan. 27 - Feb. 14, 2020	x		x				
Zamani, et al. [4]	2600000	Mar. 12 - Jun. 30, 2020	x			x		x	
Ordun, et al. [6]	23830322	Mar. 24 - Apr. 9, 2020	x				x		
Yin, et al. [3]	4919471	Apr. 1 - Apr. 14, 2020	x	x				x	

In [4], the authors proposed a dynamic content-specific LDA topic modelling technique which have shown more coherent topics than standard LDA topics on COVID-19 corpus built from publicly available Twitter data by themselves, and the topics are used as features for predicting monthly mobility and unemployment. In [5], the authors collected COVID-19 related tweets by U.S. governors and presidential cabinet members, then applied a network "Hawkes binomial topic model" to the tweet data to track sub-topics around risk, covid testing and treatment, and an influence network among government officials have been constructed.

Unlike those papers, in our work, we analysed cluster occurrence and fading time on COVID-19 tweets with pre-trained language models and agglomerative clustering approach over time. Although the whole process is simple, our method can give insights into complex tweet stream and obtain topic timeline like in [5] easily.

III. METHODOLOGY

In this section, several text mining methods used in our analysis are introduced briefly.

A. Topic Model

Topic model, a probabilistic model, is based on the assumption that a document is generated from multiple topic distributions and topics are generated from multiple word distributions. Topics are characterized by the words with higher probabilities in the cluster of words and the documents contain topics having different topics probabilities. The most typical topic modelling method is Latent Dirichlet Allocation (LDA) [7], which is a generative probabilistic model for text corpora. Graphical model of LDA is represented in Fig. 1, where D denotes the number of documents, N is number of words in a given document, α is the parameter of the Dirichlet prior on the per-document topic distributions, β is the parameter of the Dirichlet prior on the per-topic word distribution, θ_d is the topic distribution for document d, φ_k is the word distribution for topic k, $Z_{d,n}$ is the topic for the n-th word in document d, and $W_{d,n}$ is the specific word.

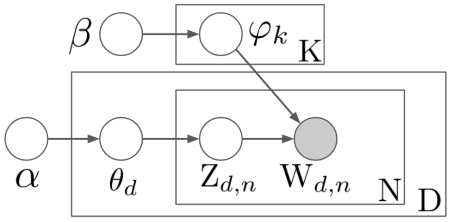

Fig. 1. Graphical Model of LDA

B. Word Embeddings

In natural language processing, representations of word are important because computers cannot process the words consisting of only strings. Classically, the frequency or co-occurrence probability of the words are used to represent a word, however the vector representation, the word embedding is used in recent years. Word2Vec [8], one of the most popular word embedding method, is used in this study.

C. BERT and Sentence-BERT

BERT is one of the revolutionary neural language representation model, which stands for Bidirectional Encoder Representations from Transformers [9]. The model learns Masked Language Model (MLM) and Next Sentence Prediction (NSP) tasks, and the last hidden layer outputs token vector that is affected by meaning of the sentence. Usually the pretrained model, trained with enormous corpus, is sufficient to use as a general-purpose language representation model. Therefore it is used in particular application with fine tuning or transfer learning.

Sentence-BERT is a modification of the pretrained BERT network which is tuned with metric learning to derive semantically meaningful sentence embeddings [10]. The sentence embeddings can be compared using cosine-similarity, but the authors noted that comparing the embeddings with negative euclidean distance is not very different to using cosine-similarity.

D. Document Clustering

In simulation experiments, the tweets are clustered by two methods: k-Means and single-linkage clustering.

k-Means [11] is the most famous clustering method that adjust positions of the cluster centroids iteratively until convergence to the final clusters. At first, $k(\leq n)$ a predefined number of cluster centers, $M = \{\mu_1, \mu_2, \ldots, \mu_k\}$ are usually chosen randomly from all data points $X = \{x_1, x_2, \ldots, x_n\}$, and data points are assigned to one of the clusters $S = \{S_1, S_2, \ldots, S_k\}$ to satisfy following equation:

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2 \tag{1}$$

In each iteration, all other points are included in an optimum way according to a predefined strategy to the proper cluster and the cluster centers are recomputed. Finally the algorithm stops when no further readjustments are noted in successive iteration.

Single-linkage clustering [12] is a member of hierarchical agglomerative clustering methods. All data points are assigned to different clusters that include only oneself firstly. These clusters are joined recursively with a link function as following

$$l(K, K') = \min_{x \in K} \min_{x' \in K'} d(x, x') \tag{2}$$

where each of K, K' is a cluster, x, x' are data points of clusters K, K' respectively, and d denotes the distance metric between the data points. In this study, the link function is upper-bounded by T as below:

$$\begin{cases} \text{link} & (l(K, K') < T) \\ \text{do not link} & (otherwise) \end{cases}$$

The value of T corresponds to a semantic boundary.

E. Clustering Index

To evaluate clustering results, various indexes are proposed, such as Sum of Squared Error (SSE), Dunn's Index, etc. In this study, Pseudo-F [13] have been used to evaluate the clustering results. The formula is as below:

$$Pseudo\ F = \frac{BSS/(K-1)}{WSS/(N-K)} \tag{3}$$

where N is the number of observations, K is the number of clusters, BSS is the between-cluster sum of squares, and WSS is the within-cluster sum of squares. The larger the Pseudo F, the more close-knit and independent are the clusters.

IV. Dataset and Exploratory Analysis

In this section COVID-19 twitter data set and the results of the analysis of twitter data by various text mining techniques are presented.

A. Dataset

COVID-19 related tweets [14] from 22:00 on January 21, 2020 to 23:00 on July 23, 2020 and random 30 tweets per hour, are collected from Twitter. Tweets that starts with "RT" i,e. Retweets have been excluded. 30679 tweets are used for the analysis.

B. Exploratory Analysis

1) Topic Modelling: At first, topic modelling by LDA have been conducted for all tweets. Tweets have been segmented into words, converted to lowercase, lemmatized, and some words and URLs are filtered out. Filtered words are in the table below;

TABLE II
Filtered Words

(1 character words)
rt
corona
virus
coronavirus
covid
covid19
covid-19

Then a word frequency dictionary have been constructed. Words that appeared in five or less tweets are also excluded, and the excluded words appeared in 60% or more tweets.

LDA model has been developed from tweets with topic numbers ranging from 2 to 50. The topic models with different topic numbers are compared with metrics of coherence and perplexity as shown in Fig.2. The higher coherence and lower perplexity indicate better LDA model. From Fig.2, considering the values of coherence and perplexity, 11 topics have been chosen as the proper number of topics from topic modelling by LDA.

Top 10 words of the topics are shown in Table III.

From the results shown in the table, it is found that there is no noticeable duplication of words in the topics, and each topic seems to be not difficult to interpret from human perspective.

C. Clustering of Tweets

Secondly, tweet clustering have been conducted. All tweets have been transformed as embedding vectors with; 1) Word2Vec; 2) BERT; 3) Sentence-BERT.

1) Word2Vec: Tweets have been preprocessed in the same way as is used for topic modelling above, then selected words that are not included in the pretrained model 'word2vec-google-news-300', are taken as the mean of the word vertices as a tweet vector. The result shows that the clusters are strongly affected by the words and the word frequencies of each cluster have less duplication, same as the result obtained by LDA.

2) BERT: Tweets have been converted to lowercase, tokenized by pretrained BERT wordpiece tokenizer, and converted to token vertices by the pretrained model 'bert-base-uncased' with Huggingface's Transformers. Tweet vector is the mean of the token vertices. The result shows that the clusters are

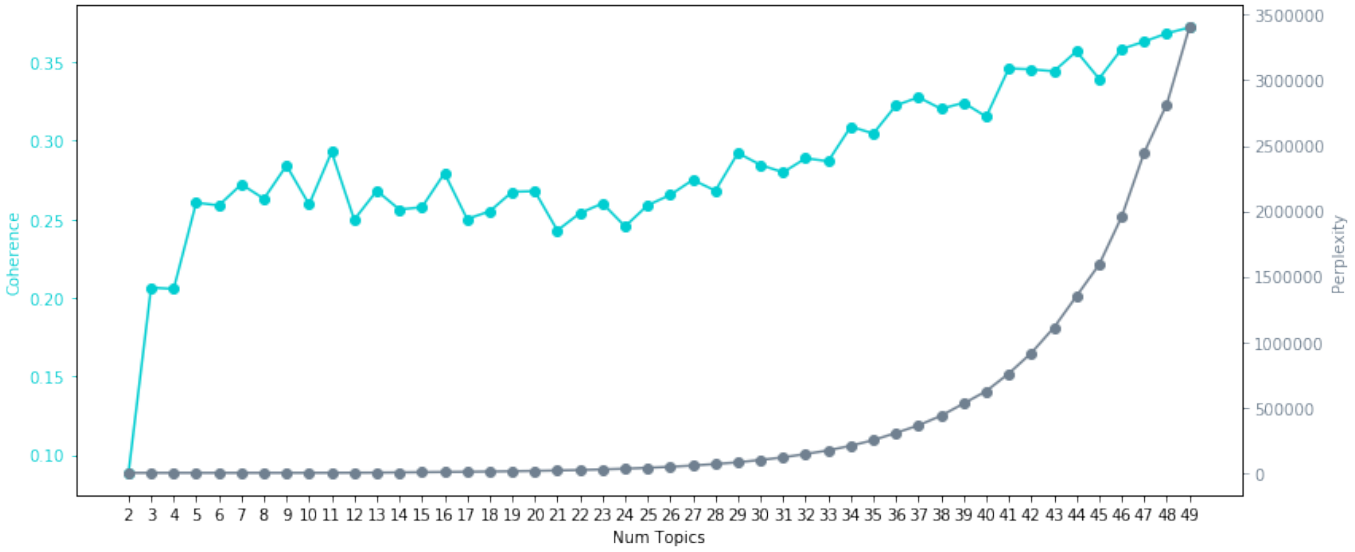

Fig. 2. Coherences and Perplexities of LDA Models

TABLE III
TOP 10 WORDS OF THE TOPICS

Topic 1	2	3	4	5	6	7	8	9	10	11
case	florida	die	trump	student	social	house	child	mask	lockdown	pandemic
death	people	pandemic	right	pandemic	people	patient	pandemic	wear	test	testing
new	die	support	pandemic	school	distancing	start	like	home	positive	july
report	big	pay	people	response	job	hospital	year	stay	week	19
number	single	look	president	india	kid	bill	time	face	month	party
day	donald	long	think	governor	business	need	election	life	..	people
record	go	economy	know	government	go	old	lose	fauci	new	health
reopen	rate	hydroxychloroquine	@realdonaldtrump	administration	black	fuck	white	safe	see	crisis
high	million	will	want	trump	get	worker	feel	people	day	world
today	republican	like	open	state	work	refuse	hear	save	ago	global

affected by the sentence and the word frequencies of each cluster have many duplication.

3) Sentence-BERT: Tweets are converted to tweet vertices by pretrained Sentence-BERT model 'bert-base-nli-mean-tokens' with Sentence Transformers framework based on Huggingface's Transformers. As a result, the clusters seem to be about midway between that of Word2Vec and BERT.

Table IV, shows the evaluation of the above three methods by using the metric $PseudoF$ with two different similarity measures. From the scores of Pseudo-F in table IV, it seems that the Sentence-BERT is the most appropriate sentence embedding method for document clustering, that is, the embedding space has more meaningful distribution than Word2Vec and BERT.

TABLE IV
PSEUDO-F

	Word2Vec	BERT	Sentence-BERT
euclidean	1036.9	2039.2	2343.7
cosine	1063.3	2014.8	2339.6

D. Dynamic Tweet Clustering

Thirdly, tweets have been clustered following the procedure below:

1) Divide tweets according to hour, day, or month
2) Embed tweets into vector space with pre-trained language model, Sentence-BERT
3) Cluster each unit time group with T upper-bounded single-linkage method
4) Join i, $i+1$ unit time clusters if the following conditions are satisfied; 1) the two are close enough to link each sample, and 2) the cluster means are not farther than L.

For experiments, the parameter T ranges from 0.1 to 0.9 in 0.1 increments for cosine distance, 6 to 18 in 2 increments for euclidean distance. L have been set to T and $T/2$.

Fig. 3 and 4 are unique plots of the cluster lifetimes in unit time corresponding to hour and day respectively. We have excluded very small clusters that have only 10 or less tweets. Result of clustering by month have not divided into multiple clusters this time. Therefore this process has not worked for monthly tweets. For all the experiments with various values of all the parameters, noise clusters are only observed for results with cosine distance and $L = T/2$. From the daily plot, it is seen that the persistent clusters appear and disappear about every 15 days, that means a COVID-19 related topic had continued about almost two weeks.

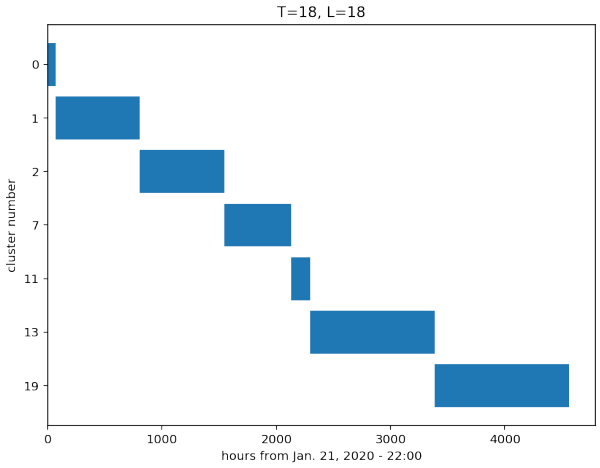

Fig. 3. Hourly Cluster Lifetimes

To analyse some of the persistent clusters such as cluster 2 (clusters are counted according to time) in Fig. 3, word frequencies have been counted. The result shows that continuous or persistent clusters have similar words such as "lockdown", "pandemic", or "distancing". Then noise clusters or independent clusters such as cluster 3 have been analysed by the same way where some unique words have been observed, for example, "business", "industry", "stay", "home", "child", "fake", "mask", etc.

Looking inside some clusters of Fig. 3, cluster 1 and 2 have negative mentions for China or Wuhan while after some passage of time such as in cluster 7, positive attitudes for the

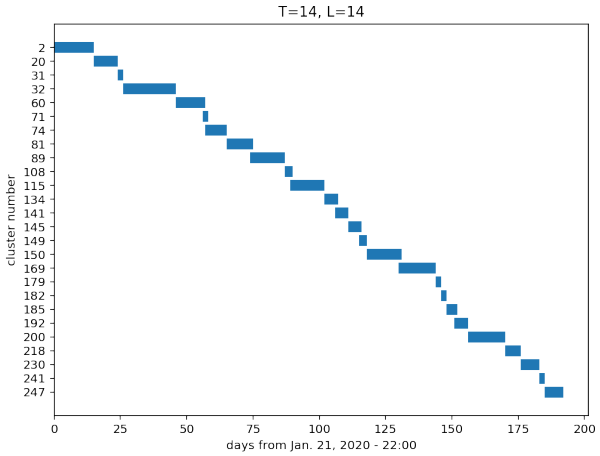

Fig. 4. Daily Cluster Lifetimes

solution of COVID-19 pandemic and daily life with the virus are found. Also looking at the clusters of the Fig. 4, cluster 115 have political or economic opinions and hypotheses of the infection and the treatment. Then, the infection status after some countermeasures and controversies for that are found in cluster 230.

V. Summary and Conclusion

In this study, COVID-19 related tweets have been collected and retweets are excluded. The rest of the tweets are then analysed by topic modelling, clustering with different sentence embeddings, and dynamic clustering.

Firstly, the topics of tweets that confirmed by LDA are well-expressed regarding what happened about COVID-19. LDA is a good tool to extract an overview of the corpora.

Then each sentence embedding has been compared by k-Means clustering results to detect which embedding space is the most appropriate to compare the tweets semantically. As a quantitative evaluation, Sentence-BERT scores the best Pseudo F value. However, it does not mean Sentence-BERT is necessarily better for formation of topic clusters. Actually, the clustering result with Word2Vec sentence embeddings have drawn similar wordclouds to the LDA topics.

Analyzing dynamic clusters is still not enough, because it seems that there is no appropriate tool to visualize the overview of sentence clusters. For example, a tool that extract a least common multiple like sentence from a cluster is needed. The forming process of dynamic clusters should also need to be reconsidered because a cluster once interrupted will not appear ever after, that means two clusters are treated as different clusters if the same cluster appear again after some interval.

This study shows that some effective method can be developed for analysis of tweet data to uncover the underlying change of information over a certain time period. We are currently working for more refinement of the techniques for extracting change of information dynamically from the analysis of tweet data.

REFERENCES

[1] Richard J. Medford, Sameh N. Saleh, Andrew Sumarsono, Trish M. Perl, and Christoph U. Lehmann. An "infodemic": Leveraging high-volume twitter data to understand public sentiment for the covid-19 outbreak. *medRxiv*, 2020.

[2] M Cinelli, W Quattrociocchi, A Galeazzi, CM Valensise, E Brugnoli, AL Schmidt, P Zola, F Zollo, and A Scala. The covid-19 social media infodemic. arxiv. *arXiv preprint arXiv:2003.05004*, 2020.

[3] Hui Yin, Shuiqiao Yang, and Jianxin Li. Detecting topic and sentiment dynamics due to covid-19 pandemic using social media, 2020.

[4] Mohammadzaman Zamani, H. Andrew Schwartz, Johannes Eichstaedt, Sharath Chandra Guntuku, Adithya Virinchipuram Ganesan, Sean Clouston, and Salvatore Giorgi. Understanding weekly COVID-19 concerns through dynamic content-specific LDA topic modeling. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 193–198, Online, November 2020. Association for Computational Linguistics.

[5] Hao Sha, Mohammad Al Hasan, George Mohler, and P. Jeffrey Brantingham. Dynamic topic modeling of the covid-19 twitter narrative among u.s. governors and cabinet executives, 2020.

[6] Catherine Ordun, Sanjay Purushotham, and Edward Raff. Exploratory analysis of covid-19 tweets using topic modeling, umap, and digraphs, 2020.

[7] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022, March 2003.

[8] Tomas Mikolov, Ilya Sutskever, Kai Chen, G.s Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26, 10 2013.

[9] J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.

[10] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. pages 3973–3983, 01 2019.

[11] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pages 281–297, Berkeley, Calif., 1967. University of California Press.

[12] Gunnar Carlsson and Facundo Mémoli. Classifying clustering schemes. *Foundations of Computational Mathematics*, 13, 11 2010.

[13] Tadeusz Caliński and Harabasz JA. A dendrite method for cluster analysis. *Communications in Statistics - Theory and Methods*, 3:1–27, 01 1974.

[14] Emily Chen, Kristina Lerman, and Emilio Ferrara. Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. *JMIR Public Health Surveill*, 6(2):e19273, May 2020.