Topic Evolution Analysis of COVID-19 News Articles

Yun Bai¹, Suling Jia¹, Lao Chen²⁺*

¹School of Economics and Management, Beihang University, Beijing, China;
²Information Office of Beihang University, Beihang University, Beijing, China
chenlao@buaa.edu.cn

Abstract. Recently, numerous media publishes various news on the latest developments every day due to the global spread of COVID-19. The news provides rich information about COVID-19 and includes a wide range of evolving topics. Our study is intended to develop a dynamic topic analysis system to monitor the evolution of the large-scale text data topics and assist with the social management and policymaking. The system expands the Dynamic Topic Model (DTM) with two modules: data sparsity computing and topic number selecting, which makes the experimental process more natural and generalizable. Data sparsity is designed to determine the length of single time slice. UCI, UMass and NPMI are introduced for choosing the optimal number of topics. This paper explores CBC news articles using DTM and captures the impact of COVID-19 on various aspects and the development of specific events. The experimental results demonstrate the effectiveness of our system for discovering and tracking the evolving topics. This system also plays an important role to improve the awareness of the public and serves as an analysis platform for government.

1. Introduction

COVID-19 broke out in China in December 2019 and caused the lockdown of Wuhan. As of February 8, 2020, more than 30,000 people have been diagnosed and at least 800 people died [1]. COVID-19 has become a major global disease and received widespread attention. People can receive news about COVID-19 through various channels. The media publishes news to state facts and express opinions, from which people learn about the spread of the epidemic, the efforts made by the government and hospitals, and how can people prevent infections in their daily lives. The online news is fresh and constantly evolving, and has a major impact on society, politics and people's lives. The dynamic topic model (DTM) can extract evolving contents from texts with temporal features. Analyzing the trends and topics provides insights into both changing opinions of society and the COVID-19 phenomenon.

In this paper, we explored over 3500 news articles of COVID-19 from Canadian Broadcasting Corporation (CBC) via the DTM. This dataset is from Kaggle, click to download. The sparsity of the dataset is considered to split the timeline into slices, and some coherence scores assist to choose the appropriate number of topics. We then capture the evolution of topics by extracting the fluctuation of certain words in some topics of each time slice. The rest of the paper is organized as follows. Section 2 introduces the theory of the main approach. Section 3 presents our experimental process and results. Finally, all work is concluded in Section 4.
2. Approach

2.1 Dynamic Topic Model (DTM)

DTM is composed of a series of latent Dirichlet allocation (LDA) models, which are assigned to each time slice. LDA treats the collection of documents in each time slice as observations, which are generated through a process of generative probability. The process contains hidden variables, which can reflect the topic structure of the text. First of all, $k$ topics $\beta_1, \beta_2, ..., \beta_k$ are drawn from a Dirichlet distribution $\text{Dir}(\eta)$, which is treated as the topic distribution. Then for each document $d$, topic proportions $\theta_d$ are obtained from a different Dirichlet distribution $\text{Dir}(\alpha)$. To generate a single word $n$, topic assignment $Z_{d,n}$ for per word is drawn from a multinomial distribution $\text{Mult}(\theta_d)$. Finally, the observed word $W_{d,n}$ can be got from the multinomial distribution $\text{Mult}(\beta_{z_{d,n}})$. In DTM, documents of the same time slice are regarded as a collection, in which the documents are exchangeable. Blei [2] believes that the content of topics evolves. A logistic normal distribution is used to model the evolving topics within each time slice. Let $\beta_{k,t}$ be the $k$-th topic of time slice $t$, which can be obtained from $N(\beta_{k,t-1}, 1\sigma^2)$. Variance $\sigma$ can be treated as the degree of change from $\beta_{k,t-1}$ to $\beta_{k,t}$.

In practice, each document is a mixture of the distribution of topics in the corpus. Only the words in the document are observed, and all parameters are unknown. Variational inference can be used to estimate parameters to make them more realistic.

2.2 Sparsity of Data

How to divide the time slice is a point we need to consider next. In general, common methods include dividing by day, month, and year. If the time slice is too long, it may cause the result that the topic evolution is not being fully excavated. Conversely, a lot of memory and speed is sacrificed for the DTM. Wang et al [3] proposed an index of data sparsity as follows, which can be used to determine the size of time slices:

\[ \text{Sparsity} = 1 - \left( \sum_t \sum_w \delta_{t,w} \right) / (VT) \]  

In formula (1), $\delta_{t,w}$ is a binary variable that judges whether the word $w$ is unique in time slice $t$. When $w$ is unique, $\delta_{t,w}$ is equal to 1, otherwise $\delta_{t,w}$ is 0. Then $\sum_t \sum_w \delta_{t,w}$ is the total number of unique words in each time slice. $V$ is the size of the vocabulary and $T$ is the number of time slices.

Considering that not all words in each time slice are used to measure topic evolution, sparsity is designed to simply measure how much the words have changed in different time slices. When the lower the sparsity is, the denser the data set, and the DTM is more effective.

2.3 Topic Coherence

Topic coherence measures the similarity between words with high scores within a topic. If the top words are highly similar, these words can be considered to support each other within the topic. Therefore, topic coherence can be used to evaluate topic models. There are three common calculation methods of topic coherence: UCI, UMass, and NPMI.

UCI is used in [4] according to the Pointwise Mutual Information (PMI):

\[ UCI(w_i, w_j) = \log \frac{p(w_i, w_j)}{p(w_i)p(w_j)} \]  

UMass is introduced in [5], which is also a pairwise score like UCI:

\[ UMass(w_i, w_j) = \log \frac{p(w_i, w_j) + 1}{p(w_i)} \]  

NPMI, also named Normalized Pointwise Mutual Information, is an extension of PMI. In NPMI, the negative logarithm is used to weight the PMI [6]:

\[ NPMI(w_i, w_j) = \frac{PMI(w_i, w_j)}{-\log (p(w_i, w_j))} \]
3. Experiment and Results

3.1 Number of Time Slices
We obtained 3534 COVID-19 related news articles from Kaggle, which are released by CBC. The dataset spans from January 8 to March 27, 2020. According to [7], DTM model is more efficient when data sparsity is around 0.2. The results in Figure 1 represent that when the period is 12 days, the sparsity is equal to 0.21. Therefore, we choose 12 to be the length of each time slice and obtain 7 time slices.

![Figure 1. Data sparsity of time span 1 to 15.](image)

3.2 Topics and Contents
It’s assumed in DTM that the number of topics in each time slice is fixed. We calculated the UCI, UMass, and NPMI under a different number of topics and determined that the optimal number of topics is 20, for the reason that all indicators show a high value, as represented in Figure 2.

![Figure 2. Three indicators under different number of topics](image)

Since each topic is the distribution over words in the corpus, we can understand the latent meaning by observing the top 10 words under each topic. Table 1 lists the 20 topics selected and the corresponding keywords at the first time slice (from Jan 8 to Jan 19) according to coherence.

| Topic | Top words |
|-------|-----------|
| Topic 1 | Nova, Scotia, province, covid-19, Newfoundland, people, Brunswick, Labrador, Strang, John |
| Topic 2 | game, people, year, church, time, athlete, covid-19, Olympic, Tokyo, decision |
| Topic 3 | school, student, university, parent, covid-19, class, online, march, child, home |
By observing and comparing the top words under different topics, we can find the difference in content among the topics. For example, most of the words in topic 1 are the names of provinces and cities in Canada. Topic 1 is conjectured to be the impact of COVID-19 on the whole of Canada.

The topic model has brought many inspiring discoveries. First, the influences of the COVID-19 epidemic include all aspects. For instance, the Tokyo Olympics (topic 2); the transformation of students’ class (topic 3); urban transportation (topic 4); government and policies (topic 5); consumption patterns (topic 7); medical supplies (topic 11); people's employment situation (topic 12); entertainment activities (topic 13); national finance (topic 18). COVID-19 is a global public health event, many countries have been affected at varying degrees (topic 14), and the international economy has also suffered (topic 9).

The results of the topic model also represent events closely related to COVID-19, such as the outbreak in Wuhan, China (topic 16); cross-border tourism leads to the spread of the epidemic (topic 17); scientists study viruses carried by humans and animals (topic 20).

### 3.3 Evolution of Topics

The topic distribution, which is the probability that each document belongs to each topic, can be calculated from the DTM. The mean value of the topic distribution of a collection is taken as the topic intensity in that period. The higher the intensity, the more of the document collection focuses on this topic. We use a heat map as Figure 3 to show the evolution of 20 topics in 7 time slices. In Figure 3, the lighter the color, the higher the topic intensity. Figure 3 illustrates that the number 15 is stronger than the others, which corresponds to Topic 16 in the previous section, due to numbering starts from 0. Below we will focus on Topic 16 and study the evolution of its content and top words.

Figure 4 lists the evolution process of the top 10 words of topic 16 in 7 time slices, from which we can find that the three words *China, coronavirus, and case* have always ranked in the top three. To show the probability variety of words, we selected *health, outbreak, Wuhan, ship, and cruise* as the line chart shown in Figure 5.
Figure 3. Heat map of the topic evolution.

Figure 4. Evolution of top 10 words of Topic 16

Figure 5. Probability variety of the 5 words

Figure 5 represents that the words health and outbreak are relatively stable, with a slight downward trend; the probability of Wuhan continues to decline, and it disappears after the fifth time slice; the
probability of ship and cruise first increases rapidly and then remains stable. The evolution of topic content reflects the development of COVID-19. The epidemic first broke out in Wuhan, China, and began to spread around the world. Many passengers on the Diamond Princess cruise ship were infected and they had to be quarantined to prevent the coronavirus from spreading further. Many Canadians on the cruise are anxiously waiting to return to homeland, according to the CBC news "Canadians wait for flight home as COVID-19 numbers rise on quarantined ship" released on February 19. This also explains why in the fourth time slice, ship and cruise reached the maximum probability.

4. Conclusion
In 2020, the outbreak of COVID-19 has become the most concerned in the world. Many people have lost their lives and global politics and economy have been affected at varying degrees. Lots of media continue to release relevant news every day to broadcast the latest situation of the epidemic. Mastering online information is essential for social management. There are many studies that analyse the topic content and evolution based on big data. This article proposes a system for analyzing the evolution of news reports topics utilizing DTM. More than 3,500 articles of CBC news were downloaded from Kaggle. The date is from January 8 to March 27, 2020, covering the time from the beginning of the outbreak to the worldwide spread. First, the sparsity of the data set is computed to determine the length of the time slice; next, after calculating UCI, UMass, and NPMI, the appropriate number of topics is selected. After the above two steps, we identified 7 time slices and 20 topics. These 20 topics inspired many influences of COVID-19, which are reflected in many aspects of life, such as public transportation, online courses, medical supplies, and purchase methods.

The analysis of topic evolution includes state and content. We use a heat map to show the development of 20 topics over time. When analyzing content evolution, Topic 16, the most obvious in the heat map, is considered. By comparing top words from different periods, we captured the incident of the Diamond Princess cruise ship.

The dynamic topic model is suitable for content mining of large-scale unstructured data. Unlike the static topic model, DTM can capture the evolution of the topic, which is vital for tracking the development of major events. Major events often attract a lot of attention, and the analysis of them is helpful for public opinion management and hot spot tracking and also has certain social significance.

References
[1] Liu, Y., et al., The reproductive number of COVID-19 is higher compared to SARS coronavirus. Journal of travel medicine, 2020.
[2] Blei, D.M. and J.D. Lafferty. Dynamic topic models. in Proceedings of the 23rd international conference on Machine learning. 2006.
[3] Wang, C., D. Blei, and D. Heckerman, Continuous time dynamic topic models. arXiv preprint arXiv:1206.3298, 2012.
[4] Newman, D., et al. Automatic evaluation of topic coherence. in Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics. 2010. Association for Computational Linguistics.
[5] Mimno, D., et al. Optimizing semantic coherence in topic models. in Proceedings of the conference on empirical methods in natural language processing. 2011. Association for Computational Linguistics.
[6] Aletras, N. and M. Stevenson. Evaluating topic coherence using distributional semantics. in Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)--Long Papers. 2013.
[7] Cao, L. and X. Tang, Topics and trends of the on-line public concerns based on Tianya forum. Journal of Systems Science and Systems Engineering, 2014. 23(2): p. 212-230.