Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Mining topic and sentiment dynamics in physician rating websites during the early wave of the COVID-19 pandemic: Machine learning approach

Adnan Muhammad Shah\textsuperscript{a,}\textsuperscript{*}, Xiangbin Yan\textsuperscript{b}, Abdul Qayyum\textsuperscript{c}, Rizwan Ali Naqvi\textsuperscript{d}, Syed Jamal Shah\textsuperscript{e}

\textsuperscript{a} Department of Management Science and Engineering, School of Management, Harbin Institute of Technology, Harbin, China
\textsuperscript{b} School of Economics and Management, University of Science and Technology, Beijing, China
\textsuperscript{c} Faculty of Management Sciences, Riphah International University, Islamabad, Pakistan
\textsuperscript{d} Department of Unmanned Vehicle Engineering, Sejong University, Seoul, Republic of Korea
\textsuperscript{e} Antai College of Economics and Management, Shanghai Jiao Tong University, China

ARTICLE INFO

Keywords:
Text mining
Topic modeling
COVID-19
LDA
Dynamics of healthcare topics
Discrete emotions

ABSTRACT

Introduction: An increasing number of patients are voicing their opinions and expectations about the quality of care in online forums and on physician rating websites (PRWs). This paper analyzes patient online reviews (PORs) to identify emerging and fading topics and sentiment trends in PRWs during the early stage of the COVID-19 outbreak.

Methods: Text data were collected, including 55,612 PORs of 3430 doctors from three popular PRWs in the United States (RateMDs, HealthGrades, and Vitals) from March 01 to June 27, 2020. An improved latent Dirichlet allocation (LDA)-based topic modeling (topic coherence-based LDA [TCLDA]), manual annotation, and sentiment analysis tool were applied to extract a suitable number of topics, generate corresponding keywords, assign topic names, and determine trends in the extracted topics and specific emotions.

Results: According to the coherence value and manual annotation, the identified taxonomy includes 30 topics across high-rank and low-rank disease categories. The emerging topics in PRWs focus mainly on themes such as treatment experience, policy implementation regarding epidemic control measures, individuals’ attitudes toward the pandemic, and mental health across high-rank diseases. In contrast, the treatment process and experience during COVID-19, awareness and COVID-19 control measures, and COVID-19 deaths, fear, and stress were the most popular themes for low-rank diseases. Panic buying and daily life impact, treatment processes, and bedside manner were the fading themes across high-rank diseases. In contrast, provider attitude toward patients during the pandemic, detection at public transportation, passenger, travel bans and warnings, and materials supplies and society support during COVID-19 were the most fading themes across low-rank diseases. Regarding sentiment analysis, negative emotions (fear, anger, and sadness) prevail during the early wave of the COVID-19.

Conclusion: Mining topic dynamics and sentiment trends in PRWs may provide valuable knowledge of patients’ opinions during the COVID-19 crisis. Policymakers should consider these PORs and develop global healthcare policies and surveillance systems through monitoring PRWs. The findings of this study identify research gaps in the areas of e-health and text mining and offer future research directions.

1. Introduction

The novel coronavirus (COVID-19), which originated from Wuhan, China, has been devastating globally. In the United States (U.S.), thousands of daily new cases have been reported since March 2020. The total number of confirmed reported cases of COVID-19 in the U.S., at this writing (June 27, 2020), exceeded 2,537,636 with 126,203 mortalities, and over 13,039,853 cases with 571,659 mortalities were reported globally [1]. The World Health Organization (WHO) has called for global measures to control the spread of the pandemic. Several countries have recommended strict preventive measures, such as travel bans, lockdowns, quarantine and isolation of clustered cases, frequent hand...

* Corresponding author.
E-mail addresses: adnanshah486@gmail.com (A.M. Shah), xbyan@ustb.edu.cn (X. Yan), abdul.qayyum@riphah.edu.pk (A. Qayyum), rizwanali@sejong.ac.kr (R.A. Naqvi), jimnikakakhail@hotmail.com (S.J. Shah).

https://doi.org/10.1016/j.ijmedinf.2021.104434
Received 21 July 2020; Received in revised form 20 February 2021; Accepted 24 February 2021
Available online 26 February 2021
1386-5056/© 2021 Elsevier B.V. All rights reserved.
wearing, and wearing facemasks in public places. Governments of various countries, including the health officials in the U.S., are now imposing more intensive steps such as social distancing to prevent/discourage gatherings in public areas [2].

With the global spread of the COVID-19 pathogen, peoples’ online activities (i.e., Facebook, Twitter, and YouTube, etc.) began to increase. This increasing trend of using social media could significantly influence individuals’ healthcare behavior. Several studies have shown peoples’ use of online media during the crisis and its impact on their attitudes and behaviors [3,4]. Social media can definitely play a key role in communicating up-to-date policies and regulations from health officials [5]. Therefore, analysis of the online data showing individuals’ behaviors and concerns about healthcare providers during COVID-19 offers unique opportunities to address public concerns and improve the quality of healthcare services [6]. Online users’ publicly available data on social media sites can be used to easily classify key concepts, sentiments, and issues that people worry about regarding the COVID-19 infection [4]. Analysis of online data can also complement conventional surveys and aid the development of efficient health interventions and policies at large [7].

Using natural language processing (NLP) to automatically mine topics and opinions from patient online reviews (PORs) has been a complicated task. Topic modeling is a statistical tool for uncovering abstract topics from a series of documents. Notably, latent Dirichlet allocation (LDA) [8] is a probabilistic statistical model that taxonomically describes its components. Focused on the comparatively simple and vigorous bag-of-words explained as text, it does not take into account the order of words and sentences. The fundamental assumption of this model is a mixture of topic terms [8]. LDA points out those different themes involved in online reviews that may be linked to one of the topics. It has been extensively used to identify topics from unstructured data in different domains, particularly in healthcare [9,10]. Due to the extensive applicability of this topic modeling approach in the biomedical domain, several improvements were made to the standard LDA. Previous work has explained the potential extensions of LDA in the medical domain, with examples that include the extraction of hidden information from electronic health records [11], different patterns mining relating to patients’ health status [12], consumer health expressions mining [13], and analysis of user-generated content (UGC) [14].

While prior research has improved the LDA model for analyzing UGC, there has been little research using it during the crisis because it is difficult to model the semantic interpretation of UGC posted during the crisis. Furthermore, the diversified distribution of topics makes it difficult to extract topics from particular documents. Researchers have conducted several studies on different pandemic crises, such as COVID-19 [3], Zika [15], and Middle East respiratory syndrome (MERS) [16], using different text mining techniques. For example, a study by Liu et al. [4] used a topic modeling approach to extract different topics from Chinese social media. Abd-Alrazaq et al. [5] used LDA algorithms to detect key topics posted by Twitter users related to the Coronavirus infection. Using structural topic modeling and network analysis, Jo et al. [2] also investigated web-based health communication trends. They analyzed what motivated peoples’ anxiety in the early stages of COVID-19 in South Korea. Stokes et al. [17] used LDA to identify peoples’ preferences and concerns regarding COVID-19 in an online discussion forum. Kleinberg et al. [18] used topic modeling to predict peoples’ emotional behaviors during the COVID-19 crisis in the United Kingdom. Using topic modeling, Orduñ [19] identified different topics from COVID-19 tweets. Cho et al. [20] also used LDA to identify the patient safety concerns about healthcare stakeholders during the pandemic period. Several other studies qualitatively analyzed Twitter data during the COVID-19 period in different countries [21-23]. Researchers have also studied prediction [6], public perceptions and behaviors [7,24,3], and emerging research trends [25]. In summary, few studies have focused on improved LDA for analyzing UGC during crises. However, earlier investigations on the improvement of LDA have overlooked the semantic framework, which has strict limitations. The semantics in any particular natural language sentence are named sentence importance, and the semantic unit communicates significance in the context of the sentence. The natural paragraph has unique semantic and pragmatic roles as the organizational structure of the document. Moreover, the content in the natural paragraph overwhelmingly concentrates on a single topic. Previous studies disclosed that semantically analyzing UGC has achieved good results [19,17,26].

In this age of digitization, many healthcare facilities face the need to quickly upgrade their information technology systems to address contemporary challenges regarding patients’ concerns during the COVID-19 crisis. Social media analytic tools constantly monitor changing public sentiments, assess public desires and concerns, and forecast real-time disease patterns [24]. Although previous studies have explored crisis-related public sentiments and discussion topics [16,18], recent investigations have shown that PORs, including specific types of emotions reflected in those reviews, play a crucial part in patients’ decision-making process [24,2]. If the opinions expressed in a review are a combination of discrete positive and negative emotions regarding the evaluation of products or services, the assessment of the products or services becomes perfect. Therefore, the authors intend to provide real insights into the impact of public health communications and assume that specific emotions appearing in the PRWs and their underlying PORs are extremely important during the current COVID-19 crisis.

Learning topic dynamics and sentiment trends from a broad corpus have gained more interest in data mining, as an extensive amount of user feedback is accessible on the Internet. However, little is known about the emerging and fading trends in doctor-patient interactions during the early wave of the current COVID-19 crisis in the U.S. Researchers need to investigate different topic dynamics and sentiment trends in PRWs regarding patients’ experiences with doctors during the COVID-19 outbreak. To cope with this challenge, an improved topic coherence-based LDA (TCLDA) model and sentiment analytic technologies were implemented. Using a mixed-methods approach (TCLDA model, manual annotation, and sentiment analytic technology), this study aims to implicitly discover the domain-specific topics and discrete emotions from a large collection of PORs crawled from three publicly available PRWs in the U.S. The data sources used in the study are different from the general social media sites such as Twitter, Facebook, and Weibo used in previous studies. The data from the PRWs (RateMDs.com, HealthGrades.com, and Vitals.com) provide information about the dynamics of these topics and sentiments to gain insight into the healthcare system during the pandemic. It is important to understand which aspects of doctor-patient interactions are rising or falling in popularity. With the implementation of the improved LDA model, several documents are assigned to the topics, and each topic is a collection of several keywords. The PORs are evaluated using patients’ quality of treatment and other aspects through these topic distributions. Given the large body of studies that apply the same research methods to identify related topics and sentiments from social media, this will help researchers to argue the significance of these topics and sentiment trends from the PORs data. Studying shifts in public opinions using the extracted topics and four fundamental discrete emotions will expose how public perceptions concerning the crisis have shifted over time.

2. Material and methods

The authors’ context is three leading, publicly available, online health rating platforms in the U.S. (RateMDs, HealthGrades, and Vitals). These platforms provide an ideal setting and ample data to explore different topics and sentiment trends regarding patients’ healthcare experiences during the COVID-19 pandemic.

A Python-based web crawler was developed to extract PORs from these platforms. This study’s final dataset includes 55,612 patient reviews of 3430 doctors collected from March 1, 2020 to June 27, 2020.
Ten types of disease specialties were chosen, including five high-rank and five low-rank diseases, based on the mortality rate and disease rank from the U.S. Centers for Disease Control and Prevention [27].

Levels of need and motivation modify a person’s behavior. According to Maslow’s hierarchy of needs theory [28], low-level needs and requirements, for instance, physiological requirements and action should be fulfilled before high-level needs. Once basic consumer needs are met, they move to a higher level of specifications. In healthcare, the patients’ needs and motivation also shift across different disease types [29]. Some physicians must meet low-level needs for low-disease severity, and others must satisfy high-level needs for high-disease severity. Patients’ traits significantly influence their degree of satisfaction with the quality of care. Therefore, the rationale behind splitting the dataset into the physician quality of care and hence expect different levels of service high-rank vs. low-rank disease categories was because patients suffering from low-rank diseases (mild diseases) tend to be more sensitive to the physician quality of care and hence expect different levels of service than those suffering from low-rank diseases (mild diseases). The varying disease conditions significantly affect patients’ health consultation decisions and opinions toward a healthcare provider [29,30]. A data split was performed to identify high-rank (32,441 of total 55,612 reviews) and low-rank (23,171 of total 55,612 reviews) disease categories. The overall proposed methodology is shown in Fig. 1.

2.1. Data analysis

After data collection, several pre-processing steps were performed to clean raw data using Stanford NLP tools. 1) Remove URLs, short words (with less than 3 characters), and stop words (a, an, the, etc.) since they do not provide valuable information for opinion mining. Stop words can be identified by recognizing words that frequently appear in various documents. For example, represent each word as \( w_i \) (i = 1... M), in document \( d \). For each word \( w_i \), calculate its frequency in document \( d \) denoted as \( f_i \). Arrange them in descending order and remove the top 20 % of words. 2) Remove additional letters from words; for instance, the word “Awesome” was substituted for “AwesomEEE.” 3) Delete question words (who, why, whom, etc.) and special characters ($, #, %, etc.) as both do not influence the study analysis. 4) To optimize the review content, delete digits from the dataset. 5) All white spaces and tabs were deleted. 6) All reviews were processed to lower case. 7) Non-standard vocabulary (spelling correction, etc.) was corrected. 8) Stanford Chunker was employed to chunk input reviews. 9) To normalize verb chunks and noun chunks, clause normalization was performed. 10) All the words in the corpus were lemmatized to their base form using the WordNetLemmatizer function and WordNet Python natural language toolkit (NLTK) (http://nltk.org). A step-wise data pre-processing is shown in Fig. 1.

A matrix for document terms was created with the term frequency-inverse document frequency (TF-IDF). TF-IDF is used to highlight the significance of a word in a document [31]. To extract an appropriate LDA topic number, a mixed-methods investigation was performed, including an improved LDA (i.e., TCLDA).

Topic modeling (LDA) is a complex text-mining approach, perfect for this study’s research objective, to understand patients’ voices by identifying emerging and fading topics on the RateMDs, HealthGrades, and Vitals rating platforms. However, the drawback of conventional LDA is that it does not take into account any of the semantic characteristics embedded within the document content. Rather than using conventional LDA to model topics at the document level, it is preferable to refine the document into various semantic topic units, which is especially appropriate for modeling text reviews. This study emphasized the LDA model [8] and its improved version (i.e., TCLDA), which has become a benchmark model for unsupervised topic identification in a text document. TCLDA was used to analyze PORs by semantically identifying different topic trends on the three different U.S.-based platforms. The proposed version is an improvement over the traditional LDA models, which explicitly analyzed the patients’ posted comments [10,17].

Let \( \alpha \) and \( \beta \) reflect the Dirichlet’s parameter prior to the per-document topic distribution and per-topic word distribution, respectively. \( \theta^{(d)} \) denotes the topic distribution for document \( d \) (e.g., a POR). \( \phi^{(z)} \) indicates the word representation for topic \( z \). \( w \) is the word, and \( D \) stands for the number of documents. \( w_{d|d} \) represents the number of words in a document \( d \), and \( T \) signifies the number of topics. The LDA corpus-generation model assumes the following mechanism for a document \( d=(w_1,\ldots,w_{d|d}) \) including \( N_d \) words from a dictionary involving \( V \) different terms, where \( w_i \) is the \( i \)th word for all \( i = 1,\ldots,N_d \) [32]. The following are the three stages of the LDA model:

1) Choose \( \phi^{(z)} \sim Dir (\beta) \).
2) Choose \( \theta^{(d)} \sim Dir (\alpha) \).
3) For each of the \( N_d \) words:
   4) Choose a topic \( z \sim Multinomial (\phi^{(z)}) \).
   5) Choose a word \( w_i \sim Multinomial \) topic \( z \): \( P(w_{i} \mid z; \phi^{(z)}) \).

Fig. 1. Overall proposed methodology.
The topics in the review text corpus can be explored by Gibbs sampling methods [33] that assign values to the unknown parameters $\theta$ and $\phi$. The coherence score was used to determine the collection of the appropriate number of topics [34]. Topic coherence estimates the uniformity of a particular topic by calculating the semantic similarity between high-score words in a topic that improve the semantic understanding of the topic [35]. For instance, words are denoted as vectors in vector space modeling, and cosine similarity is calculated between word vectors. The coherence is the average of cosine similarities [36]. The coherence model from the Gensim Python library [37] was used to compute the coherence value. The coherence score increases significantly up to 15 topics, becomes stable at 20 topics, and then trends downward after stabilization (see Fig. 2). The lines reflect the mean coherence values of three runs in which the number of topics ranged from 1 to 30. It has been noticed that it is difficult for researchers to interpret the results if only statistical methods are applied [38]. Therefore, statistical methods and manual annotations were integrated and 15 topics were selected from each disease category using Python 3.6 and LDAvis tool (pyLDAvis) [39] (the hyper-parameter and value of the mixture weight $\lambda$ is the probability of a word being generated by the latent feature topic-to-word model $0 \leq \lambda \leq 1$ and fix 30 topics with their keywords). Labels were assigned to the corresponding keywords to elaborate on the topics. The visualization of 30 topics was represented as circles in Section 3.2. The blue color reflects the topics, while the red color represents highlighting a particular topic (Topic 1 in this study). The size of the circle represents the percentage of tokens included in the corpus of the topic. These circles overlapped, and their centers are defined based on the computed topics distance [39]. LDAvis leverages the R language [40], and in particular, the shiny package [41], not just its browser-interaction functionality with $D^3$ [42], but allows users to quickly alter the topical distance calculation and multidimensional scaling algorithm to construct a global topic view.

2.2. Dynamics of topics in physician-rating websites

The authors’ proposed method extracts a set of topics from PORs. To gain more insight into these topics, it is important to find the dynamics of these topics across different disease categories during the early wave of the COVID-19 outbreak, more specifically, which topics of discussion are emerging (growing) or fading (declining) in popularity. Following the previous topic trends identification studies [43,44], a topic is categorized as an emerging topic if it is a heavily (popular) discussed topic over time. Moreover, emerging features, such as terms or phrases, are typically used to detect emerging topics. An emerging topic should therefore have a high degree of novelty and a low degree of fading. In PRWs, the topic typically exhibits a life cycle that arises at some point in time and builds to a climax while attracting more and more interest, leans towards fading, and eventually disappears or restarts. In contrast, a topic demonstrates a fading trend (less popular) if it was marginal earlier discussion but attracts less attention [43]. In general, a fading trend has less novelty value. Fig. 3 illustrates the topic life cycle.

Thus, topics of interest for patients can be analyzed based on the probability model [43]. In this research, the authors propose a novel probability and a fading probability for the collective identification of emerging and fading topics from latent topics created by a topic model. The emerging topics and fading topics were calculated along with the probability of each topic within each document, and the total probability was recorded. Topic trends were produced by adding the probability of topics for a given document and then dividing them by the total number of documents in a corpus. A higher topic strength denotes increasing popularity as an emerging topic, while a lower topic strength represents decreasing popularity.

The value of $\theta$ can be obtained from the prior discussion. The meaning of $\theta$ in a matrix for document terms is presented below:

$$
\theta_{D,k} = \begin{bmatrix}
\theta_{D_1,0} & \theta_{D_1,1} & \cdots & \theta_{D_1,k} \\
\theta_{D_2,0} & \theta_{D_2,1} & \cdots & \theta_{D_2,k} \\
\vdots & \vdots & \ddots & \vdots \\
\theta_{D_D,0} & \theta_{D_D,1} & \cdots & \theta_{D_D,k}
\end{bmatrix}
$$

Hence, column entries represent topics and row entries represent documents. The value of $\theta_{D,k}$ indicates the probability of topic $k$ given document $D$. The emerging and fading topics are defined based on topic strength (i.e., aggregate probability of topic $k$ in document $D$) as given below:

$$
\hat{\theta}_k = Ex = \frac{1}{D} \sum_{d=1}^{D} \theta_{D,k}
$$

Where $\hat{\theta}_k$ denotes the topic strength of topic $k$ and $D$ represents the document dataset.
As emerging and fading topics are based on temporal changes of topics, the dynamics of the research topics (i.e., emerging and fading topics) may be learnt using the following method:

1) Classify documents by time. The authors use week as the time unit in this paper. Suppose the document collection for week \( W_i \) is \( C_i \).

2) The probability \( \hat{\theta}_k \) of each ith-document allocated to topic \( k \), according to \( \theta \), is obtained.

3) Keep the topic strength \( \hat{\theta}_k \) as follows:

\[
\hat{\theta}_k = E_x = \frac{1}{|C_i|} \sum_{d \in C_i} \theta_k^d
\]

4) With the graphical illustration of the kth topic strength \( \hat{\theta}_k \) in week \( W_i \), the dynamics of the topic \( k \) can be seen as emerging and fading topics.

The entire process of the TCLDA algorithm for topic extraction and mining topic dynamics is defined in Appendix A.

2.3. Human annotation and testing procedure

While efforts have been made to label the topics automatically, they are imprecise and do not ensure coherence [45]. To produce meaningful topic names, each topic was labeled manually, despite the fact that it was more labor-intensive. After several rounds of discussion, topics were named following the top 10 keywords and the annotators’ consensus decisions. The detailed annotation and testing process is as follows.

In order to conduct an extrinsic assessment, the LDA-based topics were compared with the ground truth (actual topic). The authors’ trained LDA used the PORs in the corpus (80% training data and 20% testing data) to label certain reviews. The ground truth was defined by using human annotators, who manually categorized reviews from the corpus (1500 reviews in total) into a topic. The annotators were experts in the information systems domain who were familiar with the list of identified topics. They were asked to categorize each review into the list of topics assigned to them. Since three annotators coded each review, the disagreements between the three coders were addressed until all disputes had been resolved. The inter-rater agreement between any two coders was calculated using Cohen’s kappa coefficient, ranging from 0.70 to 0.84, indicating a good agreement. Therefore, the manually annotated reviews that all of the annotators agreed upon were considered the ground truth. The underlying qualitative approach of manual annotation was based on a wide variety of in-depth interpretations from human coders, which facilitated inductive, exploratory research, and the implications of theoretical methods [46]. To assess performance in a quantitative manner, topic-wise accuracy, precision, and F1 measures were calculated, which are the most common assessment metrics used in text classification. After checking TCLDA’s output on training and testing data, the trained model was used to label each review to its most probable topic.

2.4. PORs overall sentiment dynamics

Sentiment analysis is a computational and NLP-based method that focused on the analysis of people’s thoughts, emotions, and behaviors in a given text [47], and it is an important technique in social media analytics research. In this study, the authors used a machine-learning model to predict the public emotional response in PORs. This model categorized each POR into four discrete emotions in Plutchik’s wheel of emotions [48]. For a given POR, the model returned one emotion from the group of four discrete emotions, specifically, fear, anger, sadness, or joy. Following Plutchik’s wheel of emotions [48], fear-anger and sadness-joy are viewed as opposite emotions. Fear is an awful feeling that usually triggers threats or uncertainties created by conditions, whereas anger arises from ambiguities caused by others. Sadness is seen as a negative emotion usually encountered after traumatic situations beyond one’s control, and joy is a happy feeling resulting from enjoyable events that are considered assured and under control [49].

This research shows the emotional reactions of PRWs users toward the pandemic. Trends for the four discrete emotions and descriptions underlying these emotions were analyzed. The emotional connotations behind the PORs were analyzed using a sentiment analytic technology called CrystalFeel, which offered proven accuracy in recent work [50]. CrystalFeel is a machine learning technique that uses parts-of-speech, n-grams, word embedding, multiple affective lexicons, and original in-house built EI Lexicons to predict the degree of strength associated with fear, anger, sadness, and joy in the PORs. For the approach, the authors used the CrystalFeel algorithm’s emotional strength scores (quantitative values) and transformed them into labels (qualitative values) for a more precise understanding, where “fear,” “anger,” “sadness,” and “joy,” were labeled as each POR’s dominant emotion (see Algorithm 1). Pearson correlations \( r \) between emotions and weekly time periods were performed to statistically show emotional patterns over time.

Algorithm 1: Mining sentiment trends during the early wave of the pandemic

---

**Input:** Collection of online reviews from three physician rating websites;  
**Output:** Sentiment trends on physician rating websites during the study period;  
**Process:**

- emotion_class = “no particular emotion”;
- if (emotion_score > 0.52):
  - print “The public emotional trend toward COVID-19 is:”
  - print emotion_class = “joy”;
- elif (emotion_score < 0.48):
  - print “The public emotional trend toward COVID-19 is:”
  - print emotion_class = “anger”;
- if ((fear_score > anger_score) and (fear_score > sadness_score)):
  - print “The public emotional trend toward COVID-19 is:”
  - print emotion_class = “fear”;
- elif ((sadness_score > anger_score) and (sadness_score > fear_score)):
  - print “The public emotional trend toward COVID-19 is:”
  - print emotion_class = “sadness”;
- end
3. Results and discussion

3.1. Descriptive statistics

Fig. 4 shows the relationship between the number of new confirmed COVID-19 cases, deaths, and online reviews during the study period. Statistics regarding the increasing number of deaths and confirmed cases were collected from the WHO [1]. All three measures were negligible in the first week of the study period (i.e., March 7, 2020) because the virus started to spread widely in mid-March. Therefore, there was a sudden increase in these measures since the third week of the study period (i.e., March 15, 2020) because there was no lockdown before that date, and people were less careful about social distancing and wearing masks in public. However, these measures were dropped in May (week 10) due to the strict lockdown and travel bans. During the George Floyd protests in June 2020 (week 14 and onwards), these three measures surged again following Floyd’s murder by Minneapolis police. An estimated 15–26 million people participated in the “2020 Black Lives Matter” protests in the U.S., making Black Lives Matter one of the largest movements in U.S. history. During the protest, people did not care about social distancing and wearing masks in public places, and the number of infected cases increased sharply. As a result, the discussion on PRWs (PORs posting) regarding different aspects of care increased significantly during this phase of the COVID-19 crisis.

3.2. COVID-19-related topics, themes, and their dynamics

The COVID-19 pandemic has received considerable global attention. Topic modeling provides insights into what people discussed more or less frequently regarding their treatment experience with doctors and other aspects during the COVID-19 crisis.

Fig. 5 presents the topic model design, representing 30 different
includes the largest proportion of all topics. Topic 1 was chosen, and the optimistic and enthusiastic predictions that might eradicate needless these topics. Active medical strategies could be used to disseminate to interact with people utilizing Internet media during the pandemic; reflected the policy implementation regarding the epidemic control (8,441/32,441 reviews) of the high-rank disease corpus. These topics were mostly related to the treatment experience theme.

Fig. 6 is an example that presents the top ten words for topic 1, which includes the largest proportion of all topics. Topic 1 was chosen, and the system visualized the distribution of word frequencies compared to the entire corpus. Each bar displays the total frequency of each keyword and the approximate frequency in topic 1. In topic 1, the PORs mainly discussed doctor competence regarding infection diagnosis, and they mentioned virus symptoms, professional, experienced, and skills most frequently. Then, the topic label was assigned according to the content. This approach has been well demonstrated by previous studies [51,4].

Table 1 shows the top keywords for 15 topics across high-rank diseases. These keywords were based on the probability distribution against each topic’s keywords. This process is reported in previous literature [52]. There were a total of 11 emerging topics and 4 fading topics across high-rank diseases. Considering emerging topics, doctor competence regarding infection diagnosis, treatment/operational process and research, disease diagnosis and prevention, friendly staff, emergency services and trauma center, and hospital cafeteria service-scape represented a total of 50.44% (16,364/32,441 reviews) of the high-rank disease corpus. These topics were mostly related to the treatment experience theme.

Regarding emerging topics, topics such as prevention and control procedures and vaccine development represented a total of 26.01% (8,441/32,441 reviews) of the high-rank disease corpus. These topics reflected the policy implementation regarding the epidemic control measures theme. It is critical for health professionals and policymakers to interact with people utilizing Internet media during the pandemic; therefore, the executives of key management agencies, medical facilities, and methods of community regulation were also emphasized in these topics. Active medical strategies could be used to disseminate optimistic and enthusiastic predictions that might eradicate needless patient fears and ensure trust in the country’s health officials to control the virus. Two other topics, emotions toward the epidemic and the sense of the pandemic, represented a total of 10.56% (3426/32,441 reviews) of the high-rank disease corpus. These topics indicated the individuals’ attitudes toward the pandemic. Finally, the topic related to mental health represented a total of 4.56% (1479/32,441 reviews) of the high-rank disease corpus. Previous research found that both medical workers and citizens under quarantine had elevated mental health issues [53] in earlier pandemics. Such investigations can enable readers to re-center this easily overlooked domain, and therefore early intrusion can be made based on this emerging sociological development.

In contrast, fading topics deal with panic buying and daily life impact and represented a total of 2.21% (718/32,441 reviews) of the high-rank disease corpus. These details showed that the medical and social sectors were adversely affected by the COVID-19 crisis. It is also a big challenge for public health that needs worldwide collaborations. It was also observed that across high-rank diseases, topics such as communication (listen and explain), appointment process, and doctor value were less discussed on the three platforms, representing a total of 6.76% (2194/32,441 reviews) of the high-rank disease corpus. These three topics lie under the common theme of treatment process and bedside manner. The topic related to the appointment process included keywords representing the secondary impact of COVID, such as longer waiting time and closure/cancellation of a doctor’s appointment.

Table 2 presents the 11 emerging topics and 4 fading topics across low-rank diseases. Emerging topics that were converged under the treatment process and experience theme included topics such as convenient hospital location, uncomfortable environment, physician knowledge and confidence, treatment/operational process, disease diagnosis, unfamiliar and non-cooperative staff, medical examination for COVID-19, and treatment cost, representing a total of 59.06% (13687/23,171) of the low-rank disease corpus. The findings showed that PRWs transmitted this sort of health information by concentrating on identifying suspect cases and medication that could cure the patient and spread the virus. The topic related to treatment cost indicated that the ongoing COVID-19 pandemic has greatly influenced peoples’ economic conditions.

The topics under the theme for awareness and COVID-19 control measures included confirmed cases and public health measures and represented 25.47% (5903/23,171) of the low-rank disease corpus. The COVID-19 deaths, fear, and stress topic represented a total of 10.65% (2469/23,171) of the low-rank disease corpus. This suggests that PRWs have assisted public health utility because of their evolving rates, and the number of cases gives consumers intuitive feelings about the pace, momentum, and risk of viral spread. It also allows users to remain alert to viral transmission, avoid the virus, and alter their day-to-day activities.

For fading topics, the theme healthcare provider attitude toward patients during pandemic included two topics, patient visit process and medical ethics (relational conduct), which represented a total of 2.66% (617/23,171) of the low-rank disease corpus. The other two individual topics were the detection at public transportation, passenger, travel bans and warnings and materials supplies and society support during COVID-19, which represented a total of 1.18% (273/23,171) and 0.98% (227/23,171) of the low-rank disease corpus, respectively. To deal with COVID-19, the U.S. government has taken steps on the basis of these
Table 1
Emerging and fading topics in patient online reviews across high-rank diseases.

| Nature of topics | Topic number | Topic labels | Keywords (Probability) |
|------------------|--------------|--------------|------------------------|
| Emerging Topics  | Topic #1     | Doctor competence regarding infection diagnosis | Virus symptoms (.362), professional (.130), experienced (.094), skills (.088), intelligent (.074), honest (.091), symptoms (.04), breath shortness (.045), answering questions (.055), dizziness (.051), infection (.280), treatment (.091), diagnosis (.091), isolation (.087), medication (.071), spread (.065), changed the life (.061), medical procedures (.059), COVID-19 vaccine (.055), laboratory research (.041), Test shortage (.055), racist jokes (.021), making fun (.103), compensation (.076), loved one (.022), lost job (.019), afraid (.011), great job (.009), stay safe (.008), new wave (.006). |
|                  | Topic #2     | Treatment/operational process and research | Fight against the outbreak (.279), meeting (.145), fight (.087), work regulation (.081), community (.075), help (.073), opportunity (.071), service (.067), accurate diagnosis (.063), serious and responsible (.059). |
|                  | Topic #3     | Emotions toward epidemic | Rude (.37), depressed (.105), long (.087), wait (.079), mental (.071), damage (.070), nationwide (.068), less time (.066), society (.063), long (.021). |
|                  | Topic #4     | Prevention and control procedures | Cough (.189), bowed (.185), blood (.187), spread (.061), control (.059), pandemic (.058), official (.055), CDC (.051), warn (.048), report (.035). |
|                  | Topic #5     | Mental health | Problem diagnosis (.432), clinical trials (.106), phase 3 (.071), recovery (.069), immunity (.066), side effects (.065), policy (.063), infection (.045), symptoms (.044), dose-escalation (.039), Helpful (.378), friendly (.121), offer sanitizer (.091), wonderful (.089), social distance (.081), bedside manner (.075), frontline (.056), soap (.053), professional staff (.031), great staff (.025). |
|                  | Topic #6     | Disease diagnosis and prevention | 24 h (.255), quarantine (.125), ventilator (.088), pain reduction (.085), outbreak (.081), surgery (.078), reduced strain (.075), trauma (.073), special facilities (.071), specialized equipment (.069). |
|                  | Topic #7     | Vaccine development | Hygienic (.519), clean (.068), clean table (.061), delicious food (.058), covered (.055), filter (.053), water (.050), quick service (.048), fresh principles. Detecting viral infections in public conveyance networks provoked great public concern. Since the outbreak was sudden and the transmission is rapid, people in affected areas required medical materials and other necessities. Hence, public agencies should maintain material supplies and raise donations and funding to protect people from the pathogen.

As shown in Table 3, the high values of topic-wise accuracy, precision, and F1 scores suggest that a majority of the reviews can be correctly labeled on their most probable topic by using the TCLDA model.

As described in Section 2.2, the authors used the probability model and topic-strength-based approach to the PORs dataset to examine the dynamics of the topics learned from the study’s method from March 1, 2020 to June 27, 2020. Figs. 7 and 8 portray the dynamics of 30 topics (15 topics within each disease category having 10 most probable keywords in those topics as shown in Tables 1 and 2). For the high-rank disease category in Fig. 7, it can be observed that most of the topics are comparatively steady. For example, topic 6 (disease diagnosis and prevention), topic 7 (vaccine development), topic 8 (friendly staff), topic 9 (emergency services and trauma center), topic 10 (hospital cafeteria servicescape), topic 11 (sense of pandemic) and topic 15 (doctor value). Some topics indicated increasing trends as emerging topics, including topic 1 (doctor competence regarding infection diagnosis), topic 2 (treatment/operational process and research), topic 3 (emotions toward epidemic), topic 4 (prevention and control procedures), and topic 5 (mental health). Other topics, such as topic 12 panic buying, topic 13 (communication and explain), and topic 14 (appointment process), reflected decreasing trends as fading topics.

In contrast, for the low-rank disease category in Fig. 8, topics such as topic 1 (convenient hospital location), topic 2 (uncomfortable environment), topic 3 (physician knowledge and confidence), topic 4
Emerging and fading topics in patient online reviews across low-rank diseases.

| Nature of topics | Topic number | Topic labels | Keywords (Probability) |
|------------------|--------------|--------------|------------------------|
| Emerging Topics  |              |              |                        |
| Topic #1         | Convenient hospital location | Good location (.754), accessible (.055), transport (.051), perfect (.045), convenient (.031), ease (.026), center (.025), travel restrictions (.021), available (.010), hospital location (.005). |                        |
| Topic #2         | Uncomfortable environment | Noisy environment (.697), dirty rooms (.103), rush (.045), hot (.032), corridors (.030), sitting area (.028), disinfection (.025), loud (.023), cry (.010), messy wards (.007). |                        |
| Topic #3         | Physician knowledge and confidence | Hope (.411), noble (.123), trust (.092), confidence (.081), better (.075), best (.069), think (.061), assurance (.041), happy (.032), confirmed cases (.015). |                        |
| Topic #4         | Treatment/ operational process | Great treated (.432), diagnosis (.152), medication (.131), diagnosed with ease (.061), blood test (.058), antibiotics (.055), COVID-test (.041), mask (.032), hospitalization (.023), laboratory report (.015). |                        |
| Topic #5         | COVID-19 deaths, fear, and stress | Horror (.087), pain (.125), serious (.036), danger (.033), diarrhea (.027), blood (.024), kill (.021), died (.018), tension (.016), strain (.013). |                        |
| Fading Topics    |              |              |                        |
| Topic #6         | Confirmed cases | Confirmed (.487), patient (.156), pneumonia (.092), novel (.088), infection (.071), new type (.066), novel coronavirus (.031), deadly (.006), infection (.003). |                        |
| Topic #7         | Disease diagnosis | Treatment (.369), blood sugar (.145), blood pressure (.123), control (.111), insulin (.054), vaccine (.051), drug (.043), stable (.042), frequent urination (.033), weight loss (.029), hands (.210), wash (.198), soap (.101), away (.095), nh95 (.091), face (.085), mask (.081), sanitizer (.035), gloves (.018), safe (.008). |                        |
| Topic #8         | Public health measures | Staff (.256), unfriendly (.130), rude (.125), front desk staff (.121), call (.091), unhelpful (.085), ignore (.081), loud (.076), twice (.051), bother (.031). |                        |
| Topic #9         | Unfriendly and non-cooperative staff | Operation (.32), X-ray (.191), medical examination (.082), n95 (.080), clear report (.075), prescription (.073), protection (.071), laboratory report (.063), reexamine (.051), medical supply (.045), pneumonia (.041), blood test (.031). |                        |
| Topic #10        | Medical examination for COVID-19 | Expenses (.628), money (.195), registration fee (.045), waste money (.031), expensive (.026), travel bans (.023), affordability (.021), difference (.016), insurance (.012), credit (.003). |                        |

The main difference between the topic dynamics across high-rank and low-rank disease categories might be their connection with mortality. As the disease-rank is linked to physical factors (health condition) and physiological factors (pain and worry), a patient suffering from a high-rank disease spends more effort to get high-quality services from a competent doctor [29]. For instance, a patient with heart disease feels more pain and anxiety than a patient who just has influenza or a cold. As a result, they are keen on higher-quality medical services to deal effectively with physical and physiological factors [30]. As such, most emerging topics in the high-rank disease category focused on technical aspects of care, whereas most emerging topics in the low-rank disease category focused on the treatment process during COVID-19. However, fading topics for both disease categories concentrated on a different aspect of care by showing less interest in these aspects on the three PRWs.

Overall, it is interesting to see that the similarities and differences among emerging and fading topics from both disease categories are quite close to the reality of the U.S. healthcare system, which indicates that the online feedback does reflect real patient experience and genuineness of the healthcare system. Therefore, the methodology used in
this study can assist healthcare providers and hospital officials in better understanding and improving the healthcare service delivery quality by mining patients’ voices regarding a healthcare system.

3.3. Overall sentiment dynamics

PORs disclose information about individuals’ perceptions and emotions. Fig. 9 shows individuals’ emotional reaction trends to the COVID-19 crisis. It reflected the overall weekly sentiment distribution of PORs in PRWs during the study period. The findings indicated a Pearson correlation coefficient (r) value of 0.821 between the inferred sentiment strength with out-of-training sample of human annotations and of 0.711, 0.740, 0.703, and 0.716 on emotional strength in forecasting fear, anger, sadness, and joy [50]. It can also be observed that fear (blue line in
ment measures and spread, indicated by words such as “confirmed case,” “wave,” and “outbreak.” However, as the pandemic intensified, the narratives showed fear of medical and food supply shortages, articulated with the terms “supply shortage” and “sold out.”

The status of fear fell to less than one-quarter of weekly PORs from early to mid-April as the crisis developed ($r_{wp} = -0.89; P < .001$). On the other hand, PORs exhibiting anger (the second most dominant emotion represented as a brown line in Fig. 9) gradually increased from the beginning to the end of March, peaking at 27% on March 12, the day after the WHO Pandemic Declaration. Since then, the PORs with anger decreased marginally but persisted at a very high level ($r_{wp} = 0.72; P < .001$). PORs exhibiting fear and anger dramatically decreased following the declaration of the pandemic, while PORs displaying sadness (the third dominant emotion represented as green line in Fig. 9) were comparatively lower than those of other emotions but doubled after the WHO announcement ($r_{wp} = 0.85; P < .001$). The most frequent keywords regarding anger included “racist jokes” and “making fun.” Anger then moved to the conversation about the exhaustion of loneliness (sadness) that could arise because of social isolation and other factors indicated by keywords such as “quarantine,” “isolation,” “lost job,” and “deaths.” Likewise, PORs about joy (a less dominant emotion until March 31, represented as an orange line in Fig. 9), often reflecting a sense of pride, appreciation, hope, and satisfaction [50], also increased ($r_{wp} = 0.83; P < .001$). People around the globe have also seen a parallel increase of a sense of joy, incorporating gratitude, optimism, and a desire for human resilience with words such as “friendly,” “great job,” “compensation,” and “happy.” However, there was a sudden increase in joy after April 1, 2020, as the U.S. Department of Labor announced the Federal Pandemic Unemployment Compensation program of $600 per week to be distributed to individuals collecting Unemployment Compensation.  

3.4. Research strengths

There are several possible contributions of this research. First, the authors did not find any work regarding the most popular and least popular trends during the epidemic crisis [54,16]. This paper investigates different psychological reactions in PRWs as emerging and fading topics and sentiment trends regarding patients’ healthcare experiences. The findings help patients to improve perceptions, attitudes, and awareness of clinicians regarding different medical issues.

Previously, papers explicitly analyzed health-related text to mine patients’ opinions by applying the most commonly used topic modeling techniques (i.e., LDA) [5,20] and sentiment analysis approaches [47]. Generally speaking, this paper methodologically contributes to text mining techniques and offers a mixed-method approach (TCLDA model, manual labeling, and sentiment analysis tool) to more systematically and semantically analyzes domain-specific knowledge regarding COVID-19-related topics and sentiments from PORs across time. The authors assume that this will bring major technical advancement in the field of text mining.

In this study, PORs were mined across high-rank and low-rank diseases. Fear of the pandemic was especially dominant during the early wave of the crisis. The study findings agree with the results of other studies [55,24,3,56], which demonstrate that COVID-19 significantly affects the cognitive behavior of individuals. Opinion mining of the recent crisis-related content helps us to understand how the public feels and what they believe about pandemics. Various stakeholders have different opinions on the standards of treatment that are influenced by their perceptions and aspirations across various disease categories. The study results have implications for health officials regarding the need to take into account the well-being and mental health of the society and those dealing with mental illness [3]. The authors assume that the topic modeling and sentiment mining methodology used for text mining have great potential to obtain useful insights and clues from larger datasets.  

3.5. Practical implications

The findings indicate that PORs represent an emerging online asset that closely resembles the real healthcare system of the U.S. As a result, PORs could assist healthcare providers to better understand patients’ voices in terms of anger and joy to improve the quality of care.

For practice, researchers have shown that, in fact, online crisis management activities increasingly become “synchronized and interlinked” [57] and that crisis response activities in reality and online are becoming increasingly “simultaneous and intertwined” [57]. Social networking sites offer a valuable forum for the public to freely communicate and disseminate information and awareness about public health [58]. Nevertheless, social networks can also be an influential tool, particularly during a public health crisis, to increase patient satisfaction and reduce frustration with healthcare services.

Moreover, five different emerging and fading topics (communication abilities, relational behavior, professional skills and expertise, personal qualities, availability and accessibility) point toward future areas of physician training and the quality-of-service evaluation. To advance understanding of these constructs, healthcare professionals should begin specifying the different aspects of patient satisfaction and dissatisfaction.

Fig. 9. Overall emotion trends during the early wave of the COVID-19 pandemic.
### Summary Points

**Already known**
- Analyzing online data showing individuals’ behaviors and concerns about healthcare providers during the COVID-19 crisis offers unique opportunities to address public concerns and improve the quality of healthcare services.
- There is no systematic way to analyze the emerging and fading trends in doctor-patient interactions during the COVID-19 crisis in the U.S.
- Researchers need to investigate different topic and sentiment dynamics in physician rating websites to classify the key healthcare concepts, dominant opinions, thoughts, and issues.

**What we added**
- An improved topic modeling (TCLDA), manual annotation, and sentiment analytic technology were employed to implicitly discover topics and dominant emotions from a collection of patient online reviews.
- With the implementation of the TCLDA model, several documents were assigned to some topics, and each topic included the collection of several keywords.
- The authors discovered health-related topics as emerging and fading topics and dominant emotions across different disease categories during the early wave of the COVID-19 pandemic.

with doctors [13]. This allows for an improved capacity to identify and target these specific areas of satisfaction and dissatisfaction that warrant improvements in individual healthcare and address important, unmet needs. On the other hand, patients can take advantage of this publicly available online asset to choose a competent doctor who would provide the best medical services for their disease treatment.

Finally, the increasing negative emotions in the form of fear, anger, and sadness must be healed and offset by strategic communications regarding public health that aim to stabilize the psychological well-being of the people [2]. Suppose authorities fail to address public concerns efficiently. In that case, citizens could develop an increasing mistrust regarding handling of the virus that could lead to false information about the disease control and prevention measures taken by healthcare officials [4,5]. In addition, more efforts are required to formulate policies and globally develop disease diagnosis and surveillance systems through the analysis of PORs posted on PRWs. Stronger and more constructive involvement of the public on health social media is required. Governments and health officials can also “listen to patients’ voices” or monitor the PORs from the PRWs, particularly in a time of crisis, to help formulate policies and laws related to public health (e.g., social distancing and quarantine) and supply chains.

### 3.6. Limitations

This paper reveals some limitations. First, the findings of this study are based on the early stage of the COVID-19 pandemic and are therefore limited by the relatively smaller sample size (n = 55,612). Given that text-mining methods are better able to perform with larger datasets, future research can continue data collection after the pandemic in order to capture more novel topics and themes. Second, although the three platforms used for the analysis are the most popular PRWs in the U.S., future research can also investigate these emerging and fading trends in other regions where death and recovery rates were faster than in the U.S., such as the United Kingdom, Russia, Brazil, and India. Third, currently, this study only takes the COVID-19 period as a dataset. A comparison of topics between pre-COVID and the current time would be interesting, considering questions like: “Are topics changing due to COVID-19?” “Are people more positive about their doctors?” “Are there COVID-related topics in PORs?”

Collecting and comparing reviews pre-COVID and during the COVID pandemic may be an interesting avenue for future research. Fourth, the time range adopted in this work was from March 1, 2020 to June 27, 2020. Future research could compare and expand the study to multiple years by performing longitudinal topic modeling, for example, the period of March to June during the years of 2018, 2019, and 2020. There might be significant differences in patient sentiments that can be attributed to the COVID-19 pandemic. Finally, this paper raised an exciting research question regarding the popularity of healthcare topics and sentiments on PRWs. The authors believe that patients’ attention may change due to the COVID-19. Hence, it is important for readers to know what topics and sentiments were important to patients before and after the COVID-19 pandemic. The authors’ future research will address this issue.

### 4. Conclusion

The collection and analysis of PORs illustrates how U.S. patients express their views during the early period of the COVID-19 crisis. To the authors’ knowledge, this was the first study to implicitly mine different topic dynamics (i.e., emerging and fading topics) and sentiment trends (i.e., dominant and less dominant emotions) from PORs across different disease categories. A mixed-methods approach using implicit data analysis (i.e., TCLDA model, manual labeling, and sentiment analytic technology) was used to highlight 22 emerging topics and 8 fading topics as various semantic topic units across high-rank and low-rank diseases. Moreover, negative emotions were dominant during the early wave of the COVID-19 pandemic. In summary, the findings indicated that analysis of PORs using an automated text mining approach could provide exciting insights for patients regarding their opinions toward the quality of care and for practitioners to take several steps in improving the healthcare system.

### Authors’ contribution

Dr. Adnan Muhammad Shah contributed to every part of this research. Prof. Xiangbin Yan was involved in planning and supervising the research. Dr. Qayyum and Dr. Jamal participated in the data acquisition, developing the algorithm and implementation phase. Dr. Rizwan Ali verified the analytical methods and outputs, helped with all the technical details, and proofread the manuscript. All authors have discussed and contributed to the final version of the manuscript.

### Declaration of Competing Interest

The authors report no declarations of interest.
Acknowledgments

This research is supported by the National Natural Science Foundation, People’s Republic of China (No.71531013, 71729001).

Appendix A. An improved LDA topic model-based on topic coherence (TCLDA)
from helper import *
# ! pip install nltk gensim pydavis
# Ignore warnings
import warnings
warnings.filterwarnings('ignore')
# Import python libraries
import nltk
nltk.download('stopwords')
nltk.download('wordnet')
# Result of a topic model
article_topic
# Topics with keywords
topic_term
# Text preprocessing
from nltk.corpus import stopwords
from nltk.stem.wordnet import WordNetLemmatizer
import string
stopwords = set(stopwords.words('english'))
punctuation = set(string.punctuation)
lemmatize = WordNetLemmatizer()
def cleaning(article):
    one = " ".join(([i for i in article.lower().split() if i not in stopwords])
    two = " ".join([i for i in one if i not in punctuation])
    return one

df = pd.read_table('plot.tok.q.txt5.00008', names=['text'])
df.info()
df.head(15)
text = df.applymap(lambda x: 'text')
text_list = [x.split() for x in text]
len(text_list)
text_list[0]
# Add log recording the model fitting data while training
from time import time
import logging
logging.basicConfig(format='%(asctime)s : %(levelname)s : %(message)s',
level=logging.INFO, filename='running.log', filemode='w')
logging.info('running

from gensim import corpora
# Creating the term dictionary of the study corpus, where every unique
term is assigned an index. dictionary = corpora.Dictionary(doc_clean)
dictionary.save('dictionary.dict')
print dictionary
# Build corpus
# Converting list of documents (corpus) into Document Term Matrix using
dictionary prepared above.
doc_term_matrix = [dictionary.doc2bow(doc) for doc in text_list]
corpora.MmCorpus.serialize('corpus.mm', doc_term_matrix)
print len(doc_term_matrix)
print doc_term_matrix[100]
# Running LDA Model
start = time()
# Creating the object for LDA model using gensim library
lda = gensim.models.LdaModel.doc_term_matrix
lda = gensim.models.LdaModel.load('lda.model')
lda.print_topics(num_topics=15, id2word=dictionary, pass_first=True)
for i in lda.print_topics():
    for j in i:
        print j
# Save model for future use
lda.save('model.model')
# load saved model
from gensim.models import LdaModel
loading = LdaModel.load('lda.model')
print(len(loading.print_topics(num_topics=15, id2word=dictionary))

# predicting(classifying) new(existing) docs

# pass token to model
# return the probability of particular topic

# plotting
import pyDAvis.gensim
import gensim
d = gensim.corpora.Dictionary.load('dictionary.dict')
c = gensim.corpora.MmCorpus('corpus.mm')
lda = gensim.models.LdaModel.load('lda.model')
data = pyDAvis.gensim.prepare(lda, c, d)
data
pyDAvis.save_html(data, 'vis.html')
the International Working Conference on Advanced Visual Interfaces, Association for Computing Machinery, Capri Island, Italy, 2012, pp. 74–77, https://doi.org/10.1145/2254556.2254572.

[53] H. Jeong, H.W. Yim, et al., Mental health status of people isolated due to Middle East Respiratory Syndrome, Epidemiol. Health 38 (2016), e2016048, https://doi.org/10.4178/epih.e2016048.

[54] K.-W. Fu, H. Liang, et al., How people react to Zika virus outbreaks on Twitter? A computational content analysis, Am. J. Infect. Control 44 (12) (2016) 1700–1702, https://doi.org/10.1016/j.ajic.2016.04.253.

[55] S. Li, Y. Wang, J. Xue, N. Zhao, T. Zhu, The impact of COVID-19 epidemic declaration on psychological consequences: a study on active Weibo users, Int. J. Environ. Res. Public Health 17 (6) (2020) 2032, https://doi.org/10.3390/ijerph17062032.

[56] J. Xue, J. Chen, et al., Twitter discussions and emotions about the COVID-19 pandemic: machine learning approach, J. Med. Internet Res. 22 (11) (2020), e20550, https://doi.org/10.2196/20550.

[57] S.R. Veil, T. Buehner, M.J. Palenchar, A work-in-process literature review: incorporating social media in risk and crisis communication, J. Contingencies Crisis Manag. 19 (2) (2011) 110–122, https://doi.org/10.1111/j.1468-5973.2011.00699.x.

[58] J.Y. Breland, L.M. Quintiliani, K.L. Schneider, C.N. May, S. Pagoto, Social media as a tool to increase the impact of public health research, Am. J. Public Health 107 (12) (2017) 1890–1891, https://doi.org/10.2105/ajph.2017.304098.