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.

International Journal of Medical Informatics 145 (2021) 104309

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf

Track Iran's national COVID-19 response committee's major concerns using two-stage unsupervised topic modeling

Fatemeh Kaveh-Yazdy *, Sajjad Zarifzadeh

Department of Computer Engineering, Yazd University, Yazd, Iran

ARTICLE INFO

Keywords:
COVID-199
Public health crisis
Government response
News mining
Topic modeling
Topological data analysis
Sentence embedding

ABSTRACT

Background: Since the World Health Organization (WHO) declared the COVID-19 as a Public Health Emergency of International Concern (PHEIC) on January 31, 2020, governments have been enfaced with crisis for timely responses. The efficacy of these responses directly depends on the social behaviors of the target society. People react to these actions with respect to the information they received from different channels, such as news and social networks. Thus, analyzing news demonstrates a brief view of the information users received during the outbreak.

Methods: The raw data used in this study is collected from official news channels of news wires and agencies in Telegram messenger, which exceeds 2,400,000 posts. The posts that are quoted by NCRC's members are collected, cleaned, and divided into sentences. The topic modeling and tracking are utilized in a two-stage framework, which is customized for this problem to separate miscellaneous sentences from those presenting concerns. The first stage is fed with embedding vectors of sentences where they are grouped by the Mapper algorithm. Sentences belonging to singleton nodes are labeled as miscellaneous sentences. The remained sentences are vectorized, adopting Tf-IDF weighting schema in the second stage and topically modeled by the LDA method. Finally, relevant topics are aligned to the list of policies and actions, named topic themes, that are set up by the NCRC.

Results: Our results show that major concerns presented in about half of the sentences are (1) PCR lab. test, diagnosis, and screening, (2) Closure of the education system, and (3) awareness actions about washing hands and facial mask usage. Among the eight themes, intra-provincial travel and traffic restrictions, as well as briefing the national and provincial status, are under-presented. The timeline of concerns annotated by the preventive actions illustrates the changes in concerns addressed by NCRC. This timeline shows that although the announcements and public responses are not lagged behind the events, but cannot be considered as timely. Furthermore, the fluctuating series of concerns reveal that the NCRC has not a long-time response map, and members react to the closest announced policy/act.

Conclusion: The results of our study can be used as a quantitative indicator for evaluating the availability of an on-time public response of Iran's NCRC during the first three months of the outbreak. Moreover, it can be used in comparative studies to investigate the differences between awareness acts in various countries. Results of our customized-design framework showed that about one-third of the discussions of the NCRC's members cover miscellaneous topics that must be removed from the data.

1. Introduction

Late December 2020, several numbers of pneumonia cases of unknown etiology were reported by hospitals in Wuhan, Hubei, China. On December 31, the World Health Organization (WHO) office in China was informed about the cases. WHO published its first situation report on January 21, 2020, and on January 31, declared this disease as Public Health Emergency of International Concern (PHEIC). The disease is caused by a newly discovered coronavirus named Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), and the disease itself is called coronavirus disease 2019 (COVID-19) [1]. The incubation period of COVID-19 can be extended to 14 days, and the median incubation period is 4–5 days [2,3]. Findings of research groups in China showed that the overall death rate is 2–3.2 % as of February 16, 2020 [3,4].

* Corresponding author at: Department of Computer Engineering, Yazd University, Yazd, Iran.
E-mail addresses: fkavehy@stu.yazd.ac.ir (F. Kaveh-Yazdy), szarifzadeh@yazd.ac.ir (S. Zarifzadeh).

https://doi.org/10.1016/j.ijmedinf.2020.104309
Received 23 June 2020; Received in revised form 22 September 2020; Accepted 20 October 2020
Available online 4 November 2020

Although the earlier estimated mortality rate for COVID-19 was less than that of for flu, but the counted number of deaths in the week ending April 21 showed a higher rate of mortality for COVID-19 [5].

The disease spread rapidly to different countries and caused the COVID-19 pandemics on March 1, 2020. The first confirmed cases in Iran was reported on February 19, 2020, and Iran's president established Iran's National Corona Response Committee (NCRC) on February 23. Since February 24, the NCRC's spokesman, Dr. Kianoush Jahanpour MD, has presented a briefing report on the provincial and national levels two times a day and addressed the concerns and policy set up to mitigation policies. In addition, the head of the NCRC, Dr. Saieed Namaki Ph.D., and the vice deputy of NCRC that is Dr. Iraj Harirchi MD have held press conferences and interview sessions to address and de-brief the NCRC's actions and decisions.

In this article, we present the results of our research aimed at analyzing the concerns and policies made by the NCRC. While our targeted issues have been addressed by different sources and members of the NCRC, we decide to collect the required information from news posts. Then, we adopt text mining methods to extract, select, group, and analyze the underlying knowledge.

2. Backgrounds

Utilizing text mining methods to analyze disseminated information about a disease goes back to earlier 2008. Disease-related text mining researches with respect to their application can be divided into four primary groups as follows,

1 Outbreak monitoring and prediction.
2 Infodemic and misinformation detection.
3 Social/public concern detection.
4 Control Disease Centers response analyzing.

The first two groups of researches are targeted disease-related information extraction, and the remaining groups mine the socio-political information reflected the government and in-charge institutes responses and society's reactions to these policies. In the following sub-sections, we introduce these groups and briefly review the recent researches in each area.

Outbreak monitoring and prediction

Outbreak monitoring and prediction are studies under a research field called digital epidemiology [6] and go back to the time when Google released its research project Google Flu Trends (GFT) to predict the number of flu prevalence in 25 countries. GFT service predicted the H1N1 pandemic in 2009. The pandemic as a non-seasonal flu outbreak started in summer, and it was the first critical outbreak after service release. Investigations of Cook et al. [7] showed that GFT's predicted series are highly correlated with the number of prevalence registered by the U.S. Outpatient Influenza-like Illness Surveillance Network (ILINet). However, Nature reported that the predictions of GFT in the 2012–2013 flu season were two times more than the CDC's ILINet numbers [8]. In addition, Salathé [9] indicated that the private ownership of GFT as its Achilles' heel, which means that the research community cannot investigate the data collection and models independently. GFT service was shut down in 2015. Although, many research teams still uses Google Trends to predict flu [10,11], and other infectious diseases, such as dengue fever [12] and Zika [13]. Furthermore, Twitter has been actively used as a source of information required for outbreak detection in recent years [14–16].

By earlier 2020, the focus of the digital epidemiology researches shifts from other diseases to COVID-19 and outbreak detection using indices of search engines and tweets to predict the disease trends. Li et al. [17] used the Baidu search index and Weibo search queries to predict the raw number of COVID-19 lab-confirmed and suspected cases. Their results show that the search index series are highly correlated with the number of prevalence and suspected cases with 8–10 and 5–7 days lag, respectively. Kaveh-Yazdy et al. [18] used query logs of a Persian search engine to study the feasibility of disease dynamics monitoring in Iran. Qin et al. [19] analyzed the social media search index (SMSI) to predict the suspected cases of COVID-19. Their analysis reveals that among the listed keywords for COVID-19, such as dry cough, fever, chest distress, coronavirus, and pneumonia, the numbers of searches for fever and pneumonia are highly correlated with the number of suspected cases. Ayyoubzadeh et al. [20] collected the Google search index of nine keywords and the number of cases of the previous day in Iran to generate a multi-feature time series. They utilize a linear regression model as well as a long short-term memory (LSTM) neural network to predicts the incidence of the next day. Their results show that the root mean square error (RMSE) of the linear regression model is less than the LSTM model; however, the LSTM model extracts the up-down trends better, which is resulted from overfitting. Their results alert that the LSTM is not an appropriate model for short and sparse time series, while it could be reliable in data collected from longer periods.

2.1. Infodemic and misinformation detection

During the 2003 SARS outbreak, the term "Infodemic" was coined from parts of the words "Information" and "Epidemic" [21]. Infodemic addresses the massive amount of misinformation about a disease circulating in media and the web. SARS [22], MERS [23], Zika [23], and Ebola [24] outbreaks drive the waves of misinformation and rumors in the web; however, their infodemics is not comparable to the COVID-19′s one. The enormous number of posts, including misinformation published in social media and newspapers, makes public health experts worry about the quality and validity of the information that would be accessible for the public. A short while after WHO's PHEIC report, the WHO's risk communication team launched a new portal, called WHO Information Network for Epidemics (EPI-WIN) to publish valid and evident-based information about COVID-19 [25].

We should note that misinformation circulated in social media, such as Facebook, Twitter, TikTok, Pinterest, and YouTube can increase the levels of fear and anxiety of society. Furthermore, the rumors can affect the behaviors of the people and their reactions to disease control policies. For example, the CNN anticipation about the Lombardy region lockdown, which was published before the official announcement of the government of Italy, led to overcrowded trains and airports. People escaping from the Lombardy region spread the disease to the other regions and increase contagion [26]. Two important steps in fighting the infodemics are finding the sources and communicating faster than the rumors.

Cinelli et al. [26] collected the COVID-19-related posts from Twitter, Instagram, YouTube, Reddit, and Gab for 45 days using the top keywords of Google Trends. They extracted links to news outlets and analyzed their contents to classify them into two main groups, namely Reliable and Questionable. The Questionable news already includes Conspiracy-Pseudoscience, Pro-Science, and Questionable topics. Their results show that Gab is the environment that is more susceptible to misinformation dissemination. Hua and Shaw [27] collected news posts from Sina Weibo's hot search list, the coronavirus timeline in china compiled by the users of five social networks, and the web usage data log of Mob-Tech research institute. They defined five phases from January 31, 2020, to February 29, 2020, according to the government's responses to the outbreak and analyzed the reactions of the society to official policies during these phases. Their results suggest that the initial delay was justified by restricting regulation and successful communication, big data adoption, and digital technologies. Directives by China's Supreme Court on fake news publishing and the Tencent's website named "Rumors exposed website" are samples of the country-level responses to fight against the rumor and misinformation.

Erku et al. [28] addressed the role of pharmacists in COVID-19 infodemic, providing up-to-date and reliable information to their community via social media platforms. Furthermore, they indicated that

pharmacists must ensure the education and qualified homecare for individuals, suspected patients, and family members in the lockdown period even by referrals for the psychological consultant. Kouzy et al. [29] followed 14 trending English hashtags on Twitter to extract tweets related to COVID-19. They validated the information in tweets with respect to peer-reviewed sources of medical/health resources and studied the dynamics of misinformation spreading on Twitter. According to their investigation, some of the Twitter accounts are more associated with unverifiable information. Furthermore, the proportion of misinformation-included tweets to trusted ones is growing. These phenomena alert the need for early intervention by health agencies, physicians, medical associations, and even by scientific journals.

2.2. Social/public concern detection

Public concerns directly affect the attitudes and behavior of society, enfacing the diseases. Thus, analyzing concerns, anxiety, and fears of society reveals the hidden issues that are able to affect disease control responses. Nelson et al. [30] studied psychological and epidemiological concerns of society through the lens of surveys in social media. They found that the major concerns of society are hand washing, remaining in-home, and practicing social distancing. Their study showed that the level of concern depends on the age group. Issues regarding buying hand sanitizer and food, child-care, and economical consequences of the COVID-19 outbreak are the most common difficulties. Wang et al. [31] investigated the impacts of the COVID-19 outbreak on psychological health factors in Chinese society. Their survey showed that the psychological impacts of the outbreak were rated as moderate to severe by 53.8 % of people. Moreover, 16.5 % of the Chinese reported moderate to severe depressive symptoms; 28.8 % reported moderate to severe anxiety symptoms, and 8.1 % reported moderate to severe stress levels. Findings of Wang et al. [32] showed that the high quality of health information was associated with better mental health outcomes during the outbreak.

Van der Vegt and Kleinberg [33] investigated 5000 pieces of text (half of them are short texts, and the rest are long). In this experiment, male and female participants were invited to express their emotions about the SARS-COV-2 virus. They found that crucial sources of concern and anxiety for women are family and health issues. Moreover, women are more worried, anxious, and sad than men and also angrier than men. On the other hand, the main source of anxiety for men is socio-economic issues. Deng et al. [34] used selected keywords to collect Weibo's posts related to the 2013 H7N9 bird flu outbreak in China. Collected posts are partitioned into sentences, and their Tf-IDF vector representations are clustered using the *k*-means algorithm. In the end, some of the related clusters covering different aspects of the same topic are aggregated together. Tracking the topics evolved in the social media reveals that the first reaction to the first case in Shanghai was shocking, and after spreading the flu to Beijing, sleeping issues became the next primary concern. The third trend was about "Treatment and Precaution," and then the feared society started to post about 2003 "SARS" and the correlation between the new flu and the increasing number of deaths in pigs. Finally, traditional Chinese medicine attracted attention.

Lazard et al. [35] analyzed the tweets of American users during the 2014 Ebola outbreak to identify anxiety and public fears. In addition, they started to collect tweets and retweets of the offices of the Centers for Disease Control (CDC) in federal, state, and local levels. Comparing the CDC responses to the concerns raised by Twitter users shows that the CDC could not cover the lack of certain reliable information about issues such as pathogen, its spread mechanism, and fear of air travel. Lazard et al. concluded that CDC accounts must communicate with users to present guidelines with respect to their priorities. Several research groups began to analyze negative tweets to extract public concerns regarding outbreaks because it seems that concerns are more likely to be expressed in negative tweets. For example, the results of Mamidi et al. [36] showed that Zika abnormalities, neural defects, and symptoms are mostly expressed in negative tweets. Ji et al. [37] analyzed the negative personal tweets to study mental concerns to design a disease tracking framework. Their proposed system, named Epidemic Sentiment Monitoring System (ESMOS), can visualize and group user concerns and track how they evolve during the outbreak. This framework can guide disease control and prevention agencies to determine their communication priorities on-time.

2.3. Control disease centers' response analyzing

Among different aspects of the relation between text mining and public health, government responses and communications have been understudied. This research area tracks the communications of disease control and prevention agencies and their evolution while studying the reactions of the society to the announced policies. Lopez et al. [38] collected a large multi-language dataset of tweets with COVID-related keywords and hashtags to analyze the statistical properties of the data. This dataset can be used in association with a policy timeline to reveal the public reactions to precaution, measures set up, and acts. Chen et al. [39] investigated the Citizenship Engagement through Government Social Media (CEGSM) in China's earlier post-COVID-19 period. Their results suggest that parameters, such as media richness (vs. plain text), the length of the in-media response time, dialogic loop, the topic diversity of the announced news posts, and the polarity of the government response posts (Pos., Neg., and Neut.) are associated with the level of CEGSM. In another study, Liu et al. [40] used the WiseSearch database to collect news about the SARS-COV-2 virus published in China from January 1, 2020, to February 20, 2020. They performed topic modeling analysis to extract primary topics and concerns in the news posts. Their findings demonstrate a vision of the information spreading role of Chinese news media. According to the results, the news media in China lags behind the viral spread. Moreover, the importance of essential knowledge regarding personal health practices, clinical choices, and disease diagnosis are underestimated in news agency communications.

Hale et al. [41] introduced the Oxford COVID-19 Government Response Tracker (OxCGRT) index to track the stringency of government responses to COVID-19 systematically. This index is defined based on the 18 unique indicators (ordinal, numerical, and textual), which are manually collected from the official reports and news from the web. These indicators cover geographical, economical, and political aspects of the evolving responses made by governments. They indicate that the OxCGRT, as a bridging the gap stringency index, can be used to evaluate the efficacy of decisions that have been made or will be made in the future.

3. Data

Telegram is the most popular social media among Iranians in a way that more than 50 million Telegram users are Iranians (more than 56 % of Telegram users) [42]. People use Telegram channels for education, reading news, and its public groups for social discussion, as well as its encrypted private chat for messaging. Thus, we decided to extract news posts from the official channels of the most popular news agencies, as well as private popular news re-publishing channels to analyze the evolution of the government responses. The list of news sources includes 20 telegram channels which are divided into three groups, as follows[1] ;

- Official Telegram channels of top 10 news agencies in Alexa 50 top sites in Iran.
- Official channels of top 8 news agencies selected from Buzdid.ir as a news-aggregator website.
- Two private Persian news channels with more than 900 K followers.

[1] List of news channels and number of posts are covered in Appendix A.

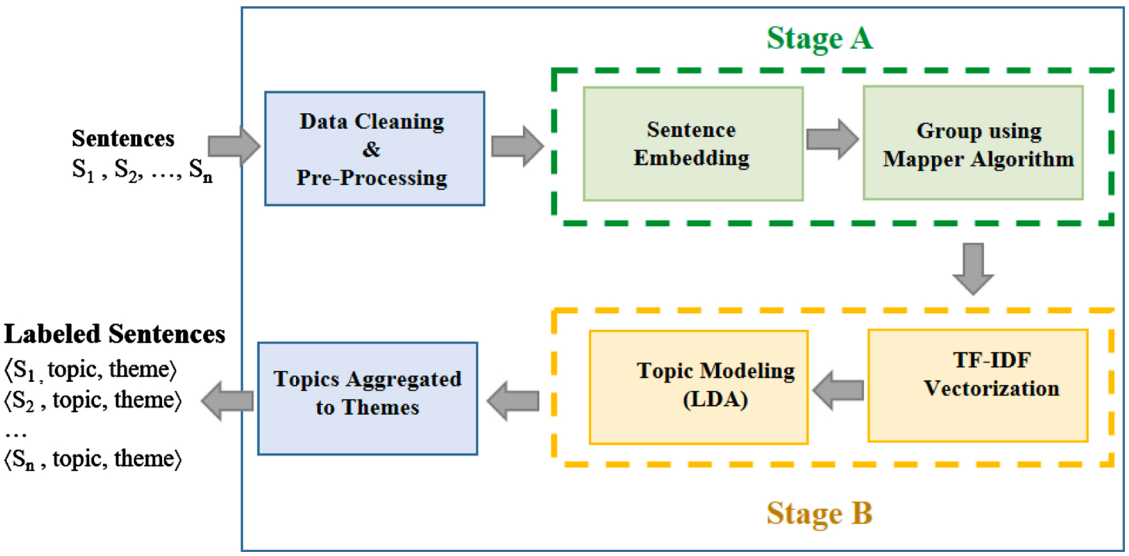

Fig. 1. The schematic view of the concern extraction framework.

All of the news posts from February 1, 2020 (00:00') to May 10, 2020 (23:59') are collected and parsed to extract the textual content, publish date-time, and channel IDs. Multimedia contents, such as photo, video, GIF files, hashtags, channel IDs, and Emoji symbols are omitted from the posts. Empty posts, i.e., posts including only multimedia files, are removed from the dataset. The total number of collected posts is 2,411,761, and the filtered dataset contains 198,824 posts. The last step in data pre-processing is character unification. In this step, Arabic characters and Latin punctuation marks are replaced by their corresponding Persian characters.

Posts quoted at least one sentence from speeches and press conferences delivered by the head, the vice deputy, and the spokesman of Iran's NCRC are selected. In the next step, the posts are split into sentences to generate the sentence dataset. The dataset contains 45,209 sentences that are likely to express concerns, actions, and decisions of NCRC.

4. Materials and methods

The prerequisite of extracting the government or public concerns is topic modeling and tracking. It can be defined as the process of extracting the abstract topics underlying the contents of documents. The first generation of topic modeling methods included statistical techniques such as Latent Semantic Indexing (LSI) [43], Probabilistic Latent Semantic Analysis (PLSA) [44], and Latent Dirichlet Allocation (LDA) [44]. The parameter set and structure of LDA make it ideal for different optimizations and improvements. There are many variations of the LDA model, such as Correlated Topic Model (CTM) [45], Hierarchical LDA (HLDA) [45], supervised LDA (sLDA) [46], relational Topic Model (rTM) [47], and Markov Topic Model (MTM) [48]. While word embedding models evolve the language processing, the LDA leverages the rich representation provided by such models. Word embedding-based LDA with Gaussian mixture model [49], LDA2Vec [50], and word embedding augmented LDA [51] are samples of word embedding-based topic modeling techniques. Their successful records motivated researchers to adopt them in health concern tracking as well. Lazard et al. [35] and Kim et al. [52] utilized the LDA-based public concern tracking in the 2014–2015 season Ebola outbreak, and Glowacki et al. [53] analyzed the CDC's communications in response to the public concerns after the spreading of Zika in Florida. The rapid rate of COVID-19 spread alerts the global community to respond promptly to medical and mental concerns to accelerate the action of disease control [39]. In this way, Dong et al. [54], Stokes et al. [55], and Liu et al. [40] employed topic modeling techniques to detect public concerns and health communications.

4.1. The concern detection framework

In this article, a two-stage framework is devised to extract the major concerns of the NCRC. This framework receives the pre-processed/cleaned dataset and analyzes it in two stages. Each row of the input data includes news agency, date, and a sentence that probably expresses NCRC's concerns. In the first stage (i.e. stage A), sentences are vectorized by a sentence embedding, and a group of similar sentences covering the NCRC's concerns are selected to be re-processed in the next stage. The second stage (i.e. stage B) has the same structure in which a data vectorization method generates sentence representations, and then, the underlying topics are extracted using the LDA. The schematic view of the proposed framework is illustrated in Fig. 1.

Although the selected posts include quotes made by the members of the NCRC, additional processes are required to extract sentences that really indicate concerns of the NCRC. During the press conferences, the spokesman and other members of the committee address different issues related to the COVID-19 pandemics, such as expressing condolences to other countries and medical aids NCRC received from the health department of the Ministry of Defense. Due to the term similarity of the sentences expressing concerns and those related to peripheral issues, we design a two-stage framework. The first stage divides the domain into two sets: one extensive set of the concern-addressing sentences, and the remaining sentences. Sentences of the first set are re-vectorized in a way that they express their underlying topics with associated words. The second stage generates human-understandable topics. Finally, the topics are re-considered by a human expert to be aggregated with respect to the list of concerns. The topic aggregation is popular for aligning the detected topics with the enlisted response concerns (similar to [40,55]).

We should here explain about the order of methods used in our framework. To shed light on the topic modeling issues that occurred in this problem, we present an example we found during our investigations. Suppose that sentences are vectorized using two different schemas, i.e., Tf-IDF[2] and sentence embedding in parallel. Then, the sentences addressing an issue related to "Hospitals and Medical Centers" and "Closure of Schools/Universities" are labeled. The vectors are embedded in a 2D space using a parametric dimension reduction

[2] The popular vector representation mostly used with LDA topic modeling.

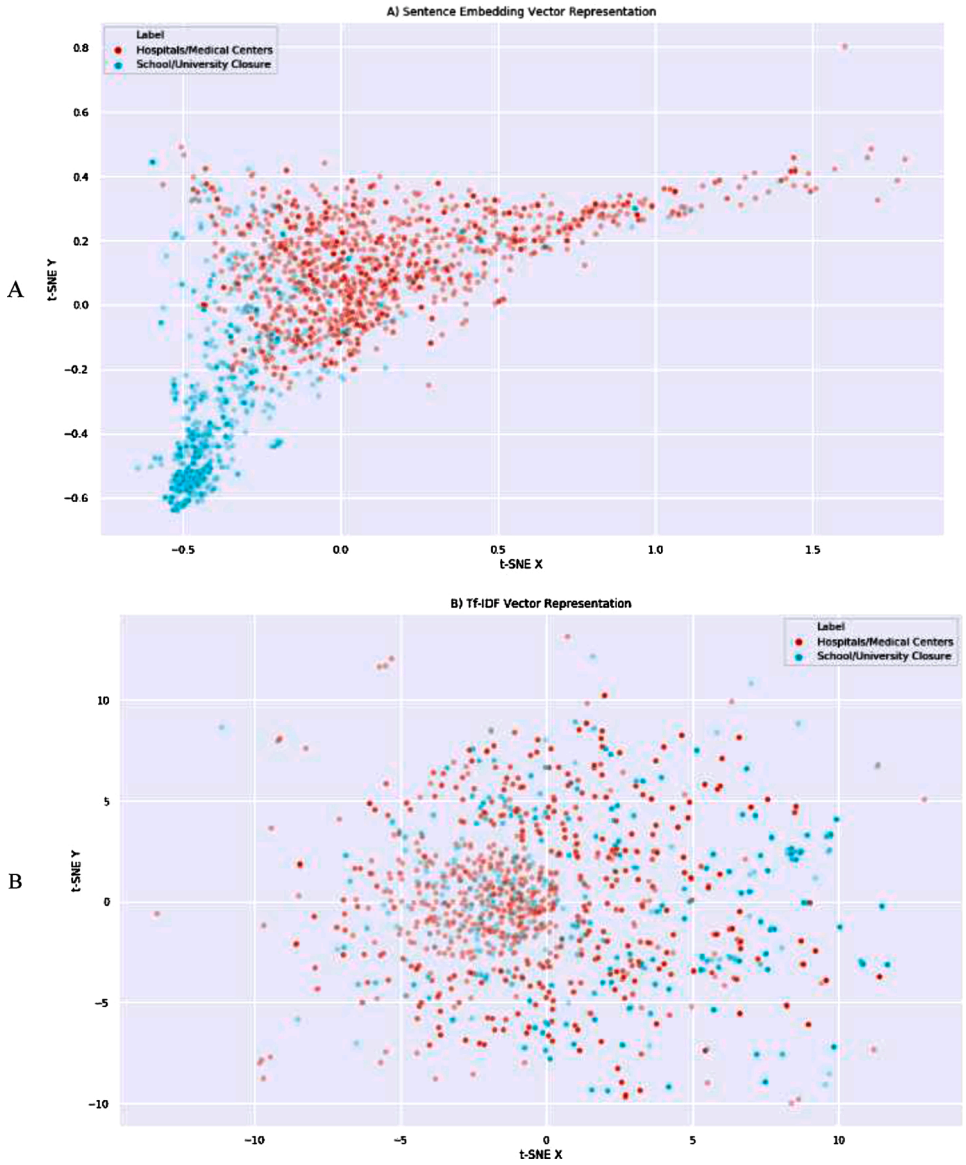

Fig. 2. Subplots A and B show the 2D scatter plot of the sentence vector spaces generated by sentence embedding and Tf-IDF representation, respectively.

method, called t-Distributed Stochastic Neighbor Embedding (t-SNE) [56]. Subplots A and B of Fig. 2 demonstrate the distribution of the classes in two different vector spaces that are sentence embedding and Tf-IDF, respectively. As it can be seen in Fig. 2, the sub-topics (means various concerns under the one general topic, i.e., COVID-19) are better separable in the sentence-embedding space. In summary, the sentence embedding vector space model leverages the semantically-rich representation of the sentences, which makes it capable of separating sentences with common terms better than the Tf-IDF representation.

Table 1
List of public Mapper packages.

Package Name	License	Developer(s)	Link
Python Mapper	GPL	Daniel Müllner	http://danifold.net/mapper/
KeplerMapper	MIT	Hendrik Jacob van Veen	https://github.com/scikit-tda/kepler-mapper
Sakmapper	MIT	Nathaniel Saul Sakellarios Zairis	https://github.com/szairis/sakmapper
TDA Mapper	GPL	Pault Pearson	https://github.com/paultpearson/TDAmapper

4.2. Stage A

Stage A includes two steps: (1) sentence embedding, and (2) sentence grouping. The whole pre-processed news posts are used for learning word representation. Among various word embedding packages, we selected Facebook's FastText package [57], which is fast and accurate for sparse datasets. Although BERT embedding [58] is shown to be very successful, but training it on our data, which is a small-sized sparse Persian corpus, does not show promising results. Instead, FastText trained in less than an hour shows reliable results. The FastText is fed by the whole corpus to generate embedding vectors. The 150-dimensions word embedding vectors, as well as the selected sentences, are given to the FastText to build a sentence embedding representation. Embedded sentences are represented by 150-dimensions vectors as well.

In the next step, sentences must be clustered or grouped based on their embedded semantics. Sentences quoted by the NCRC's members express similar topics scattered around the COVID-19 and SARS-COV-2 viruses. We know that a major proportion of the sentences express NCRC's concerns, but the exact definition of these groups of sentences or their distribution is unknown. Furthermore, the target group of sentences and the peripheral group may overlap with each other due to long

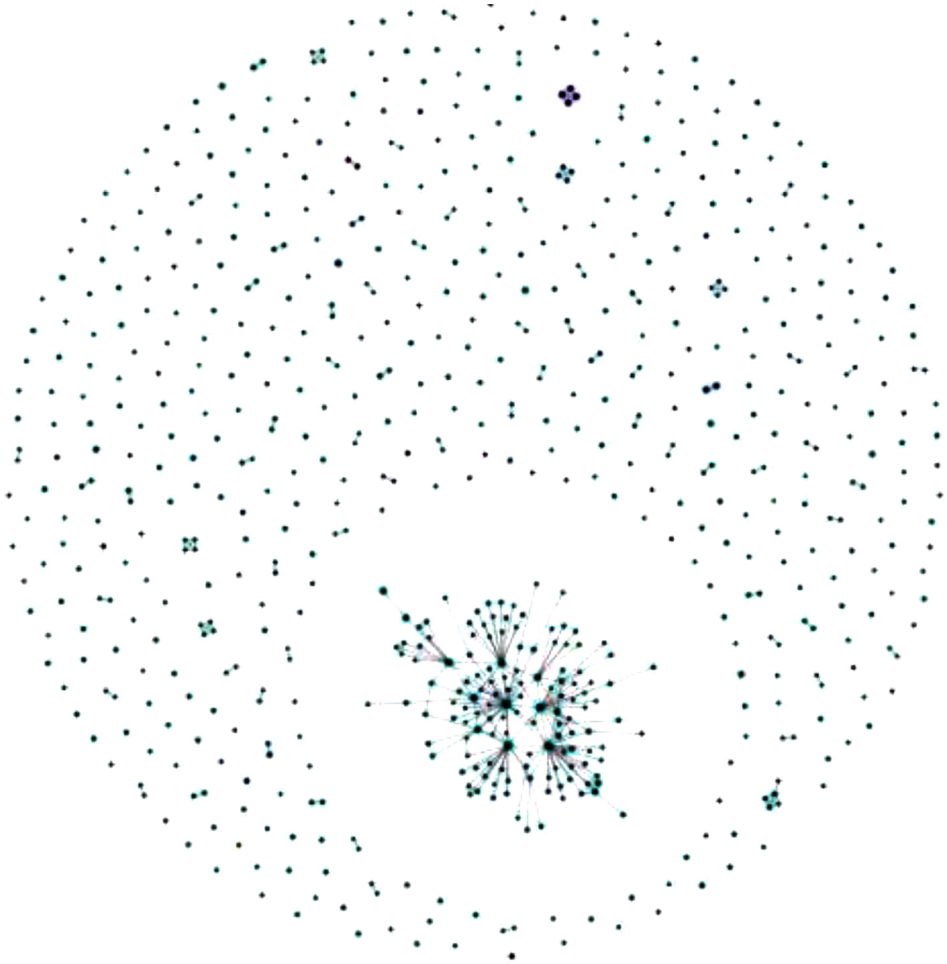

Fig. 3. Plot of the Mapper graph's results in which there are several connected components.

sentences expressing a concern and a peripheral topic at the same time. One of the algorithms that supports overlap grouping of data points with respect to their structure is the Mapper Algorithm introduced by Singh et al. [59], which reduces the high dimensional datasets into simplicial complexes. The low-dimensional representation is not accurate; however, it saves the topological shape of the data and is less sensitive to the selected distance metric. The Mapper algorithm has been applied to various problems such as 3D object recognition [59], bioinformatics [60], data visualization [61], and cancer detection [62]. The Mapper algorithm is applied to the 150-dimensions vectors to group them. There exist different packages for the Mapper algorithm, listed in Table 1.

We use the KeplerMapper package [63], which adopts a clustering algorithm to group the data objects. In this research, the selected clustering algorithm is DBSCAN with cosine distance metric, which is tuned by epsilon of 0.05. The number of hypercubes of the Mapper is set to 10. The generated graph has 708 nodes and 293 edges. Furthermore, this graph includes several connected components; however, there are 11, 799 singleton data points that are not clustered with other samples. Fig. 3 shows the Mapper graph, including one major connected component and several smaller and single-node components. We should note that the single-nodes in this graph are clusters of data samples. The mentioned singleton samples which are not grouped under any of the clusters are not shown in this figure. Even though the Mapper graph uncovers the topological structure of the dataset, it cannot be used to extract topic models. The result of this stage tells us which sentences cannot be grouped with any other sentences. Since the set of addressed concerns of the NCRC are limited, the sentences that have not shared meaning with any other cluster do not express a concern. Consequently, we remove 11,799 singleton sentences from the dataset before passing the vectors to stage B.

4.3. Stage (B)

Word embedding-based topic models are benefited from the rich semantic representation provided by word embedding while probabilistic and matrix factorization methods, such as LDA, PLSA, and NMF models, are more human-readable [64]. In our research, we need to align the extracted topics with the list of NCRC's concerns; thus, we have to adopt a human-understandable topic modeling technique to obtain the topics. In this work, we use LDA as one of the well-implemented probabilistic topic modeling methods. In the first step of the Stage (B), a Tf-IDF vectorizer generates Tf-IDF of the sentences generated by stage A. The Tf-IDF is a vector-based weighting schema used in information retrieval and text mining and is defined based on the multiplication of two elements, which are term frequency (*Tf*) and inverse of document frequency (*IDF*) [65]. More precisely, the Tf-IDF of term t in document d is computed as follows

$$Tf - IDF_{t,d} = Tf_{t,d} \times IDF_t \tag{1}$$

where $Tf_{t,d}$ is the number of times term t occurs in document d, and IDF_t addresses the logarithm of the inverse of document frequency. The simplest form of IDF_t is defined as

$$IDF_t = \log_2\left(\frac{N}{df_t}\right) \tag{2}$$

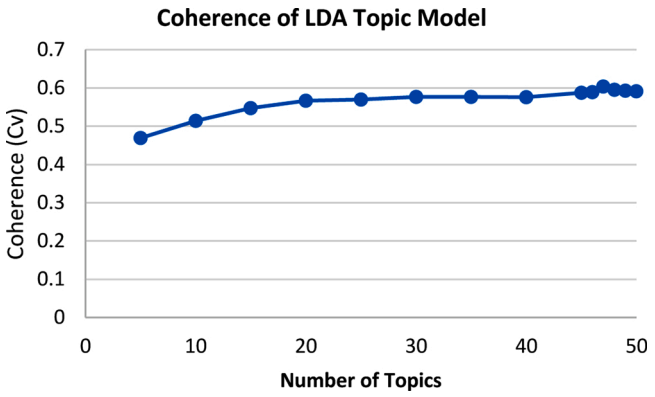

Fig. 4. Coherence score of topics of the LDA models with different number of topics.

in which df_t is the total number of documents that include t in a corpus of N documents. Manning et al. [65] listed several variations of Tf-IDF; however, the above-mentioned version is the most popular one. The LDA model is fed with the Tf-IDF weighted vectors. While Petz et al. [66] noted the efficacy of cleaned text in opinion mining, we removed a predefined list of Persian stop-words and terms which are repeated in more than 80 % of documents from the vectorizer's input.

The LDA model is a hierarchical Bayesian model that uses a bag of words representation of documents to represent them in a semantic space with a lower number of dimensions. This model is fed with the vector representation of documents to calculate the probability of classes for each document. In this study, we use Gensim's implementation of the LDA with the Tf-IDF model [67]. In the next step, the number of topics for the LDA model must be determined. Finding the appropriate number of topics is studied under the model coherence analysis. Topic coherence measures guarantee the understandability of the topics. Roder et al. [68] analyzed seven coherence measures and compared their ability and run-time. Amongst these measures, we select the C_v measure which beats other measures. The best number of topics is determined with respect to the coherence of the underlying topics of the LDA model. Fig. 4 shows the coherence scores for the different number of topics (n). The best score happens when $n = 47$. Thus, we use the LDA with $n = 47$ to extract understandable topics. In this model, each topic is a group of documents represented in a higher-dimensional space of the same size as vectors. The 2D demonstration for the intertopic space of the trained LDA model is demonstrated in Fig. 5, which is visualized using the LDAvis package [69].

4.4. Topic aggregation

There is a limited number of expressed concerns and set up actions enlisted by Abdi [70]. Subsequently, we have to aggregate the topics that cover different aspects of the same action or concern. Our list of concerns and actions include eight themes as follows,

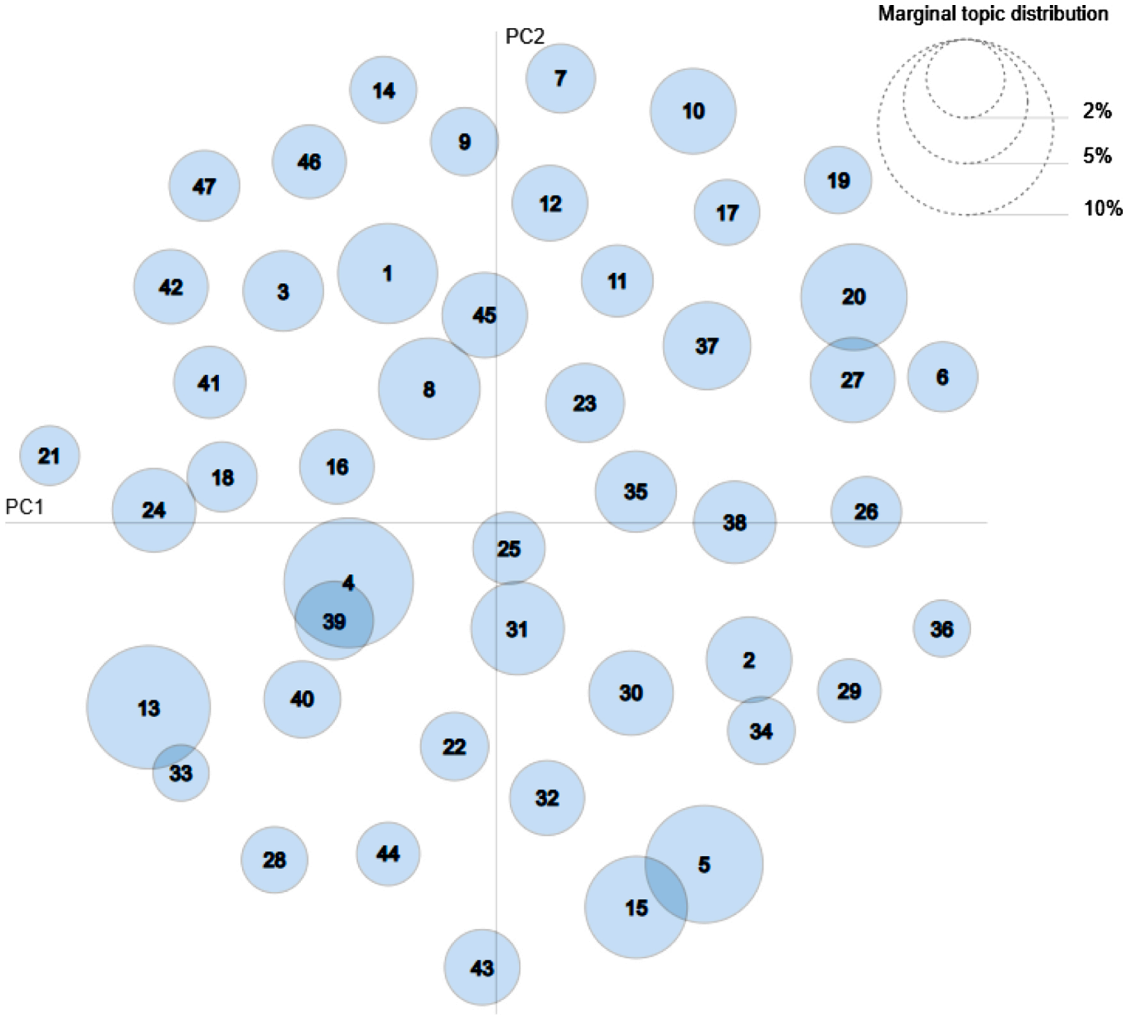

Fig. 5. Intertopic distance map of the LDA model with n = 47 topics.

Table 2
Number of merged topics and sentences to form each theme.

#	Theme	Number of Topics	Number of Sentences	Rank
1	Increase awareness about washing hands, hygiene productions, facial masks	5	4806	3
2	Smart social distancing and working/ business regulations	7	4253	5
3	PCR laboratory test, COVID-19 diagnosis, and screening	3	5931	1
4	Lack of adequate medical infrastructure, equipment, PPE and pressure on health system	4	3505	6
5	Intra-provincial travel and down-town traffic restrictions (cancel pilgrimage, congregational prays, and Jumu'ah prayers)	5	1942	8
6	Briefing the national and provincial status (Number of confirmed, susceptible and recovered cases)	8	2841	7
7	Closure of schools, academic institutes and canceling national and international exams	4	5401	2
8	Miscellaneous topics	11	4731	4

1 Increasing awareness about washing hands, hygiene productions, facial masks.
2 Smart social distancing and working/business regulations.
3 PCR laboratory test, COVID-19 diagnosis, and screening.
4 Lack of adequate medical infrastructure, equipment, PPE and pressure on health system.
5 Intra-provincial travel and down-town traffic restrictions (canceling pilgrimage, congregational prays, and Jumu'ah prayers).
6 Briefing the national and provincial status (Number of confirmed, susceptible and recovered cases).
7 Closure of schools, academic institutes and canceling national and international exams.
8 Miscellaneous topics.

The first seven themes indicate the concerns or responses of the NCRC, and the last theme covers issues, such as thanking medical care personnel. We reviewed the top 20 words of each topic to assign it to the closest theme. Table 2 shows the number of topics and sentences merged to form themes.

5. Results

The largest number of topics merged for a theme belongs to the miscellaneous theme, which is resulted from higher diversity of the miscellaneous topics addressed by NCRC's members. Among the COVID-related themes, the test and screening theme attracts more attention. Ministry of Health and Medical Education (MHME) of Iran collaborates with the Ministry of ICT to deploy an online mobile screening application just ten days after the first confirmed prevalence. This application plays the role of an in-home triage recommending users with respect to their body temperature and symptoms. This application helps the MHME to manage the mental stress of the mass crowd and control the number of medical care requests to prevent overcrowding in hospitals. Promoting the population to use this application and presenting the results of analyzes on the App's data is a major part of the press conferences of the NCRC. The second most important concern is the education system closure. The latest weeks of winter to the middle weeks of spring is an important period in Iran's educational year because of the conducted undergraduate and graduate entrance exams. Duo to the COVID-19 crises, all kindergartens, schools, academic institutes, and universities were closed province by province. The closure decisions were made in a period from February 22 to March 1 with respect to the status of the host province. Increasing awareness by addressing the handwashing protocols, alcoholic-based hand sanitizers, and facial masks is the third most discussed concern in public press conferences and interviews. We demonstrate the number of sentences expressed for each theme from February 1 to May 10 using an area graph (Ref. Fig. 6).

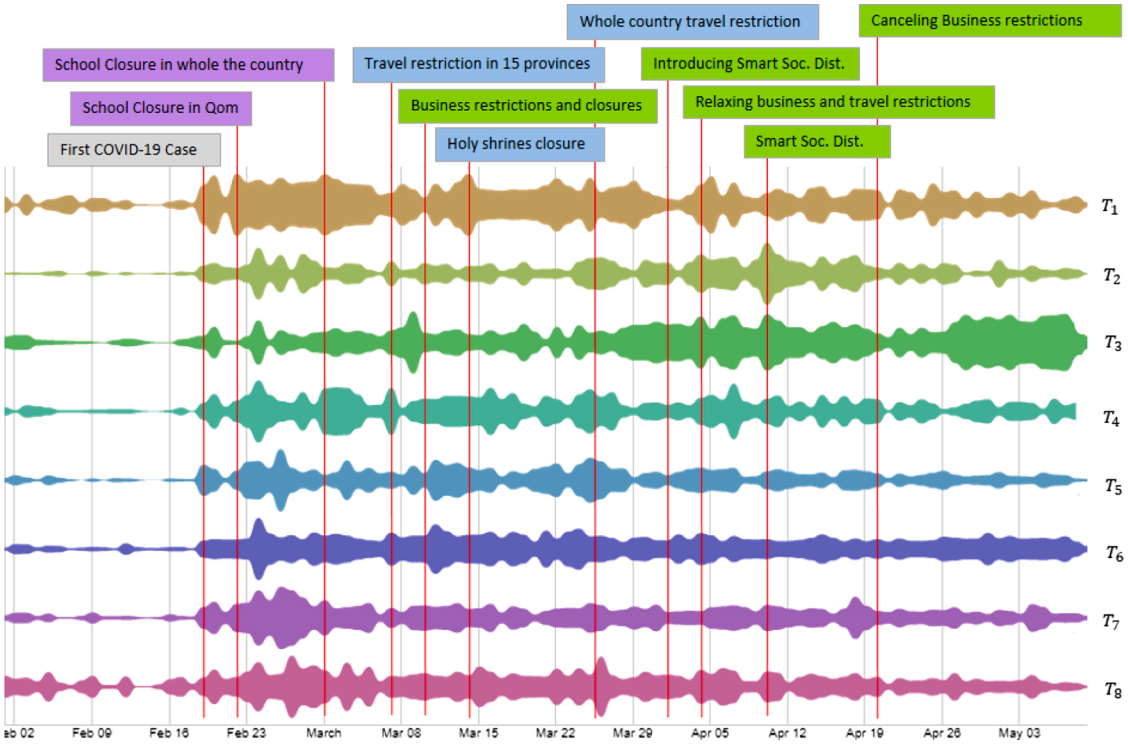

Fig. 6. The area graph demonstrates the changes in the number of addressing different themes. This graph is annotated with the timeline of the preventive actions in Iran.

The area graph of Fig. 6 is annotated with the timeline of Iran's preventive actions. The first COVID-19 confirmed case in Iran is reported on February 19, 2020, in Qom. The first reaction to this news was school closure in Qom on February 22. Schools and universities around the country were consequently closed. The whole education system was shut down on March 1, 2020. The widest area (eq. 20.6 %) in the related theme (i.e., T_7) is spotted between February 22 and March 1 which means that in this period the most important concern of the NCRC was decreasing the rate of spread by avoiding the contacts in the educational system. Moreover, this time window is the period in which the NCRC's members focused on awareness acts (ref. T_1). We should note that the area under T_1 is about 16.8 % of the total area of this theme.

The Persian new year ceremony, called Nowruz, is celebrated on March 19 (March 20 in leap years). Although schools and universities were closed to decrease the spreading risk, families started to plan their trips to low-risk provinces of the country. Subsequently, the next controlling action was setting up travel restriction in 15 provinces with orange and red status codes on March 7. Business closure and working hour decrement regulations were set on March 10. Holy shrines in Qom, Mashhad, and Shiraz were closed on March 14, and Jumu'ah prayers throughout the whole country were canceled on the same day. During the period that started from March 9 to March 17, the NCRC focused on two major issues, that are awareness about handwashing and sanitization (eq. 11.8 % of the total area under T_1), and decreasing trips and pilgrimages (eq. 14.2 % of the total area under T_5). Meanwhile, the NCRC emphasized on the increasing number of confirmed and death cases in the national and provincial level (eq. 13.7 % of the total area under T_6) to indicate the pressure on national health systems and medical care personnel.

On April 1, the minster of MHME introduced the national smart social distancing plan, which aimed at relaxing the businesses and traffic restrictions. This plan is designed to be a pilot program to study the feasibility of a nation-wide re-opening plan. The business and traffic restrictions were relaxed on April 4, and the whole plan was run on April 10. In this period, the major concerns were evolved to mass media plan briefing, emphasizing on the necessity of testing and screening for susceptible cases (eq. 11.9 % of the total area under T_5). NCRC tried to convince the people with COVID-like symptoms to stay at their homes and rely on the recommendations of the NCRC's certified screening App. Meanwhile, people who received such testing recommendations must avoid attending the public areas and drive to the nearest testing center directly. After canceling all the restrictions applied to businesses on April 20, NCRC's members focused on testing and screening.

6. Discussion

Analyzing the behavior of governments and centers for disease control has been done during previous outbreaks, such as SARS [22], Zika [53], and Ebola [35]. These studies indicate that on-time response and answering the questions raised in the web promptly decrease the probability of misinformation dissemination in social media. Furthermore, governments and disease control centers can fight rumors and conspiracy theory posts by active monitoring of social media [71]. Utilizing text mining techniques to extract health-related information for disease surveillance, misinformation detection, and public concern detection is widely studied; however, analyzing the concerns reflected in the news is not studied well. De Coninck et al. [72] addressed the forgotten role of the news media as a powerful legacy media that enables us to influence people's behaviors. With respect to the reliability of news sources for society, we aimed at extracting and tracking the changes in concerns of Iran's NCRC in this research. We collect news posts, including quotes made by members of the NCRC, and then group them to select a major part of the sentences covering similar topics. This stage helps us to remove sentences about peripheral issues that do not reflect the NCRC's concerns or acts. The selected sentences are clustered using a topic modeling method with maximum coherence, and then, the assigned topics are aligned to the list of concerns and actions, enlisted by Abdi [70]. In the last step, the number of news labeled by each theme is visualized. Our findings show that NCRC members make an official statement upon actions on the day that are set up. Even though NCRC is not lagged behind, it is not considered to be an on-time reaction. According to the recommendations of Liu et al. [40], governments have to react ahead of policy set up to maximize their impacts. The themes' time series are very fluctuating, which means that the NCRC's members express their concerns regarding different issues under the influence of the closest action or concern. One of the questionable practices found in this research is the late reaction of the NCRC's to test and tracking issues. Although screening is started on the earliest days of the outbreak in Iran, the test and tracking cycle is considered late.

Le et al. [66] studied information notified as demanded information by the diverse socioeconomics groups in Vietnam during the COVID-19 lockdown. Results of their surveys show that "updated news about pandemics", "disease's symptoms", "updated news about the outbreak", and "notices on how to prevent the disease" are the most requested topics. The least demanded information is "notices on travel". Our findings show that the top four topics covered in Iran's NCRC press conferences are "lab test, screening, and diagnosis", "closure of schools and educational institutes", "preventive awareness", and "pressure on the health system". Similarly, the topic with minimum interestingness level is "information on travel restriction." Although the list of demanded topics in Vietnam and Iran are different, but they assert that major information required to be safe are notices on preventive practices and symptoms/diagnosis.

7. Conclusion and future works

In this article, we used a two-stage framework to group, select, and cluster the sentences expressing concerns of Iran's NCRC. Our framework leverages the ability of the Mapper algorithm to discriminate the sentences covered the peripheral topics from the target sentences. The target sentences are clustered by adopting a latent Dirichlet analysis topic model. Topics addressed the different aspects of one topical theme are aggregated together. The results reveal the fast pace of NCRC reactions. Hence it would be expected from the disease control trustees to react ahead of the actions and policies. Early responses of governments addressed in news media prepare the society to behave in a manner that

Summary Points

3 More than 2,400,000 news posts are used to generate the dataset included the quotes of Iran's NCRC[31] members.
4 The major concerns of the committee are (1) PCR lab. test, diagnosis, and screening, (2) Closure of the education system, and (3) awareness actions about washing hands and facial mask usage.
5 Among the concerns, intra-provincial travel and traffic restrictions, as well as briefing the national and provincial status, are under-presented.
6 The timeline of concerns shows that although the announcements and public responses are not lagged behind the events, but cannot be considered as timely.
7 This study reveals that the NCRC has not a long-time response map, and members react to the closest announced policy/act.

Table A1
List of news channels and number of posts.

#	News Channel	Number of Raw Posts	Number of Remained posts (after Data Cleaning)
1	Akhbar-e-Fori	158,164	13,523
2	Asr-e-Iran	291,439	22,427
3	Borna News	68,843	13,966
4	Eghtesad Online	107,097	8334
5	Fararu	111,729	5096
6	Fars	173,467	8257
7	Ilna	298,861	29,477
8	Irna	10,838	774
9	Isna	153,672	10,604
10	Jam News	16,516	2172
11	Khabar Online	252,972	15,325
12	Khabar Sarasari	53,017	8158
13	Mashregh News	79,790	164
14	Mehr	24,753	9664
15	Namnak	9657	1455
16	Parsineh	129,097	9994
17	Shoma News	119,619	6010
18	Tabnak	71,996	9547
19	Tasnim	138,602	10,791
20	YJC News	141,632	13,078

mitigates the risk of diseases. We plan to expand our research to analyze the impacts of the expressed concerns on people's behaviors through the lens of their social media posts and tweets.

Author contributions

Author 1: Fatemeh Kaveh-Yazdy
Conceived and designed the analysis
Collected the data
Contributed data or analysis tools
Performed the analysis
Wrote the paper
Other contribution
Author 2: Sajjad Zarifzadeh
Conceived and designed the analysis
Performed the analysis
Wrote the paper
Other contribution

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRediT authorship contribution statement

Fatemeh Kaveh-Yazdy: Conceptualization, Methodology, Software, Validation, Formal analysis, Data curation, Writing - original draft, Writing - review & editing, Visualization, Project administration. **Sajjad Zarifzadeh:** Conceptualization, Validation, Writing - original draft, Writing - review & editing, Visualization, Resources.

Declaration of Competing Interest

The authors report no declarations of interest.

Acknowledgment

None.

[3] National COVID-19 Response Committee (NCRC)

Appendix A

Table A1.

Appendix B. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.ijmedinf.2020.104309.

References

[1] WHO, Naming the Coronavirus Disease (COVID-19) and the Virus That Causes It, World Health Organization Portal, 2020 (accessed May 01, 2020), https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it.
[2] S.A. Lauer, et al., The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application, Ann. Intern. Med. 172 (9 May) (2020) 577–582, https://doi.org/10.7326/M20-0504.
[3] W. Guan, et al., Clinical characteristics of coronavirus disease 2019 in China, N. Engl. J. Med. 382 (18) (2020) 1708–1720, https://doi.org/10.1056/NEJMoa2002032.
[4] Z. Wu, J.M. McGoogan, Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72 314 cases from the chinese center for disease control and prevention, JAMA 323 (13) (2020) 1239–1242, https://doi.org/10.1001/jama.2020.2648.
[5] J.S. Faust, C. del Rio, Assessment of deaths from COVID-19 and from seasonal influenza, JAMA Intern. Med. (2020), https://doi.org/10.1001/jamainternmed.2020.2306.
[6] M. Salathé, et al., Digital epidemiology, PLoS Comput. Biol. 8 (7) (2012) 1–3, https://doi.org/10.1371/journal.pcbi.1002616.
[7] S. Cook, C. Conrad, A.L. Fowlkes, M.H. Mohebbi, Assessing Google Flu Trends Performance in the United States during the 2009 Influenza Virus A (H1N1) Pandemic, PLoS One 6 (8) (2011) 1–8, https://doi.org/10.1371/journal.pone.0023610.
[8] D. Butler, When google got flu wrong, Nature 494 (2013), https://doi.org/10.1038/494155a.
[9] M. Salathé, Digital epidemiology: what is it, and where is it going? Life Sci. Soc. Policy 14 (1 January) (2018) 1, https://doi.org/10.1186/s40504-017-0065-7.
[10] M. Kang, H. Zhong, J. He, S. Rutherford, F. Yang, Using google trends for influenza surveillance in South China, PLoS One 8 (1) (2013) 1–6, https://doi.org/10.1371/journal.pone.0055205.
[11] S. Kandula, J. Shaman, Reappraising the utility of google flu trends, PLoS Comput. Biol. 15 (8) (2019) 1–16, https://doi.org/10.1371/journal.pcbi.1007258.
[12] A. Husnayain, A. Fuad, L. Lazuardi, Correlation between Google Trends on dengue fever and national surveillance report in Indonesia, Glob. Health Action 12 (1) (2019) 1552652, https://doi.org/10.1080/16549716.2018.1552652.
[13] Y. Teng, et al., Dynamic forecasting of zika epidemics using google trends, PLoS One 12 (1) (2017) 1–10, https://doi.org/10.1371/journal.pone.0165085.
[14] A. Alessa, M. Faezipour, Flu outbreak prediction using twitter posts classification and linear regression with historical centers for disease control and prevention reports: prediction framework Study, JMIR public Heal. Surveill. 5 (2 June) (2019) e12383, https://doi.org/10.2196/12383.
[15] S. Masri, et al., Use of Twitter data to improve Zika virus surveillance in the United States during the 2016 epidemic, BMC Public Health 19 (1) (2019) 761, https://doi.org/10.1186/s12889-019-7103-8.
[16] S. Yousefinaghani, R. Dara, Z. Poljak, T.M. Bernardo, S. Sharif, The assessment of Twitter's potential for outbreak detection: Avian influenza case study, Sci. Rep. 9 (1) (2019) 18147, https://doi.org/10.1038/s41598-019-54388-4.
[17] C. Li, L.J. Chen, X. Chen, M. Zhang, C.P. Pang, H. Chen, Retrospective analysis of the possibility of predicting the COVID-19 outbreak from Internet searches and social media data, China, 2020, Eurosurveillance 25 (10 March) (2020), https://doi.org/10.2807/1560-7917.ES.2020.25.10.2000199.
[18] F. Kaveh-Yazdy, A.-M. Zareh-Bidoki, Search engines, news wires and digital epidemiology: presumptions and facts, Int. J. Med. Inform. 115 (July) (2018) 53–63, https://doi.org/10.1016/j.ijmedinf.2018.03.017.
[19] L. Qin, et al., Prediction of number of cases of 2019 novel coronavirus (COVID-19) using social media search index, Int. J. Environ. Res. Public Health 17 (7 March) (2020), https://doi.org/10.3390/ijerph17072365.
[20] S.M. Ayyoubzadeh, S.M. Ayyoubzadeh, H. Zahedi, M. Ahmadi, S.R. Niakan Kalhori, Predicting COVID-19 incidence through analysis of google trends data in Iran: data mining and deep learning pilot Study, JMIR public Heal. Surveill. 6 (2 April) (2020) e18828, https://doi.org/10.2196/18828.
[21] B. Zimmer, Infodemic': when unreliable information spreads far and wide, Wall Street J. 05 (March) (2020).
[22] Institute of Medicine (US) Forum on Microbial Threats, Learning From SARS: Preparing for the Next Disease Outbreak, National Academies Press, Washington (DC), 2004.
[23] J.S. Kwaak, MERS, rumors spread in South Korea, Wall Street J. 05 (June) (2015).
[24] S.O. Oyeyemi, E. Gabarron, R. Wynn, Ebola, Twitter, and misinformation: a dangerous combination? BMJ 349 (2014) https://doi.org/10.1136/bmj.g6178.
[25] J. Zarocostas, How to fight an infodemic, Lancet 395 (10225 Febuary) (2020) 676, https://doi.org/10.1016/S0140-6736(20)30461-X.

[26] M. Cinelli, et al., The COVID-19 social media infodemic, Arxiv (March) (2020). Accessed: May 27, 2020. [Online]. Available: http://arxiv.org/abs/2003.05004.
[27] J. Hua, R. Shaw, Corona virus (COVID-19) 'Infodemic' and emerging issues through a data lens: the case of China, Int. J. Environ. Res. Public Health 17 (7 March) (2020), https://doi.org/10.3390/ijerph17072309.
[28] D.A. Erku, et al., When fear and misinformation go viral: pharmacists' role in deterring medication misinformation during the 'infodemic' surrounding COVID-19, Res. Social Adm. Pharm. (May) (2020), https://doi.org/10.1016/j.sapharm.2020.04.032.
[29] R. Kouzy, et al., Coronavirus goes viral: quantifying the COVID-19 misinformation epidemic on twitter, Cureus 12 (3 March) (2020) e7255, https://doi.org/10.7759/cureus.7255.
[30] B.W. Nelson, A.K. Pettitt, J. Flannery, N.B. Allen, Rapid assessment of psychological and epidemiological correlates of COVID-19 concern, financial strain, and health-related behavior change in a large online sample, PsyArXiv (April) (2020), https://doi.org/10.31234/osf.io/jftze.
[31] C. Wang, et al., A longitudinal study on the mental health of general population during the COVID-19 epidemic in China, Brain Behav. Immun. 87 (July) (2020) 40–48, https://doi.org/10.1016/j.bbi.2020.04.028.
[32] C. Wang, et al., Immediate psychological responses and associated factors during the initial stage of the 2019 coronavirus disease (COVID-19) epidemic among the general population in China, Int. J. Environ. Res. Public Health 17 (5 March) (2020) 1729, https://doi.org/10.3390/ijerph17051729.
[33] I. van der Vegt, B. Kleinberg, Women worry about family, men about the economy: gender differences in emotional responses to COVID-19, Arxiv (April) (2020). Accessed: May 26, 2020. [Online]. Available: http://arxiv.org/abs/2004.08202.
[34] L. Deng, B. Xu, L. Zhang, Y. Han, B. Zhou, P. Zou, Tracking the evolution of public concerns in social media, ACM International Conference Proceeding Series (2013) 353–357, https://doi.org/10.1145/2499788.2499826.
[35] A.J. Lazard, E. Scheinfeld, J.M. Bernhardt, G.B. Wilcox, M. Suran, Detecting themes of public concern: a text mining analysis of the centers for disease control and prevention's Ebola live Twitter chat, Am. J. Infect. Control 43 (10 October) (2015) 1109–1111, https://doi.org/10.1016/j.ajic.2015.05.025.
[36] X. Ji, S.A. Chun, Z. Wei, J. Geller, Twitter sentiment classification for measuring public health concerns, Soc. Netw. Anal. Min. 5 (1) (2015) 13, https://doi.org/10.1007/s13278-015-0253-5.
[37] X. Ji, S.A. Chun, J. Geller, Monitoring public health concerns using twitter sentiment classifications, IEEE International Conference on Healthcare Informatics (ICHI 2013) (2013) 335–344, https://doi.org/10.1109/ICHI.2013.47. Sep.
[38] C.E. Lopez, M. Vasu, C. Gallemore, Understanding the perception of COVID-19 policies by mining a multilanguage Twitter dataset, Arxiv (March) (2020). Accessed: May 28, 2020. [Online]. Available: http://arxiv.org/abs/2003.10359.
[39] Q. Chen, C. Min, W. Zhang, G. Wang, X. Ma, R. Evans, Unpacking the black box: how to promote citizen engagement through government social media during the COVID-19 crisis, Comput. Human Behav. 110 (2020) 106380, https://doi.org/10.1016/j.chb.2020.106380.
[40] Q. Liu, et al., Health communication through news media during the early stage of the COVID-19 outbreak in China: digital topic modeling approach, J. Med. Internet Res. 22 (4 April) (2020) e19118, https://doi.org/10.2196/19118.
[41] T. Hale, A. Petherick, T. Phillips, S. Webster, Variation in government responses to COVID-19, Blavatnik Sch. Gov. Work. Pap. 31 (2020).
[42] M. Iqbal, Telegram revenue and usage statistics (2020), Business Apps 24 (April) (2020).
[43] C.H. Papadimitriou, H. Tamaki, P. Raghavan, S. Vempala, Latent semantic indexing: a probabilistic analysis, Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (1998) 159–168, https://doi.org/10.1145/275487.275505.
[44] T. Hofmann, Probabilistic latent semantic indexing, SIGIR Forum 51 (2 August) (2017) 211–218, https://doi.org/10.1145/3130348.3130370.
[45] D.M. Blei, J.D. Lafferty, Correlated topic models, Proceedings of the 18th International Conference on Neural Information Processing Systems (2005) 147–154.
[46] D.M. Blei, J.D. McAuliffe, Supervised topic models, in: Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 3-6, 2007, 2007, pp. 121–128 [Online]. Available: http://papers.nips.cc/paper/3328-supervised-topic-models.
[47] J. Chang, D.M. Blei, Relational topic models for document networks, in: Proceedings of 12th International Conference on Artificial Intelligence and Statistics (AISTATS), vol. 5, 2009, pp. 81–88 [Online]. Available: http://proceedings.mlr.press/v5/chang09a.html.
[48] C. Wang, B. Thiesson, C. Meek, D.M. Blei, Markov topic models, in: Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics, {AISTATS} 2009, Clearwater Beach, Florida, USA, April 16-18, 2009, vol. 5, 2009, pp. 583–590 [Online]. Available: http://proceedings.mlr.press/v5/wang09b.html.
[49] R. Das, M. Zaheer, C. Dyer, Gaussian LDA for topic models with word embeddings, Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (2015) 795–804, https://doi.org/10.3115/v1/P15-1077. Jul.
[50] C.E. Moody, Mixing dirichlet topic models and word embeddings to make lda2vec, Arxiv (May) (2016). Accessed: May 29, 2020. [Online]. Available: http://arxiv.org/abs/1605.02019.
[51] M. Shi, J. Liu, D. Zhou, M. Tang, B. Cao, WE-LDA: a word embeddings augmented LDA model for web services clustering, Proceedings of 24th International Conference on Web Services (ICWS 2017) (2017) 9–16, https://doi.org/10.1109/ICWS.2017.9. Sep.
[52] E.H.-J. Kim, Y.K. Jeong, Y. Kim, K.Y. Kang, M. Song, Topic-based content and sentiment analysis of Ebola virus on Twitter and in the news, J. Inf. Sci. 42 (6) (2016) 763–781, https://doi.org/10.1177/0165551515608733.
[53] E.M. Glowacki, A.J. Lazard, G.B. Wilcox, M. Mackert, J.M. Bernhardt, Identifying the public's concerns and the Centers for Disease Control and Prevention's reactions during a health crisis: an analysis of a Zika live Twitter chat, Am. J. Infect. Control 44 (12 December) (2016) 1709–1711, https://doi.org/10.1016/j.ajic.2016.05.025.
[54] M. Dong, X. Cao, M. Liang, L. Li, G. Liu, H. Liang, Understand research hotspots surrounding COVID-19 and other coronavirus infections using topic modeling, medRxiv (2020), https://doi.org/10.1101/2020.03.26.20044164.
[55] D.C. Stokes, A. Andy, S.C. Guntuku, L.H. Ungar, R.M. Merchant, Public priorities and concerns regarding COVID-19 in an online discussion forum: longitudinal topic modeling, J. Gen. Intern. Med. (May) (2020) 1–4, https://doi.org/10.1007/s11606-020-05889-w.
[56] L. van der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (2008) 2579–2605 [Online]. Available: http://www.jmlr.org/papers/v9/vandermaaten08a.html.
[57] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Trans. Assoc. Comput. Linguist. 5 (2017) 135–146, https://doi.org/10.1162/tacl_a_00051.
[58] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1, 2019, pp. 4171–4186, https://doi.org/10.18653/v1/N19-1423. Jun.
[59] G. Singh, F. Memoli, G. Carlsson, Topological methods for the analysis of High dimensional data sets and 3D object recognition, Symposium on Point Based Graphics (2007) 91–100, https://doi.org/10.2312/spbg/spbg07/091-100.
[60] A. Holzinger, M. Dehmer, I. Jurisica, Knowledge discovery and interactive data mining in bioinformatics - State-of-the-Art, future challenges and research directions, BMC Bioinform. 15 (6) (2014) I1, https://doi.org/10.1186/1471-2105-15-S6-I1.
[61] S. Liu, D. Maljovec, B. Wang, P. Bremer, V. Pascucci, Visualizing high-dimensional data: advances in the past decade, IEEE Trans. Vis. Comput. Graph. 23 (3 March) (2017) 1249–1268, https://doi.org/10.1109/TVCG.2016.2640960.
[62] Q.-T. Bui, B. Vo, H.-A.N. Do, N.Q.V. Hung, V. Snasel, F-Mapper: a fuzzy mapper clustering algorithm, Knowledge-Based Syst. 189 (2020) 105107, https://doi.org/10.1016/j.knosys.2019.105107.
[63] H.J. van Veen, N. Saul, KeplerMapper (January) (2019).
[64] C.E. Moody, Introducing Our Hybrid lda2vec Algorithm, Stichfix, 2016 (accessed May 25, 2020), https://multithreaded.stitchfix.com/blog/2016/05/27/lda2vec/#topic=38&lambda=1&term=.
[65] C.D. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval, Cambridge University Press, New York, NY, USA, 2008.
[66] H.T. Le, et al., Demand for health information on COVID-19 among vietnamese, Int. J. Environ. Res. Public Health 17 (12 June) (2020) 4377, https://doi.org/10.3390/ijerph17124377.
[67] R. Rehurek, P. Sojka, Software framework for topic modelling with large corpora, Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, May (2010) 45–50.
[68] M. Röder, A. Both, A. Hinneburg, Exploring the space of topic coherence measures, Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (2015) 399–408, https://doi.org/10.1145/2684822.2685324.
[69] C. Sievert, K. Shirley, LDAvis: a method for visualizing and interpreting topics, Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, Jun. (2014) 63–70, https://doi.org/10.3115/v1/W14-3110.
[70] M. Abdi, Coronavirus disease 2019 (COVID-19) outbreak in Iran: actions and problems, Infect. Control Hosp. Epidemiol. 41 (6 June) (2020) 754–755, https://doi.org/10.1017/ice.2020.86.
[71] X. Han, J. Wang, M. Zhang, X. Wang, Using social media to mine and analyze public opinion related to COVID-19 in China, Int. J. Environ. Res. Public Health 17 (8 April) (2020) 2788, https://doi.org/10.3390/ijerph17082788.
[72] D. De Coninck, L. D'Haenens, K. Matthijs, Forgotten key players in public health: news media as agents of information and persuasion during the COVID-19 pandemic, Public Health 183 (2020) 65–66, https://doi.org/10.1016/j.puhe.2020.05.011.

Fatemeh Kaveh-Yazdy, Ph.D., Senior Data Scientist, at Yazd University. Her research interests include information extraction, data mining, and machine learning.

Sajjad Zarifzadeh, Ph.D., Assistant professor at Department of Computer Engineering, Yazd University, Yazd, Iran. His research interests include big data, data analysis, internet services, and network security.