Journal of Biomedical Informatics 121 (2021) 103862

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Special Communication

Tracking and analysis of discourse dynamics and polarity during the early Corona pandemic in Iran

Fateme Jafarinejad *, Marziea Rahimi, Hoda Mashayekhi

Faculty of Computer Engineering, Shahrood University of Technology, Shahrood 3619995161, Iran

ARTICLE INFO

Keywords
Social media analysis
COVID-19
Latent Dirichlet allocation
Dynamic topic modeling
Data stream analysis
Concept drift

ABSTRACT

It has not been long since a new disease called COVID-19 has hit the international community. Unknown nature of the virus, evidence of its adaptability and survival in new conditions, its widespread prevalence and also lengthy recovery period, along with daily notifications of new infection and fatality statistics, have created a wave of fear and anxiety among the public community and authorities. These factors have led to extreme changes in the social discourse in a rather short period of time. The analysis of this discourse is important to reconcile the society and restore ordinary conditions of mental peace and health. Although much research has been done on the disease since its international pandemic, the sociological analysis of the recent public phenomenon, especially in developing countries, still needs attention. We propose a framework for analyzing social media data and news stories oriented around COVID-19 disease. Our research is based on an extensive Persian data set gathered from different social media networks and news agencies in the period of January 21-April 29, 2020. We use the Latent Dirichlet Allocation (LDA) model and dynamic topic modeling to understand and capture the change of discourse in terms of temporal subjects. We scrutinize the reasons of subject alternations by exploring the related events and adopted practices and policies. The social discourse can highly affect the community morale and polarization. Therefore, we further analyze the polarization in online social media posts, and detect points of concept drift in the stream. Based on the analyzed content, effective guidelines are extracted to shift polarization towards positive. The results show that the proposed framework is able to provide an effective practical approach for cause and effect analysis of the social discourse.

1. Introduction

The global pandemic of COVID-19 has caught us by surprise. From its unexpected ability to spread and survive, which makes it different from many other pandemics of our era, has arisen a sense of anxiety and confusion among people and even authorities. This can lead to low levels of public compliance to governments' advisories and thus their success in controlling and defeating COVID19. On the other hand, the world-wide coronavirus spread and its massive impact across all sectors of society, have caused an explosion of coronavirus-related discussions in social media, and similarly an abundance of related news articles. These contents can reflect the real-time reactions of the people to the daily notifications of new developments, advisories and fatality statistics. This vast amount of data and the severity of the disease's effect on our day-to-day life has prompted the interest of many researchers worldwide to investigate the potentials of analyzing this data to better understand the disease itself and the people's reactions. This knowledge can in-turn lead to more realistic decisions, effective advisories and appropriate actions by authorities or social activist in their effort to control the disease.

News and Social media contents have proved to be a reliable source of information in better understanding similar but more limited outbreaks like Influenza [1] or SARS [2] and other health-related issues. For example, twitter has been used in surveillance of Lyme disease in Ireland [3], and to discover the influential groups that are the spreading sources of health information about diseases like Zika [4] or Ebola [5]. Similar analysis has been performed on Facebook and Instagram posts [6,7].

In this paper, we provide an analysis of the behavior of social media users in reaction to COVID-19-related news and status. We propose a framework for analyzing social media data and news stories oriented around the COVID-19 disease. Our main goal is to reveal the changes that has occurred in people's mentality regarding the COVID-19 through the discourse in social media, and if these changes are on par with the changes in the news article contents. We also want to know the correspondence between these changes and the decisions made by

* Corresponding author.
E-mail addresses: jafarinejad@shahroodut.ac.ir (F. Jafarinejad), marziea.rahimi@shahroodut.ac.ir (M. Rahimi), hmashayekhi@shahroodut.ac.ir (H. Mashayekhi).

https://doi.org/10.1016/j.jbi.2021.103862
Received 15 October 2020; Received in revised form 24 May 2021; Accepted 30 June 2021
Available online 3 July 2021

authorities. These changes can be tracked by investigating the topics discussed by people in different durations after the COVID-19 outbreak, which leads us to the probabilistic topic modeling [8] and its dynamics. Another question we want to answer is whether the changes are positive or negative. For answering this part of our research, we use data stream analysis techniques to detect concept drifts [9] in the polarity.

In the proposed framework, topic modeling is used to reveal the probable changes in topics of discourse during the four months from late January to the end of April. In a probabilistic topic model, a topic is a distribution over the vocabulary words, and each document is a mixture of topics. The mixture of topics is a low-dimensional representation of the high-dimensional text documents, and the topics are more informative compared to the single words [8]. This is one advantage of the probabilistic topic models that makes them suitable to be used for analysis of text data in the manner we aim for. Latent Dirichlet Allocation (LDA) [10] can be considered as the most established topic model. The topics provided by LDA during each slot of our timeline can provide us with a general interpretation of subjects discussed by the people.

Moreover, detecting changes in these topics, during the mentioned timelines, encouraged us to dig further with more focused timelines using the Dynamic Topic Model (DTM) [11]. DTM is an LDA-based topic model that keeps track of topics over time and examines the evolution of topics in a sequence of time-ordered documents. The order and arrangement of the documents reflects the evolving set of topics. This arrangement is critical in detecting topic evolution in different time-dependent applications. Thus, DTM is a suitable technique for addressing our question about the changes in topics during the mentioned period, in a less biased and more focused manner than LDA.

The next step is to investigate the changes in the polarity of people's opinions and discourse using the data stream analysis techniques. Concept drift occurs when the statistics of data distribution in a stream changes. Techniques for detection of concept drift can be applied based on different probabilistic and machine learning methods. Hoeffding's Inequality Based Drift Detection Method (HDDM) [9] is a concept drift detection method based on the Hoeffding's Bound, which is shown to be effective on different datasets [12]. Due to its good performance in detecting drift in data streams [12], this method is used to analyze the social network data in this paper.

Social media posts have been previously used to gain new insights regarding many aspects of COVID-19 [13-15]. However, our work is different from the existing literature in two aspects, namely including the methodological approach and the employed dataset. Our research questions, mentioned above, and the process we employ to utilize text analysis and data mining techniques for addressing the questions are significantly different from the literature

Moreover, to the best of our knowledge, no previous study has analyzed the social discourse and its dynamics in Iran, during the early pandemic stages. We base our research on a large data set of Persian language tweets and news published mostly by Iranians from January to April 2020. The results reported in this paper show the gradual change of discourse in terms of topics and polarity, which are investigated in depth to offer realistic analysis on public views and behavior. In addition, we discuss the benefits and shortcomings of the applied algorithms and techniques from a methodological view.

The structure of this article is as follows: In the second section, we will review the literature on the subject. The introduction of probabilistic topic modeling, dynamic topic modeling, and the concept drift detection method are discussed in section three. The fourth section presents the architecture of the proposed framework. In section five, the results obtained from applying the framework are analyzed. Section six summarizes the results obtained in the previous section to provide a general analysis of the data. Finally, in the seventh section, we will provide the conclusion and suggestions for future work.

2. Literature review

As mentioned before, in this paper, an analysis of COVD-19 related social media posts and news articles is provided. In this analysis, the aim is to answer a series of questions regarding the changes in people opinions and reactions, and the polarity of these changes. To conduct the analysis and answer the mentioned questions, the LDA, DTM, and concept drift detection methods are employed in a systematic manner, which is presented as a framework. In this section, we review the studies that have used social media posts to provide new insights regarding different aspects of COVID-19 specially those using similar algorithms utilized in this paper.

Considering the propagation of information in social media through the discourse of COVID-19, and analyzing its negative/positive effect is an important view of research in COVID-19-related topics. To understand the destructive effect of spreading false news on the peace of society and the people's mental health, some researches devoted to the problem of detecting fake news and misinformation [16-20]. Samuel et al. [21] addressed the issue of propagation of situational information in social media in COVID-19 epidemic. Applying NLP methods on Weibo[1] data, COVID-19-related information are categorized into seven types of situational information. Using published COVID-19 daily time-series data of US confirmed cases, Marmarelis, et al. [22] proposed a method which identifies interactions between three groups of people: "Susceptible", "Infectious" and "Recovered/Removed" cases. The dynamic inter-relationships of these fractions are represented with first-order differential equations and by a Riccati equation. Applying this approach on the data results in the decomposition of the epidemic time-course into five Riccati modules representing major infection waves until June 18th, 2020.

Twitter data have been used to analyze facets of social distancing across US [13]. Another research [23] investigates how COVID-19-related issues have circulated on Twitter using network analysis techniques. Jim Samuel et al. [24] used naïve Bayes and logistic regression for sentiment classification of coronavirus-related tweets and shown that naïve Bayes is more suitable for classifying shorter tweets. Another work [25] has reported the use of deep learning for sentiment analysis of coronavirus-related tweets. The difference in the reactions of citizens from different countries to the coronavirus pandemic and to the subsequent actions taken by their respective goverments has also been investigated through sentiment analysis of Twitter data [26]. Another study [27] is performed on the posts generated by people with suspected or laboratory-confirmed coronavirus infection on Weibo to analyze their distribution and other epidemiological characteristics. Word embedding in combination with statistical analysis techniques are used in a study [28] to analyze the tone of officials' tweets as either alarming or reassuring, and to capture the response of people based on reposting and retweeting information on twitter. The idea of using big-data and artificial intelligence techniques to detect and predict the COVID-19 pandemic cases has been investigated in a work [29] by Agbehadji, et al. This work provides a review of the computing models that can be adopted to enhance the performance of the mentioned tasks. Dimitrov et al. [30] introduce a publicly available knowledge base of more than eight million coronavirus-related tweets with extracted meta data, entities, hashtags, user mentions and sentiments.

Topic modeling has also been used to analyze social media content with several different perspectives. For example, Liu et al. [31] and Stokes et al. [32] use the generated topics from social media to analyze people's reactions to COVID-19. In another example [14], LDA is used to reveal the important topics of people's comments in Reddit. The authors used LDA to extract the topics, and used deep learning for sentiment classification of the comments. In another example [17], LDA is simply

[1] Sina Weibo is a major Chinese micro blogging social media available at https://www.weibo.com

applied on a corpus of tweets discussing social-distancing as a result of coronavirus pandemic to extract the most important topics in the corpus. Another research [18] uses a type of dynamic event modeling method to track evolving sub-topics around risk, testing and treatment in the tweets of US government officials. Ordun et al. [19] applied LDA to tweets of mostly English language, and analyzed the distribution of extracted topics through time by considering the frequency of topics in one-minute time series of tweets. Then, by applying change-point detection techniques, they investigated if people are paying attention to the U.S. government briefings about coronavirus by comparing the position of the fluctuations against the time of those briefings. Topic model is also used in combination with Qualitative Thematic Analysis to analyze public discourse and sentiment regarding older adults in COVID-19 on Twitter after the coronavirus outbreak [15]. Trend analysis and thematic analysis of Weibo are conducted in [20]. In this work, the unsupervised BERT model is adopted to classify sentiment categories and TF-IDF is used to summarize the topics of posts. Italian tweets during covid-19 event are used in [33] to detect/track topics and construct a topic graph. In this paper, a term-frequency analysis in some time slots is pipelined with nutrition and energy metrics for computing hot terms. Social information of users is also used as a measure of quality of tweets information. In [14] topic modeling is used to uncover various issues related to COVID-19 from public opinions in social media. Furthermore, LSTM recurrent neural network is used for sentiment classification of COVID-19 comments. These data are used to understand issues surrounding COVID-19 and to guide related decision-making. In [34] an analysis of the COVID-19 patients' situations and their relationships is designed. Combining the semantic web of the geographic knowledge graph and the visual analysis model of geographic information, this work is useful in community prediction, discovering patients' relationships, the analysis of the spatiotemporal distribution of patients, and the prevention and control of high-risk groups.

Beijnon et al. [35] applied dynamic topic models to understand user perceptions of smart watches on short documents of Reddit. Few researches devoted to the problem of applying DTM on heterogeneous inputs. Mele et al. [36] proposes a new DTM model called discrete DTM, which is able to detect and track multiple events between heterogeneous news streams. Despite the widespread use of DTM in other applications, its use in disease discourse analysis has not been as common as LDA. Breen et al. [37] used DTM and LDA methods to track the discourse and review sentiments on ability of Pre-Exposure Prophylaxis on HIV diesis prevention.

Data steam analysis techniques have also been scarcely used to address some issues regarding COVID-19. Suprem and Pu [38] designed a streaming toolkit for consuming and processing streaming data, and used it to collect a multilingual large-scale dataset of COVID-19 tweets. Their dataset exhibits concept drift and despite applying any drift detection methods, visual assessment shows that social discussion on the disease has changed multiple times after January 2020. In another study, Pu et al. [39] tried to capture new, previously unknown, information from evolving social media content, and distinct verifiable facts from misinformation and disinformation. For the latter, they leveraged the data from authoritative sources as a filtering basis for other gathered information.

In this paper, we aim to analyze social media data to investigate people's reaction to coronavirus-related news along with the changes in topics discussed in the news articles, and in response to authorities' decisions and advisories. To conduct our investigation, we use the LDA topic model, DTM method, and concept drift detection methods in a systematic manner, which is presented as a framework. Despite some similarities mainly in using topic modeling, our work is different with the existing literature. Our perspective is different because we are try to address different question and are investigating different population. In addition, some of techniques we are using have not been utilized in analyzing coronavirus-related data.

3. Preliminaries

In order to analyze the discourse and mentality change of Iranian people in the Corona crisis, three general steps of probabilistic topic modeling, dynamic topic modeling and concept drift detection have been used in our proposed framework. In this section we briefly introduce each of these methods, and references for further study.

3.1. Topic modeling

LDA [10] assumes each document d_m is a mixture of topic distributions or in the other words a distribution over topics. This distribution is denoted by θ_m which follows a Dirichlet distribution with parameter α. Each word w_{mn} in the document is derived from topic z_{mn}. Each topic is a distribution over words denoted by φ_k which follows a Dirichlet distribution with parameter β. Fig. 1 (a) shows the graphical representation of LDA. LDA works based on the word co-occurrences in the documents, and does not emphasize local word relationships.

3.2. Dynamic topic modeling

In a dynamic topic model, we assume that data is partitioned on a time basis (e.g. daily). In other words, in contrast to static topic models, in dynamic topic models the order and arrangement of the documents reflects the evolving set of topics. The documents of each temporal section are modeled with a K-component topic model in such a way that the topics related to the documents of each time slice, t, are evolved from the topics of the previous time slice, $t-1$. For a model with K topics and V terms, consider $\beta_{\mathrm{t,k}}$ as a V-vector for parameters of the kth topic at time$t$ following a Dirichlet distribution. Moreover, to represent the relationship between two distributions at consecutive time intervals $t-1$ and t, the normal distribution is used as shown in the Eq. (1) [11]:

$$\beta_{k,t}\,\beta_{k-1,t} \sim N(\beta_{k-1,t}, \sigma^2 I) \tag{1}$$

Hence, in a dynamic topic model, we have a set of topic models that are sequentially chained to each other. The graphical representation of this model is shown in Fig. 1 (b).

3.3. Concept drift detection

A data stream is an infinite sequence of data, generated from various sources such as social networks, bank transactions, and sensor networks. Concept drift in data stream occurs when the pattern and distribution of data is unstable and varies over time. If this change is also accompanied by the emergence of new classes, concept evolution has occurred. Presence of drift can affect the underlying properties of classes. By identifying these changes, better and more accurate decisions can be made in terms of the data behavior. Concept drift is divided into four categories based on how it occurs, namely abrupt, gradual, incremental, and recurring.

HDDM [9] is a concept drift detection method, which is shown to be effective on different datasets [12]. HDDM monitors performance of the base learner using probability inequalities. To provide theoretic guaranties in change detection, all random variables are assumed to be independent, univariate, and bounded. It provides boundaries for both false positive and false negative rates. HDDM has two variants, HDDM-A and HDDM-W, which use the moving average and the weighted moving average values, respectively. Due to its good performance in detecting drift in data streams [12], this method is used to analyze the social network data in this paper.

4. Proposed framework

In this section, we will introduce the employed dataset and the proposed framework for data analysis. We provide a general structure of

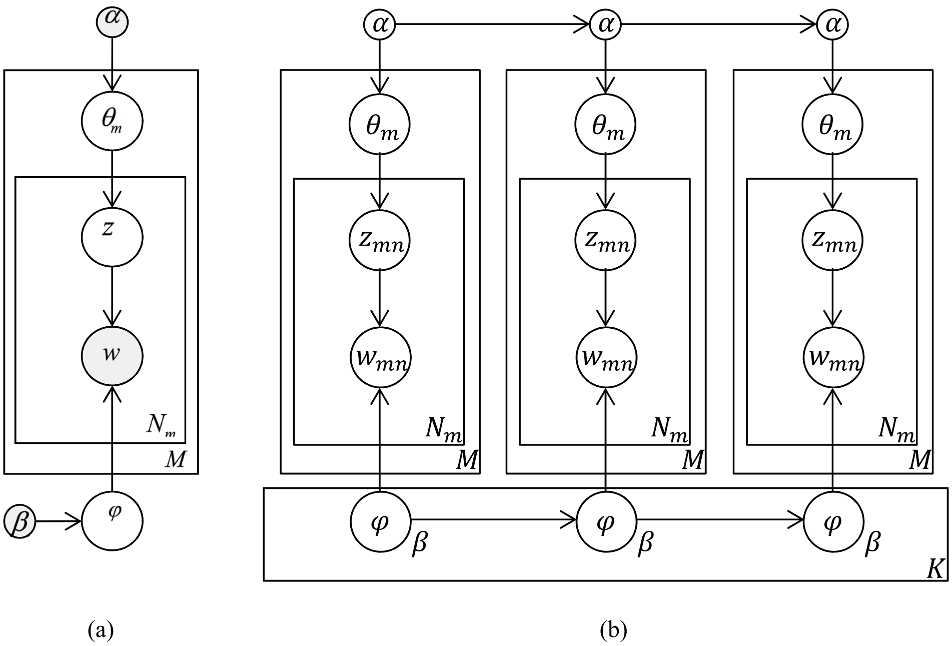

Fig. 1. The plate diagrams for (a) LDA [10], and (b) DTM [11].

our proposed framework and then in the next section, we introduce each element of the framework in detail.

4.1. Dataset

We have used a dataset[2] provided by the Cognitive Science and Technologies Council in Iran. This dataset consists of posts and comments from Instagram[3], Telegram[4] and Twitter[5] published during the first months of the coronavirus outbreak in Iran. It also includes Persian news articles from different online news agencies, released on the same period. The dataset consists of 22 million coronavirus-related posts and news article, all in Persian language, from which 6000 randomly selected posts have manually been tagged for their polarity as negative and positive. The statistics of the data is summarized in Table 1. The data covers the social network comments related to the disease from January 21 to April 29, 2020 which roughly corresponds to four months of the Solar Hijri (SH) calendar; “Bahman”, “Esphand”, “Farvardin”, and one third of “Ordibehesht”, as stated in Table 2. Whenever applicable, we divide the dataset timeline into the four mentioned months, numbered as months 1–4, respectively. As the new year in SH calendar usually occurs on March 21st, the four months correspond to the two last months of year 1398, and the first two months of year 1399 in the SH calendar.

Table 1
Dataset statistics.

	Number of entries	Maximum length	Average length
news	1,047,317	30,981	448.87
Telegram	12,156,077	1512	90.78
Instagram	4,919,839	849	77.72
Twitter	4,165,177	147	27.88

Table 2
The dataset timeline divided into four months according to the SH calendar.

	Solar Hijri calendar	Gregorian calendar
Month 1	Bahman (the whole month), 1398	January 21 to February 19, 2020
Month 2	Esphand (the whole month), 1398	March 20 to April 19, 2020
Month 3	Farvardin (the whole month), 1399	March 20 to April 19, 2020
Month 4	Ordibehesht (the first 10 days), 1399	April 20 to April 29, 2020

4.2. Architecture of the proposed framework

The structure of the proposed framework for tracking and analyzing the discourse and its polarity consists of three different steps: topic modeling, dynamic topic modeling, and concept drift detection. These steps and their conceptual interactions are outlined in Fig. 2. Topic modeling and dynamic topic modeling methods are unsupervised methods directly employ the collected data. The concept drift detection method requires data polarity and therefore uses tagged data. Dynamic topic modeling and concept drift detection methods work on sequential data sorted by time.

Finally, analysis and results obtained from the three steps will be combined, and a comprehensive analysis of the impact of the discourse and mentality of individuals in Iranian society from the Corona crisis will be obtained.

5. Results and analysis

In order to analyze the results of applying the framework more accurately, in this section we will provide the results obtained from each of the three steps separately. In the next section, we will summarize these results and provide a general analysis.

5.1. Results of topic modeling

Topic modeling is used to investigate the changes in subjects of social network posts and news articles in reaction to COVID-19 during four months of the analysis timeline. The aim is to obtain a general idea of the changes and confirm if they are meaningful in relation to COVID-19. In

[2] The dataset can be requested at https://covidchallenge.cogc.ir/
[3] https://www.instagram.com/
[4] https://telegram.org/
[5] https://twitter.com/

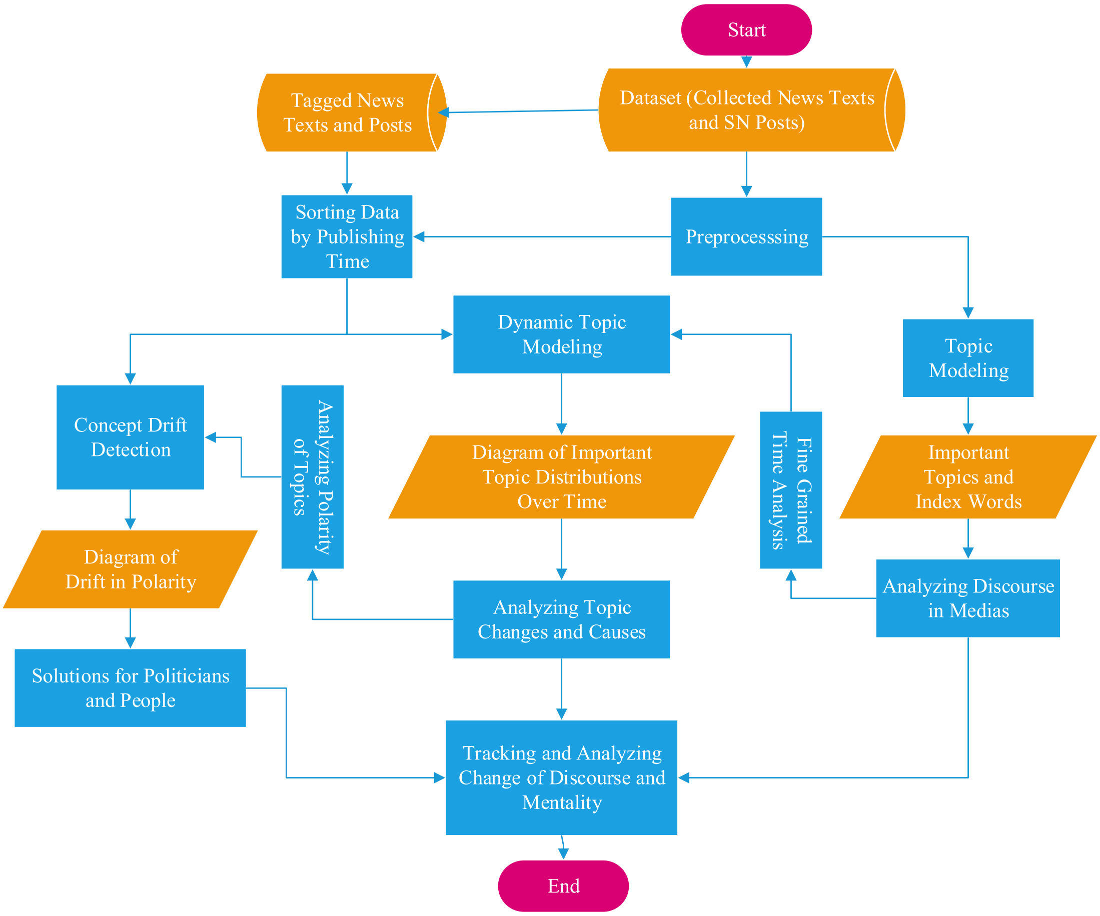
Start
Tagged News Texts and Posts
Dataset (Collected News Texts and SN Posts)
Sorting Data by Publishing Time
Preprocesssing
Dynamic Topic Modeling
Topic Modeling
Concept Drift Detection
Analyzing Polarity of Topics
Diagram of Important Topic Distributions Over Time
Fine Grained Time Analysis
Important Topics and Index Words
Diagram of Drift in Polarity
Analyzing Topic Changes and Causes
Analyzing Discourse in Medias
Solutions for Politicians and People
Tracking and Analyzing Change of Discourse and Mentality
End

Fig. 2. The Proposed analysis framework.

order to implement this model, the Gensim[6] library and pyLDAvis[7] have been used. Note that, all the experimental results are obtained on a PC, running with Windows 10, Intel(R) Core(TM) i7-8700 k CPU @ 3.7 GHz and 64 GB RAM.

The data of each social network and the news are partitioned into the four segments corresponding to the SH months. LDA is not very successful on short texts and works better when the input documents are longer [40]. For extracting higher quality topics, we have modified the social networks data by merging every 50 posts of each part. The news articles however, are long enough and thus were left unchanged in this stage. Five topics are extracted from each part using the LDA topic model.

In the results, topics are shown by an ordered list of their most important words. The coherence and coverage of each topic are also reported. Coherence is evaluated by the UMass [41,42] metric which returns negative values and when closer to zero indicates a more coherent topic. The UMass topic coherence is defined by Equation (2) where $C(k, T^k)$ is the coherence of topic k which is specified by an ordered list of its I most probable words denoted by $T^k = (t_1^k, t_2^k, ..., t_I^k)$. $N(t_j^k)$ is the number of documents the word t_j^k has appeared in, for at least one time and $N(t_i^k, t_j^k)$is the number of documents containing both t_i^k and t_j^k.

$$C(k, T^k) = \sum_{i=2}^{I} \sum_{j=1}^{i-1} log \frac{N(t_i^k, t_j^k) + 1}{N(t_j^k)} \tag{2}$$

Coherence of a topic shows how related to each other the most probable words of the topic are. In other words, how meaningful the topic is according to human perception. The coverage of each topic shows what proportion of the data (each month) is occupied by the most probable words of the topics. One of the five topics in every partition is usually occupied with frequent but not tightly related words and can be considered incoherent.

Let us consider some examples of the extracted topics. Table 3 contains the five topics of each month extracted from Instagram posts. We have highlighted the topics that are strictly related to COVID-19. The intensity of the colors is an indicator of the total coverage of such topics in each month.

During the first month of our timeline, topics 1 and 5 are clearly and strictly related to COVID-19. Topic 2 represents “Coronavirus in China” while topic 5 corresponds to “Coronavirus in Iran”. The first word in this topic is “Qom” which is the Iranian city in which the first case of coronavirus infection was reported. Topic 4 also contains words that could be considered as related to COVID-19. Such words cover half of the 10 most probable words and thus this topic is considered as another coronavirus related topic. Topics 1 and 3 are unrelated to COVID-19. The other topic, as mentioned before is incoherent. All these topics are considered coherent as they have near-zero values. In this month, the COVID-19-relatd topics, together, occupy 49.6 percent of the whole corpus.

During the second month of the timeline, Coronavirus-related topics hover around health care. In this month, the number of such topics decreased to two topics while in the third month it has diminished to one topic. However, Topics 1 and 4 in this month, which are the most related to COVID-19, occupy 51 percent of the whole posts. Thus, this is still the main discourse. Another line of discourse is, not surprisingly, the new year, which arrives, with spring, at the end of this month. Its related

[6] https://radimrehurek.com/gensim

[7] https://pyldavis.readthedocs.io

Table 3
The topics generated for each month from Instagram posts.

Month #1			
Topic	Most probable words of the topic	Coherence	Coverage
1	Imam, Yamani, Mahdi, era, Ahmadalhassan, use, messenger, water, date	−0.25	26.5
2	virus, Corona, China, disease, prevention, healthcare, spread, Iran, people, infected, Wuhan	−0.12	26.3
3	laugh, video clip, Valentine, love, dance, romantic, mother, jock, luxury, song	−0.14	22.3
4	Iran, China, people, vote, virus, cost, Islamic, corona, republic, year	−0.16	17.2
5	Qom, Corona, suspected, vote, primary, university, test, results, tests, news	−0.06	10.4
Month #2			
Topic	Most probable words of the topic	Coherence	Coverage
1	people, person, country, that, spread, disease, declaration, attention, Islamic, face	−0.20	30
2	Iran, Tehran, love, new year, send, year, video clip, Iranian, laugh, cost	−0.16	25.9
3	One, year, that, head, water, life, that, Imam, become, into	−0.11	21.8
4	Corona, overcome, home, virus, stay, quarantine, Iran, Coronavirus, staying, Qom	−0.15	21.7
5	Leadership, Revolution, Corona, Iran, guardian, Leader, celebrities, gifts, virus, Independence	−2.24	0.5
Month #3			
Topic	Most probable words of the topic	Coherence	Coverage
1	love, laugh, video clip, Corona, Tehran, Irani, quarantine, girl, jock, romantic	−0.25	32
2	that, date, use, them, people, will, water, perform, title	−0.16	29.4
3	Corona, virus, overcome, Iran, health care, people, mask, medical, disease, treatment	−0.23	20.2
4	shipment, competition, sell, Toman, purchase, page, order, number, online	−0.04	16.4
5	gifts, *(in Arabic:)* Saudi, Kuwait, explorer, in, Riyadh, Jeddah, Iraq, follow, Bahrain	−0.91	2
Month #4			
Topic	Most probable words of the topic	Coherence	Coverage
1	Corona, laugh, video clip, jock, Tehran, Iranian, dance, live, quarantine, girl	−0.01	30.4
2	shipment, cost, purchase, order, Toman, skin, hour, competition, sell, online	−0.14	30.4
3	that, date, use, anniversary, era, attention, will, Yamani, work	−0.92	23.6
4	Corona, virus, defeat, Iran, people, mask, Coronavirus, health care, quarantine, home	−0.13	15
5	revolution, leader, leadership, revolutionary, Ghalibaf, election, combative, fight, guardian, great	−0.14	0.6

words have appeared in topics 2 and 3. The other remaining topic may be considered incoherent. The value of coherence for this topic is −2.24 which is very low considering the other values. An interesting observation is that the topics of the first month were affected by issues such as the virus origin and spread. But, from the second month, gradually words such as water, home, quarantine, stay, mask, etc. appear more frequently in the topics, which shows that people are becoming aware of the prevention mechanisms. This highlights that the process of public awareness must be taken more seriously [43].

In the third month, economy-related and health care-related topics are the common discourse and the only COVD-19-related topic covers only 20 percent of the whole posts. This is a significant decrease in-relation to the last month. In this month, topic 1 is about entertainment and topic 4 about shopping which covers 48.4 percent of the whole posts.

During the last month, topic 4 is the only topic which is strictly related to Coronavirus. However, the words of this topic have less coverage on the whole dataset which is the lowest among the four considered months. The word "Corona" which is the most-used word in Iran to refer to COVID-19, is the first word in topic 1. However, this topic is dominated by words that are unrelated to Coronavirus, and no other coronavirus-related word appears in the first 30 words of this topic and thus this topic is not about COVID-19 and represents the entertainment discourse.

Topics 1 and 2 are related to entertainment and online shopping, respectively. These topics cover 60.8 percent of the posts in this month. Such topics cover only 22.3 percent of the posts in the first month. Words related to entertainment and shopping have also appeared in the topics of the second month, but they are not dominating the topics. Contrary to that, in the last two months such words are among the most probable words of the corpus and cover over 40 and 60 percent, respectively. Though the gradual rise of miscellaneous topics may decrease the level of public stress caused by the social media contents [44], as the social media mostly present personal opinions, the real condition of the disease may be neglected because of misinformation [16] caused by this considerable decrease in coronavirus-related topics. This may encourage people to escape to the safety of collective ignorance which reduces the compliance to protocols. Accordingly, a subsequent peak in number of infected cases in Iran is observed during the last month[8].

Topics extracted from News articles go through similar changes. However, while coronavirus-related topics have more coverage in news articles, among them, some are not strictly about coronavirus itself but are about the disease's effects in the society and economy. Table 4 shows

Table 4
The coronavirus-related topics of both news articles during the entire timeline. Each topic is denoted by a pair of (month #, topic #).

News Articles			
Topic	Most probable words of the topic	Coherence	Coverage
1, 3	people, virus, China, Corona, Infected people, spread, life, world, infected, dangerous, Hubei, Chinese, victims, death, fatal	−1.15	17.9
1, 4	health care, ministry, turism, medical, country, quarantine, students, education, health, Iran, deputy minister, Namaki (*the current minister of health*), suspected, sciences	−1.11	16.3
1, 5	mask, use, diseases, system, respiratory, body, disease, immunity, symptoms, human, drugs, hands, material, viruses, vaccine	−0.46	11.7
2, 1	province, people, city, Corona, disinfecting, county, university, science, Gilan, hygiene, citizens, mask, traffic, municipality, centers	−0.95	25.2
2, 4	virus, disease, use, home, consume, Corona, people, water, hand, life, them, advise, body, observance, coronavirus	−0.93	18.2
3, 2	province, plan, headquarters, Tehran, distribution, activity, execution (*of a plan*), sanitary, observance, distance, social, disinfecting, units, system, communities of practice	−0.83	25.4
3, 4	people, infected, patients, disease, COVID, virus, infection, health care, Corona, infected people, treatment, medical, hospital, death	−1.65	13.1
4, 1	province, education, headquarters, online, face, county, observance, program, distribution, activity, insurance, knowledge, sanitary, network, Corona	−0.98	34.0
4, 4	people, virus, disease, Corona, 19, hygiene, COVID, infected, infected people, patients, infection, toll, china, medical, death	−0.73	14.0
4, 4	Corona, virus, defeat, Iran, people, mask, Coronavirus, health care, quarantine, home, health, medical, disease, statistics, stay	−0.13	15

[8] https://covid19.who.int/region/emro/country/ir

the coronavirus-related topics in the news. In this table, topics are numbered using an ordered pair where the first denotes the month in our timeline and the second is the topic id in that month.

In the first month, there are three coronavirus-related topics which together cover 45.9 percent of the whole corpus (news articles). Among these three, topic 3 (1, 3), which has the most coverage, is about disease origin and fatality, topic 4 (1, 4) is about the measures taken by ministry of health to control the disease, and topic 5 (1, 5) is about the symptoms and function of COVID-19. In the second month, two topics are related to coronavirus and cover 43.4 percent of the entire corpus. Among these topics, topic 4 seems to be about the spread of COVID-19 in the province of Gilan, which was one of the first provinces of Iran that was caught in the first coronavirus wave. This province, because of its unique cuisine and beautiful sentries created by the temperate broadleaf forests and Caspian Sea, is a very popular destination for tourists all over the year. This was considered as one of the main causes for the sudden raise in the infections and Gillan's capital was one of the first cities to consider a quarantine (lock-down). Topic 4 is about the advisories to prevent infection. In the third month, there are two topics relate to coronavirus which cover 38.5 percent of the corpus. The last month also has two topics which together cover 48 percent of the corpus.

As observed, the coverage of the coronavirus-related topics decreases from the first to third months. However, in the last month a rise on the coverage of such topics happens, most of which is shouldered by topic 1 with coverage of 34 percent. This topic seems to be about coronavirus-related problems. It seems that, when after two months of very strict measures the disease was still spreading albeit less rapidly, it is more important to think long term about dealing with its side-effects. Therefore, journalists focus more on discussing the related problems and the effectiveness of the taken measures.

It is clear that the news topics discuss more serious and specific matters. For example, topic 5 in the first month is about the symptoms and function of the disease or topic 2 in the third month is about the measures taken to control the disease and its spread. On the other hand, the news topics are usually more negative. For example, while the people speak about defeating the disease, the journalists are speaking about death toll, especially in the last two months. This is depicted in topics 4, in both the third and fourth month of our timeline. The reason why some people fail to follow the COVID-19 advised protocols, may be the fact that they mostly refer to the social network and family and friends with whom they communicate a lot on these platforms [45]. As this content may be less negative compared to the news articles, the necessity of following the protocols is less perceived. With some half-measures, control of the outbreak will be less effective [46], and as reported by WHO, a new wave of infections appeared in the last month.

News articles also contain topics that are unrelated to coronavirus. Table 5 contains such topics. In this table, similar to Table 4 most of the topics hover around economic or political subjects. For example, topic two in the first month (1, 2) is related to the petroleum price or topic 3 in the second month is about the Iranian parliament (Islamic Consultative Assembly) election which happened in that month.

Looking at Tables 3-5, the coherences of topics in news articles are generally worse than Instagram posts. According to our experiments and observation, constructing LDA with more topics on news article will give us topics with better coherences that are comparable to Instagram topics. It reinforces our other observation suggesting that journalist discuss problems with more divers and specific topics.

The mentioned results confirm that changes in topics through the four consecutive months of our timeline are meaningful in relation to COVID-19. However, one may wonder if a simpler method can lead us to the same findings and interpretations. We consider two other approaches for comparison, namely looking at the 10 most frequent words in each month, and also words with higher TF-IDF[9] values. The results are shown in Tables 6 and 7, respectively.

Table 5
The topics of news articles that are unrelated to coronavirus.

News Articles			
Topic	Most probable words of the topic	Coherence	Coverage
1, 1	Participation, people, mobile, team, Iran, fair, snow, revolution, national, matches	−0.65	31.2
1, 2	decrease, dollar, petroleum, percent, China, price, market, increase, America, Corona	−1.07	22.9
2, 2	America, Iran, countries, china, war, team, Italy, Trump, game, Europe	−1.27	21.9
2, 3	assembly, Islamic, president, consultative, headquarters, session, people, service, education, country	−1.37	19.8
2, 5	Market, petroleum, price, decrease, bank, million, sell, dollar, economic, payment	−0.94	14.9
3,1	that, knowledge, online, life, team, home, conditions, education, space, students	−0.97	28.7
3, 3	America, government, economy, economic, million, decrease, market, petroleum, bank, price	−1.19	21.9
3, 5	Islamic, assembly, revolution, imam, Iraq, people, hadrat (*an honorific*), indigents, leader, political	−0.93	11
4, 2	government, economic, assembly, country, Iran, policies, force, republic, political, law	−1.68	19.4
4, 3	international, monetary, fund, America, Trump, team, league, football, united, states	−0.64	18.7
4, 5	market, Toman, price, billion, percent, million, decrease, production, rate, car	−0.91	13.9

The most frequent words shown in Table 6 are mostly the same for each month and so cannot capture any meaningful changes in the timeline. Here, contrary to topic modeling, we are just looking at frequencies and not co-occurrences which are the key to many text analysis techniques. On the other hand, the mentioned topical relationships are derived from higher levels of word co-occurrences and not only the first level as is done in constructing word-word co-occurrence matrices [8,47]. In other words, topic models focus on the topical relationships of words which can clarify the meaning and therefore topics are more meaningful than single words. The first words in Table 7 are also mostly the same for each month. However, some rare words have also been surfaced in every topic but their contexts are not clear and they are scattered all over the place and cannot bring coherent meanings and insights and to minds.

We have also compared LDA to other clustering methods to show that it can outperform base and state of the art methods for term clustering. For evaluating the methods, we again use the U-Mass coherence metric. As simplistic base model we have used k-means to cluster the documents and then selected the most frequent words of each document. We call this base method, "most probable words in document clusters (MPDC) term clustering". And also we have used the bag of concepts (BOC) model [47] which, simply put, derives the embedding vectors for

Table 6
The most frequent words of each month for instagram corpus.

Month	Most frequent words (TF)
1	Corona, virus, China, Iran, disease, that, healthcare, day, country, people
2	Corona, virus, Iran, Coronavirus, people, defeat, home, disease, Tehran, will
3	Corona, virus, Iran, home, defeat, quarantine, will, stay, Tehran, year
4	Corona, Iran, virus, defeat, home, will, quarantine, Tehran, stay, laugh

[9] Term Frequency-Inverse Document Frequency (TF-IDS) is a popular method of weighting words in a corpus

Table 7
The words with largest TF-IDF scores for instagram corpus.

Month	Words with the largest TF-IDF scores
1	Corona, virus, China, Iran, Yamani, Ahmadalhassan, disease, COVID, that, bassinet, system (*as in solar system*), healthcare
2	Corona, virus, Majalepezeshk (*a medical magazine*), China, Iran, Yamani, Ahmadalhassan, glutton, disease, COVID
3	Corona, virus, Iran, COVID, nurse, *(names of channels):*Malakebash, Drbasiri, Karbalayiiha, Toosansanaat, Sirtwsir
4	Corona, COVID, Iran, virus, defeat, home, will, quarantine, Yamani, Borazjan (*name a city*)

the words in the corpus and then use them to cluster the words. The word vectors are derived using word2vec [48] embedding method. Table 8 contains the results of these evaluations for the LDA, MPDC and BOC.

The UMass coherence in Table 8 is obtained over the 10 most probable words of each cluster. As discussed before, for the UMass coherence, the less is the better. The best coherences in the paper are written in bold. Almost in every case, the best coherences belong to LDA. MPDC selects the most frequent words of each document cluster and obviously because of their high frequencies and their occurring in the same document clusters, these words would be highly co-occurred. UMass works by considering how co-occurred the most probable words of each topic (word clusters here) are and therefore it is expectable that MPDC would yield good coherence results but still LDA is better. The worst results are achieved by BOC which is also expectable because it works by clustering the word2vec vectors which is based on local context and therefore loses a big proportion of co-occurrence information [49,50] which will leads to worse coherence scores.

The results conveyed in this part of our research confirm that topic modeling is able to reveal the changes in people's opinions through extracting the topics they have discussed in social networks. During the first three months of the timeline, COVID-19 dominates the extracted topics from social media. However, it loses the center of attention and fades into other topics like online shopping and entertainments during the last month. The topics extracted from news articles go through similar changes. However, generally speaking, they are more bitter and serious than social media posts. The news articles generally convey real-world information, while the social media mostly present personal opinions and experiments. The latter can lead to misinformation, which in turn may reduce compliance with protocols [16].

5.2. Results of dynamic topic modeling

In the previous section, we find the important issues in each of the media using LDA. In this section, using dynamic topic modeling, we will be able to examine the change of topics in each of the media in a more detailed way. Moreover, we can analyze the reason of change of topic distribution according to the type and time of the changes. Generally speaking, these causes can be due to changes in policies of politicians or occasional events such as Nowruz. In order to implement this model, the dtmvisual[10] library written on the Gensim library has been used. This library offers more features for graphical representation of word distributions or topic distributions than the dtm module in the Gensim library.

Table 8
The coherence evaluations for the word clusters constructed using LDA, MPDC and BOC during the four months of our time line on Instagram and news articles.

Month #1 Data	LDA	MPDC	BOC
Instagram	**−0.15**	−0.19	−2.67
News	**−0.79**	−0.81	−1.98
Month #2			
Data	LDA	MPDC	BOC
Instagram	**−0.57**	−0.61	−2.18
News	−1.09	**−0.96**	−2.88
Month #3			
Data	LDA	MPDC	BOC
Instagram	**−0.32**	−0.35	−2.48
News	**−0.97**	−0.99	−2.33
Month #4			
Data	LDA	MPDC	BOC
Instagram	**−0.27**	−0.39	−2.06
News	**−0.96**	−1.01	−1.82

In addition to the usual preprocessing in topic modeling (normalization, deletion of station words and deletion of rare words), documents must be arranged by publishing time. Moreover, the number of documents of each time slice must be calculated so that it can distinguish between documents of different time slices and calculate its parameters separately in each time slice. We merge 10 posts published in the same day to apply the method correctly to short texts (such as Twitter texts). 4-day time intervals are used to display changes in the subject more smoothly and visibly. Of course, after this step, smaller intervals (for example, hourly) can also be used to model topic evolutions of 4-day intervals in more detail. However, we omitted this further fine grained topic modeling due to the fact that the rate of discourse change around the issue of Corona is not so fast.

The results of the implementation of the dynamic topic model on news texts, and three social networks (Telegram, Twitter and Instagram) posts are shown in Fig. 3. In this figure, the horizontal axis represents the analysis timeline (as mentioned, January 21 - April 29, 2020). The vertical axis represents the distribution of each subject over time, or $\varphi_{k,t}$ (distribution of the topic k on time-slice$t$). Due to the long running time of the dynamic modeling process, only the results of the 3 most important topics have been reported. Top features of each of the topics of data sources and the coherence of the topics are listed in Table 9. High memory consumption in DTM makes us to run the model on sample of data (twenty percent for each source). Evaluation of the results of DTM, using coherence metric (using UMass technique) as represented in the Table 9, suggests that despite this limitation, topics are coherent.

Note that, diagrams of DTM results represent the evolution of topics over time, which further indicates the amount of public attention to the issue corresponding to the specified topic over time. The rise or fall of a topic distribution over a period of time indicates a decrease or increase in attention of public or news to the issue. Furthermore, a sharp decrease or peak in the frequency of a keyword at a particular time can be a signal for analyzing and finding the cause of a problem. Therefore, these diagrams can be interpreted as measurable signals drawn from the data to inform public health either through active monitoring or analyzing retrospectively.

Fig. 3 (a) shows the results of dynamic topic modeling on corona-related news texts. Considering the existence of keywords such as disease, health, medicine, prevalence in top most frequent words of topic #1, it can be said that this topic represents the discourse on the corona virus from a medical perspective. In topic #2, the existence of words such as production, market, oil, economy, along with quantities such as the corona, leads us to the conclusion that we are dealing with an Corona-related economic issue in this topic. Topic #3, with keywords such as the America, Parliament, Government, Labor, Society, next to the word Corona, indicates that we are dealing with a socio-political discourse around this issue.

According to the interpretation of topics using index words and the diagram of Fig. 3 (a), it can be said that top-most important issues in news texts are medical, political, and economic issues, respectively. At the end of April, we are witnessing the culmination of the debate over health issues and its eventual defeat in the face of economic issues. This change in the discourse of news texts fully reflects the decisions of politicians in Iranian society at this time. At this time, with the end of the Nowruz holiday, an important decision is being made regarding the

[10] https://github.com/GSukr/dtmvisual

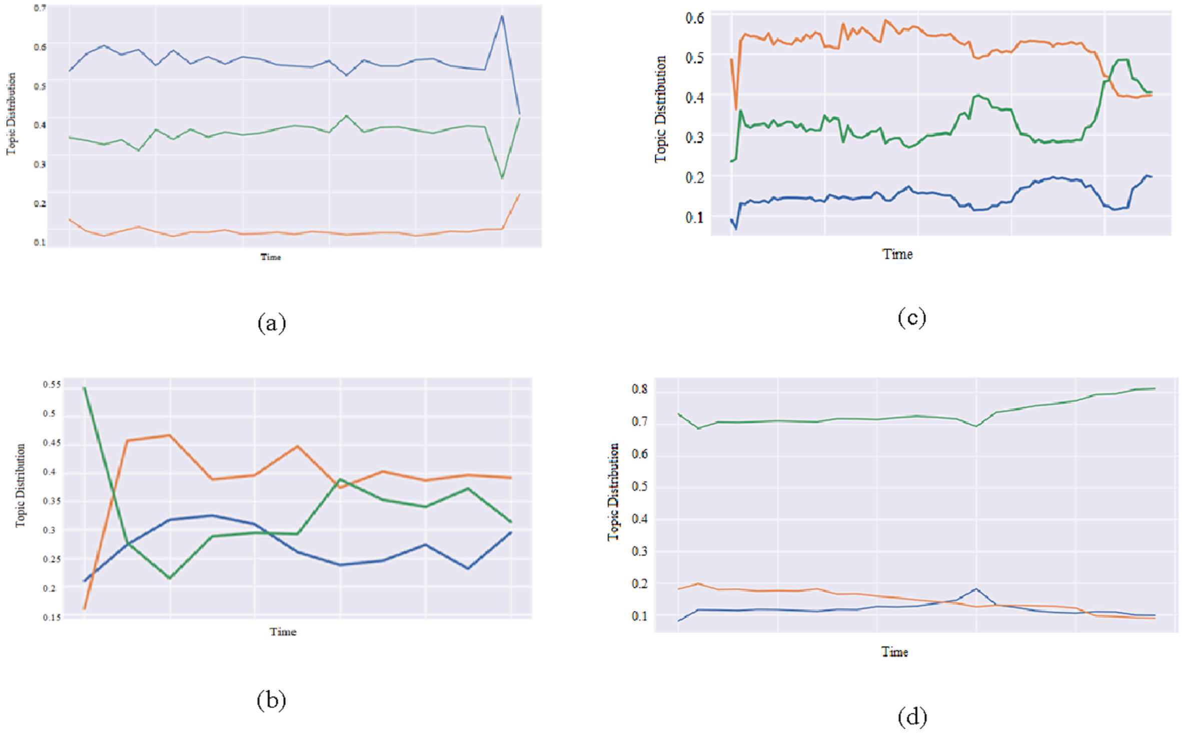

Fig. 3. Results of DTM on different data sources: a) news, b) telegram, c) twitter, and d) Instagram posts.

Table 9
Most frequent keywords of topics of different data sources and their coherence.

News		
Topic	Most probable words of the topic	Coherence
1	Corona, virus, disease, province, person, health, country, medicine, prevalence, report	−0.84
2	Corona, year, production, decrease, market, thousand, country, virus, oil, economic	
3	Corona, Iran, people, circumstances, appointment, America, labor, parliament, government, society	
Telegram		
Topic	Most probable words of the topic	Coherence
1	Virus, corona, China, health, announcement, Iran, infected, suspected, news, Wuhan, bat	−0.61
2	Corona, virus, Iran, health, patient, infected, treatment, china, borders, first, hospital	
3	Corona, virus, infected, china, city, children, diseases, food, age, history, cardiac, pulmonary	
Twitter		
Topic	Most probable words of the topic	Coherence
1	Corona, Iran, virus, people, America, outbreak, China, person, struggle, IRGC	−0.55
2	Corona, virus, Qom, people, infected, health, prevalence, hospital, country, death, quarantine, mask	
3	Corona, torment, people, day, disease, Muhammad, world, God, Tehran, Imam, hospital	
Instagram		
Topic	Most probable words of the topic	Coherence
1	Corona, Iran, GOD, failure, year, quarantine, stay, Nowruz, happy, 99, spring, health	−0.43
2	Corona, virus, people, safety, health, treatment, mask, home, hospital, quarantine, news	
3	Corona, Iran, laugh, clips, love, send, people, movies, follow, books	

continuation of quarantine or the start of intelligent distancing phase by reopening of production-economic centers. According to the discussion of this issue, first the concerns lead to the intensification of medical debates and finally, with this decision, the economic debates around this issue will be highlighted.

Fig. 3 (b) shows the result of dynamic topic modeling on telegram posts. In Topic 1, the existence of more specific words such as bat, wuhan, and china, along with more general words such as corona, which are present in all topics, leads us to conclude that this discourse is on the origin of the virus (spatial and substantive). Topic 2, with keywords such as virus, infected, patient, treatment, hospital, expresses medical issues around Corona. Topic 3 shows some of the social concerns surrounding Corona by stating the prevalence of the virus in Iranian cities, and among different age groups and at-risk groups with a history of the disease. The order of importance of these three topics in the Telegram social network is as medical issues, social concerns and the origin of the virus.

As shown in Fig. 3 (b), the first topic that discusses the bio-spatial origin of the corona virus peaked at second half of February. The third topic, which is one of the most important concerns of the people, is the top issue in late January, when people were informed of the existence of this virus; while two other topics are still emerging in the public discourse. In early March, as Nowruz approaches, this topic again becomes the top-most important subject of corona-centered discourses. Heading towards the last days of the SH year, this topic may have become a source of stress caused by the disease [44]. Along with the advice to stay home, the situation has led to more people paying attention to prevention protocols resulting in a declining corona trend after two weeks, as reported by WHO. Still, the main cause of such decline is the new year holidays which promotes social distancing.

Topic 2, which talks about medical issues, was in the emergence phase until early-February. However, since then it has been at the forefront of popular discourse for a long time. According to Telegram activists this is the most important issue that should be discussed and informed, too. Although, sometimes, political and social issues have partially overshadowed this therapeutic issue, but with the end of the occasional events, the issue of therapy has returned to the top of discourse issues.

Fig. 3 (c) shows the results of dynamic topic modeling on Twitter posts. Examining the keywords in each of these three topics, it can be said that the three most important topics in people's discourse on Twitter have been about medician, religion and politics, respectively. The first issue, by including words like America, struggle, the IRGC, China, can indicate a discourse on political issues. The second issue with words like health, hospital, prevalence, mask, quarantine, prevention can be considered as a topic around medical issues. Finally, the third issue with words like God, Muhammad, Imam could show a topic with a religious discourse.

Given the above topics and the distribution of these topics, which are shown in Fig. 3 (c), it can be said that from the perspective of Twitter activists, medical issues related to the spread of the virus were more important than political and religious issues around this crisis. This distribution has always been included in the discourse of the people. Of course, at some point in time related to the end of April and the arrival of the holy month of Ramadan, religious debates, even in the field of corona, have become more prominent than other debates related to corona.

Fig. 3 (d) shows evolution of three most important topics on Instagram posts. According to the key words in each of the topics, words such as Nowruz, Happy, Spring in the first topic indicate discourse around Nowruz celebration. Moreover, words such as God, health, Nowruz express beautiful wishes at the beginning of the new solar year. The keywords in the second topic, health, safety, hospital, mask, treatment, indicate the medical issues of Corona. As this issue is among the three most important issues in texts of the other three media, so it is in Instagram social network. Although, a surprising point in this social network is that this issue is not the most important topic in discourses of this media. Of course, public attention to this issue has been decreased in late-April in other medias as well. The keywords in top-most important topic in Instagram, as the third topic, such as clips, laugh, movies, follow, books indicate that this is an entertainment-related topic.

In order to provide a general analysis of all data sources, we first randomly select a set of data from each source. In this selection, due to the very short length of Twitter, Instagram and Telegram posts compared to the length of news texts, we first concat several posts on the social network that belong to one day. The number of connecting posts on Twitter, Telegram and Instagram are 20, 10 and 10 documents, respectively. Then we randomly select some data from each source in a manner that finally, the number of input documents to the model by each of the input sources is equal. The results of this modeling are shown in Fig. 4. Top-10 most frequent keywords for each of the three most important topics are represented in Table 10.

According to the results obtained from the general data of all four resources, the issue of health-care and entertainment (topic #3) is at the top of the topics of discourse around the coronavirus. Other issues (political topics and statistics of patients) in this section, which includes a summary of previous analyzes, are much less important than the topic #1. In general, it can be said that a large percentage of discussions about the corona virus have looked at the issue of Corona virus from a medical perspective or entertainment to distract thoughts from the corona issue.

Table 10
Top-10 most frequent keywords for most important topics of the whole dataset obtained by DTM.

Topic	Most probable words of the topic
1	Virus, Corona, China, People, Disease, Country, Iran, Outbreak, Health, America, Chinese
2	Corona, Virus, Hospital, Iran, Day, Flood, Water, China, Contact, Earthquake
3	Corona, Virus, Prevention, Disease, Laughter, Mashhad, Iran, Prevalence, Clip, Attention

In order to compare the results obtained from dynamic topic modeling with static topic modeling, we must note that the goal of both procedures was to observe the dynamics of discourse. This goal is achieved spontaneously by dynamic topic modeling. In static topic modeling, it has been done by performing monthly intervals and running the LDA model independently on different months. Topic independence of different time intervals in the LDA model can find topics more freely and therefore more meticulously each month. Following DTM procedure, the subjects of different periods are considered as dependent on the subject of the previous time so that the change of the subject can be obtained and seen automatically. Of course, considering that the change of subject distribution between time intervals in this model can be investigated spontaneously and not by human, this model has been implemented on smaller four-day intervals.

Despite all this, the results obtained from dynamic and static topic modeling have been complementary. For example, both methods confirm the increase in discourse on economic issues related to the corona in May and among news texts. In both methods, the dominant discourse in different months and different input sources was the discourse on medical issues, which, of course, sometimes due to events and occasions, this issue is slightly marginalized and after that event or occasion, has returned to the forefront of popular discourse and news.

To further emphasis on the results obtained from LDA and DTM methods, we examine the evolution of four keywords (Corona, Nowruz, Shopping, production) corresponding to different medical, social, entertainment, economic topics, (resp.) in news data and in Instagram posts as a social network. The results are represented in Fig. 5.

As illustrated in Fig. 5, the word “Corona” is a frequent word in both data source in the whole timeline (however its rate of use is decreased in 15th interval –of 4-day intervals- which corresponds to new year celebration in Iran. The word Nowruz just peaked in 15th interval in both data source. However its frequency in Instagram is more than its frequency in news. On the other hand, the word “Production” peaked its rate of use in news before 15th interval (of 4-day intervals), but it peaked its rate of use in Instagram after 15th interval. This confirms to some extent the results of LDA about The impact of economic discourse on social media from this discourse in the news (however we see only one

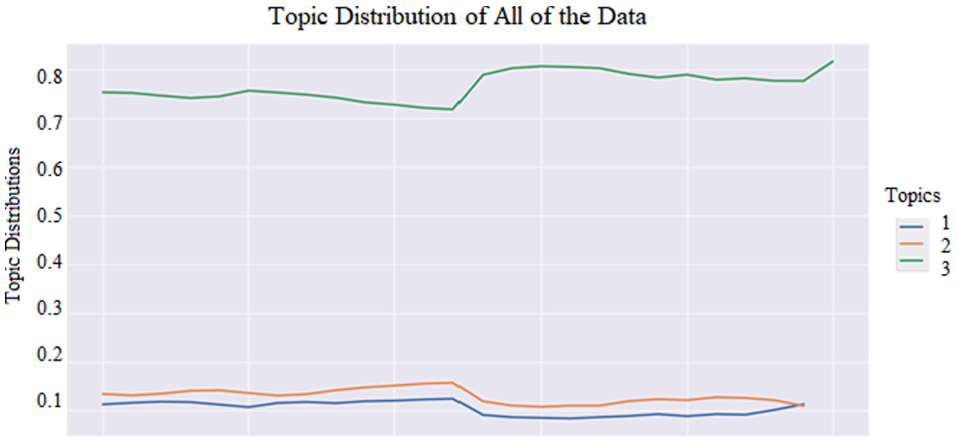

Fig. 4. Topic Distribution of all media as a Result of Dynamic Topic Model.

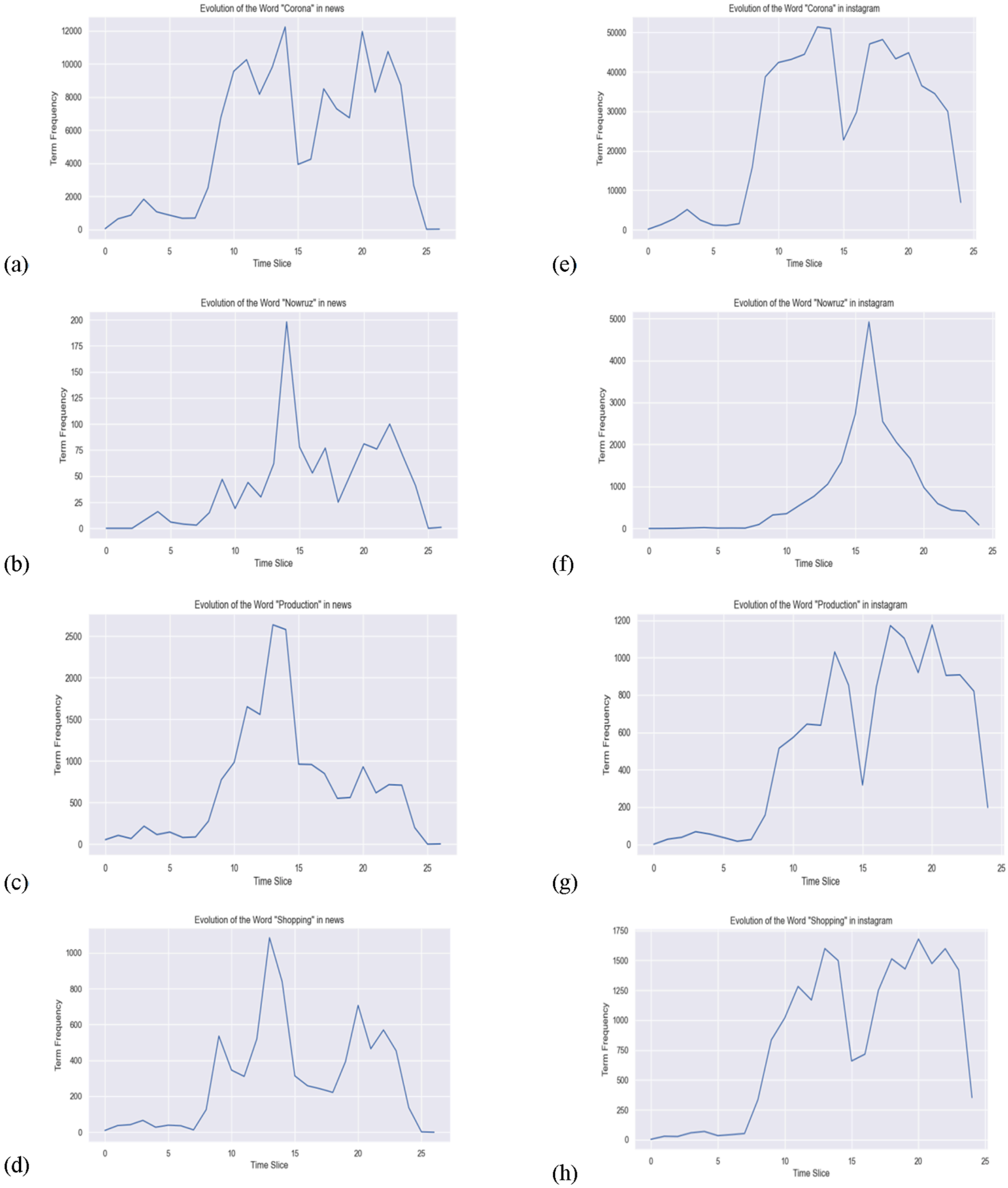

Fig. 5. Evolution of term frequency of different keywords in news and Instagram. Parts a, b, c, and d: term frequency of words corona, nowruz, production, and shopping, respectively in News data. Parts e, f, g, and h: term frequency of words corona, nowruz, production, and shopping, respectively in Instagram posts.

keyword of economic-related topic and there is another peak for this word in instagram before the 15th interval). The word "Shopping" (as a keyword corresponding more to the entertainment topic) is more frequent in Instagram than in the news in the whole timeline. This result confirms the results obtained from DTM and LDA that discourse about entertainments is more popular in the social networks than in the news.

5.3. Results of concept drift detection

After investigating the important topics and the changes in their distributions using the previous methods, we examine the polarity of discourse and its alternations using concept drift detection methods. Using the HDDM_A as a concept drift detection method, we can identify and report the time points of change in the polarity of public discourse. Therefore, we can examine the data and polarity of them before and after drift points to analyze the cause of change in mentality and polarity of public discourse. Hence, drift point can be addressed as another important signal that informs analysts about possible cause of some important public events. We use the scikit-multiflow[11] implementation of the algorithm to obtain the results. Note that, the concept drift is directly applied on the individual polarity values and because of that, normalization was not required by the algorithm. Changes in polarity of records during the period of study (January 21-April 29, 2020) are shown in Fig. 6. As observed, initially the negative polarity is dominant, and this dominance continues with variant intensity over the period. Fig. 7, which shows the average polarity of every 100 data records, confirms the previous observation. As exposed, the published materials on the corona disease are mostly negative, but have a mild tendency towards neutral polarity as time goes on. Therefore, adjusting this attitude and orienting it towards the positive polarity is of great importance for improving the morale and behavior of the society.

To further investigate concept drift in the data, the HDDM-A algorithm is executed on the records with an assigned polarity label. As the concept drift is directly applied on the individual polarity values, normalization is not required by the algorithm. In Table 11 we mention some cases of the concept drift detected by the algorithm, along with the actual sample content around the drift. Due to the dominance of negative and neutral polarity, the concentration is on cases that polarity is oriented towards positive.

Despite the predominance of negative polarity shown in Fig. 7, positive attempts are made to change this practice as observed in Table 11. According to the content of Table 11, these efforts can be classified according to the source of the content as follows.

Authorities: In official notifications, respecting the following items is necessary to promote the public opinion, increase the morale of the society, and its positive attitude towards governance actions:

- Emphasizing the sincere efforts of individuals and organizations, improving trust in the efforts of the governing body, justify public opinion on important governance decisions.
- Strengthening national spirit, holding safe social ceremonies, constructive slogans, characterizing ordinary people confronting the disease.
- Upgrading the required medical and non-medical facilities, exposing justice in access to facilities, Articulating health issues effectively and frequently.
- Refraining from expressing news with a negative tone, combining unfavorable news with patriotism, and religious expressions.

People: Members of the community can change the course of public views and opinions for the better by publishing the following: hopeful messages, humor, personal experiences and positive memories, content on occasional events, spreading slogans, building characters, and republishing positive messages.

After analyzing the results of drift detection in the dataset, we are prompted to share some of the methodological experiences we had while dealing with concept drift detection. Most well-known and cited concept drift methods have restricted their evaluation to closed data sets, including the retrospectively cleaned data on events. The more complex changes in reality (like the COVID-19 pandemic) do not appear in the training data, and therefore the methods fall short in detecting real-world concept drifts with high accuracy. The real datasets contain concept drifts, which may be apparent in different granularities of data. Gradual global changes may exist despite the local abrupt drifts. Many concept drift methods do not analyze data in this respect, and detecting local drifts diverts them from considering long-term incremental changes. This is particularly due to resetting the detection parameters after detecting a drift [9]. In our analysis, despite detecting local drifts, the long-term gradual shift from negative to neutral polarity went undetected. In a separate experiment, where we manually summarized polarity of neighboring data to decrease the granularity, HDDM fell short to detect the drift.

Many well-known concept drift methods are restricted to univariate data, which hinders them from considering other data features simultaneously. In addition, the class is usually binary and multi-class problems are less investigated [12]. In general, it is required that the concept drift methods deal with heterogeneous features, and consider externally computed features like the topic. Another concern is the imbalanced data streams, in which some classes are more prevalent. In our dataset more than 40% of the posts were tagged as neutral, while the positive class contained slightly larger than 20% of the data. The concept drift detection techniques may be biased in such conditions. The last point to consider here is that the outcome of concept drift algorithms should have more detailed description of the distribution under change and the direction of change. This information is vital in analyzing real-world datasets in different applications, for both explaining the results and verifying them.

6. Summary and general analysis

Our goal was to provide a framework for highlighting the social discourse and its dynamics in Iran during the first four months of the COVID-19 outbreak. We aimed to determine how the appropriate events and happenings affect the social discourse, discover and understand socio-cultural realities based on analyzing the prevailing attitude of the Iranian society towards the Corona issue, and their attention to scientific, religious, health and medical subjects. Previous studies have confirmed the large effect of social media on social awareness, mental health, and behavior after the outbreak [5,7,16,44]. One study found out that despite the stay-at-home orders issued by the organizations, the role of friends and families should be emphasized to promote compliance to the protocols [45]. As such relationships are mainly realized in social media, systematic and structural analysis of its content is very important [34].

Three general steps have been used in a framework to achieve the mentioned goal. Topic modeling is shown to be effective in organizing online social contents for a deeper understanding of social thinking and behavior [17,31,32]. However, analyzing the dynamics in the thematic structure of such content is less studied [18,19]. We achieve this goal by first applying LDA on temporal partitions of the dataset, which showed important and consequent issues, and the approach of the people and officials to the corona crisis in each of the media. Important topics on Instagram in late January included rise of Corona in Iran and China. In February, discussions on New Year-related topics and health recommendations are bold. Important topics of March are about economy, and health advice. Finally, in April, entertainment and shopping are two important topics in discourse of Instagram activists. According to WHO reports, two peaks of COVID-19 infections are observed in Iran during the target period of current study, on March 23rd and June 1st, 2020. The first wave occurs at the beginning of the second month of the study. Referring to the major topics in the first month, the origins and rise of COVID-19 along with political issues are the main lines of discourse. Eventually, the health recommendations and prevention mechanisms come into the highlight. The dedicated contents to COVID-19 related health issues such as preventing the disease, washing the hands, properly use a mask, strengthen the immune system etc. are highly are receiving millions of views in this period [51]. Continuation of the disease, along with its unprecedented physical side effects and death tolls, and the announced preventive protocols, which are all circulated in social media discussions, may increase the stress level of individuals [45]. All of these conditions, along with the annual new year holidays, may have resulted in a more strict compliance to the preventive protocols specially social distancing, which reduced the infections after the initial peak of March 23rd.

In the news, the discourse about the outbreak of the Corona virus has been very colorful; from the function of the disease and its fatality and

[11] https://scikit-multiflow.readthedocs.io

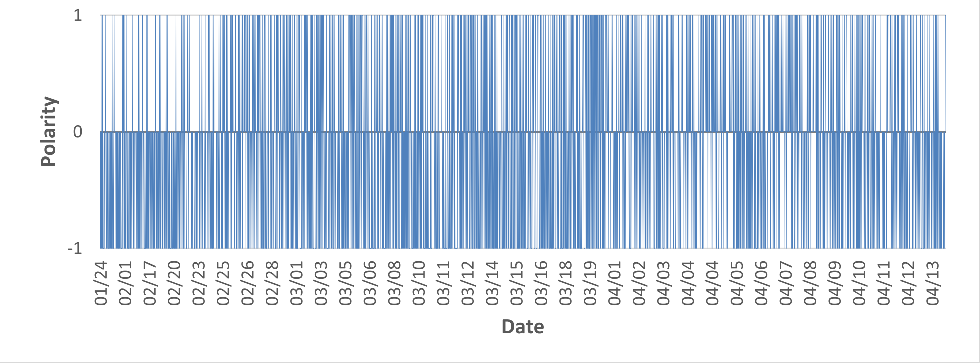

Fig. 6. The polarity during the target period.

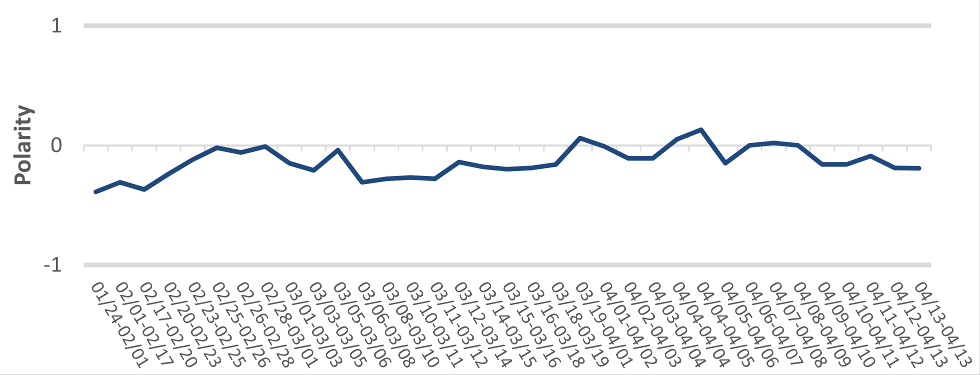

Fig. 7. The moving average over every 100 data records.

origin, to the government measures against it and preventive advisories. Other topics too have more diversity in news article. In January, petroleum, politics and election, the topics of online education, petroleum price and economics- are among the most important topics in the news discourse. From February onwards, election is also added to this list. The changes of discourse in the first two month is similar in news article and social media, however in the last two month people's discourse on social networks has not been in line with news article, while people discourse turn to issues such as entertainment and shopping, journalists are discussing about deaths and hospitals. The shift in social attention from health-related issues to the side effects of the disease, political events, and miscellaneous subjects, is apparent from the topics extracted from both social media and news in the third and fourth month under study. A detailed analysis of this phenomenon beside the decline in number of infections after the new year holidays, can clearly detect a decrease in level of social sensitivity to comply to preventive protocols. The effect is soon seen as the disease experiences its second wave in Iran in early June 2020.

In the second step, using dynamic topic modeling, we examined in more detail how the distribution of topics in different media had been changed over time. Thus, in addition to showing how the discourse changed, by having the time and type of topic changes, we were able to arrive at the reasons for the discourse change. For example, in news articles, in late April, economic issues intensify over medical issues, which could be due to the decision of the authorities to reopen the production centers, and the promised financial support [52]. In the telegram, with the beginning of Nowruz and the usual family visits, social concerns in the topics of people's discourse become very prominent. With the start of April and the holy month of Ramadan and fasting, religious topics have emerged on Twitter. The point to consider on Instagram is that, unlike the previous three media outlets, where medicine is the most important issue, entertainment and shopping issues are more important, and medical issues are of secondary importance. This variation in contents of social media platforms is also observed in the literature [5]. Putting these results next to the results of LDA topics, it can be said that at the beginning of the 100-day study period, attention is paid to margins such as the origin of the corona virus. Next, issues e.g. health advices, medical reports, risk groups, economy, and education are highlighted. At the end of this period, marginal issues such as entertainment and shopping come to the fore.

Afterwards, we focus on the polarity of the discourse and change of it during the studying period. Using the HDDM-A method, we detect changes in the concept of polarity across all media. The high percentage of data with negative polarity indicates that the subjects studied using the two methods of LDA and DTM, including medical, political, and social, are often appeared in a negative light. Of course, given the emerging and unknown corona phenomenon, this is not far from the expectation [7,20]. What is important is to strengthen social solidarity and national spirit through some strategies [5,28]. To find some solutions for this using only the provided data, positive drifts from negative/neutral posts are identified as the key content. Examination of the contents causing such drifts, leads to guidelines for both society and authorities, which are thoroughly discussed in Section 5.3.

Bringing all of these analyzes together, it should be borne in mind that the prevalence of corona as an unknown, emerging and pervasive phenomenon has caused a lot of stress in society [44]. Stress is a natural mental reaction, and no typical individual can claim the opposite. The partial downward slopes of the polarity chart can also indicate this increasing stress. Of course, if this stress encourages positive behaviors such as compliance to the protocols and monitoring personal health, it is beneficial and constructive as long as it does not exceed rational behavior and does not become pathological. Interestingly, these positive behaviors are very common in the topics extracted in February and March. Therefore, in could be said that a reasonable the stress caused by

Table 11
The results of exploring concept drift in the dataset.

Date	Change type	P, N, Ne*	Sample content (translated from Persian)	The key subject in change
Feb 5	N, Ne→P	N	The #Corona virus only affects the Chinese, because they eat all kinds of meat…	P: access to facilities
		P	The first special unit to control the corona virus was put into operation at the Afghan-Japanese Hospital in Kabul…	P: empathy and emotions
		P	When a monster makes distance between people and you have to embrace your loved ones from a distance…	
Feb 24	N, Ne→P	N	In my opinion, if you close the doors of the shrine, you have betrayed the holy site of Islam	P: Memorization and humor
		P	Let me share a memory about this new virus and the ways of transmission…	
Feb 26	Ne→P	P	If we are quarantined, we can make the environment happy like this…	P: Strengthening the spirit
		P	Army readiness to disinfect cities…	P: Building trust and access to facilities
		P	Our dear Iran has always nurtured zealous and selfless youth…	
		P	The Kuwaiti government minister went to the pharmacy anonymously and asked for a mask. The manager said that he did not have a mask…	P: Strengthening patriotism
Mar 1	N→P	N	… But it has hired Persian-speaking staff to introduce Iran as the cause of the corona outbreak in the form of the Saudi International Media …	P: Health advice P: Strengthening the spirit
		P	… If possible, if we do not have the necessary work, we should stay at home, and if we have to attend public places, there are definitely health tips …	
		P	Beautiful dance of a member of the hospital medical staff in Rasht Hospital…	
Mar 6	N, P→Ne	P	We defeat Corona with our high spirits	P: Strengthening the spirit
		N	Serious warning about bleach for corona disease	Ne: National news
		Ne	Larijani demanded that members of the National Corona Management Headquarters be sent to the provinces involved	
Mar 15	N→P	N	I hope we can travel carefree and happy soon.	N: Recalling previous conditions
		P	… In such a critical moment, these students came to the aid of us and the patients,	P: Characterization
March 19, 20	P→N	P	This year we could not buy fish because of Corona. Instead, I painted it and set it on the Haft Sin table. If you like, draw it too.	P: Positive advice N: Negative expression of news
		N	I want to start with my condolences because we lost many loved ones to the Corona virus	
Apr 4	Ne→P	Ne	Images of 65 specialists and medical staff of the country due to the Coronary disease, while serving their people with the least facilities…	P: Gratitude P: access to facilities
		P	… appreciated the activities of the municipality and the wisdom of the mayor of the center of Mazandaran in preventing the spread of the corona virus	
		P	Production of a mask for the deaf by a student in Kentucky, USA	
Apr 9	P→ Ne	P	A different celebration in the main and side streets of Hashtgerd with the cooperation of the municipality and the Islamic Council of the city	P: Strengthening the spirit and happiness
		Ne	Is heat the killer of the Corona virus?	
Apr 13	Ne→ P	Ne	Putin: W use military resources to fight the Corona if necessary	P: Slogans
		P	We defeat Corona	

* P: Positive, N: Negative, Ne: Neutral polarity, respectively

Covid-19 has led to positive behaviors in people in the community. Over time, the polarity diagram undergoes a serious change of concept and the slope of the diagram becomes increasing. This can in turn be an impact of the temporal decline in the infection rate, and may show prevalence of a neutral feeling in people of the community about the virus. This situation has a dual effect. In one hand, it can improve the mental health of individuals, which had experience large tensions after the outbreak. On the other hand, it can in turn reduce the level of stress that people are experiencing, and decrease the compliance to the protocols. Given the unknown nature of the virus and persistence of the disease, the neutral sentiment and its behavioral consequences may trigger undesirable disease conditions. This actually happened in the second wave of the disease as described earlier. This point is also confirmed in the results obtained from LDA, and DTM, where hot topics are changed from issues such as health care, medical subjects, and social impacts, to entertainment and shopping. The relationship between social topics and sentiment, is also confirmed in the literature [14,20].

Putting it altogether, there is a strict requirement for health authorities to monitor the current disease statistics, the current hot topics in online contents especially social media, and the social sentiment trend. Any change towards positive or negative in these cases should raise an alarm for preventive actions. Special attention should be paid to the differences found in the themes of contents in each social platform in current and other studies [5,23]. Our findings show that among the most important actions is promoting the social discourse on health issues, preventive protocols, and the realistic condition of the disease [15,53,54]. Actually, as long as the disease condition persist in a wide scale, such topics should remain in the spotlight.

7. Conclusions and future work

In this research, we proposed a framework to analyze the social discourse in Iran after the COVID-19 pandemic. A comprehensive dataset collected from various social media networks and news agencies was used for the analysis. In the proposed framework, we employed the Latent Dirichlet Allocation (LDA) model and dynamic topic modeling to extract the important topics as temporal subjects in the period under study. The extracted topics were analyzed, and their dynamics was traced to the underlying events and adopted policies. This analysis was performed both in general and individually for every social platform. We further investigated changes in polarization of the publicly published material, and by inspecting the contents around the drift points, derived practical guidelines on how to shift the polarization towards positive. The results exposed that the community is largely affected by the pandemic phenomenon, health-care, entertainment, medical, and economic issues are among the important concerns of the community. In addition, the general polarization of the discourse is mostly negative with a gradual shift in direction of neutral and positive. The observations reveal that the government and official decisions highly influence the social morale, and considering several guidelines in decision making and announcements can moderate the emotional state of the community. In future, the document influence modeling can be used to further assess the impact of news material on the emerging topics in social discourse. Moreover, using other dynamic clustering models in intervals around the influential events can assist in an improved semantic analysis of the subject changes. In addition, using social attributes of the published material such as the publisher and following information, can further boost the quality of cause and effect analysis for changes in social discourse.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Here we would like to present our thanks to the Social Networks Laboratory of University of Tehran[12] and the Natural Language Processing Laboratory of Shahid Beheshti University[13] for providing us with the dataset used in this research. The data is provided through the National Call for Recognition and Analysis of Discourse and Behavior of Iranian Society in the Corona Crisis. We also thank Dr. Maryam Khodabakhsh, and Mr. Shaban Gerduii for their guidance in summarizing the analysis from a psychological perspective. We thank Mr. Mustafa Baluchzehi for assisting in the concept drift detection analysis.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jbi.2021.103862.

References

[1] M.J. Paul, M. Dredze, D. Broniatowski, Twitter Improves Influenza Forecasting, PLoS Curr 1–12 (2014), https://doi.org/10.1371/currents.outbreaks.90b9ed0f59bae4ccaa683a39865d9117.

[2] T.R. Berry, J. Wharf-Higgins, P.J. Naylor, SARS wars: An examination of the quantity and construction of health information in the news media, Health Commun 21 (2007) 35–44, https://doi.org/10.1080/10410230701283322.

[3] J.S.P. Tulloch, R. Vivancos, R.M. Christley, et al., Mapping tweets to a known disease epidemiology; a case study of Lyme disease in the United Kingdom and Republic of Ireland, J Biomed Informatics X 4 (2019), 100060, https://doi.org/10.1016/j.yjbinx.2019.100060.

[4] S. Vijaykumar, G. Nowak, I. Himelboim, Y. Jin, Virtual Zika transmission after the first U.S. case: who said what and how it spread on Twitter, Am J Infect Control 46 (2018) 549–557, https://doi.org/10.1016/j.ajic.2017.10.015.

[5] H. Liang, I.C.H. Fung, Z.T.H. Tse, et al., How did Ebola information spread on twitter: Broadcasting or viral spreading? BMC Public Health 19 (2019) 1–11, https://doi.org/10.1186/s12889-019-6747-8.

[6] J.P.D. Guidry, Y. Jin, C.A. Orr, et al., Ebola on Instagram and Twitter: How health organizations address the health crisis in their social media engagement, Public Relat Rev 43 (2017) 477–486, https://doi.org/10.1016/j.pubrev.2017.04.009.

[7] E.K. Seltzer, E. Horst-Martz, M. Lu, R.M. Merchant, Public sentiment and discourse about Zika virus on Instagram, Public Health 150 (2017) 170–175, https://doi.org/10.1016/j.puhe.2017.07.015.

[8] D.M. Blei, L. Carin, D. Dunson, Probabilistic topic models, IEEE Signal Process Mag 27 (2010) 55–65, https://doi.org/10.1109/MSP.2010.938079.

[9] I. Frías-Blanco, J.D. Campo-Ávila, G. Ramos-Jiménez, et al., Online and Non-Parametric Drift Detection Methods Based on Hoeffding's Bounds, IEEE Trans Knowl Data Eng 27 (2015) 810–823, https://doi.org/10.1109/TKDE.2014.2345382.

[10] D.M. Blei, A.Y. Ng, M.T. Jordan, Latent dirichlet allocation, Adv Neural Inf Process Syst (2002) 1–8.

[11] D.M. Blei, J.D. Lafferty, Dynamic topic models, in: In: ACM International Conference Proceeding Series, 2006, pp. 113–120.

[12] R.S.M. Barros, S.G.T.C. Santos, A large-scale comparison of concept drift detectors, Inf Sci (Ny) 451–452 (2018) 348–370, https://doi.org/10.1016/j.ins.2018.04.014.

[13] J. Kwon, C. Grady, J.T. Feliciano, S.J. Fodeh, Defining Facets of Social Distancing during the COVID-19 Pandemic Twitter Analysis, J Biomed Inform (2020), https://doi.org/10.1016/j.jbi.2020.103601, 2020.04.26.20080937.

[14] H. Jelodar, Y. Wang, R. Orji, S. Huang, Deep Sentiment Classification and Topic Discovery on Novel Coronavirus or COVID-19 Online Discussions: NLP Using LSTM Recurrent Neural Network Approach, IEEE J Biomed Heal informatics 24 (2020) 2733–2742, https://doi.org/10.1109/JBHI.2020.3001216.

[15] X. Xiang, X. Lu, A. Halavanau, et al., Modern Senicide in the Face of a Pandemic: An Examination of Public Discourse and Sentiment About Older Adults and COVID-19 Using Machine Learning, Journals Gerontol Ser B XX:1–11. (2020), https://doi.org/10.1093/geronb/gbaa128.

[16] P. Bastani, M.A. Bahrami, COVID-19 Related Misinformation on Social Media: A Qualitative Study from Iran, J Med Internet Res (2020), https://doi.org/10.2196/18932.

[17] S.N. Saleh, C.U. Lehmann, S.A. McDonald, et al., Understanding public perception of coronavirus disease 2019 (COVID-19) social distancing on Twitter, Infect Control Hosp Epidemiol 2019 (2020) 1–8, https://doi.org/10.1017/ice.2020.406.

[18] H. Sha, M. Al Hasan, G. Mohler, P.J. Brantingham, Dynamic topic modeling of the COVID-19 Twitter narrative among U.S. governors and cabinet executives. (2020) arxive 2–7.

[19] C. Ordun, S. Purushotham, E. Raff, Exploratory Analysis of Covid-19 Tweets using Topic Modeling, UMAP, and DiGraphs. (2020) arxive.

[20] T. Wang, K. Lu, K.P. Chow, Q. Zhu, COVID-19 Sensing: Negative Sentiment Analysis on Social Media in China via BERT Model, IEEE Access 8 (2020) 138162–138169, https://doi.org/10.1109/ACCESS.2020.3012595.

[21] L. Li, Q. Zhang, X. Wang, et al., Characterizing the Propagation of Situational Information in Social Media During COVID-19 Epidemic: A Case Study on Weibo, IEEE Trans Comput Soc Syst 7 (2020) 556–562, https://doi.org/10.1109/TCSS.2020.2980007.

[22] V.Z. Marmarelis, Predictive Modeling of Covid-19 Data in the US: Adaptive Phase-Space Approach, IEEE Open J Eng Med Biol 1 (2020) 207–213, https://doi.org/10.1109/OJEMB.2020.3008313.

[23] H.W. Park, S. Park, M. Chong, Conversations and medical news frames on twitter: Infodemiological study on COVID-19 in South Korea, J Med Internet Res 22 (2020), https://doi.org/10.2196/18897.

[24] J. Samuel, G.G.M.N. Ali, M.M. Rahman, et al., COVID-19 public sentiment insights and machine learning for tweets classification, Inf 11 (2020) 1–22, https://doi.org/10.3390/info11060314.

[25] K. Chakraborty, S. Bhatia, S. Bhattacharyya, et al., Sentiment Analysis of COVID-19 tweets by Deep Learning Classifiers—A study to show how popularity is affecting accuracy in social media, Appl Soft Comput J 97 (2020), 106754, https://doi.org/10.1016/j.asoc.2020.106754.

[26] A.S. Imran, S.M. Daudpota, Z. Kastrati, R. Batra, Cross-Cultural Polarity and Emotion Detection Using Sentiment Analysis and Deep Learning on COVID-19 Related Tweets, IEEE Access 8 (2020) 181074–181090, https://doi.org/10.1109/access.2020.3027350.

[27] C. Huang, X. Xu, Y. Cai, et al., Mining the characteristics of COVID-19 patients in china: Analysis of social media posts, J Med Internet Res 22 (2020) 1–11, https://doi.org/10.2196/19087.

[28] H.R. Rao, N. Vemprala, P. Akello, R. Valecha, Retweets of officials' alarming vs reassuring messages during the COVID-19 pandemic: Implications for crisis management, Int J Inf Manage 55 (2020), 102187, https://doi.org/10.1016/j.ijinfomgt.2020.102187.

[29] I.E. Agbehadji, B.O. Awuzie, A.B. Ngowi, R.C. Millham, Review of big data analytics, artificial intelligence and nature-inspired computing models towards accurate detection of COVID-19 pandemic cases and contact tracing, Int J Environ Res Public Health 17 (2020) 1–16, https://doi.org/10.3390/ijerph17155330.

[30] D. Dimitrov, E. Baran, P. Fafalios, et al. TweetsCOV19 - A Knowledge Base of Semantically Annotated Tweets about the COVID-19 Pandemic. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management. ACM, New York, NY, USA, (2020) pp 2991–2998.

[31] Q. Liu, Z. Zheng, J. Zheng, 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 (2020), https://doi.org/10.2196/19118.

[32] D.C. Stokes, A. Andy, S.C. Guntuku, et al., Public Priorities and Concerns Regarding COVID-19 in an Online Discussion Forum: Longitudinal Topic Modeling, J Gen Intern Med 35 (2020) 2244–2247, https://doi.org/10.1007/s11606-020-05889-w.

[33] Santis E. De, A. Martino, A. Rizzi, An Infoveillance System for Detecting and Tracking Relevant Topics From Italian Tweets During the COVID-19 Event, IEEE Access 8 (2020) 132527–132538, https://doi.org/10.1109/ACCESS.2020.3010033.

[34] B. Jiang, X. You, K. Li, et al., Interactive Analysis of Epidemic Situations Based on a Spatiotemporal Information Knowledge Graph of COVID-19, IEEE Access 1 (2020), https://doi.org/10.1109/ACCESS.2020.3033997.

[35] B. Beijnon, T. Ha, S. Kim, J.H. Kim, Examining user perceptions of smartwatch through dynamic topic modeling, Telemat Informatics (2017), https://doi.org/10.1016/j.tele.2017.05.011.

[36] I. Mele, S.A. Bahrainian, F. Crestani, Event mining and timeliness analysis from heterogeneous news streams, Inf Process Manag 56 (2019) 969–993, https://doi.org/10.1016/j.ipm.2019.02.003.

[37] P. Breen, J. Kelly, T. Heckman, S. Quinn, Mining Pre-Exposure Prophylaxis Trends in Social Media, in: In: 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), 2016, pp. 214–221.

[38] A. Suprem, C. Pu, EDNA-Covid: A Large-Scale Covid-19 Tweets Dataset Collected with the EDNA Streaming Toolkit. arxive (2020).

[39] C. Pu, A. Suprem, R.A. Lima, Challenges and Opportunities in Rapid Epidemic Information Propagation with Live Knowledge Aggregation from Social Media. arxive (2020).

[40] J. Qiang, Z. Qian, Y. Li, et al., Short Text Topic Modeling Techniques, Applications, and Performance: A Survey, IEEE Trans Knowl Data Eng 14 (2020), https://doi.org/10.1109/tkde.2020.2992485.

[12] https://ece.ut.ac.ir/en/%D8%B4%D8%A8%DA%A9%D9%87-%D9%87%D8%A7%DB%8C-%D8%A7%D8%AC%D8%AA%D9%85%D8%A7%D8%B9%DB%8C1

[13] http://en.sbu.ac.ir/Faculties/ComputerEngineering/Pages/Natural-Language-Processing-(NLP)-Lab.aspx

[41] M. Röder, A. Both, A. Hinneburg, Exploring the Space of Topic Coherence Measures. In: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. ACM, New York, NY, USA, (2015) pp 399–408.

[42] D. Mimno, H.M. Wallach, E. Talley, et al. Optimizing semantic coherence in topic models. EMNLP 2011 - Conf Empir Methods Nat Lang Process Proc Conf (2011) 262–272.

[43] B. Honarvar, K.B. Lankarani, A. Kharmandar, et al., Knowledge, Attitudes, Risk Perceptions, and Practices of Adults Toward COVID-19: a Population and Field-Based Study from Iran, Int J Public Health 65 (2020) 731–739, https://doi.org/10.1007/s00038-020-01406-2.

[44] A. Shokri, G. Moradi, B. Piroozi, et al., Perceived stress due to COVID-19 in Iran: Emphasizing the role of social networks, Med J Islam Repub Iran 34 (2020) 55, https://doi.org/10.34171/mjiri.34.55.

[45] T. Paykani, G.D. Zimet, R. Esmaeili, et al., Perceived social support and compliance with stay-at-home orders during the COVID-19 outbreak: evidence from Iran, BMC Public Health 20 (2020) 1650, https://doi.org/10.1186/s12889-020-09759-2.

[46] M.R. Ghadir, A. Ebrazeh, J. Khodadadi, et al. The COVID-19 Outbreak in Iran; The First Patient with a Definite Diagnosis. Arch Iran Med 23 (2020) 503–504. https://doi.org/10.34172/aim.2020.48.

[47] H.K. Kim, H. Kim, S. Cho, Bag-of-concepts: Comprehending document representation through clustering words in distributed representation, Neurocomputing 266 (2017) 336–352, https://doi.org/10.1016/j.neucom.2017.05.046.

[48] T. Mikolov, I. Sutskever, K. Chen, et al., Distributed representations ofwords and phrases and their compositionality, in: In: Advances in Neural Information Processing Systems, 2013, pp. 1–9.

[49] E.H. Huang, R. Socher, C.D. Manning, A.Y. Ng, Improving Word Representations via Global Context and Multiple Word Prototypes. (2012) 873–882.

[50] Z. Rahimi, M.M. Homayounpour, Expert Systems with Applications Tens-embedding : A Tensor-based document embedding method, Expert Syst Appl 162 (2020), 113770, https://doi.org/10.1016/j.eswa.2020.113770.

[51] M. Peyravi, M. Ahmadi Marzaleh, N. Shamspour, A. Soltani, Public Education and Electronic Awareness of the New Coronavirus (COVID-19): Experiences From Iran, Disaster Med. Public Health Prep. 14 (2020) e5–e6.

[52] R. Salimi, R. Gomar, B. Heshmati, The COVID-19 outbreak in Iran, J Glob Health 10 (2020) 10365, https://doi.org/10.7189/jogh.10.010365.

[53] H. Liu, Z. Chen, J. Tang, et al., Mapping the technology evolution path: a novel model for dynamic topic detection and tracking, Scientometrics (2020), https://doi.org/10.1007/s11192-020-03700-5.

[54] M. Angeline, Y. Safitri, A. Luthfia, Can the Damage be Undone? Analyzing Misinformation during COVID-19 Outbreak in Indonesia, in: In: 2020 International Conference on Information Management and Technology (ICIMTech), 2020, pp. 360–364.