Journal of Artificial Intelligence and Data Mining (JAIDM), Vol. 11, No. 2, 2023, 315-329.

Shahrood University of Technology

Journal of Artificial Intelligence and Data Mining (JAIDM)
Journal homepage: http://jad.shahroodut.ac.ir

Research paper

Coronavirus Incidence Rate Estimation from Social Media Data in Iran

Fahimeh Hafezi and Maryam Khodabakhsh*

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

Article Info

Article History:
Received 15 February 2023
Revised 29 March 2023
Accepted 31 May 2023

DOI:10.22044/jadm.2023.12723.2426

Keywords:
Coronavirus, Social media, Semantic similarity, Fear emotion, Hope feeling, Incidence rate estimation.

**Corresponding author: M_khodabakhsh@shahroodut.ac.ir (M. Khodabakhsh).*

Abstract

Coronavirus disease as a persistent epidemic of acute respiratory syndrome posed a challenge to global healthcare systems. Many people have been forced to stay in their homes due to unprecedented quarantine practices around the world. Since most people used social media during the Coronavirus epidemic, analyzing the user-generated social content can provide new insights and be a clue to track changes and their occurrence over time. An active area in this space is the prediction of new infected cases from Coronavirus-generated social content. Identifying the social content that relates to Coronavirus is a challenging task because a significant number of posts contain Coronavirus-related content but do not include hashtags or Corona-related words. Conversely, posts that have the hashtag or the word Corona but are not really related to the meaning of Coronavirus and are mostly promotional. In this paper, we propose a semantic approach based on word embedding techniques to model Corona and then introduce a new feature namely semantic similarity to measure the similarity of a given post to Corona in semantic space. Furthermore, we propose two other features, namely fear emotion and hope feeling to identify the Coronavirus-related posts. These features are used as statistical indicators in a regression model to estimate the new infected cases. We evaluate our features on the Persian dataset of Instagram posts, which was collected in the first wave of Coronavirus, and demonstrate that the consideration of the proposed features will lead to improved performance of the Coronavirus incidence rate estimation.

1. Introduction

In 2019, a respiratory illness caused by Sars-COV-2 was reported in Wuhan, China and was called Coronavirus (Covid 19) by the World Health Organization (WHO). Since Coveid_19 has spread rapidly around the world and conflicts have arisen in more countries, WHO has declared it a global epidemic in March 2020. There was also an outbreak of the virus in Iran [1]. This epidemic has created widespread social and economic disorders and becomes an important threat to human lives around the world [2]. Coronavirus disease with a high infection rate has provided a lot of problems for the health-care systems around the world. Increasing medical facilities requires public health organizations to forecast future cases of infection reliably. Different governments proposed prompt mitigatory actions such as lockdown, school closures, and social disengagements to decelerate the spread. Therefore, an accurate prediction model can help governments to adopt the right preventive policies and attracts many researchers [3]. Forati and Ghose [3] in 2021 explored the relation between geo-localised Twitter data and county-level COVID-19 incidence rates in the USA. They found that most epidemic peaks were accompanied by peaks in coronavirus-related online activity, and that counties that saw more fake news being shared struggled the most to implement necessary restrictions. This led them to conclude that misinformation did affect health-related behaviours.

In this paper, we are interested specifically in establishing a model to predict the Coronavirus incidence rate by using social media. Social media are accessible information-sharing platforms where many people share their experiences and thoughts publicly with their friends and seek information about health online. It is not just users who can benefit from social media during an outbreak of a new disease by seeking out symptoms, preventative measures, and treatment-related information, but also health organizations do so by developing and disseminating accurate educational materials [4]. Therefore, users' posts containing health information on social media are a valuable source of information for not only conducting analyses ranging from the evaluation of public emotion responses during disease outbreaks [5, 6] to the surveillance of public attitudes [7, 8] but also detection or prediction of new cases [9, 10]. Given that public health responses require early detection of outbreaks before they become widespread and having a prediction system can help with managing disease emergencies, making rapid policy decisions, and implementing intervention strategies, researchers have already extensively explored the possibility of using socially shared health information to build predictive models by analyzing the messages published on social media about disease signs and symptoms, transmission methods, and death reports.

Recent works in the prediction of Coronavirus outbreaks have tried to keep track of the volume of content that is relevant to Coronavirus and its relation to official cases or deaths. The primary method for retrieving Coronavirus-related content among all content that is shared on social media relies on keyword search [1, 10]. For instance, Gharavi *et al*. [10] filtered tweets based on two words "cough" and "fever". Although this method has shown impressive performance, it suffers from the lexical chasm problem which means that it ignores Coronavirus-related content without the determined keywords. Also, there is no guarantee that every content including the determined keywords will relate to Coronavirus. This paper focuses on predicting the Coronavirus outbreak referred to as Coronavirus incidence rate estimation based on the content published on Instagram, which is a popular platform among Iranian users. Our research's goal is to determine whether given posts published during the Coronavirus outbreak can be an indicator to estimate the Coronavirus incidence rate. Thus, we are focusing on and investigating two fundamental ideas in our work:

1. Given that retrieving based on term frequency would not be a suitable method for determining Coronavirus-related posts, we attempt to benefit from the impact of neural information retrieval techniques where the post can be retrieved as while it does not overlap in terminology with the chosen keywords. Our hypothesis is based on considering the semantic association between the posts and Corona keywords. This is accomplished by representing Corona using well-established word embedding techniques such as FastText [11] or Word2vec [12]. As a result, it is possible to determine whether a post is about Coronavirus or not and provide a link between posts and Corona-related keywords when the problem of the vocabulary mismatch occurs.

2. Another hypothesis is based on the fact that the outbreak of Coronavirus as an infectious disease could influence people's emotions and feelings [13]. For instance, fear was a predominant emotion during the outbreak [5] and users who are reporting content about Coronavirus usually display signs of this emotion on their posts. Therefore, feeling and emotion detection of the published content could be used as a good indicator for Coronavirus-related posts.

Our approach to estimate the Coronavirus incidence rate is building a regression model that predicts the new infected cases based on the number of posts that really relate to Coronavirus. By considering the above-mentioned hypothesis, we introduce three classes of features namely semantic similarity, fear emotion, and hope feeling to retrieve Coronavirus-related posts that each feature captures a specific aspect of posts. As we will show in the experiments, each feature can be effective in identifying Coronavirus-related posts and achieving better performance of estimation.

Our main innovations are:

- We introduce a set of novel features to retrieve the posts that relate to Coronavirus.

- In order to obtain a more accurate estimate, we demonstrate how each of these features can be incorporated into a prediction process.

- We evaluate our proposed features on a Persian dataset that is systematically collected from Instagram during the first wave in Iran, examine our results from a variety of viewpoints, and highlight areas where features can considerably enhance performance.

The remaining sections of the paper are as what follows: We review pertinent research on disease surveillance through social media, and then present the related works on the Coronavirus outbreak in the part after that. The problem definition and the explanation of the details of our proposed approach are covered in Sections 3 and 4, respectively. Section 5 presents our experiments, our dataset, baselines, and our findings. The paper is concluded in Section 6.

2. Related Works

This section begins by reviewing prior work on disease surveillance through social media, and then ends by presenting the related works on the Coronavirus outbreak.

2.1. Disease surveillance through social media

Social media has infiltrated the health-related domains. Today, many traditional public health activities have taken advantage of social media, such as health education, health promotion, and disease surveillance just to name a few. For instance, due to their ability to eliminate the physical barriers that typically prevent access to healthcare resources and support, social media are increasingly being used in public health education [14]. The US Centers for Disease Control and Prevention (CDC) used its Facebook page during the 2009 H1N1 outbreak to inform the public about the illness and the value of vaccination. Social media has shown its ability to enhance people's access to public health surveillance information by looking for timely and trustworthy data on the spread or severity of influenza, Ebola, Middle East respiratory syndrome Coronavirus, and Dengue virus [15-17]. Aramaki *et al.* [15] made an effort to use Twitter to track the occurrence of flu. The strong correlation between expected and actual results shows that tweets on Twitter may accurately reflect the actual incidence of events. Bodnar *et al.* [16] used well-known regression models such as Linear, multivariable, and SVM, to assess disease outbreaks from tweets. They applied regression methods to the total number of tweets that contain at least one keyword associated with a particular disease, in this case, the" flu," and came to the conclusion that even when irrelevant tweets and randomly generated datasets were used, regression methods could still reasonably predict the prevalence of a disease. Odlum *et al.* [17] used the tweets mentioning Ebola to track the dissemination of information, the early diagnosis of a sickness, and public perceptions and attitudes. To better understand the spread of disease and information during the outbreak, they used time series analysis and geographic visualization. They discovered that the number of tweets surged three to seven days after major news events.

2.2. **Coronavirus outbreaks**

Tsao *et al.* [2] surveyed several studies about using online social media platforms as it relates to Coronavirus and divided them into three categories including (1) spreading pandemic information rapidly via social media, (2) efforts made by public health groups to produce and disseminate documents informing the public with correct information, and (3) the tracking of public opinion, mental health, and case detection during the pandemic. The studies in the third category showed that data from social media was beneficial in systematic monitoring and provided useful insight to detect or predict Coronavirus outbreaks [9, 10].

Li *et al.* [18] retrieved trend data from Google Trends, Sina Weibo Index, and Baidu Search Index by searching "coronavirus" and "pneumonia" to study the prediction of Coronavirus cases at an early stage. They dissected the information utilizing lag correlation and observed that there is a most extreme relationship between trend data with the watchword coronavirus and the number of cases diagnosed at eight to twelve days before an expansion in affirmed Coronavirus cases in the three platforms. O'Leary [19] employed regression to show the number of Coronavirus cases and deaths by three factors of Wikipedia page views, the number of tweets, and Google Trends looks for coronavirus and COVID-19.

Yousefinaghani *et al.* [9] conducted an investigation to identify waves in the number of confirmed cases in the United States and Canada using Google searches online and Twitter data. Gharavi *et al.* [10] utilized the frequencies of the watchwords, for example, "coughing" and" fevers" on Twitter for model examination, with the finding that the genuine flare-up time can be five to nineteen days sooner than authoritatively announced. Before and during the Coronavirus epidemic, they looked at Twitter data at the state level across the United States for the most typical symptoms of Coronavirus, such as "cough" and "fever". They performed a regression analysis to find the connection between the number of tweets containing "cough" and "fever" and officially reported cases. As of late review [1] planned to foresee the frequency of Coronavirus in Iran by information that was gotten from the Google Trends site. Analysts utilized linear regression and long short-term memory (LSTM) to assess the quantity of positive Coronavirus cases and found that the best factors other than earlier day rate incorporated the frequency of searching on antiseptic topics, hand sanitizer, and handwashing. The aim of researchers in [20] was to explore whether COVID-19 incidence rates differed between counties according to their socioeconomic characteristics using a wide range of indicators. They trained gradient boosting models to predict the age-standardized incidence rates with the macrostructures of the counties and used SHapley Additive exPlanations (SHAP) values to characterize the 20 most prominent features in terms of negative/positive correlations with the outcome variable. In another work [21], the researchers introduced a method to compute the real diagnostic rate and the real incidence of COVID-19 in each European country, testing whether the key hypothesis of the method is fulfilled and, if slightly off, whether it would affect all countries in the same way. In other words, they provided a recipe for policymakers that they have shown to be correct, unbiased across countries, and useful to make inter-country comparisons, provided the evolution and prognosis of the disease in a patient is not strongly dependent on socio-economic factors, and only on age, sex and previous clinical history.

Although researchers are interested in Twitter to analyze the content about the Coronavirus pandemic, there are few studies on Instagram [6, 22-24]. The researchers in [6] used Instagram as a reliable source to analyze the response of social network users as far as various aspects including topic detection, sentiment analysis, emotions, and geo-temporal characteristics. They showed that the dominant sentiment reactions on social media are neutral, while the most discussed topics by social network users are health issues.

Jafarinejad *et al.* [24] proposed a framework for analyzing the behavior of social media users in reaction to Coronavirus-related news and status. They used LDA, as the most established topic model to reveal the probable changes in topics of discourse. Their research is based on an extensive Persian data set gathered from different social media platforms and news agencies during the first wave in Iran. Another work on Persian data was done by Niknam *et al*. [22] who tried to characterize the representation of public health information related to Coronavirus posted on Instagram in Iran. They found that A total of 23 subjects rose up out of the examination of the posts containing “Corona” and “Covid-19” keywords including training and caring, general prevention guidelines, epidemiology and statistics, healthy diet, hygiene, and lifestyle just to name a few.

We would like to remark that our work sets itself apart from the current body of literature as our main goal is to propose an approach to estimate the Coronavirus incidence rate (the new infected case) based on three features that capture the specific characteristics of Instagram posts in Persian. One of the downsides of the most relevant literature to us [1, 10] in predicting the Coronavirus incidence rate is that it filtered the content based on Coronavirus related words. This method suffers from the lexical chasm problem which means that it ignores Coronavirus-related content without the determined keywords. Also, there is no guarantee that every content including the determined keywords will relate to Coronavirus. So, we attempt to benefit from the impact of neural information retrieval techniques to propose features based on considering the semantic association between the posts and Corona keywords. We will show that by using such feature, the performance of estimation will significantly improve over the existing state-of-the-art baseline.

3. Problem Definition

According to the recent studies [25], trending topics on social media can quickly alter in response to happenings in the real world and social media content generated by users can be seen as triggers for different event predictions such as disease outbreaks [26]. Our main objective is to investigate how user-generated contents can be used to get instant feedback regarding the incidence of Coronavirus. We believe that user-generated contents that are related to Coronavirus can be a symptom for tracking its changes and incidence rate during the main Coronavirus period. We refer to user-generated content on Instagram as posts and define it formally as follows:

Definition 3.1 (Post). A post, p = (text, time) is duple where p.text and p.time are the post content and it’s posting time, respectively.

It is common practice to represent the post content, p.text, as a bag of words included in it. Now, based on the fact that user’s express content about Corona by adopting words that are implicitly or explicitly related to Corona, we formalize Corona as follows:

Definition 3.2 (Corona). Corona is a bag of discriminative words related to Corona, denoted as, Corona = $\{w_1, w_2, ..., w_n\}$.

As already discussed in [9, 10], it is feasible to determine the set of such words that discriminate towards Corona. In order to detect Coronavirus-related content, [10] gathered tweets on the 2020 COVID-19 pandemic that mention the cough and fever that are the most typical Coronavirus symptoms and are geolocated in the United States. In [9], the application programming interface (API) for Twitter Premium Search was utilized to retrieve tweets with Coronavirus symptoms and treatment advice that were uploaded from the predetermined regions. In our work, in order to determine such words, we use snowballing to identify a set of hashtags related to Corona as shown in the first row of Table 1. The process of building the first seeds is completely automatic. We automatically built the initial set of hashtags by searching the keyword of "کرونا" (Corona) in https://keywordtool.io/. We then retrieve a set of Instagram posts that have these hashtags attached and review them to ensure that they are in fact Coronavirus-related posts. There are many cases where posts such as:

"پیشگیری از **ویروس کرونا** اسپری های ضدعفونی کننده سطوح
و دست بدون حساسیت برای همه افراد از بین برنده ۹۹,۹۹ درصد
انواع میکروب‌ها و ویروس‌ها و باکتری‌ها خوشبو کننده هوا با رایحه
های مختلف **سفارش محصول** دایرکت ... #کروناویروس #ایران
#ویروس کرونا #آنفولانزا# ضدعفونی #اسپری #قرنطینه #اصفهان
#تبریز #شمال #قم"

"Prevention of Coronavirus Antiseptic surface and hand disinfectant sprays for all people, eliminate 99.99% of all types of germs, viruses and bacteria. Air freshener with different fragrances. Direct **product order**... **#Coronavirus** #Iran #Corona Virus #Flu #Interfection #Spray #Quarantine #Isfahan #Tabriz #North #Qom"

Are retrieved since they contain pertinent hashtags, but they do not concern Corona. We eliminate such posts and use TFIDF of terms to determine words unique to Corona. Based on this procedure, we choose the top-10 words, as indicated in the second row of Table 1.

After determining the top-10 words, the word embedding techniques are utilized to find words that are similar to the set of top-10 words semantically. Our goal is to find words that are closest to the set of top-10 words based on their similarity in the embedding space. Words with similar semantic or syntactic features tend to be near together in this space, demonstrating the semantic and syntactic significance of the semantic space's geometric properties. There are various models for learning word embeddings including Word2vec [12], Glove [27], and FastText [11]. On this foundation, Corona is eventually constructed by incorporating the top-10 words determined in the preceding stage and the words that are the most semantically related to the set of top-10 based on one of the three word embedding techniques as shown in the last row of Table 1.

Given the fact that the Coronavirus-related posts are very strong indicators to build a reliable predictive model that can estimate the rate of Coronavirus incidence in the time interval of Coronavirus, we define our problem as follows:

Definition 3.3. (Coronavirus Incidence Rate Estimation). Let T be the time interval of Coronavirus. Given a set of posts at time interval T denoted as P^T, our problem is to establish a Coronavirus incidence rate estimator, CIRE, that takes P^T as input and estimates the Coronavirus incidence rate.

Our proposed approach to deal with addressing the challenge characterized in Definition 3.3 comprises of two sections: post modeling and Coronavirus incidence rate estimation, in which the result of the initial part turns into the input of the subsequent one. Figure 1 shows an overview of the proposed approach to estimate the incidence rate of Coronavirus. As seen in the figure, our approach consists of two main layers. The first layer is responsible for modeling Corona as a bag of discriminative words related to Corona. To this end, a set of hashtags related to Corona is used to retrieve the Instagram posts that include these hashtags attached. Then the top-10 words unique to Corona are selected by TF-IDF method. After determining the top-10 words, the word embedding techniques are utilized to find words that are similar to the set of top-10 words semantically.

In the second layer, our aim is to estimate the incidence rate of Coronavirus by training a regression using three features of semantic similarity, fear emotion, and hope feeling from daily posts. In order to build semantic similarity, and based on the produced word vectors by neural word embeddings and an initial set of seed words representing Corona, we model posts and Corona as a collection of word vectors. In the next step, the similarity of Corona and an input post is computed by using the cosine similarity measure. Also, we employ a BERT-based classifier to extract the fear emotion and hope feeling for the given posts.

Based on three features, we extract statistical indicators as the number of Coronavirus-related posts, the number of posts with the fear emotion, and the number of posts with the hope feeling. Then we train a regression with any of three statistical indicators as its input and the number of infected cases as the output.

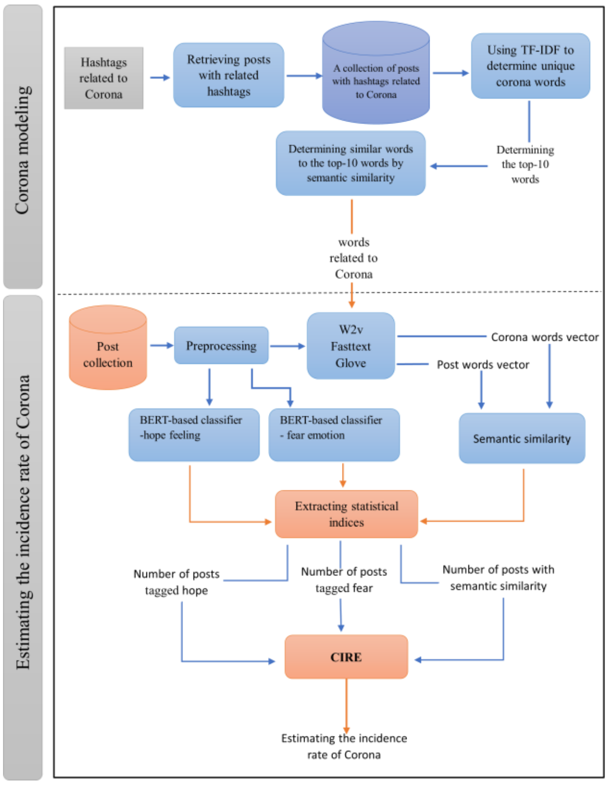
Corona modeling
Estimating the incidence rate of Corona
Hashtags related to Corona
Retrieving posts with related hashtags
A collection of posts with hashtags related to Corona
Using TF-IDF to determine unique corona words
Determining the top-10 words
Determining similar words to the top-10 words by semantic similarity
words related to Corona
Post collection
Preprocessing
W2v Fasttext Glove
Corona words vector
Post words vector
BERT-based classifier -hope feeling
BERT-based classifier - fear emotion
Semantic similarity
Extracting statistical indices
Number of posts tagged hope
Number of posts tagged fear
Number of posts with semantic similarity
CIRE
Estimating the incidence rate of Corona

Figure 1. Overview of the proposed approach.

4.1. Post-modeling

Our goal of post modeling is to extract features from posts in order to determine whether the posts mention Corona or not. We introduce three features:

1) Semantic similarity feature and 2) fear emotional feature and 3) hope feeling feature.

Double spacing. Table 1 provides samples of the appropriate type sizes and styles to use.

4.1.1. Semantic similarity

As discussed in [28-30], it is a common approach to represent the post content as a bag of words as well as Corona, and matching Corona and posts will therefore occur naturally. The basic idea of this research is that the more the frequency of Corona words in the post is, the more post content is related to Corona; the main drawback of the representation based on the bag of words is that it cannot address issues like vocabulary mismatch. A Significant number of posts are not retrieved due to the fact that they have no Corona words while their contents are related to Coronavirus.

Table 1. Bag of words for Corona.

Corona-related hashtags:	#کرونا ویروس(#Corona virus), # کرونا در ایران (#Corona in Iran), # کرونارا شکست میدهیم (#We will defeat Corona), # کرونا ویروس (#Corona Virus), # کرونا ایران (#Corona Iran), # کرونا بی (#Corona B), # کرونا بدر (#Corona Badr), # کرونا درایران (#Corona in Iran), # کرونا در زندان (#Corona in Prison), # کرونا دلتا (#Corona Delta), #کرونا در ایران (#Corona in Iran), # کرونا در همیننزدیکیست (#Corona is nearby)	
Most co-occurring terms (top-10)	کرونا (Corona), ویروس (Virus), ایران (Iran), مردم (People), کروناویروس (Coronavirus), بیماری (Disease), شکست (Defeat), شیوع (Outbreak), دست (Hand), خانه (Home)	
Most semantically similar terms to the top-10	**Word2vec**	کرونا ویروس (Corona virus), ویروس (Virus), قرنطینه (Quarantine), کوید (Covid), کووید (Coovid), کویید (Coviid), ایران (Iran), کشور (Country), شهروندان (Citizens), کرونا در ایران (Corona in Iran), کرونا را شکست (Defeat Corona), درخانه (At Home), بیماری (Disease), ابتلا (Infection), شیوع ویروس (Outbreak of virus), بهداشت (Health), وزارت (Ministry), کرونادست (Corona hand), دست (Hand)
	Glove	#کرونا ویروس(#Coronavirus), # ویروس (#Virus), قرنطینه (Quarantine), خانه (Home), شکست (Defeat), کرونا (Corona), دست (Hand), پیشگیری (Prevention), کویید (Coviid), ستاد (Headquarters), وزارت (Ministry), بهداشت (Health), درخانه (At Home), واگیر (Contagious), کشور (Country), نگرانی (Concern), روز (Day), درمان (Treatmen), شناسایی (Identification), پزشکی (Medical)
	FastText	کروناویرورس (CoronaVirurs), ویروس (Virus), شیوویروس (Shivirus), کروناویرووس (Coronaviruus), کروناویرس (Coronavirs), ویرووس (Viruus), ویروسچین (Viruschin), ککرونا (Cacrona), کروناوییروس (Coronaviirus), خطربیماری (Disease risk), بیماریی (Disease), روندشیوع (Prevalence trend), درخانه (At Home), وزیربهداشت (Minister of Health), روزهفته (Weekday), روندبیماری (Disease trend), مردمها (Peoples), اایران (Iiran), دست (Hand), شیوع (Prevalence)

For instance, the following post is related to Coronavirus while its content does not overlap in terminology with the top-10 words related to Corona:

"حزب کمونیست حاکم بر #چین اعلام کرده که در دو روز اخیر
پیش از ۱۳ هزار مورد **ابتلای** مجدد به #**کووید ۱۹ مشاهده** کرده
و استان شانگهای رو تا اطلاع ثانوی به شدیدترین نوع **قرنطینه**
برده"

"The ruling Communist Party of #China has announced that it has seen 13,000 cases of **#Covid19 reinfection** in the last two days, and has taken Shanghai province into the most severe type of **quarantine** until further notice"

In the above example, one could see the words that are similar to top-10 words semantically such as *"Covid19"*, *"quarantine"* and *"infection"*. Therefore, in this research, we are looking for semantic features that can be used to identify the concepts related to Coronavirus that users use in their posts. For this purpose, we introduce the semantic similarity feature that computes the similarity between posts and Corona in semantic space and define it formally as follows:

Definition 4.1 (Semantic similarity). Given *Corona* and *p*, semantic similarity is the maximum cosine of the word embedding pairs in the Corona and p denoted as:

$$SemSim(Corona, p)= \max_{w \in Corona,\, w' \in p.text}\{cos\,((v(w), v(w')))\} \tag{1}$$

In this case, function v computes the word w's vector using the techniques of word embedding such as FastText, Glove, and Word2vec. In semantic similarity, we first compute the pairwise similarities between all posts and Corona word embeddings and then aggregate them by a maximum function as a typical method to predict similarity [31]. We assume that the maximum similarity captures the least amount of cost needed to match the post to Corona.

Based above definition, the content of the post, p.text, is considered as a bag of word embeddings that can be obtained from word embedding techniques that take into account the content's syntactic and semantic aspects and so demonstrate their effectiveness for recognizing Coronavirus. Such techniques are beneficial due to consider semantically association between the spaces of post and Corona words and therefore can fill the gap between these two spaces when an observation is made about the vocabulary mismatch problem.

We assume that if Corona's words are accurately and correctly chosen, users will utilize terms or words close to Corona's representation in the embedding space. It is possible to compute the distance between each post and Corona under consideration based on the knowledge that Corona can be located based on the location of its words inside the embedding space and that a post can also be located in the same way in the same space.

4.1.2. Fear emotion

The outbreak of an epidemic could influence users' emotions, such as fear and anger have been observed as a result of the Ebola outbreak [32]. Recent studies show that living with the risk of becoming infected with Coronavirus was permeated with fear of death in the first wave of Coronavirus. As discussed in [5, 33], fear could be mentioned as the predominant emotion of people in the Coronavirus period. Su and colleagues [5] used psycholinguistic variables extracted from Weibo posts to investigate public emotional reactions to Coronavirus and found that fear is the most common emotion of users in the Coronavirus period. Murray *et al*. [33] used information from Reddit to assess the Coronavirus outbreak from the standpoint of personal experiences. The authors of this study used Reddit data to produce significant insights on disease symptoms as well as feelings experienced throughout the outbreak and found that fear is prevalent in the Coronavirus period. Given the fact that fear is the most important emotional response of people to Coronavirus, we can use it as a discriminative feature to distinguish Coronavirus-related posts. Our intuition behind this feature is that the posts in which the Coronavirus-related concepts are expressed with fear have strong discriminative power for Coronavirus incident rate estimation. We define the fear emotion feature as follows:

Definition 4.2 (Fear emotion) fear emotion for a given post p is a binary function that takes *p.text* as input and returns true when the fear emotion exists in *p.text.*

One could see the emotional extraction from posts as a text classification task [34] conducted by extracting linguistic features. Kleinberg *et al*. [34] developed a real-world worry dataset based on the emotional responses of UK residents to Coronavirus and then applied linear regression to predict the reported emotional values (i.e., fear, anxiety, worry, sadness) based on TF-IDF and part-of-speech (POS) features extracted from text. Due t the recent advancements in deep learning as well as large pre-trained language models, some researchers used BERT-based [35] classifiers in order to analyze the users' emotions in the Coronavirus period [18, 36]. For instance, Li *et al*. [18] tried to build a multi-label emotion classifier based on BERT and study the evolution of public emotions towards the COVID-19 through the Twitter tweets collected.

In our work, we employ a BERT-based classifier to extract the fear emotion for a given post. In other words, our solution to build the binary function is to train a BERT-based classifier that takes the content of the posts as inputs and returns a value of '1' when the fear emotion exists in the posts' content. The reason we adopt Bert for a binary classifier is that Bert has significantly enhanced the majority of natural language processing (NLP) tasks like emotion classification and demonstrated its potency in learning generic semantic representations [37]. Fear emotion classification using BERT includes two phases: pre-training for language understanding, and finetuning for the fear emotion classification task. In the pre-training phase, the unsupervised objectives are considered at the word level, such as next-word prediction, permutation, or masking strategy in order to understand language. We then extend the model by adding a fully connected layer and a SoftMax layer and fine-tune it on the fear emotion classification task by performing supervised training on a fear emotion dataset.

4.1.3. Hope feeling

Studies [38, 39] have shown that there is hope as a positive feeling about a crisis such as illness. Coronavirus community researchers [40, 41] came to the same conclusion that in addition to negative emotional responses such as fear that were common during the first wave of Coronavirus, the feeling of hope was also evident in the users' posts, although hope feeling was less prevalent than other negative emotional responses. Hope was communicated corresponding to the build of a vaccine that users could take their regular day-to-day life back to normal. Following examples are the post of the users who expressed hope for building a vaccine:

"راستش تنها **امید** منم برای درمان بیماری ناشی از کروناویروس دانشمندان ایران زمین هستند دانشمندان ما ازتمام دنیا برترند همینطور که پزشکان و پرستاران ما سرترند **امیدوارم** به زودی خبر **ساخت واکسن** و درمان این بیماری را از طرف دانشمندان خودمون بشنویم"

"Honestly, my only **hope** for the treatment of the disease caused by the Coronavirus is the scientists of Iran. Our scientists are superior to the rest of the world, just as our doctors and nurses are superior, **I hope** we will soon hear from our scientists about **making a vaccine** and treating this disease"

"به **امید** خبرهای خوب و رسیدن **واکسن** به کشورمون"

"**Hoping** for good news and the arrival of the **vaccine** in our country."

The promising phrase "we will defeat Corona." was exhibited in the posts of the Iranian users during the first wave of Coronavirus. Based on the mentioned cases, we think that hope feeling could be a strong discriminative feature to identify the Coronavirus-related posts. Similar to the fear emotion feature in the previous subsection, we consider a BERT-based binary classifier in order to identify the hope feeling feature for a given post *p*.

4.2. Coronavirus incidence rate estimation

Let *T* be the time interval of Coronavirus. After modeling each published post in *T* based on three features of semantic similarity, fear emotion, and hope feeling, we need to build a predictive model, CIRE, based on machine learning algorithms to estimate the incidence rate in the time interval of Coronavirus. In order to accomplish this, we rely on temporal Coronavirus-related posts in order to predict the incidence rate. Recent studies have already shown that user-generated content on social media which is related to Coronavirus and is surrounding the Coronavirus pandemic is a powerful tool that provides invaluable crowd-sourced near real time data for estimating the Coronavirus incident rate and has a direct relationship with the number of infected cases [9, 10]. Therefore, to express the temporal dynamics of the number of infected cases, we divide the time interval T into shorter discrete time intervals (that is one day in our case) and extract statistical indicators in these time intervals using the posts which are made in each time interval separately.

Given P^T as the published posts in the time interval of Coronavirus and more specifically, for each time interval t: $1 \leq t \leq T$, we first extract three *statistical indicators* in that time interval corresponded to the semantic similarity, fear emotion, and hope feeling features namely as the number of Coronavirus-related posts, the number of posts with the fear emotion, and the number of posts with the hope feeling respectively and then predict the Coronavirus rate incidence based on three *statistical indicators* for that time interval *t*.

Our solution for establishing *CIRE* is to train a regression with any of three *statistical indicators* as its input and the number of infected cases as output.

Gharavi *et al*. [10] utilized a linear regression in order to connect between posts surrounding the Coronavirus pandemic expressing the most common symptoms of Coronavirus and officially reported positive cases at the state level in the US. In order to establish *CIRE*, we also adapt Support Vector Regression (SVR) [42] as a classical supervised learning technique that has been proven to be an effective tool in real-value function estimation. One of SVR's advantages is the power to treat high-dimensionality data which is efficiently achieved through kernel functions.

5. Experiments

Our experiments' main goal is to assess how well our proposed approach for estimating the Coronavirus incidence rate works. In order to do this, we first study which word embedding techniques would be most effective for estimating the Coronavirus incidence rate. Finally, we determine whether the proposed features enhance the performance in comparison with the baseline in the Coronavirus incidence rate estimation task. Our proposed approach is shown in Algorithm 1-3.

Algorithm 1. Corona modeling.

```
Input: H: The set of hashtags related to Corona.
Output: Corona_related_words = {w1, w2, …., wn}.
Corona Modeling (H) {
    Data = a set of the posts that have at least one hashtag of H;
    Top_10_words = TF-IDF (Data) [0:10];
    Corona_related_words = Select the most similar words to the top_10_ words by Word2Vec, FastText or Glove and build {w1, w2, …., wn};
    return Corona_related_words
}
```

Algorithm 2. Extraction of semantic similarity feature between post and Corona-related words.

```
Input: Post, Corona_related_words = {w1, w2, …., wn}.
Output: Semsim.
Semantic Similarity (Post, Corona_related_words) {
    Semsim = 0;
    for each w ∈ Corona-related words:
        for each w′∈ post.text:
            max = cos ((v(w), v(w′))) where v is the vector(.) of the word by a neural word embedding method such as Word2Vec
            if max > Semsim: Semsim = max;
    return Semsim
}
```

Algorithm 3. Estimating the incidence rate of Corona.

```
Input: P = <p1, p2 … >: A sequence of daily posts ordered by publish time.
Output: Daily Corona incidence rate.
CIRE (P) {
    Number_of_posts_related_to_Corona = 0;
    Number_of_posts_with_fear = 0;
    Number_of_posts_with_hope = 0;
    for each pi in P:
        Number_of_posts_with_fear = Number_of_posts_with_fear + Bert classifier for fear emotion (pi);
        Number_of_posts_with_hope = Number_of_posts_with_hope + Bert classifier for hope feeling (pi);
        Semsim = SemanticSimilarity (pi, Corona_related_words);
        if (Semsim > theta):
            Number_of_posts_related_to_Corona=Number_of_posts_related_to_Corona++;
    Daily_Corona_incidence_rate= SVR (Number_of_posts_with_fear, Number_of_posts_with_hope, Semsim);
    return Daily_Corona_incidence_rate
}
```

5.1. Dataset

In our experiments, we employed four corpus of Instagram posts. The first publicly available corpus includes 4,919,839 Persian language posts which were collected from January 21, 2020, to April 21, 2020. The posts in this corpus first were preprocessed by deleting English letters, emojis, and all Persian stop words and then utilized for training the vectors representing the words of the posts and Corona as explained earlier using the Gensim library [43].

The second corpus is a labeled Persian dataset consisting of 2,067 Coronavirus-related posts that are either labeled with *"fear"* or *"no-fear"*. The posts of this dataset were reviewed by the experts in order to determine whether the concepts related to Coronavirus are mentioned or discussed in the posts with fear or not. This corpus was developed to train the BERT-based classifier to extract the fear emotion from the posts. Like this dataset, we benefited from another labeled dataset of Persian posts that included instances of *"hope"* as well as negative samples.

The performance of BERT-based classifiers over two labeled datasets in terms of precision, recall, and fscore are shown in Table. As seen in the table, the precision of both BERT-based classifier is higher than random when keeping in mind that we are performing a 2-class prediction task and therefore a random classifier would perform at a 50% precision rate compared to the current precision of the BERT-based classifier for fear emotion and hope feeling which is 57% and 64% respectively.

Table2. Performance of BERT-based classifiers over two labeled datasets.

	Precision	Recall	Fscore
BERT-base classifier (fear)	0.57	0.81	0.67
BERT-base classifier (hope)	0.64	0.91	0.74

Finally, in order to assess the proposed approach, we collect posts containing the keyword "کرونا" (Corona) from the beginning of February 20, 2020, to June 20, 2020. In other words, the last corpus consists of Persian posts that the users published during the first wave of Coronavirus in Iran. Furthermore, we preprocessed the posts of this corpus like the first corpus. The details of the four corpora are shown in Table 3.

5.2. Experimental setup

We modeled both posts and Corona as a bag of words and then obtained each word vector from Word2vec, Glove, and FastText techniques that we implemented using Gensim [43]. In order to run the Word2vec, Glove, and FastText models, the parameter settings of the Gensim were used in default and the layer size was set to 100, 200, 300 and 400. The window size is set to 5 which means the words with a frequency of fewer than 5 times are removed.

Furthermore, to implement the BERT-based classifiers, we employed the pre-trained BERT [35] trained on the top 104 languages with the largest Wikipedia, often known as Bert-base-multilingual-Cased. Also we used a threshold, θ, to compute the number of Coronavirus-related posts in the time interval *t*. In other words, we excluded the posts whose semantic similarity to Corona was lower than θ. In our experiments, θ was set to 0.4, 0.5, 0.6, and 0.7. Also to implement CIRE, we used SVR implemented in Sklearn[1] with the "RBF" kernel. As mentioned before, we divide the time interval of the Coronavirus outbreak, *T*, into shorter discrete time intervals of one day in order to express the temporal dynamics of the number of infected cases which means that the parameter of t is set to one day in our experiments.

Table 3. Specification of the corpus.

		Label	Number of posts	Application
(1)	**Publicly corpus**	-	4,919,839	Train word embedding techniques
(2)	**Gold-standard dataset**	Fear	101	Train BERT-based classifier to extract the fear emotion
		No-fear	1,966	
(3)	**Gold-standard dataset**	Hope	191	Train BERT-based classifier to extract the hope feeling
		No-hope	1,882	
(4)	**Corona dataset**	-	761,434	The assessment of our proposed Approach

5.3. Baselines and evaluation metrics

Numerous studies have sought to monitor the amount of user-generated content and link it to actual cases [1, 10]. For instance, the authors in a recent work [10] believed that Twitter content as one of the informal sources could be analyzed to detect and track Coronavirus outbreaks. Before and during the Coronavirus epidemic, they looked at Twitter data at the state level across the United

[1] https://scikit-earn.org/stable/modules/generated/sklearn.svm.SVR.html

States for the most typical symptoms of Coronavirus, such as “cough” and “fever”. They performed a regression analysis to find the connection between the number of tweets containing “cough” and “fever” and officially reported cases. So, we adopted reference [10] as the state-of-the-art baseline for comparative analytics. Just that we used the number of posts containing "کرونا"(Corona) instead of “cough” and “fever” keywords for both our approach and the baseline in the experiments.

In terms of evaluation metrics, we compare the performance of our proposed approach with the baseline using the coefficient of determination [44], commonly known as R-squared (or R2). In calculating an R-squared, we use a 3-fold cross-validation strategy for our approach evaluation as well as the baseline.

5.4. Results and findings

In order to estimate the Coronavirus rate incidence, we obtained the daily new infected cases of Coronavirus (daily incidence) from the beginning of February 20, 2020, to June 20, 2020, in Iran from the website IRANOPENDATA[2] which published official statistics on Coronavirus in Iran. In other words, our goal is to predict the new infected cases of Coronavirus by using Instagram posts of Iranian users.

5.4.1 Efficiency of the word embedding techniques

The purpose of this subsection is to investigate which word embedding techniques are impactful on *CIRE* performance. To this end, we do some experiments in which the semantic similarity feature has been used in measuring the statistical indicator of the number of Coronavirus-related posts. We report the results of the experiments in terms of R-squared in Table 2-6 when the time interval t has been set to one day. In this experiment, we aimed to examine the association of the number of daily posts that are similar to Corona concepts semantically with the new daily cases of Coronavirus and benefited from three-word embedding techniques namely Word2vec, Glove, and FastText for semantic representation of the posts as well as Corona.

Based on Table 2-6, we first observe that word embedding techniques play an important role in the final performance of the *CIRE*. As seen in the table, depending on which word embedding technique is chosen for the semantic similarity feature, the overall effect on estimating performance may differ. Our first observation is that FastText provides poorer performance against Word2vec as well as Glove in estimating the number of new infected cases. We have a 5% better correlation when Word2vec or Glove is adopted in determining the semantic similarity instead of Fastext.

We further study the effectiveness of various thresholds on the performance of *CIRE*. We find that measuring the statistical indicator of the number of Coronavirus-related posts based on the threshold of 0.5 leads to the highest performance over Glove and FastText except for Word2vec. For instance, when using the Glove technique, a threshold of 0.5 shows 29%, 32%, 31%, and 29% performance over all embedding sizes. Similarly, when FastText is used, an improvement of 17%, 25%, 26%, and 28% are observed on all embedding sizes. In contrast, the threshold of 0.7 provides the best performance of the estimation over all embedding sizes for the Word2vec embedding technique.

Another important observation based on the results in Table 4. is that the embedding size could be effective on the final performance of the CIRE. We can conclude that the best performance is equal to 32% for both Word2vec and Glove and 28% for FastText when the embedding size has been set to 200 and 400, respectively, in estimating the daily new cases.

5.4.2. Impact of the proposed features

In this subsection, we aim to explore the impact of the proposed features on the new Coronavirus cases estimation task. We expect that the performance is improved when semantic similarity, fear emotion, and hope feeling are adopted in estimating the Coronavirus incidence rate. In our experiments, we compare the performance of the state-of-the-art baseline [10] with our proposed estimator when semantic similarity, fear emotion, and hope feeling features are used for computing the statistical indicators and training SVR regression. The results are reported for each feature separately in Table 5. Also, based on the findings in the previous subsection, we select the best case of each word embedding technique to report the performance.

[2] https://iranopendata.org/dataset/official-corona-statistics-in-iran-patients-deaths-recovered/resource/d119344f-88a6-4351-9e19-33919f025e3b3

The first observation is that the semantic similarity feature provides higher performance in comparison with the baseline method [10]. As mentioned before, the baseline used only the number of daily posts containing the “Corona” word as the feature which leads to poor estimation of the new infected cases.

Table 2. Performance of CIRE in terms of R- squared (R^2) based on Word2vec.

Word2vec		Embedding size			
		100	200	300	400
θ	**0.4**	0.29	0.27	0.27	0.27
	0.5	0.28	0.28	0.28	0.26
	0.6	0.18	0.21	0.22	0.22
	0.7	0.32	0.32	0.30	0.31

Table 3. Performance of CIRE in terms of R- squared (R^2) based on Glove.

Glove		Embedding size			
		100	200	300	400
θ	**0.4**	0.28	0.28	0.28	0.29
	0.5	0.29	0.32	0.31	0.29
	0.6	0.28	0.24	0.17	0.14
	0.7	0.15	0.11	0.14	0.14

The primary cause we can identify for this is that there may be no Coronavirus-related content in the posts containing the “Corona” word such as promotional posts which are consistently focused on selling disinfectant products. From the semantic similarity perspective, the posts containing the “Corona” word that is really related to Coronavirus are considered and irrelevant ones are removed which leads to noticeable improvement over the baseline. The FastText technique sees the lowest improvement while the Word2vec and Glove techniques show similar degrees of impact of 5%.

Second, we observe that the fear emotion feature leads to an improvement of 1% over the baseline. In order to explore the impact of the fear emotion, we use the number of daily posts that contain the “Corona” word and express the users’ fear as the only feature to train SVR in our experiments. Based on the studies [5, 33], one can find that the fear emotion is expressed when the users mentioned Coronavirus. Given the results of experiments that fear emotion has an insignificant impact on estimating the new infected cases, we can conclude that there is no reason that the posts containing the “Corona” word and expressing fear are really Coronavirus-related content.

The last observation is that there is no improvement over the baseline when the number of daily posts that contain the “Corona” word and reflect the hope feeling of users is adopted as the only independent variable to train SVR. In such cases, the presence of the hope feeling feature could lead to a lower desirability indicator and result in a poorer estimation performance.

Table 4. Performance of CIRE in terms of R-squared (R^2) based on FastText.

FastText		Embedding size			
		100	200	300	400
θ	**0.4**	0.24	0.17	0.18	0.18
	0.5	0.17	0.25	0.26	0.28
	0.6	0.20	0.19	0.19	0.17
	0.7	0.16	0.19	0.17	0.18

Table 5. R-squared comparison between our proposed estimator and the baseline inspired by [10].

			R-squared
	Baseline		0.27
Our proposed approach	Semantic similarity	Word2vec	0.32
		GloVe	0.32
		FastText	0.28
	Fear emotion		0.28
	Hope feeling		0.21

To further analysis, we have conducted experiments to explore the impact of previous time intervals on the performance of CIRE. Given that the semantic similarity feature exhibits the greatest performance improvement over the baseline in comparison to other features, we select the semantic similarity feature to perform the experiments. In these experiments, we benefit from the number of Coronavirus-related posts in a previous time interval as the independent variable to predict the new infected cases by SVR. In other words, the posts that are related to Coronavirus semantically and published in previous days are considered as the statistical indicators in the experiments. The performance of SVR is reported in Table 6 in terms of R-squared when Word2vec with the threshold of 0.7 and embedding size of 100 is adapted in determining the semantic similarity. Based on the results, we can observe that the performance of CIRE drops when longer time intervals are considered to extract the statistical indicators. Also, when the number of Coronavirus-related posts that were generated, last day is used to estimate today's new infected cases, the performance of CIRE is similar to the case that the number of daily Coronavirus-related posts is used as input feature (32%).

Finally, the residual plots of the baseline and the best variation of our proposed approach when the statistical indicator is based on the semantic similarity feature are shown in Figure 3 and Figure 2. As shown in the figures, the regression model is not appropriate for estimating the daily incidence rate of coronavirus for both baseline and our proposed approach.

Table 6. Performance of SVR in terms of R- squared (R2) based on Word2vec.

Time intervals	W2v (100) (0.7)
1 day before	0.33
3 days before	0.20
5 days before	0.10
7 days before	0.001

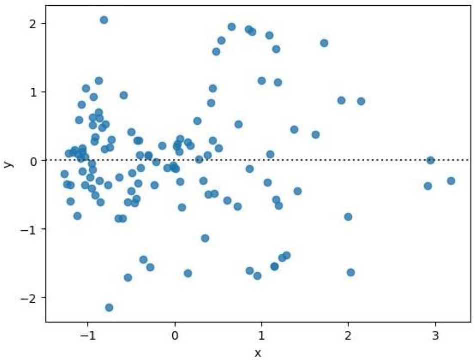

Figure 2. Residual plots of the method with semantic similarity.

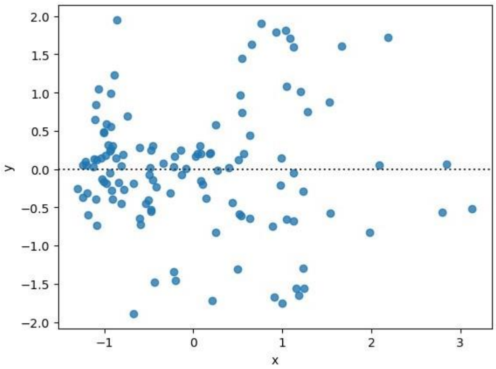

Figure 3. Residual plots of the baseline method

6. Conclusion

In this paper, we proposed an approach for the estimation of the Coronavirus incidence rate on microblogging platforms, most specifically Instagram in Iran. We have improved the state-of-the-art baseline for estimating the Coronavirus incidence rate on Instagram by retrieving Coronavirus-related posts based on three features namely semantic similarity, fear emotion, and hope feeling. The first feature calculates the Cosine similarity between embedding-based representations of Corona and a post. Two other features indicate whether fear emotion or hope feeling is expressed in a given post or not. Based on these features, we determined the number of daily posts that are related to Coronavirus and used it as a dependent variable in training a regression model to estimate the Coronavirus incidence rate. We have shown that the regression learned using this variable outperforms the existing baseline in terms of R-squared. Among the three features, semantic similarity shows the strongest performance improvement on the baseline.

References

[1] S.M. Ayyoubzadeh, S. M. Ayyoubzadeh, H. Zahed, M. Ahmadi, and S. R. N. Kalhori, “Predicting COVID-19 Incidence Through Analysis of Google Trends Data in Iran: Data Mining and Deep Learning Pilot Study,”

JMIR Public Health Surveill, vol. 6, no. 2, pp. e18828, 2020.

[2] S.-F. Tsao, H. Chen, T. Tisseverasinghe, Y. Yang, L. Li, Z. A. Butt, “What social media told us in the time of COVID-19: a scoping review,” *The Lancet Digital Health*, vol. 3, no. 3, pp. e175-e94, 2021.

[3] A. M. Forati and R. Ghose, “Geospatial analysis of misinformation in COVID-19 related tweets,” *Applied Geography*, vol. 133, pp. 102473. 2021.

[4] K. Rudra, A. Sharma, N. Ganguly, and M. Imran, “Classifying and summarizing information from microblogs during epidemics,” *Information Systems Frontiers*, vol. 20, no. 5, pp. 933– 948, 2018.

[5] Y. Su, P. Wu, S. Li, J. Xue, and T. Zhu, “Public emotion responses during covid-19 in China on social media: An observational study,” *Human Behavior and Emerging Technologies*, vol. 3, no. 1, pp. 127–136, 2021.

[6] E. Abdukhamidov, F. Juraev, M. Abuhamad, S. El-Sappagh, T. AbuHmed, “Sentiment Analysis of Users&rsquo; Reactions on Social Media during the Pandemic, ” *Electronics*, vol. 11, no. 10, pp. 1648, 2022.

[7] A. Abd-Alrazaq, D. Alhuwail, M. Househ, M. Hamdi, Z. Shah, *et al*. “Top Concerns of Tweeters During the COVID-19 Pandemic: Infoveillance Study,” *Journal of Medical Internet Research*, vol. 22, no. 4, pp. e19016, 2020.

[8] A. Al-Rawi, M. Siddiqi, R. Morgan, N. Vandan, J. Smith, C. Wenham, “COVID-19 and the Gendered Use of Emojis on Twitter: Infodemiology Study,” *Journal of Medical Internet Research*, vol. 22,no. 11, pp. e21646, 2020.

[9] S. Yousefinaghani, R. Dara, S. Mubareka, S. Sharif, “Prediction of COVID-19 Waves Using Social Media and Google Search: A Case Study of the US and Canada,” *Frontiers in Public Health*, vol. 9, pp. 359, 2021.

[10] E. Gharavi, N. Nazemi, and F. Dadgostari, “Early Outbreak Detection for Proactive Crisis Management Using Twitter Data: COVID-19 a Case Study in the US,” *arXiv preprint arXiv:2005.00475*, 2020.

[11] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, “Enriching Word Vectors with Subword Information,” *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135-46, 2017.

[12] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *Proceedings of Workshop at ICLR*, 2013.

[13] F. Amiri, S. Abbasi, and M. Babaie Mohamadeh, “Clustering Methods to Analyze Social Media Posts during Coronavirus Pandemic in Iran,” *Journal of AI and Data Mining*, vol. 10, no. 2, pp. 159-69, 2022.

[14] M. Stellefson, S. R. Paige, B. H. Chaney, J. D. Chaney, "Evolving Role of Social Media in Health Promotion: Updated Responsibilities for Health Education Specialists,” I*nternational Journal of Environmental Research and Public Health*, vol. 17, no. 4, 2020.

[15] E. Aramaki, S. Maskawa, and M. Morita, “Twitter catches the flu: detecting influenza epidemics using Twitter,” in *Proceedings of the 2011 Conference on empirical methods in natural language processing*, pp. 1568-1576, 2011.

[16] T. Bodnar and M. Salathé, “Validating models for disease detection using twitter,” in *Proceedings of the 22nd International Conference on World Wide Web*, pp. 699-702, 2013.

[17] M. Odlum and S. Yoon, “What can we learn about the Ebola outbreak from tweets?” *American journal of infection control*, vol. 43, no. 6, pp. 563-71, 2015.

[18] C. Li, L. J. Chen, X. Chen, M. Zhang, C. P. Pang, and H. Chen, “Retrospective analysis of the possibility of predicting the COVID-19 outbreak from Internet searches and social media data, China, 2020,” *Eurosurveillance*, vol. 25, no. 10, pp. 2000199, 2020.

[19] D. E. O'Leary and V. C. Storey, “A Google–Wikipedia–Twitter model as a leading indicator of the numbers of coronavirus deaths, ” *Intelligent Systems in Accounting, Finance and Management*, vol. 27, no. 3, pp. 151-8, 2020.

[20] G. Doblhammer, C. Reinke, and D. Kreft, “Social disparities in the first wave of COVID-19 incidence rates in Germany: a county-scale explainable machine learning approach,” *BMJ open*, vol. 12, no. 2, pp. e049852, 2022.

[21] M. Català, D. Pino, M. Marchena, P. Palacios, T. Urdiales, Cardona P-J, *et al*. “Robust estimation of diagnostic rate and real incidence of COVID-19 for European policymakers,” *PLoS One*, vol. 16, no. 1, pp. e0243701, 2021.

[22] F. Niknam, M. Samadbeik, F. Fatehi, M. Shirdel, M. Rezazadeh, and P. Bastani, “COVID-19 on Instagram: A content analysis of selected accounts,” *Health Policy and Technology*, vol. 10, no. 1, pp. 165-73, 2021.

[23] D. Amanatidis, I. Mylona, I. Kamenidou, S. Mamalis, and A. Stavrianea, “Mining textual and imagery instagram data during the COVID-19 pandemic,” *Applied Sciences*, vol. 11, no. 9, pp. 4281, 2021.

[24] F. Jafarinejad, M. Rahimi, and H. Mashayekhi, “Tracking and analysis of discourse dynamics and polarity during the early Corona pandemic in Iran,” *Journal of Biomedical Informatics*, vol. 121, pp. 103862, 2021.

[25] F. Abel, Q. Gao, G.-J. Houben, and K. Tao, “Analyzing temporal dynamics in twitter profiles for personalized recommendations in the social web,” *Proceedings of the 3rd international web science conference*, pp.1-8, 2011.

[26] K. Lee, A. Agrawal, and A. Choudhary, “Forecasting influenza levels using real-time social media streams,” *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, pp. 409-414, IEEE, 2017.

[27] J. Pennington, R. Socher, C. D. Manning, “Glove: Global vectors for word representation,” *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532-1543, 2014.

[28] C. Baydogan *et al.* “Deep-Cov19-hate: A textual-based novel approach for automatic detection of hate speech in online social networks throughout COVID-19 with shallow and deep learning models,” T*ehnički vjesnik*, vol. 29, no. 1, pp. 149-156, 2022.

[29] F. Es-Sabery, K. Es-Sabery, J. Qadir, B. Sainz-De-Abajo, A. Hair, B. Garcia-Zapirain, and I. De La Torre-D´ıez, “A MapReduce opinion mining for COVID-19-related tweets classification using enhanced ID3 decision tree classifier,” *IEEE Access*, vol. 9, pp. 58706-39, 2021.

[30] A. M. U. D. Khanday, Q. R. Khan, and S. T. Rabani, “Identifying propaganda from online social networks during COVID-19 using machine learning techniques,” *International Journal of Information Technology*, vol. 13, no. 1, pp. 115-122, 2021.

[31] T. Kenter and M. De Rijke, “Short text similarity with word embeddings,” *Proceedings of the 24th ACM international on conference on information and knowledge management*, pp. 1411-1420, 2015.

[32] J. M. Shultz, J. I. Cooper, F. Baingana, M. A. Oquendo, Z. Espinel, B. M. Althouse, *et al*, “The role of fear-related behaviors in the 2013–2016 West Africa Ebola virus disease outbreak,” *Current psychiatry reports*, vol. 18, no. 11, pp. 1-14, 2016.

[33] C. Murray, L. Mitchell, J. Tuke, and M. Macka, “Symptom extraction from the narratives of personal experiences with COVID-19 on Reddit,” *MAISON Workshop Proceedings of the 15th International AAAI Conference on Web and social media (ICWSM)*, 2020.

[34] B. Kleinberg, I. van der Vegt, M. Mozes, “Measuring emotions in the covid-19 real world worry dataset,” *Proceedings of the 1st Work-shop on NLP for COVID-19 at ACL*, 2020.

[35] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Hu-man Language Technologies*, vol. 1, pp. 4171-4186, 2018.

[36] M. Y. Kabir and S. Madria, “EMOCOV: Machine learning for emotion detection, analysis and visualization using COVID-19 tweets,” *Online Social Networks and Media*, vol. 23, pp. 100135, 2021.

[37] A. Chiorrini, C. Diamantini, A. Mircoli, D. Potena, “Emotion and sentiment analysis of tweets using BERT,” *EDBT/ICDT Workshop*s, 2021.

[38] J. Wang and L. Wei, “Fear and hope, bitter and sweet: Emotion sharing of cancer community on twitter,” *Social Media+ Society*, vol. 6,no. 1, pp. 2056305119897319, 2020.

[39] J. G. Myrick, A. E. Holton, I. Himelboim, B. Love, “# Stupidcancer: exploring a typology of social support and the role of emotional expression in a social media community,” *Health communication*, vol. 31, no. 5, pp. 596-605, 2016.

[40] J. G. Myrick and J. F. Willoughb, “A mixed methods inquiry into the role of Tom Hanks’ COVID-19 social media disclosure in shaping willingness to engage in prevention behaviors,” *Health Communication*, vol. 37, no. 7, pp. 824-32, 2022.

[41] C. A. Mousing, D. Sørensen, “Living with the risk of being infected: COPD patients' experiences during the coronavirus pandemic,” *Journal of clinical nursing*, vol. 30, no. 11-12, pp. 1719-29, 2021.

[42] M. Awad, R. Khanna, M. Awad, and R. Khanna, Support vector regression. Efficient learning machines: Theories, concepts, and applications for engineers and system designers, pp. 67-80, Springer nature, 2015.

[43] R. Řehůřek and P. Sojka, “Software framework for topic modelling with large corpora, ” *University of Malta*, 2010.

[44] A. Andreas, C. X. Mavromoustakis, G. Mastorakis, S. Mumtaz, J. M. Batalla, E. Pallis, “Modified machine learning Techique for curve fitting on regression models for COVID-19 projections,” *2020 IEEE 25th international workshop on computer aided modeling and design of communication links and networks (CAMAD)*, vol. 6, no. 2, pp. e18828, 2020.

برآورد نرخ بروز کرونا از داده‌های رسانه‌های اجتماعی در ایران

فهیمه حافظی و مریم خدابخش*

گروه مهندسی کامپیوتر، دانشکده مهندسی کامپیوتر، دانشگاه صنعتی شاهرود، شاهرود، ایران.

ارسال ۲۰۲۳/۰۲/۱۵؛ بازنگری ۲۰۲۳/۰۳/۲۹؛ پذیرش ۲۰۲۳/۰۵/۳۱

چکیده:

بیماری کروناویروس به عنوان یک اپیدمی مداوم سندرم حاد تنفسی چالشی را برای سیستم‌های مراقبت‌های بهداشتی جهانی ایجاد کرده است. از آن‌جایی که اکثر مردم در طول همه‌گیری ویروس کرونا از رسانه‌های اجتماعی استفاده می‌کردند، تجزیه و تحلیل محتوای اجتماعی تولید شده توسط کاربران می‌تواند بینش جدیدی برای ردیابی تغییرات و وقوع آن‌ها در طول زمان باشد. یک حوزه فعال در این فضا، پیش‌بینی موارد آلوده جدید از محتوای اجتماعی تولید شده توسط کرونا است. شناسایی محتوای اجتماعی مرتبط با ویروس کرونا یک کار چالش برانگیز است زیرا تعداد قابل توجهی از پست‌ها حاوی محتوای مرتبط با کرونا هستند اما هشتگ یا کلمات مرتبط با کرونا را شامل نمی‌شوند. برعکس، پست‌هایی که دارای هشتگ یا کلمه کرونا هستند اما واقعا به معنای کرونا مرتبط نیستند و بیشتر جنبه تبلیغاتی دارند. در این مقاله، ما یک رویکرد معنایی مبتنی بر تکنیک‌های جاسازی کلمه برای مدل‌سازی کرونا پیشنهاد می‌کنیم و سپس یک ویژگی جدید به نام شباهت معنایی را برای اندازه‌گیری شباهت یک پست داده‌شده به کرونا در فضای معنایی معرفی می‌کنیم. علاوه بر این، ما دو ویژگی دیگر یعنی هیجان ترس و احساس امید را برای شناسایی پست‌های مرتبط با کرونا پیشنهاد می‌کنیم. این ویژگی‌ها به عنوان شاخص‌های آماری در مدل رگرسیونی برای تخمین موارد آلوده جدید استفاده می‌شود. ما ویژگی‌های خود را در مجموعه داده فارسی پست‌های اینستاگرام، که در موج اول کرونا جمع‌آوری شده است، ارزیابی می‌کنیم و نشان می‌دهیم که در نظر گرفتن ویژگی‌های پیشنهادی منجر به بهبود عملکرد تخمین نرخ ابتلا به کرونا می‌شود.

کلمات‌کلیدی: رسانه‌های اجتماعی، شباهت معنایی، هیجان ترس، احساس امید، تخمین نرخ بروز.