# Analysis of Public Agenda during Covid-19 Pandemics Based on Turkish and English Tweets Using Nonnegative Matrix Factorization and Hypothesis Testing

**Mustafa Yavas[1], Aysun Guran[1, *], Yeliz Ekinci[2]**
*[1]Department of Computer Engineering, Dogus University,*
*34775 Istanbul, Turkey*
*[2]Department of Management Information Systems, Istanbul Bilgi University,*
*34060 Istanbul, Turkey*
*adogrusoz@dogus.edu.tr*

***Abstract*—In this study, Turkish and English tweets through Twitter Application Program Interface (API) between 1-31 January 2021 are analyzed with respect to Covid-19. The collected tweets are preprocessed, labeled with the Vader Sentiment library, and then analyzed by topic modeling with Nonnegative Matrix Factorization. The analysis show that the most frequently mentioned word is "vaccine/aşı" after "Covid". The topics modelled in the study are grouped into themes and the themes are seen to be similar in both languages, which means that the Turkish and world agenda are not very different in terms of themes in pandemics. Moreover, hypothesis tests are conducted to understand whether language and time period are related to sentiment class. The results show that the Turkish people are more neutral about the Covid-19 issue than other people in the world during the given period of time. Moreover, independent of the language, there are more negative and neutral tweets in the first half of January 2021, whereas there are more positive tweets in the second half of the month. To the best of our knowledge, this is the first study to analyze Covid-19 related tweets in two languages to compare the local and global agendas using topic modeling, sentiment analysis, and hypothesis testing methods.**

***Index Terms*—COVID-19; Chi-square analysis; Sentiment analysis; Topic modeling; Twitter.**

## I. INTRODUCTION

Since its initial outbreak in Wuhan, China, Covid-19 has spread throughout the world and has become a global pandemic. The World Health Organization (WHO) recognized the disease as a pandemic on 11 March 2020 [1]. The pandemic has affected many people with increasing morbidity and mortality [2]. To control the spread of the virus, several countries have implemented quarantine measures, travel bans, and restricted people's movement. In parallel, schools have been closed and many people have become unemployed due to the closures of the companies and recession economics. During these facts, activity on social media platforms such as Facebook, Twitter, and YouTube increased, since the users could express their concerns, opinions, and feelings about the pandemic in this way [1]. Furthermore, due to the fact that it is difficult to get reliable information during a pandemic, people usually try to get information on the Web or social media [3], [4]. Therefore, people have used the Web and social media to express themselves and to get information more than ever during the Covid-19 pandemic. This fact has resulted in huge data during this period.

Although classical surveys are useful to understand public health viewpoints, social media is increasingly used to discuss and share views on pandemics [5]. Twitter has especially become an important platform for obtaining public opinion and is widely used in public health research [6]. From this point on, it is obvious that using the huge data on Twitter will help us to understand the public opinion, feelings, and attitudes about the Covid-19 pandemic. The results of the analysis of these data can help policy makers and healthcare professionals to identify primary issues so that they can address them more appropriately [7]. Microblog text topic discovery and sentiment analysis are two methods that can be used in public opinion analysis [3], [8].

In this study, Turkish and English Covid-19 related tweets were collected for the time period between January 1 and January 31, 2021. We can see if there are significant differences between Turkey and the world by analyzing tweets in these two languages, assuming that English is the most widely used language on Twitter.

In this study, our first aim is to explore the attitudes of people by sentiment analysis. Our second aim is to analyze the relationships between sentiment values and time/language using hypothesis testing. Our third aim is to understand and compare the public opinion of Turkey and the world during the Covid-19 pandemic, using topic modeling.

After the preprocessing phase, sentiment analysis is performed on the tweets. To understand whether the attitudes of the people change over time, we apply the sentiment analysis for two different time periods. The first

Manuscript received 23 March, 2022; accepted 14 June, 2022.

research period covers the dates between 1 January 2021 and 15 January 2021. The second research period includes the dates between 16 January 2021 and 31 January 2021. We have seen that there are some differences between the Turkish and the English datasets over time in terms of sentiment analysis. Therefore, we apply a Chi-square test to determine whether there is a relationship between the sentiment values of the tweets (positive, negative, neutral) and the language (Turkish, English). The results of the analysis show that language is significantly related to the sentiment class. We also apply a Chi-square test to determine whether there is a relationship between sentiment values (positive, negative, neutral) and the time period (first half and second half of January). This relationship is also significant. Additionally, we apply topic modeling for the tweets in given two time periods; however, we have not seen any difference between the topics in these periods. Therefore, in this study, we report the results of the topic modeling analysis for the full month of January 2021 for both languages. We group the topics as themes, and it is seen that the public opinion in Turkish and English tweets is very similar, with small differences.

The literature is very scarce on studies that compare the agenda of different languages/countries, which can be seen by the summary of the related works in Section II. Moreover, related works also show that there are no studiesthat statistically test the relationships between sentiment values of posts/news and languages using hypothesis testing. Similarly, the literature lacks studies that analyze the relationship between the sentiment values of posts/news and the time period. These facts and gaps in the existing literature motivate us to conduct a study to understand the similarities and differences between the public sentiment and agenda of Turkey and the world about the Covid-19 pandemic in a specific period of time.

To the best of our knowledge, this is the first study to analyze Covid-19 related tweets in two languages (both Turkish and English in this study) to compare local and global agendas using topic modeling, sentiment analysis, and hypothesis testing methods.

We present our study in this paper with six more sections, including related studies, data preparation, topic modeling, hypothesis tests, experimental results, and conclusions.

## II. Related Works

In this section, an extensive summary of studies on social media data related to Covid-19 is conducted. Table I summarizes these studies. The social media data resource, the language of posts, the volume of data, the time period of the data collected, and the key findings are given in the table.

Table I shows that sentiment analysis and topic modeling methods are the most widely used techniques in social media analysis to understand the public agenda. O'Callaghan, Greene, Carthy, and Cunningham [17] state that nonnegative matrix factorization (NMF) has some advantages over latent Dirichlet allocation (LDA) methods. These advantages are: fewer parameter choices - allowing the method to be less complex - and the ability of identifying more coherent topics than LDA. Hence, NMF will be used in this study as well.

As can be seen in Table I, the vast majority of studies on Covid-19 analyze English-language social media platforms, but few examine foreign language platforms. Furthermore, none of the studies analyzes the relationship between the sentiment values of posts/news and languages, and none of them analyzes the relationship between the sentiment values and the time period. These analyses will be helpful to understand whether public opinion changes among languages and time. Therefore, there is a gap in the current literature in this area.

There are some studies that analyze only Turkish social media and news. One of them is the study of Sarıman and Mutaf [18]. They present a study written in Turkish language and identify the most common words of positive and negative tweets. Nazlı, Kocaömer, Beşbudak, and Köker [19] use a content analysis approach to analyze reactions of Turkish Twitter users to Covid-19 once the first positive case in Turkey was announced on 11 March 2020. Konakçı, Uran, and Erkin [20] use 160 news articles published between 11 March 2020 and 11 April 2020. Using a media context analysis method, they evaluate the news in the newspapers. They point out that 56.9 % of the news are informative and 6.3 % of them are advisory. 95 % of the news contain information on preventive or protective methods against Covid-19, 77.5 % of them mention complementary and alternative medicine methods. These studies also do not compare the Turkish and world agenda. Therefore, there is a gap in this area in the literature, not only for Turkey and world, but also for each country and the world.

To the best of our knowledge, this is the first study to analyze Covid-19 related tweets in Turkish an English to compare the world and Turkey's agendas using topic modeling, sentiment analysis, and hypothesis testing methods.

## III. Proposed Methodology

In this section, the proposed methodology is explained. To summarize the methodology, an overview of the framework is given in Fig. 1.

As is seen from Fig. 1, there are three main phases: data preparation, hypothesis tests, and topic modeling. These steps are explained in the following sections.

### *A. Data Preparation*

In this study, Turkish and English Covid-19 related tweets were collected from 1 January 2021 to 31 January 2021 using the Python programming language with Tweepy Application Program Interface (API). During the data collection phase, tweets containing the hashtags specified in Tables II and III were downloaded. Totally, 27,467 and 100,000 tweets are gathered for the Turkish and English languages, respectively.

Raw tweets without preprocessing phases appear to be highly unstructured and may contain redundant information. To handle these issues, preprocessing of tweets is performed by considering multiple steps. User names, stopwords, URLs, HTML tags, hashtags, and retweets are removed from tweets.

TABLE I. SUMMARY OF STUDIES ON SOCIAL MEDIA DATA RELATED TO THE COVID-19 OUTBREAK.

| **Papers** | **Social media platforms/ Languages** | **Dataset** | **Time period** | **Highlights of the papers** |
|---|---|---|---|---|
| Chandrasekaran, Mehta, Valkunde, and Moustakas, 2020 [1] | Twitter/English | 13.9 million tweets | 1 January 1 2020 to 9 May 2020 | The study gathers tweets between before and after the announcement of the Covid-19 disease. By using the LDA topic modeling method and the VADER sentiment analysis tool, the main themes and topics of the tweets are identified. They also explore the change and variation of related emotions in the process. |
| Abd-Alrazaq, Alhuwail, Househ, Hamdi, and Shah, 2020 [2] | Twitter/English | 167,073 tweets | 2 February 2020 to 15 March 2020 | Using the LDA topic modeling method, 12 different topics are identified and the topics are grouped into 4 themes (the origin of Covid-19; the sources of the virus; the impact of Covid-19 on people, countries, and the economy; and ways to reduce the risk of infection). |
| Li, Xu, Cuomo, Purushothaman, and Mackey, 2020 [3] | Weibo/Chinese | 115,299 posts | 23 December 2019 to 30 January 2020 | A significant positive correlation is reported between posts collected from the Chinese microblogging site Weibo and the number of reported cases within China and Hubei province. |
| Shen, Chen, Luo, Zhang, Feng, and Liao, 2020 [4] | Weibo/Chinese | 15 million posts | 1 November 2019 to 31 March 2020 | Human annotators label 11575 posts as an ingroup sick post and outgroup sick post. By training the annotated posts with different supervised machine learning classifiers, the “sick posts” are identified. These posts significantly predict daily case counts. |
| Sarker, Lakamana, Hogg-Bremer, Xie, Al-Garadi, and Yang, 2020 [7] | Twitter/English | 499,601 tweets | From late February to early April | Detect 203 Tweeter users who have a positive Covid-19 test result. When examining the symptoms mentioned by the users, the most common (fever/pyrexia, cough, body ache/pain, fatigue, headache, dyspnea) are reported. |
| Tao *et al*., 2020 [8] | Weibo/Chinese | 15,900 posts | 31 December 2019 to 16 March 2020 | Examined 15,900 Weibo posts on the supply and demand of dental services. It is reported that dental services, dental needs treatment, and home oral care information have changed dynamically during the pandemic. |
| Wahbeh, Nasralah, Al-Ramahi, and El-Gayar, 2020 [9] | Twitter/English | 10,096 tweets | 1 December 2019 to 1 April 2020 | The Covid-19 related tweets shared by 119 medical professionals are taken into consideration. Using a mixed method approach, eight topics are identified (actions and recommendations, fighting misinformation, information and knowledge, healthcare system, symptoms and illness, immunity, testing, and infection and transmission). |
| Budhwani and Sun, 2020 [10] | Twitter/English | 193,862 tweets | 9 March 2020 to 25 March 2020 | The number of tweets shared on Twitter about the “Chinese virus” or “China virus” is analyzed. A noticeable increase in the frequency of tweets shared on these issues has been reported. |
| Rufai and Bunce, 2020 [11] | Twitter/English, French, German, Italian, Japanese | 203 tweets belong to G7 country leaders | 17 November 2019 to 17 March 2020 | It is reported that out of 203 tweets, 166 tweets are “Informative”, 19 tweets are “Morale-boosting”, and 14 tweets are “Political”. |
| Alshalan, Al-Khalifa, Alsaeed, Al-Baity, and Alshalan, 2020 [12] | Arabic tweets/Arabic | 1 million tweets | 27 January 2020 to 30 April 2020 | The study analyzes Arabic tweets. According to the results, the rate of hate tweets is 3.2 %. 71.4 % of hate tweets contain low levels of hate. Most of the Covid-19 hate tweets are reported to be posted by people in Saudi Arabia. By applying the NMF topic modeling algorithm, seven topics are extracted. It is pointed out that most of the identified topics are related to hate speech against China and Iran. |
| Jang, Rempel, Roth, Carenini, and Janjua, 2021 [13] | Twitter/English | 123 million tweets | 28 January 2020 to 11 May 2020 | Explored public reactions and concerns regarding Covid-19 tweets from Canada. LDA topic modeling and aspect-based sentiment analysis methods are applied on the dataset. The results are interpreted with the public health professionals. |
| Chang, Monselise, and Yang, 2021 [14] | Twitter/English | 60.32 million tweets | March 2020 to June 2020 | The weekly list of the most discussed Covid-19 topics on Twitter is monitored. Two topic modeling algorithms are proposed: (Rolling-Online NMF) and (Sliding-ONMF). Some important themes (government policy, economic crisis, Covid-19-related updates, Covid-19-related events, prevention, vaccines and treatments, and Covid-19 testing) are extracted. |
| Meaney *et al*., 2022 [15] | Clinical text data/English | 382,666 primary care progress notes | 1 January 2017 to 31 December 2020 | Applying NMF topic modeling to 382,666 clinical text data, 50 topics are identified for policymakers to understand the holistic impacts of the COVID-19 pandemic on the primary healthcare system and community/public health. |
| Feng and Zhou, 2022 [16] | Twitter/English | 650,563 unique geo-tagged tweets | 25 January 2020 to 10 May 2020 | Conducting sentiment analysis steps, public emotions are measured. Tweet volumes are normalized based on COVID-19 case and death numbers. Finally, ten topics are summarized with LDA-based topic modeling. |

Also, repeating tweet messages are eliminated from both datasets. At the end of this process, the total number of Turkish tweets is reduced from 27,467 to 24,576, and the number of tweets in English is reduced from 100,000 to 90,104.

After the preprocessing phase, sentiment analysis is performed on the tweets. For this process, Valence Aware Dictionary and Sentiment Reasoner (VADER), which is a lexicon and rule-based sentiment analysis tool, is used. VADER works in conjunction with the NLTK Python package. Based on a lexicon, VADER returns a dictionary of sentiment scores in each of four categories (negative, neutral, positive, and compound). The compound score is calculated by summing up the valence scores of each word in the lexicon and normalized between -1 (most extreme negative) and +1 (most extreme positive) [21]. Table IV and Table V show positive ($P^{+}$), negative ($N^{-}$), neutral ($N^{0}$) scores, and compound dictionary rating values (C) of Turkish and English tweets. The last columns of the tables represent the sentiment labels of tweets (1: positive, 0: negative, 2: neutral).

We apply all the experiments for two different time periods. The first research period covers the dates between 1 January 2021 and 15 January 2021. The second research period includes the dates between 16 January 2021 and 31 January 2021. The following figures show the rate of positive, negative, and neutral tweets of Turkish and English datasets.

As can be seen from Fig. 2, the positive and neutral tweets increased in the second half of January 2021, in the Turkish dataset. Similarly, in the English dataset, the ratio of the positive tweets increased, while keeping the neutral tweets ratio unchanged (see Fig. 3).

Another result that can be derived by comparing Fig. 2 and Fig. 3 is that in the English dataset, the percentages of both the positive and negative tweets are higher than those of the Turkish dataset. The ratio of the neutral tweets in the English dataset is, accordingly, lower.

### *B. Topic Modeling*

Topic modeling approaches can be very useful in highlighting the various latent concepts of tweets. For our problem, a topic can be thought of as a cluster of similar tweets that talk about the same subject. To obtain the main topics in the corpus, we apply nonnegative matrix factorization (NMF) on the tweet-term matrix ($V$) [22]. We choose NMF because it is an unsupervised topic modeling algorithm that factorizes the given tweet-term matrix ($V$) into two matrices ($H$, $W$) that have no negative elements. Non-negativity provides more interpretable results.

Once tweets are represented with vectors using the term frequency and the inverse document frequency (tf-idf), we put all tweets together in a tweet-term matrix ($V^{twxt}$), which has dimensions of the number of tweets in the corpus ($tw$) by the number of terms ($t$). NMF decomposes the matrix $V^{twxt}$ into the product of two lower-rank matrices $H^{twxk}$, $W^{kxt}$

$$V^{twxt} \approx H^{twxk} W^{kxt}. \tag{1}$$

Equation (1) can be formulated by the following Frobenius norm optimization problem

$$\min \left\| V - HW \right\|_F^2, s.t. H \geq 0, W \geq 0. \tag{2}$$

Since (2) is not convex in $H$ and $W$ together, Lee and Seung [23] propose iterative multiplicative update rules to minimize this objective function.

The matrix $H$ represents the tweets ($tw$) in the topic space ($k$). Each column of $H$ depicts the degree of membership of each tweet in a given topic. On the other hand, the matrix $W$ gives the combinations of words that describe each topic.

To determine the number of topics, some quantitative ways are needed to evaluate whether the topics are interpretable and meaningful. We use the Topic Coherence-Word2Vec (TC-W2V) metric presented by [24] to detect the number of topics. TC-W2V measures the coherence between terms assigned to topics based on Word2Vec, which is a prediction-based word embedding method that can learn semantic relationships among nearby words within documents [25].

After applying the NMF topic modeling algorithm for different numbers of topics ($k$), we calculate the average TC-W2V for each $k$ value. And for both datasets, we use the $k$ with the highest average TC-W2V value to train a final NMF model.

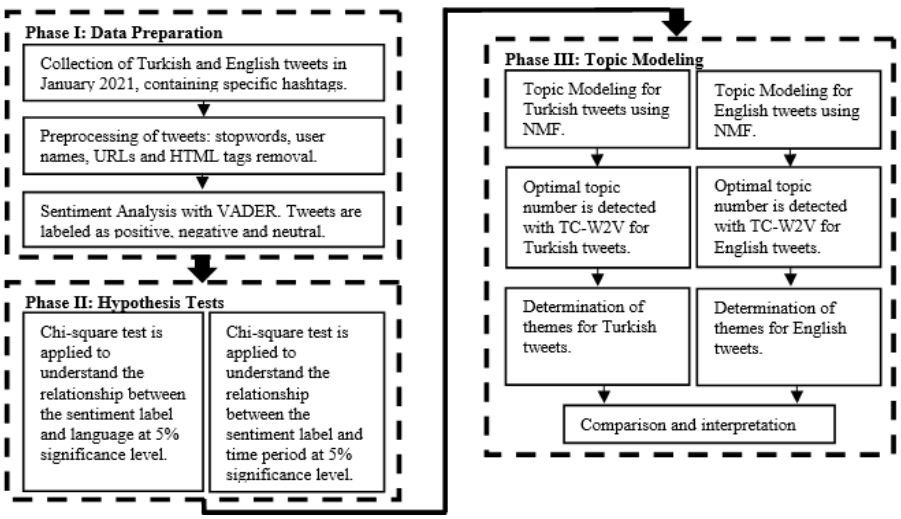


Fig. 1. Framework of the proposed methodology.

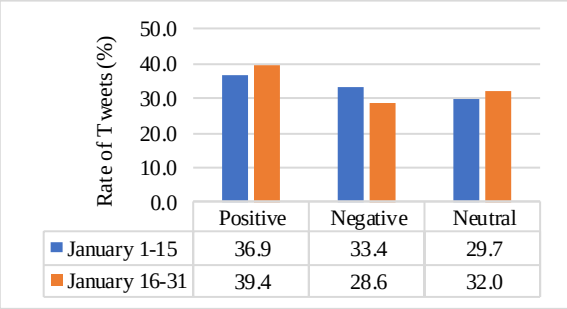


Fig. 2. The rate of the positive, negative, and neutral tweets of the Turkish dataset for two time periods.

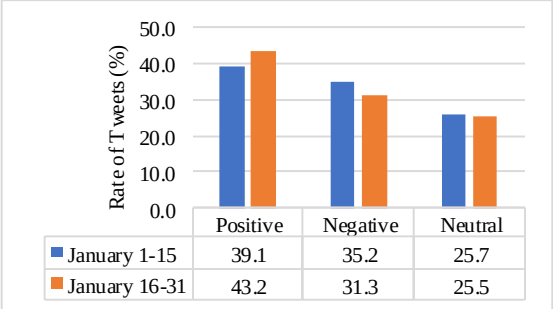


Fig. 3. The rate of the positive, negative, and neutral tweets of the English dataset for two time periods.

TABLE II. NUMBER OF THE COLLECTED COVID-19 TURKISH TWEETS CONTAINING SPECIFIED HASHTAGS.

| Time Period | Number of Tweets | Hashtags |
|---|---|---|
| 1-15.01.2021 | 14,366 | #covid19, #covid-19, #covid_19, #covid, #corona, #coronavirus, #ncov, #SAR-COV-2, #pandemic, #Korona, #koronavirüs, #coronaTürkiye |
| 16-31.01.2021 | 13,101 | |
| Total Number of Tweets | 27,467 | |

TABLE III. NUMBER OF THE COLLECTED COVID-19 ENGLISH TWEETS CONTAINING SPECIFIED HASHTAGS.

| Time Period | Number of Tweets | Hashtags |
|---|---|---|
| 1-15.01.2021 | 50,000 | #covid19, #covid-19, #covid_19, #covid, #corona, #coronavirus, #ncov, #SAR-COV-2, #pandemic |
| 16-31.01.2021 | 50,000 | |
| Total Number of Tweets | 100,000 | |

TABLE IV. EXAMPLES OF VADER SENTIMENT SCORES OF TURKISH TWEETS.

| Tweets | C | $N^-$ | $N^0$ | $P^+$ | S |
|---|---|---|---|---|---|
| COVID 19 hayatlara mal oldu, milyonları hasta etti ve küresel ekonomiyi yıkıcı bir resesyona sürükledi. | -0.214 | 0.219 | 0.655 | 0.126 | 0 |
| Çin Dışişleri Bakanı Wang Yi, Afrika ziyaretini tamamladı. COVID 19 salgınına göğüs geren ziyaret, özellikle Çin'in Afrika ile derin dostluğunu sürdürmeye verdiği önemi göstermiştir. | 0.687 | 0.0 | 0.832 | 0.168 | 1 |
| Sağlık Bakanlığı'nın 81 ile gönderdiği genelgede belirttiği kuralların uygulandığı koronavirüs aşılama odaları ilk kez görüntülendi. | 0.0 | 0.0 | 1.0 | 0.0 | 2 |

TABLE V. EXAMPLES OF VADER SENTIMENT SCORES OF ENGLISH TWEETS.

| Tweets | C | $N^-$ | $N^0$ | $P^+$ | S |
|---|---|---|---|---|---|
| Same thing happened to my dad unfortunately he didn't make it. Some people just don't know that COVID 19 is real and it can kill you or your loved ones so they just party and go anywhere as if there's no pandemic. | -0.347 | 0.166 | 0.712 | 0.122 | 0 |
| Last Thursday the head of Germany's vaccine regulator described the OAZ Covid 19 vaccine as excellent, there had been some debate around the best usage pattern but still the efficacy remains outstanding. | 0.858 | 0.0 | 0.74 | 0.26 | 1 |
| Dr. Brita talked to about the health system's streamlined scheduling for first and second doses of the COVID vaccine. | 0.0 | 0.0 | 1.0 | 0.0 | 2 |

## IV. RESULTS

### A. *Hypothesis Test Results for Sentiment Analysis*

In this study, one of the aims is to understand public opinion by analyzing Turkish and English tweets. Therefore, we will be able to see if there are significant differences between Turkey and the world agenda, assuming that English is the most widely used language on Twitter. We conducted a Chi-square analysis to determine the profile of the sentiment of tweets in terms of language and time period. Therefore, we derive the following hypothesis:

H1: There is a relationship between the sentiment values (positive, negative, neutral) of the tweets and the language (Turkish and English).

As can be seen in Table VI (to use an equal number of tweets in the hypothesis test, we selected 24000 tweets from each language randomly; therefore we have 48000 tweets in both hypothesis tests in Tables VI and VII), the results of the analysis show that language (Pearson chi-square = 121.866, p = 0.000, at the $p < 0.05$ significance level) is significantly related to the sentiment class. The cross-tabulation matrix shows that the negative and positive tweets are more in English, while the neutral tweets are more in Turkish. This can be interpreted as the Turkish people are more neutral about the Covid-19 issue than the world.

In addition, we apply a Chi-square test to determine if there is a relationship and dependence between sentiment values and the time period. Therefore, we derive the following hypothesis:

H2: There is a relationship between the sentiment values (positive, negative, neutral) of the tweets and the time period (first half and second half of January, 2021).

As can be seen from Table VII, the analysis results show

that the time period (Pearson chi-square = 155.269, p = 0.000, at the p < 0.05 significance level) is significantly related to the sentiment class. The cross-tabulation matrix shows that there are more negative and neutral tweets in the first half of January 2021, while there are more positive tweets in the second half of the month. The proportions given in Figs. 2 and 3 confirm this hypothesis as well. We see that the proportion of positive tweets increases in both languages in the second half of the month. The reason may be the increased use of vaccines as time passes.

TABLE VI. SENTIMENT LABEL AND LANGUAGE CROSS-TABULATION MATRIX.

| | Language | | Test Results | | | |
|---|---|---|---|---|---|---|
| **Sentiment Label** | **English** | **Turkish** | **N** | **$X^2$** | **Phi** | **P** |
| Negative count | 8183 | 7694 | 48000 | 121.866 | .050 | .000 |
| Neutral count | 5784 | 6849 | | | | |
| Positive count | 10033 | 9457 | | | | |

*Note: **N** = Number of samples, **$X^2$** = Pearson Chi-square, **Phi** = A measure of effect size, **P** = Probability of obtaining the test statistic.*

TABLE VII. SENTIMENT LABEL AND TIME PERIOD CROSS-TABULATION MATRIX.

| | Time Period | | Test Results | | | |
|---|---|---|---|---|---|---|
| **Sentiment Label** | **1-15 Jan** | **16-31 Jan** | **N** | **$X^2$** | **Phi** | **P** |
| Negative count | 8278 | 7599 | 48000 | 155.269 | .57 | .000 |
| Neutral count | 6646 | 5987 | | | | |
| Positive count | 9076 | 10414 | | | | |

*Note: **N** = Number of samples, **$X^2$** = Pearson Chi-square, **Phi** = A measure of effect size, **P** = Probability of obtaining the test statistic.*

### *B. Experimental Results on Topic Modeling*

We apply the topic modeling for the two time periods mentioned above. However, we have not seen any difference between the periods. Therefore, we report the results of the topic modeling analysis in this section for January 2021 (full month).

First, we analyze the frequency of Turkish and English keywords, which is given in Table VIII for the 15 most frequent keywords (the frequencies are given in parentheses). It can be seen from the results that *covid* and *vaccine (aşı)* are the most frequently used keywords in both Turkish and English. The pandemic (*pandemi* in Turkish) takes fifth and sixth place in both languages. Another parallel result is that *news* (*haber* in Turkish) takes the same place in both lists (they are 13th). It is interesting that the keywords *quarantine* and *mask* are not in the English list, while the Turkish translations of them, *karantina* and *maske* words, are in the Turkish list. However, a more negative word *death* is in the list of English corpus, while the Turkish translation of it, *ölüm*, is not.

We use the TC-W2V score to determine the number of topics. Topic coherence measures the consistency of each topic by measuring the semantic similarity between words with high scores on a topic. Words are represented with Word2Vec vectors, and semantic similarity is calculated as the cosine similarity between word vectors [25]. The coherence is the arithmetic mean of these similarities [24]. During the application of Word2Vec algorithm, the vector length is set to 100, the window size is determined as 5, and the words occuring at least 5 times are taken into consideration.

In the literature, there is no consensus on the optimal number of topics, regarding the coherence values. However, reviewing the studies on topic modeling related to Covid-19, we have seen that meaningful topics and themes are generated in a number of around 15 [26].

Hence, we calculate the coherence scores of 4 to 20 topics, and as can be seen from Fig. 4, k = 15 gives the highest coherence scores for both Turkish and English datasets (0.5999 for Turkish and 0.3127 for English). Therefore, the results of the topic modeling analysis give 15 topics for each corpus.

TABLE VIII. LIST OF THE MOST FREQUENT KEYWORDS IN TURKISH AND IN ENGLISH DATASETS.

| **Frequency of the Turkish keywords** | **Frequency of the English keywords** |
|---|---|
| covid (1677) | covid (5270) |
| aşı (1205) | vaccine (2677) |
| koronavirüs (882) | case (1562) |
| evdekal (680) | new (1551) |
| pandemi (593) | death (1291) |
| korona (509) | pandemic (1257) |
| sağlık (456) | coronavirus (1210) |
| vaka (385) | people (1198) |
| karantina (352) | health (1022) |
| corona (306) | day (837) |
| ocak (295) | vaccination (831) |
| coronavirus (251) | state (738) |
| haber (233) | news (721) |
| sayısı (232) | year (712) |
| maske (225) | time (708) |

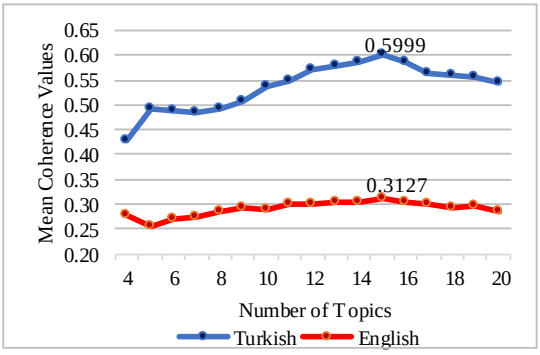


Fig. 4. Coherence values for different k (# of topics) values for Turkish and English.

We divide these 15 topics into different themes to better study them. By this approach, the 15 topics modeled from the Turkish dataset are classified into 6 main primary themes and are shown in Table IX.

TABLE IX. TURKISH TOPIC GROUPINGS AS THEMES.

| **Theme 1: Covid cases and deaths in Turkey and in the world** |
|---|
| *Topic 01:* covid, coronavirus, corona, olarak, olan, koronavirus, dünya, var, tedavi, salgını<br>*Topic 05:* vaka, günlük, yüzde, artış, themachinecorona, test, aktif, fark, Brezilya, İngiltere<br>*Topic 11:* kişi, hayatını, kaybetti, saatte, testi, türkiyede, pozitif, kovid, kişinin, çıktı<br>*Topic 15:* sayısı, vaka, vefat, dünya, ocak, geçti, hasta, milyonu, kişi, toplam |

| **Theme 2: Vaccines** |
|---|
| *Topic 02:* aşı, var, coronavac, sinovac, olan, doz, çin, fahrettinkoca, kovid, kişi<br>*Topic 13:* türkiye, corona, coronavac, genelinde, ocak, aşılandı, şimdiye, sondakika, tablosu, okul<br>*Topic 14:* aşısı, çin, sinovac, pfizer, biontech, coronavac, doz, olan, kovid, acil |
| **Theme 3: Declarations from the Turkish Health Ministry and prevention methods** |
| *Topic 03:* koronavirüs, ocak, tamamı, tablosu, kovid, sondakika, kapsamında, haber, fahrettinkoca, tip<br>*Topic 09:* sağlık, bakanı, koca, fahrettin, bakanlığı, fahrettinkoca, kurulu, bilim, açıkladı, hasta |
| **Theme 4: Local news** |
| *Topic 10:* korona, corona, virüs, coronavirus, koronavirus, haber, kovid, samsun, maske, karantina<br>*Topic 04***:** haberleri, tekirdağ, edirne, çerkezköy, trakya, kapaklı, kırklareli, com, trakyaflashaber, evdekal |
| **Theme 5: General news** |
| *Topic 07*: pandemi, döneminde, salgın, var, coronavirus, sürecinde, eğitim, okul, haber, meb<br>*Topic 08:* evdekal, ocak, cumartesi, fenerbahce, olsun, galatasaray, iddaa, taburespor, besiktasinmacivar, karantina<br>*Topic 12:* sokak, haber, kayseri, eğitim, gündem, ekonomi, kayserispor, kaza, trafik, covet |
| **Theme 6: Prevention** |
| *Topic 06:* maske, mesafe, temizlik, tedbirlerine, valimiz, sağlıkicinhepimizicin, icisleri, koordinesinde, denetimler, Oktay |

Table IX shows that there are 4 topics under theme *Covid cases and deaths in Turkey and in the world*. These topics include keywords about the Covid tests, positive cases, deaths in Turkey and around the world, and countries such as England and Brazil exist among the keywords. Another theme is named *Vaccines*. Keywords like "sinovac", "Pfizer", "biontech", "vaccine", and "vaccinations" are included in the topics. Since our dataset is collected during January 2021, vaccinations had started before this time period in Turkey; thus, the keywords include the following types of tweet. *Declarations from the Turkish Health Ministry* theme is composed of three topics. The Turkish Health Ministry makes regular declarations in Turkey; hence, people tweet about it regularly. The *Local news* theme is about coronavirus news from different cities of Turkey. Theme 5 is, in fact, a theme that includes keywords in tweets about several types of daily bulletins and agenda in Turkey. There are keywords about education, schools, sports, economics, pandemics, accidents, and traffic. Finally, the *Prevention* theme includes only one topic, which consists of keywords about tweets like "distancing", "mask", and "hygiene".

Similarly, the 15 topics modeled from the English are classified into themes, and it can be seen by comparing Table IX and Table X that both the theme numbers and names are very similar. Theme 1 is named *Covid cases and deaths* in the English dataset. "Tests", "cases", "deaths", "numbers", and "reports" are the main keywords that make up this theme. The second theme is *Vaccines*, similar to Turkish dataset. The main difference between the two datasets is that, here, only "Pfizer" is seen as the vaccine type. Theme 3 is named as *News and declarations from the President.* This theme has some differences from the Turkish results. One of them is the change of the President seat in USA in January 2021. Another one is the government relief plan, since the USA citizens were waiting for this plan for a while. Theme 4 is different from the Turkish dataset, it is about *new variants*. Keywords such as "new", "variant", "Africa", "south", and "update" compose this theme. Instead of *local news*, *new variants* of the virus show up as a theme in the English dataset, which makes sense. Theme 5 in English dataset has a similar theme to the Turkish dataset, *general news*-based tweets. Unsurprisingly, the keywords in the tweets on mental health, help, service systems, workers, and schools make up this theme. The last theme is *prevention*, which includes keywords such as "mask", "spread", "social", "distancing", and "hand".

TABLE X. ENGLISH TOPIC GROUPINGS AS THEMES.

| **Theme 1: Covid cases and deaths** |
|---|
| *Topic 03*: case, total, active, today, covid, confirmed, county, reported, number, update<br>*Topic 09:* death, total, covid, reported, rate, today, toll, report, number, related<br>*Topic 15:* test, positive, tested, testing, week, result, negative, pcr, quarantine, today |
| **Theme 2: Vaccines** |
| *Topic 02:* vaccine, dos, pfizer, dose, effective, novavax, first, trial, distribution, rollout<br>*Topic 05:* people, died, many, virus, dying, million, vaccinated, still, like, need<br>*Topic 14:* vaccination, state, county, first, week, site, dos, update, resident, appointment<br>*Topic 12:* get, need, vaccinated, help, back, want, let, please, going, know |
| **Theme 3: News and declarations from the President** |
| *Topic 01:* covid, via, news, patient, due, biden, help, lockdown, trump, update<br>*Topic 06:* pandemic, business, since, world, global, response, help, biden, work, way<br>*Topic 11***:** president, plan, american, relief, administration, joe, order, republican,government, official |
| **Theme 4: New Variants** |
| *Topic 07:* new, variant, report, york, strain, virus, coronavirusupdate, gmt, coronaviruspandemic, update<br>*Topic 04:* news, variant, say, virus, south, effective, novavax, trial, africa, strain |
| **Theme 5: General news** |
| *Topic 08:* health, public, care, county, department, mental, worker, official, service, system<br>*Topic 10:* time, need, help, like, home, school, know, work, please, many |
| **Theme 6: Prevention** |
| *Topic 13:* mask, wear, wearing, face, spread, social, hand, distancing, distance, stay |

## V. DISCUSSION

Healthcare professionals and governments should be aware of the fact that Twitter data can be used to understand public opinion, perceptions and emotions about the Covid-19 pandemic. It should be noted that perceptions and emotions differ among countries, which should be addressed by policy makers in different countries, as noted in related works [14]–[16]. These differences lead the countries to behave differently. Existing studies also show that perceptions and emotions change in time in a similar way. The literature states that vaccinations, especially, lead to this change during the Covid-19 pandemic [14], [15]. Our study also supports these findings.

The findings of the topic modeling analysis in this study are in parallel with the literature in terms of the words and themes most commonly used [2], [12]–[15]. Specifically, vaccines, news about pandemics, new variants, and prevention themes are common in Covid-19 topic modeling-based studies. However, there is no study that analyzes the relationship between sentiment values of posts and languages/time period using hypothesis tests. Therefore, a discussion on the comparison between our study and existing studies is difficult in this context.

## VI. CONCLUSIONS

The aims of this study can be explained threefold. One of them is to understand the similarities and differences between the public opinion of Turkey and the world. After conducting an extensive literature review, we have seen that there are very few studies that compare the agenda of different languages/countries. Additionally, there are no studies that analyze the relationship between the sentiment values of posts/news and languages. Similarly, there are no studies that analyze the relationship between the sentiment values of posts/news and the time period. In this study, to fill these gaps in the literature, Turkish and English tweets are collected using Twitter hashtags related to the Covid-19 disease between 1-31 January 2021. Sentiment analysis is conducted by labeling the tweets using the Vader Sentiment package.

We apply a Chi-square test to determine if there is a relationship and/or dependence between the sentiment values of the tweets (positive, negative, neutral) and the language (Turkish, English). The results of the analysis show that language is significantly related to the sentiment class (at the $p < 0.05$ significance level). This means that the number of positive/negative/neutral tweets differs depending on the language. The interpretation of the cross-tabulation matrix shows that the Turkish people are more indifferent to the Covid-19 issue than the rest of the world. Specifically, the number of neutral tweets in Turkish is 16 % more than the neutral English tweets (Table VI). Additionally, we apply a Chi-square test to determine if there is a relationship between the sentiment values of the tweets and the time period (first half and second half of January 2021). The findings indicate that the period of time is significantly related to the sentiment class (at the $p < 0.05$ significance level), hence the number of positive/negative/neutral tweets differs depending on the period of time. The cross-tabulation matrix shows that the proportion of positive tweets increases in both languages in the second half of the month. Specifically, while the numbers of negative and neutral tweets decrease in the second half of the month, positive tweets increase by 14.7 % (Table VII). The reason may be the increased use of vaccines as time passes.

The keywords most frequently used are also analyzed and it is seen that, in both languages, the word most frequently mentioned is “aşı/vaccine” after “Covid”. The topics modeled by NMF in the study are grouped into themes such that the 15 topics modeled from the datasets are classified into 6 main primary themes. Surprisingly, the themes have occurred to be very similar in both languages, which means that the Turkish and world agenda are not very different in terms of pandemics. However, there are some different themes on the Turkish and world agenda. Tweets about new variants, USA president election, and the US relief plan take place on the world agenda, which is different from the Turkey agenda. On the other hand, some local news is a theme on the Turkish agenda.

As a future work, this study can be widened by examining a larger period of time. Moreover, the methods used in this study can be used to compare the agenda and opinion of different countries. Therefore, the methodology proposed in this study can be updated using data of longer time periods and greater number of tweets in different languages.

## CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

## REFERENCES

[1] R. Chandrasekaran, V. Mehta, T. Valkunde, and E. Moustakas, “Topics, trends, and sentiments of tweets about the COVID-19 pandemic: Temporal infoveillance study”, *Journal of Medical Internet Research*, vol. 22, no. 10, p. e22624, 2020. DOI: 10.2196/22624.

[2] A. Abd-Alrazaq, D. Alhuwail, M. Househ, M. Hamdi, and Z. Shah, “Top concerns of tweeters during the COVID-19 pandemic: Infoveillance study”, *Journal of Medical Internet Research*, vol. 22, no. 4, p. e19016, 2020. DOI: 10.2196/19016.

[3] J. Li, Q. Xu, R. Cuomo, V. Purushothaman, and T. Mackey, “Data mining and content analysis of the Chinese social media platform Weibo during the early COVID-19 outbreak: Retrospective observational infoveillance study”, *JMIR Public Health and Surveillance*, vol. 6, no. 2, p. e18700, 2020. DOI: 10.2196/18700.

[4] C. Shen, A. Chen, C. Luo, J. Zhang, B. Feng, and W. Liao, “Using reports of symptoms and diagnoses on social media to predict COVID-19 case counts in mainland China: Observational infoveillance study”, *Journal of Medical Internet Research*, vol. 22, no. 5, p. e19421, 2020. DOI: 10.2196/19421.

[5] S. Yousefinaghani, R. Dara, S. Mubareka, A. Papadopoulos, and S. Sharif, “An analysis of COVID-19 vaccine sentiments and opinions on Twitter”, *International Journal of Infectious Diseases*, vol. 108, pp. 256–262, 2021. DOI: 10.1016/j.ijid.2021.05.059.

[6] L. Sinnenberg, A. M. Buttenheim, K. Padrez, C. Mancheno, L. Ungar, and R. M. Merchant, “Twitter as a tool for health research: A systematic review”, *American Journal of Public Health*, vol. 107, no. 1, pp. e1–e8, 2017. DOI: 10.2105/AJPH.2016.303512.

[7] A. Sarker, S. Lakamana, W. Hogg-Bremer, A. Xie, M. A. Al-Garadi, and Y.-C. Yang, “Self-reported COVID-19 symptoms on Twitter: An analysis and a research resource”, *Journal of the American Medical Informatics Association*, vol. 27, no. 8, pp. 1310–1315, 2020. DOI: 10.1093/jamia/ocaa116.

[8] Z.-Y. Tao *et al.*, “Nature and diffusion of COVID-19–related oral health information on Chinese social media: Analysis of tweets on Weibo”, *Journal of Medical Internet Research*, vol. 22, no. 6, p. e19981, 2020. DOI: 10.2196/19981.

[9] A. Wahbeh, T. Nasralah, M. Al-Ramahi, and O. El-Gayar, “Mining physicians’ opinions on social media to obtain insights into COVID-19: Mixed methods analysis”, *JMIR Public Health and Surveillance*, vol. 6, no. 2, p. e19276, 2020. DOI: 10.2196/19276.

[10] H. Budhwani and R. Sun, “Creating COVID-19 stigma by referencing the novel coronavirus as the “Chinese virus” on Twitter: Quantitative analysis of social media data”, *Journal of Medical Internet Research*, vol. 22, no. 5, p. e19301, 2020. DOI: 10.2196/19301.

[11] S. R. Rufai and C. Bunce, “World leaders’ usage of Twitter in response to the COVID-19 pandemic: A content analysis”, *Journal of Public Health*, vol. 42, no. 3, pp. 510–516, 2020. DOI: 10.1093/pubmed/fdaa049.

[12] R. Alshalan, H. Al-Khalifa, D. Alsaeed, H. Al-Baity, and S. Alshalan, “Detection of hate speech in COVID-19–related tweets in the Arab region: Deep learning and topic modeling approach”, *Journal of Medical Internet Research*, vol. 22, no. 12, p. e22609, 2020. DOI: 10.2196/22609.

[13] H. Jang, E. Rempel, D. Roth, G. Carenini, and N. Z. Janjua, “Tracking

COVID-19 discourse on twitter in North America: Infodemiology study using topic modeling and aspect-based sentiment analysis", *Journal of Medical Internet Research*, vol. 23, no. 2, p. e25431, 2021. DOI: 10.2196/25431.

[14] C.-H. Chang, M. Monselise, and C. C. Yang, "What are people concerned about during the pandemic? Detecting evolving topics about COVID-19 from Twitter", *Journal of Healthcare Informatics Research*, vol. 5, no. 1, pp. 70–97, 2021. DOI: 10.1007/s41666-020-00083-3.

[15] C. Meaney, M. Escobar, R. Moineddin, T. A. Stukel, S. Kalia, B. Aliarzadeh, T. Chen, B. O'Neill, and M. Greiver, "Non-negative matrix factorization temporal topic models and clinical text data identify COVID-19 pandemic effects on primary healthcare and community health in Toronto, Canada", *Journal of Biomedical Informatics*, vol. 128, art. 104034, 2022. DOI: 10.1016/j.jbi.2022.104034.

[16] Y. Feng and W. Zhou, "Work from home during the COVID-19 pandemic: An observational study based on a large geo-tagged COVID-19 Twitter dataset (UsaGeoCov19)", *Information Processing & Management*, vol. 59, no. 2, art. 102820, 2022. DOI: 10.1016/j.ipm.2021.102820.

[17] D. O'Callaghan, D. Greene, J. Carthy, and P. Cunningham, "An analysis of the coherence of descriptors in topic modeling", *Expert Systems with Applications*, vol. 42, no. 13, pp. 5645–5657, 2015. DOI: 10.1016/j.eswa.2015.02.055.

[18] G. Sariman and E. Mutaf, "Covid-19 sürecinde Twitter mesajlarinin duygu analizi", *Euroasia Journal of Mathematics Engineering, Natural and Medical Sciences*, vol. 7, no. 10, 2020. DOI: 10.38065/euroasiaorg.149.

[19] A. K. Nazli, C. Kocaömer, M. Beşbudak, and N. E. Köker, "Understanding the initial reactions of turkish twitter users during the Covid-19 pandemic", *The Turkish Online Journal of Design, Art and Communication*, vol. 11, no. 1, pp. 20–41, 2021. DOI: 10.7456/11101100/002.

[20] G. Konakci, B. N. O. Uran, and O. Erkin, "In the Turkish news: Coronavirus and "alternative & complementary" medicine methods", *Complementary Therapies in Medicine*, vol. 53, art. 102545, 2020. DOI: 10.1016/j.ctim.2020.102545.

[21] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text", in *Proc. of the Eighth International AAAI Conference on Web and Social Media*, 2014, vol. 8, no. 1.

[22] A. Güran, N. G. Bayazit, and E. Bekar, "Automatic summarization of Turkish documents using non-negative matrix factorization", in *Proc. of 2011 International Symposium on Innovations in Intelligent Systems and Applications*, 2011, pp. 480–484. DOI: 10.1109/INISTA.2011.5946121.

[23] D. D. Lee and H. S. Seung, "Learning the parts of objects by nonnegative matrix factorization", *Nature*, vol. 401, pp. 788–791, 1999. DOI: 10.1038/44565.

[24] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures", in *Proc. of the Eighth ACM International Conference on Web Search and Data Mining*, 2015, pp. 399–408. DOI: 10.1145/2684822.2685324.

[25] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality", arXiv preprint, 2013. DOI: 10.48550/arXiv.1310.4546.

[26] S. Boon-Itt and Y. Skunkan, "Public perception of the COVID-19 pandemic on Twitter: Sentiment analysis and topic modeling study", *JMIR Public Health and Surveillance*, vol. 11, no. 4, p. e21978, 2020. DOI: 10.2196/21978.

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution 4.0 (CC BY 4.0) license (http://creativecommons.org/licenses/by/4.0/).