Turkish Journal of Electrical Engineering and Computer Sciences

Volume 33 | Number 3 Article 11

5-23-2025

Vaccine hesitancy in Türkiye: A natural language processing study on social media

SEMİH SARI

ULYA BAYRAM

Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2025) 33: 392 – 409

doi:10.55730/1300-0632.4133

Vaccine hesitancy in Türkiye: A natural language processing study on social media

Semih SARI[1], **Ulya BAYRAM**[2,*]

[1]Department of Computer Engineering, Faculty of Engineering, Çanakkale Onsekiz Mart University, Çanakkale, Turkiye

[2]Department of Electrical and Electronics Engineering, Faculty of Engineering, Çanakkale Onsekiz Mart University, Çanakkale, Turkiye

Received: 17.07.2024 • **Accepted/Published Online:** 19.04.2025 • **Final Version:** 23.05.2025

Abstract: Vaccine hesitancy is a significant public healthcare problem that is threatening everyone worldwide. Vaccine hesitancy has become more ingrained in Turkish society, mainly through social media. Unfortunately, reflections of this hesitancy are preventable deaths or permanent disabilities. Because of the uncontrolled spread of misinformation and disinformation on social media, Türkiye is facing a future health crisis. As a step towards preventing this crisis, our main objective is to use the power of artificial intelligence techniques on Turkish social media posts to detect antivaccine posts. Through this study, it will be possible to raise awareness about the importance of vaccines in Turkish society, strengthen Türkiye's defense mechanism against potential epidemics, and ensure that our society exchanges information in a healthier digital environment. We collected and cleaned a novel Turkish social media dataset, resulting in 3778 posts. Then, we used a baseline machine learning method, logistic regression, popular machine learning methods, support vector machines, and XGBoost to detect antivaccine thoughts and misleading information from Turkish social media posts. Further, we included transformers that changed the natural language processing domain. Evaluations are conducted using a multilingual BERT and two models specifically trained for recognizing Turkish texts, such as BERTurk. Results showed that transformers can separate Turkish social media posts with antivaccine beliefs from other posts with a 75.9% Area Under the ROC curve rate.

Key words: Natural language processing, machine learning, transformers, vaccine hesitancy, Turkish texts, social media

1. Introduction

Vaccine hesitancy and any related misinformation (unknowingly spreading false information) or disinformation (knowingly spreading false information with an agenda) spread about this topic causes a decrease in global vaccination rates, yielding the reemergence of previously eradicated diseases via decades of vaccination efforts. These reemerging diseases are inducing deaths and disabilities worldwide. The spread of epidemic diseases due to vaccine hesitancy around the world is also a prominent threat to the future of our country, Türkiye. According to the T.C. Ministry of Health report, measles disease was eradicated in Türkiye by 2006 thanks to the vaccination program. However, the situation changed in 2008 when an Iraqi student in Ankara was diagnosed with measles, and the disease spread to three healthcare workers[1]. According to the same report, three refugees in İstanbul in 2009 and seven other non-Turkish citizens in 2010 were diagnosed with measles. As a result,

*Correspondence: ulya.bayram@comu.edu.tr

[1]https://hsgm.saglik.gov.tr/tr/surveyanslar/kizamik-kizamikcik.html

between 2012-2013, 8500 measles cases were identified in Türkiye as the disease spread to unvaccinated Turkish citizens, and these numbers continue to increase to this day. Consequently, decades of financial and emotional efforts of the Turkish vaccination program to prevent measles became futile, and today, Turkish citizens are at risk of contracting an eradicated disease despite having been vaccinated against it.

The spread of false information is accelerating due to the widespread usage of social media [1], causing the prevalent upsurge of vaccine hesitancy [2]. For example, during the COVID-19 pandemic, the claim "COVID-19 vaccines contain microchips" on social media acquired many believers[2]. Regardless of the explanations from scientists, the false information becomes a "common knowledge" without any proof. Not restricted to COVID-19, these social media frenzies cause the spread of diseases and put the lives of the public and healthcare workers at risk. In addition, anxiety, depression, and an increased tendency towards suicide caused by the spread of false information are other factors one must consider [3]. Thus, it is essential to intercept these misinformation and disinformation attempts on social media regarding vaccinations. In our study, we propose a significant step towards a preventative measure through the automatic detection and moderation of antivaccine thoughts on social media.

Turkish disinformation or misinformation detection employing natural language processing (NLP) techniques has become a popular research area. Various disinformation detection studies conducted in Türkiye have achieved success rates ranging from 80% to 98% [4–7]. Yet, when the main research focus is to detect vaccine hesitancy in Türkiye, the existing studies in the literature have some limitations: (1) The studies focus on binary classification between two classes and are conducted on easy-to-classify data, but in a real-world scenario there are more than two classes and are many factors such as noise, irony, and indecision, among others. (2) The existing studies exclusively use tweet data from X, formerly known as Twitter. (3) They do not focus on healthcare-related issues. Literature review shows that the few studies focusing on Turkish disinformation detection in healthcare mainly rely on surveys, statistics, and analyses [8]. The study presented in this paper provides solutions to all three limitations.

The current study concentrates on vaccine hesitancy-related disinformation or misinformation detection using NLP techniques on Turkish social media data for the first time in the literature. Considering real-life social media usage behavior in Türkiye, we focus on a three-class classification problem instead of a straightforward binary task: distinguishing antivaccine, provaccine, and neutral posts. Therefore, this study goes beyond a simple sentiment analysis and moves toward stance detection, aiming to identify more nuanced opinions, such as support, opposition, skepticism, or neutrality. Plus, we take a step towards public health and citizen awareness by, for the first time, automatically detecting vaccine hesitancy and related misleading information in Türkiye. Contributions are as follows:

1. Instead of the popularly used X (former Twitter) as the data domain, posts from four distinct topics under the Ekşi Sözlük Turkish forum are collected and manually labeled, resulting in a novel data collection. Ekşi Sözlük is one of the most popular social media platforms in Türkiye, where many individuals from various age groups and backgrounds have been using for many years to share news, information, personal views, seeking or providing help, among others[3].

2. Instead of developing a simple, binary classification task, we tackle a more demanding challenge by focusing on a three-class classification problem. Multiclass classification offers a more comprehensive

[2]https://www.milliyet.com.tr/galeri/son-dakika-dunya-bunu-konusuyor-koronavirus-asisiyla-cip-6380172
[3]https://eksisozluk.com/

and nuanced understanding of public opinion towards vaccine hesitancy of Turkish citizens. Performing a binary classification alone neglects the equally important perspective of those who support vaccines and those who are neutral.

3. The detection of Turkish “antivaccination” thought with NLP has not been studied before.

4. Through comparing the performance of baseline, previously and currently popular different machine learning methods, as well as two transformer models, we cover a wide range of solutions.

In the following section, we summarized existing misinformation/disinformation detection studies in Turkish and other languages. Then, we explain the data collection process and the contents of the novel vaccine hesitancy dataset with a detailed analysis. Subsequently, selected machine learning and NLP methods are justified, followed by the results, discussions, and conclusions.

2. Background

Studies in Table 1 show that there are predominantly two types of approaches to misinformation or disinformation detection: fake news and disinformation detection from social media. Those that focus on fake news detection define this task as the classification of news according to their extent of truthfulness [1]. A study from 2021 attempts to detect false information within digital news published in English between 2016 and 2018 using the Long Short Term Memory (LSTM) method [9]. LSTM is a Recurrent Neural Network (RNN) where sequence information (i.e. time domain) is critical in decision-making. To convert texts into numerical features, they use a pretrained word embedding bundle named GloVe, consisting of vector representations for words where the vectors of values are only meaningful in the embedding space. Another study uses RNNs and Convolutional Neural Networks (CNN) with the same objective [10]. They compare the results of these two methods against the popular transformer techniques on a dataset of 44,898 fake and factual news in English and find that the former deep learning methods provide higher classification performance than the contemporary transformers. Likewise, another false news detection study combines RNN and CNN methods and obtains a CNN BiLSTM model where the features are also the formerly mentioned pretrained GloVe embeddings [11]. An alternative study employs Knowledge Graphs (KG) for the fake news detection task in English while experimenting with LSTMs and transformer models like SentRoBERTa [12].

The task of false information detection from news articles also received attention from Turkish researchers. A study uses Logistic Regression (LR) to detect false news in Turkish after automatically translating the English news articles to Turkish [13]. They report that removing stopwords does not improve their performance. Another study directly collects Turkish news from two websites, one with genuine Turkish news (hurriyet.com) and another with ironic and made-up Turkish articles (zaytung.com) [14]. After collecting approximately 2000 texts, they experimented with Random Forest (RF), Bidirectional Encoder Representations from Transformers (BERT), Naïve Bayes (NB), Support Vector Machines (SVM), Multinomial NB, LR, and a transformer pretrained in Turkish named BERTurk.

Studies that aim to detect false information from social media posts also receive extensive attention. One study collects tweets from X and posts from a Chinese microblogging website named Weibo and employs SVM, Decision Trees (DT), LSTMs, and Gated Recurrent Units (GRU) to detect false information [15]. Another study collects tweets in English posted during the COVID-19 pandemic and automatically translates them into Turkish [4]. After manually labeling those containing disinformation, they fine-tune BERTurk for classification. Instead of translating data from one language to Turkish, some studies directly utilize Turkish tweets. A study collects Turkish tweets, manually labels those with fake information, and experiments with BERT, BERTurk, and CNNs to automatically detect such posts [6]. Another study collects and labels Turkish data to automatically detect real news and disinformation from the X posts about soccer players [7]. They analyze the topics using the Latent Dirichlet Allocation (LDA) method and then extract Term Frequency-Inverse Document Frequency (TF-IDF) features for classification. They also experiment with word2vec embedding features, collect classification performance from K-Nearest Neighbors (K-NN), SVM, RF, GRU, and LSTM methods, and achieve success rates reaching 80%.

Three common themes in Table 1 are significant. Primarily, as the previous researchers also stated [11], studies that employ social media data mainly rely on tweets from X (Twitter). Yet there are many other platforms, perhaps even more popular than X in Türkiye. Second, Turkish NLP studies on false information detection do not focus on healthcare problems [8]. The only Turkish healthcare-related study present in Table 1 collects tweets from March to August 31st in 2021 with 'aşı' (vaccine) and 'covid-19' hashtags [5]. After utilizing SVM, XGBoost, RF, and Soft Voting techniques on these tweets, they conduct sentiment analysis to detect positive, negative, and neutral feelings and present the population's feelings towards the COVID-19 vaccine rather than presenting a framework to classify false information about vaccines. Third, most studies in Table 1 perform binary classification. However, thoughts or opinions posted on social media do not always support one side of an argument; some individuals stay neutral or undecided.

Table 1. A survey of the NLP literature on disinformation/misinformation detection. EN: English, TR: Turkish, CN: Chinese data. "EN to TR" means the data was translated from English to Turkish.

Data	Language	Task	Methods	Score	Study
News between 2016-2018	EN	Binary	GloVe, LSTM	91% F1	[9]
44,898 fake and real news	EN	Binary	Transformers, RNN, CNN	89% F1	[10]
Buzzface Dataset, FakeNewsNet, Twitter data	EN	Binary	GloVe, CNN Bi-LSTM	98% F1	[11]
FakeNewsNet	EN	Binary	Extratreeclassifier, LSTM, SentRoBERTa, KG	88%, 78% F1	[12]
Fake news from Twitter and Weibo	EN, CN	Binary	SVM, DT, LSTM, GRU	89%, 95% F1	[15]
20,800 fake news	EN to TR	Binary	LR	93% accuracy	[13]
Tweets	EN to TR	Binary	BERTurk	98% accuracy	[4]
412,588 tweets from the COVID-19 pandemic	TR	3-class	tf-idf, SVM, XGBoost, RF, Soft Voting	90% accuracy	[5]
Fake and real news	TR	Binary	BERT, BERTurk, CNN	94% accuracy	[6]
Fake news from Teyit.org, and soccer-related tweets	TR	Binary	tf-idf, word2vec, LDA, KNN, SVM, RF, RNN, GRU, LSTM, BiGRU, BiLSTM	90% F1	[7]
4,459 News from Zaytung and Hurriyet	TR	Binary	NB, RF, SVM, NBM, LR, BERTurk	99% accuracy	[14]

Instead of purchasing Turkish tweets from X, we selected another highly popular Turkish social media environment as the data source. Additionally, instead of restricting the investigation to a specific type of vaccine like COVID-19, the data spans the general vaccine hesitancy of Turkish social media users. Meanwhile, we account for those who do not choose a side from the vaccine hesitancy arguments through multiclass classification.

3. Data

Due to the lack of available Turkish corpora on vaccine hesitancy, the present study starts with constructing a novel textual dataset. The process follows the same outline as many other NLP studies: using web scraping tools to collect texts, removing short texts from the collection, manually labeling them, and conducting various analysis techniques to review the content. In the following, these steps are explained in more detail to ensure the repeatability of the present study.

3.1. Collection and cleaning

Among the most significant parts of the current study is the novel dataset. Rather than following the previous literature, the present research does not utilize tweets from X as data. Instead, possibly the oldest social media platform in Türkiye that is still active, called Ekşi Sözlük, is used. After a careful search among the topics with acceptably many posts (e.g., above 100 posts), we found four topics related to vaccine hesitancy present in Table 2:

- Topic titled "şaka maka aşı yaptırmayanların haklı çıkması" translates as "turns out those who did not get vaccinated were right," expressing the hesitancy towards vaccines. Since the topic emerged after the COVID-19 pandemic, it is safe to assume that posts under this topic mostly refer to the COVID-19 vaccines.
- Topic "covid aşısı olmuyorum" translates as "i'm refusing the covid vaccine."
- The "mrna aşıların seneler sonraki yan etkileri" translates as "the future side-effects of mrna vaccines," also referring to COVID-19 vaccines.
- The title "aşı karşıtlığı" translates as "vaccine hesitancy," predating the COVID-19 pandemic, therefore contains posts related to the general vaccine hesitancy.

To assemble all posts under the selected four topics, we used the BeautifulSoup library for web scraping[4]. This Python library allows efficient parsing and extraction of content from HTML. After fetching the HTML structure of the target pages of the topics, we wrote the relevant scripts to parse the content, extract the posts, and construct a dataset.

Social media data are prone to contain too much noise. Posts of short responses that do not contain enough words to represent opinions and those that contain only links to other websites or topics are a few examples of the mentioned noise. Therefore, after collecting the social media posts through web scraping, posts that did not contain words other than the website links and posts that only linked to another topic were deleted. Further, we removed posts with less than four words. The number of posts per topic before and after the cleaning process is present in Table 2. Since previous NLP studies have shown that operations such as stemming, lemmatization, and deletion of stopwords do not make a big difference, they are excluded in the data cleaning [13, 16]. We used lemmatization only during the data analysis.

3.2. Labeling

Assembling a novel dataset for misinformation/disinformation detection requires labeling. Two annotators jointly labeled each post with one of four labels: antivaccine, provaccine, neutral, and undecided. A third annotator independently labeled the posts without seeing the initial labels. The agreement ratio between these

[4] https://www.crummy.com/software/BeautifulSoup/

Table 2. The selected topics from eksisozluk.com website, the number of posts before and after data cleaning, and word count statistics after cleaning.

Topic	Before	After	Word counts (mean ± stdev)	Word counts (max)
şaka maka aşı yaptırmayanların haklı çıkması	971	925	55 ± 65	531
covid aşısı olmuyorum	479	423	59 ± 90	1,092
mrna aşıların seneler sonraki yan etkileri	1,127	1,071	64 ± 98	1,523
aşı karşıtlığı	1,446	1,359	88 ± 158	3,798

two labeling processes is 87%. Posts marked as "antivaccine" contain thoughts that oppose the idea of being vaccinated against diseases. Those marked as "pro-vaccine" support the vaccines and the science behind them. Posts marked as "neutral" do not prefer a side or provide an opinion. The "undecided" posts support both antivaccine and provaccine views and cannot choose a side within a post. Since the scope of the present study does not include detecting undecided individuals, we excluded the latter posts.

One could argue that manually annotating social media posts is challenging since the process is prone to the subjectivity of opinions. To simplify the labeling process, we experimented with Turkish sentiment detection methods. Among the many available fine-tuned Turkish transformers, we selected two contemporarily popular models [17][5][6], each capable of detecting positive, negative, and neutral sentiments. While the former two sentiments are irrelevant to the labels of pro and antivaccine, the neutral class is relevant. Hence, we tested if sentiment detection methods can auto-label neutral posts accurately. The first test visualized the confusion matrix of the two models in Figure 1, where Model 1 refers to BounTi and Model 2 refers to the other method. It shows that one of the models classified most of the posts with negative sentiment, while the other model classified the same data as positive. To enumerate the agreement between the two models, we used Cohen's Kappa scores [18]. Computed as:

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

where P_o is the observed agreement between annotators, P_e is the expected agreement by chance, a κ score close to -1 shows disagreement, close to 1 shows agreement, and close to 0 shows randomness. From the results in Figure 1, the above equation returned 0.035 Cohen's Kappa score. Closeness to zero highlights the randomness. This finding signifies a few critical points: Due to the lack of large Turkish corpora, fine-tuned Turkish sentiment detection models are not ready for annotation frameworks yet. Also, considering the nuances in a language like irony, jokes, and idioms, automatically labeling social media posts without a human in the loop is currently out of reach.

According to Table 3, the number of antivaccine posts per topic is in the minority compared to supportive or neutral posts. Therefore, the following supervised NLP experiments will occur on an imbalanced dataset. The labeled dataset is available for academic use on demand.

3.3. Analysis

Data analysis can provide critical insights into the content and shared contexts of thousands of texts, which would otherwise remain hidden. So, we conducted various data analysis techniques to unveil patterns, distributions, and relationships from our novel dataset. The first analysis showed the distribution of labels per topic. In

[5]https://huggingface.co/akoksal/bounti
[6]https://huggingface.co/emre/turkish-sentiment-analysis

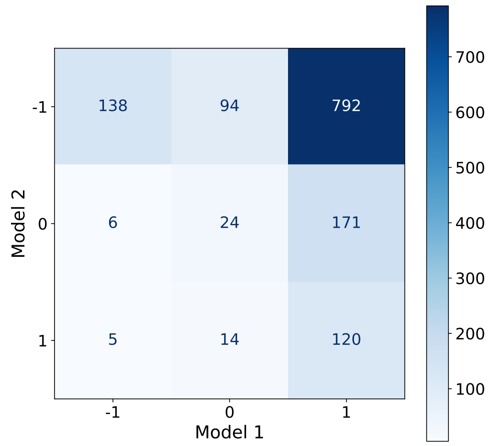


Figure 1. Confusion matrix showing the agreement between two Turkish sentiment detection models. -1: Negative, 0: neutral, +1: positive sentiments.

Table 3. The number of labels per topic.

Topic	**Provaccine**	**Antivaccine**	**Neutral**
şaka maka aşı yaptırmayanların haklı çıkması	368	157	316
covid aşısı olmuyorum	214	34	148
mrna aşıların seneler sonraki yan etkileri	374	100	486
aşı karşıtlığı	660	202	414

Table 3, regardless of the topic of choice, actual antivaccine thoughts are not as prevalent as one might expect from the title: The number of neutral posts and pro-vaccine posts are in the majority compared to the antivaccine posts. For example, the “şaka maka aşı yaptırmayanların haklı çıkması” topic that expresses those who chose not to receive the COVID-19 vaccination has 19% of its posts defending the title. The “aşı karşıtlığı” topic’s 16% of posts oppose all vaccines. Meanwhile, the “mrna aşıların seneler sonraki yan etkileri” topic that expresses mRNA vaccines will have side effects in the future years, has 10% of its content, and the remaining “covid aşısı olmuyorum” topic has almost 9% of its labels belonging to the antivaccine class. Hence, we observe that among the users of this social media domain, despite the strong opinions in the titles, antivaccine supporters are either in the minority or are hesitant to express their viewpoints openly.

The subsequent analysis in Figure 2 shows the word count distributions of the labeled posts normalized by the number of posts per topic. The distributions reveal two main trends: the “aşı karşıtlığı” and “mrna aşıların seneler sonraki yan etkileri” topics have similar distributions with similar mean values, suggesting that topics have parallel patterns of word use. Also, the distributions follow Poisson with long tails, which shows that posts have many rare words. Meanwhile, the remaining distributions are closer to Gaussian, indicating that the post lengths are near the average. Considering the labels, the antivaccine group tends to write long posts, and the pro-vaccine group has a range of lengths, but the neutral group has short posts.

The statistics of unique words are also meaningful. The number of different words in a text can be a reliable measure of the scholarly level of its author [19]. Thus, through whisker plots (also known as box plots), Figure 3a is obtained. The selected social media platform does not have maximum or minimum word limits for posts. Therefore, the number of unique words the classes have is comparable regardless of the class imbalance. Median lines of the three boxes in Figure 3a show that the neutral group, independent of its high number of

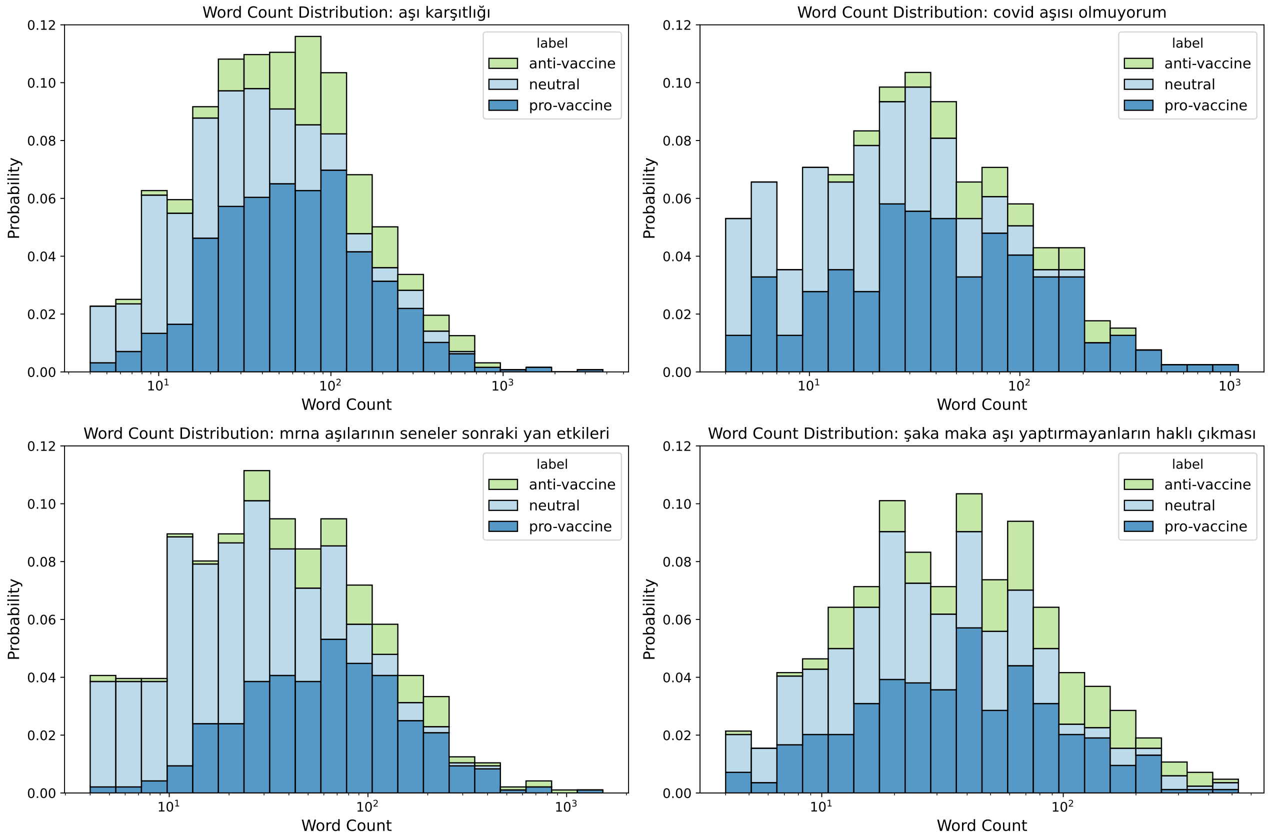
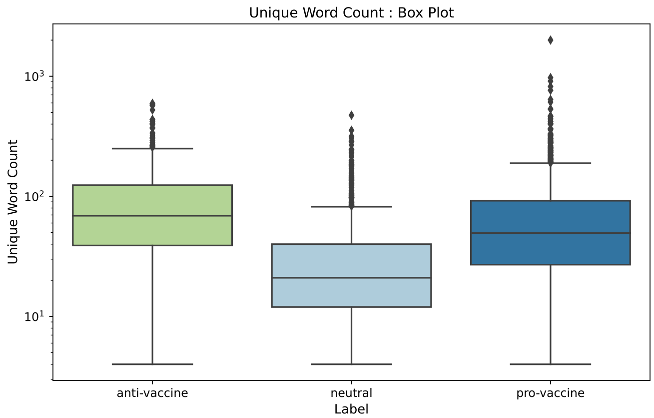


Figure 2. Distribution of labeled posts' word counts per topic.

posts, uses the smallest vocabulary. Meanwhile, the antivaccine group has a slightly richer vocabulary than the pro-vaccine group, which is significant because the number of posts in the antivaccine group is the smallest. Further, the extremes above the boxes express that the pro-vaccine group has some posts with a rich vocabulary. Yet, the extremes in the antivaccine group are not much higher than the median value, which shows that the group's members usually use similar words to express their opinions.

We also used word clouds to visualize the most frequent words in our dataset. Since stopwords (the common words that are not meaningful alone but are essential to tie the words in a sentence) would overwhelm the word clouds due to their high frequency, we removed them before constructing the clouds. However, existing Turkish stopword lists assembled from formal texts could not be helpful for social media posts with informal language. Thus, we created a new stopwords list specific to the social media platform we selected. Since the main subject of this research is vaccine hesitancy, certain common words related to these are excluded (e.g., the word "aşı" that means vaccine, and "covid"). Also, we removed the words in the topic titles. Further, we lemmatized every word to reduce noise, and the resulting word clouds are present in Figure 3b. All four clouds have the words "yok" (translates to no or absent) and "zaman" (time). The cloud of "mrna aşılarının seneler sonraki yan etkileri" topic has the words "biontench," "dna," and "kanser" (cancer) among the prevalent ones, indicating the conspiracy theories that mention how this vaccine will alter the DNA and cause cancer in the future. The word cloud of "aşı karşıtlığı" topic has the words "cahil" (ignorant), "salgın" (epidemic), and



(a) Unique word usage statistics per class.

(b) Word clouds of the four topics.

Figure 3. Two distinct analyses of the present data collection.

”pandemi” (pandemic), showing how the COVID-19 pandemic dominated even the general subject of vaccine hesitancy arguments. Subsequently, the cloud of “şaka maka aşı yaptırmayanların haklı çıkması” mentions “doz” (dosage), “kalp” (heart) and “kalp krizi” (heart attack), demonstrating how the COVID-19 vaccine hesitancy is related to the fear of heart attack the vaccine may induce. The last word cloud belongs to the topic titled “covid aşısı olmuyorum” which has the word “olma” (do not get it), showing how the posts were more aligned with the statement “if you do not want the covid vaccine, then do not get it”.

4. Methods

As visible from Figure 4, we evaluate two NLP strategies to reveal the ideal approach for the multiclass classification of opinions towards vaccine hesitancy. The first strategy uses a combination of n-gram features, feature selection, and a wide range of machine-learning methods, where one of them is a baseline. The second strategy uses the transformer family, which has changed the NLP literature since its first release. Since our dataset has 3,778 posts, splitting the collection into training and test sets and then relying on the performance obtained from the small test set is unlikely to demonstrate generalizable performance. Evaluating the performance in a generalizable way for smaller-scale datasets is through cross-validation [20]. So, we chose to use the k-fold cross-validation framework. This method divides the dataset into k-equal pieces after randomly shuffling the samples and assuring the equivalent class imbalance in each piece. Then, within a loop with k times, each piece becomes the test set once, and the rest comprises the training and validation sets [21]. For the k value, we selected five as the appropriate number of splits, so each test set is large enough. In the following, we describe our methods and parameter choices for repeatability.

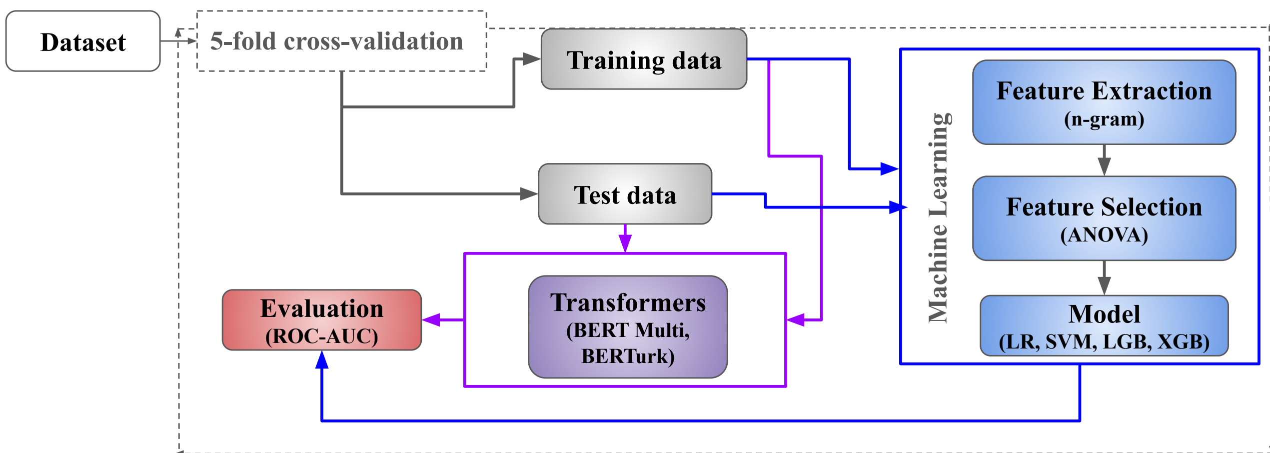


Figure 4. The framework followed in the present study.

4.1. Machine learning methods

Feature extraction: Due to their simplicity and past popularity [16], n-grams are selected as the feature extraction method (i.e. to convert texts into numerical information). The first n-gram feature is unigram, where each word is a feature, and the feature value is the frequency of that word in every document. The other feature is bi-grams, which are consecutive word pairs within a sentence, and the feature value is their frequency in each document. There can be many n-grams depending on the choice of n value. However, the researchers should be aware that as the n value increases, the possibility of finding nonzero feature values will be rare. Since having many zero values in a feature space is not ideal for machine learning performance [22], n = 2 is selected. That means the present feature space includes only unigram and bigram features. We removed rare words, e.g., words that exist less than ten times in the entire training set, from the unigram set to eliminate rare words like typos. Likewise, we removed infrequent bigrams based on whether they occurred more than twice in the complete training set. Thus, we obtained around 7900 n-grams from the training set at each cross-validation fold. Yet, this number of features is still too many for a machine-learning feature space.

Feature engineering: Feature engineering involves some processing steps after extracting features. Accordingly, we applied standard scaling on the feature space to balance the magnitudinal difference between the frequent unigram and infrequent bi-gram feature values by subtracting the mean of each feature from every occurrence x_i, divided by its standard deviation $x_i^{'} = \frac{x_i - \mu_i}{\sigma_i}$ [23]. Subsequently, to estimate which features help more in classifying three classes, we used ANOVA tests following a common framework [24], so we reduce the feature space by selecting only the most helpful features. The ANOVA process helps us determine features with high variance between classes. We sorted the features from highest to lowest variation and experimented with different numbers of top features using the LR method in a 5-fold cross-validation on the training data (no test data leakage is allowed). Then, we selected the number of features resulting in the highest mean ROC-AUC score for that fold and trained the models on the same feature set. So, all of the machine learning results are comparable. We conducted all of the experiments in Python 3.10 and used the scikit-learn library for machine learning methods unless we express otherwise [23]. As parametric machine learning methods can behave unexpectedly when improper values are set [22], we employed the optuna library on the training set data and used the optimum parameters in the experiments [25].

Logistic regression: LR is a statistical technique used to analyze the relationship between a categorical dependent variable and one or more independent variables [26]. It uses the maximum likelihood with iterative processes for parameter estimation. We employed the LR method with its default parameters as the baseline approach.

Support vector machines: SVM is developed for classification tasks and later extended to other applications such as regression [27]. It performs classification by determining the optimum hyperplane between two classes in a feature space with a kernel function. Then, it classifies the test set using the determined hyperplane. Before the popularity of deep learning methods, SVMs were the most popular classification method due to their high performance in the literature [16, 19]. So, we selected SVM for its popularity in the disinformation literature in Table 1. We chose a nonlinear kernel (rbf) for its ability to capture complex relations and conducted a parameter search to find the remaining values.

LightGBM: LightGBM is an advanced gradient boosting algorithm that has improved the speed and accuracy of classification methods compared to the traditional methods [28]. It employs techniques such as Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) [29]. LightGBM uses a histogram-based approach to determine splitting points and focuses on high-gradient samples, resulting in faster training and better prediction performance. We selected this approach for its success in recent years and used parameter search on the training set to determine the optimum values.

XGBoost: Like LightGBM, XGBoost is also popular in recent years due to its efficiency and accuracy [30]. Thanks to its structure with many decision trees and the parallelization to multiple GPUs, it can scale up to billions of samples. Therefore, we used this parametric method after conducting a parameter search on the training set.

4.2. Transformers

Transformers rely on their ability to calculate mathematical operations in parallel. They capture semantic relationships between words in big data and achieve state-of-the-art performance on various NLP tasks. One significant advantage they have is the lack of need for preprocessing large textual datasets or performing feature extraction. Instead, they prefer raw textual data. Starting with Bidirectional Encoder Representations from

Transformers (BERT) [31], transformer methods gave birth to various text classification methods such as RoBERTa, DistilBERT, and XLM-RoBERTa[7]. BERTs consist of layers. They adopt bidirectional training of the transformer models using the masked language model and the sentence prediction tasks, allowing them to capture contextual information from both directions for all layers [32]. While the early layers capture information at the expression level, the intermediate layers create a hierarchy of language features that progress from the surface to the syntactic and semantic levels [33]. BERT has demonstrated superior performance in complex tasks such as classification, prediction, and opinion mining [34]. Their success has sparked great interest in the machine learning community and opened new doors for language understanding research [31]. So, we selected them for the current text classification problem to compare with the machine learning results to answer the question: Is increasing the model complexity necessary for the current task?

Among the available Turkish BERT models, we chose two prominent ones. The first one is "bert-base-multilingual-uncased" pretrained in multiple languages, including Turkish [35]. The second one, which also has the BERT base structure, is only pretrained on Turkish texts and is called BERTurk [36]. Hence, we compare the performance of these two transformers in capturing thoughts and ideas from Turkish social media posts. We add a fully connected linear layer to these pretrained BERT base models as in Figure 5. This classification layer has three neurons, each learning to detect one of the three categories: antivaccine, provaccine, and neutral.

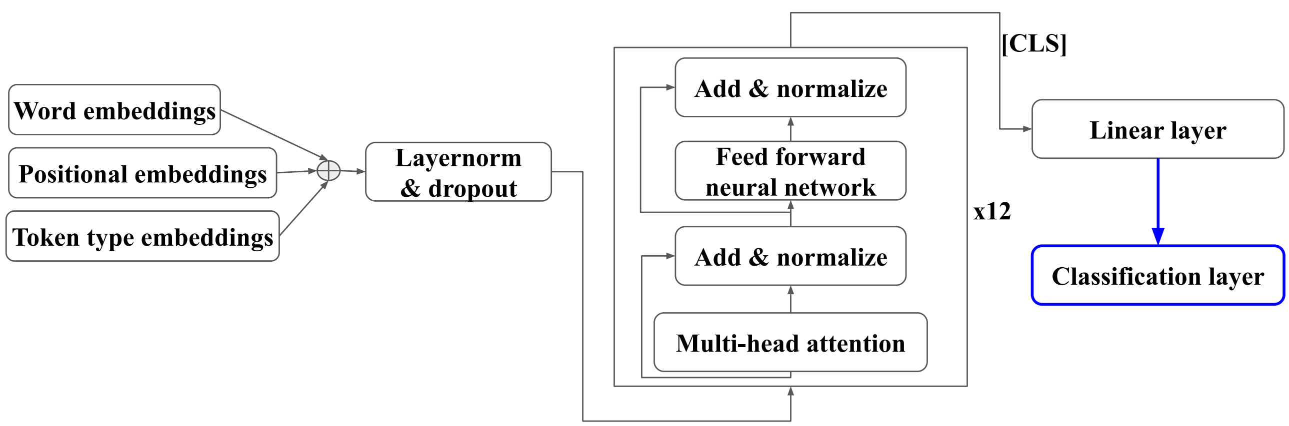


Figure 5. The BERT base framework and the added classification layer.

During the implementation, we used the AdamW optimizer, known for its effectiveness in fine-tuning transformer-based models by preventing overfitting through weight decay, alongside the CrossEntropyLoss function ideal for multiclass classification. We set the maximum token length to 256 and padded the shorter posts. A small initial learning rate (0.00005) is selected to allow stable fine-tuning convergence where the rate changes dynamically as the training progresses. Gradient clipping is applied to avoid unstable updates caused by exploding gradients. We set the batch to 32 posts, early stop patience to 5, and the number of epochs to 100. At each fold, we used 20% of the training set for validation. We experimented with two fine-tuning scenarios. First, we froze all layers of the pretrained models and tuned the classification layer to test their capacity to detect the Turkish antivaccine context. Second, we fine-tuned all layers to compare with the first experiment.

[7]`https://huggingface.co/`

4.3. Performance evaluation

There are many approaches to measuring classification performance. We considered the following factors: (1) multiclass classification and (2) imbalanced dataset. Metrics like accuracy would favor predictions for the majority classes, masking poor detection of minority classes. Metrics like precision and recall also depend on the chosen probability threshold, commonly 0.5, for classification, which might be arbitrary or suboptimal for imbalanced classes. ROC-AUC provides a robust measure of classifier performance across various decision thresholds by plotting the true positive rate (TPR) against the false positive rate (FPR), practical in real-world applications where a fixed decision threshold may not always be practical [37]. By macro-averaging, we prevent the dominant influence of more frequent classes in the imbalanced scenario. To ensure robustness, we reported the mean and standard deviation of ROC-AUC scores across five-fold cross-validation, offering a reliable estimate of generalization performance. Additionally, we employed confusion matrices to determine in what proportions the three classes get mixed during classification [38].

5. Experiments and findings

The Algorithm 1 summarizing the current study shows that we evaluated the machine learning methods for their ability to capture differences between vaccine hesitancy, vaccine support, and neutral thoughts in Turkish social media posts. We ensured that test data never leaks into the models or the training data or parameter selection, which is one of the essential things to consider when conducting machine learning studies on social media data [39].

Algorithm 1 The complete machine learning experiments' algorithm.

y : Labels per post
X : Features per post
$n_{folds} = 5$
for $i = 1 : n_{folds}$ **do**
  X_{test}, y_{test} : Test set of the current fold i
  X_{rest}, y_{rest} : Remaining set of the current fold i
  Ensure ($X_{test} \wedge X_{rest} = \emptyset$)
  Ensure ($X_{test} \vee X_{rest} = X$)
  $X_{rest}[K] \leftarrow$K best features using 5-fold cross-validation and logistic regression on X_{rest}, y_{rest}
  for Each machine learning classification method $Model()$ **do**
    Find optimum classifier parameters with optuna using X_{rest}
    Train the optimum model $Model(X_{rest}, y_{rest})$
    $y_{predicted} \leftarrow Model(X_{test})$
    Collect ROC-AUC score by comparing $y_{predicted}$ and y_{test}
  end for
end for
For each model, report the mean and standard deviation ROC-AUC scores of n_{folds}

The mean values of the five-fold cross-validation results of the machine learning experiments and their standard deviations are in Table 4. The baseline LR shows an acceptably high score of 73.5% AUC on the challenging multiclass classification task on imbalanced data. It is noteworthy that LR performs slightly better than the rbf SVM method with 72.7% AUC, which used to be one of the most popular classifiers in the NLP problems of the past [16]. Meanwhile, the two contemporarily popular machine learning methods, LightGBM and XGBoost, deliver higher scores on this complex problem, whereas XGBoost provides a higher AUC of

75.4%. All the standard deviations are around 1%, which shows that different test sets/folds of cross-validation did not affect the classification decision, proving the generalizability of the findings.

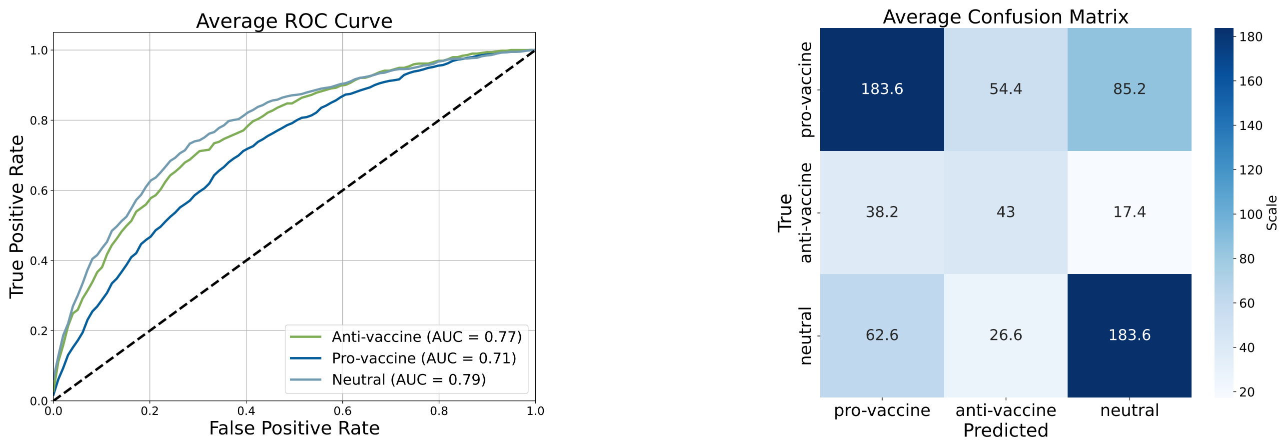
Table 4. Machine learning and NLP methods' classification performance on distinguishing Turkish vaccine hesitancy, vaccine support, and neutral thoughts.

	Methods	ROC-AUC scores (%)
Machine Learning	LR	73.5 ± 1.8
	SVM	72.7 ± 1.2
	LightGBM	74.3 ± 1.3
	XGBoost	75.4 ± 1.6
Transformers, Last Layer	BERT Multilingual	70.1 ± 0.8
	BERTurk	75.1 ± 0.8
Transformers, All Layers	BERT Multilingual	71.0 ± 2.1
	BERTurk	**75.9 ± 1.0**

The same framework in Algorithm 1 is utilized for the transformers experiments, which are at the heart of every NLP research in recent years. However, unlike the machine learning methods, transformers directly work on the raw textual data. Thus, we did not employ the feature engineering parts in Algorithm 1 in the following experiments. At each fold, a transformer model is fine-tuned from scratch using the X_{rest} samples of the current fold and then used to classify the current X_{test}. Results in Table 4 display two different transformer results. First, we froze the pretrained BERT model parameters and fine-tuned only the added classification layers to evaluate how much context the Turkish pretrained models deliver. According to the results of these "tuning only the classification layers" experiments, the BERTurk model returned much higher performance than the multilingual BERT. The score of 75.1% is almost as high as the previous XGBoost model's performance. Meanwhile, these BERT experiments have insignificant standard deviations between five folds with 0.8% ROC-AUC score, confirming the generalizability of the findings.

In the last experiment, we fine-tuned all transformer layers compared to the previous findings of fine-tuning only the last layer. Yet, despite tuning all layer weights, the multilingual BERT's performance only improved by 1% while the standard deviation increased to 2%. This result indicates that the pretrained multilingual BERT model is not a good choice for capturing contexts from Turkish social media data, and fine-tuning all layers is not enough to fill the gap. Meanwhile, the results of the BERTurk model reveal the opposite outcomes. According to Table 4, the highest performance among all experiments belongs to the BERTurk after fine-tuning all layers with the training data. Further, the standard deviation between the five folds is only 1%. Therefore, BERTurk not only outperforms the multilingual BERT model but also outperforms all of the methods.

A ROC-AUC plot visually demonstrates how well the classifier balances sensitivity (true positive rate) and specificity (false positive rate) for each class. To understand BERTurk's success, Figure 6a shows the method's averaged (out of the five-fold cross-validation framework) ROC-AUC plots per class. The AUC score of the neutral class is 79%, while the remaining two classes have 71% AUC scores, respectively. This finding could suggest that the neutral class has more distinct contextual patterns that make it easier for the model to identify. To better understand which classes got mixed up during the classification with BERTurk, we plot averaged (out of the five-fold cross-validation framework) confusion matrix in Figure 6b. The matrix rows correspond to actual labels, and the columns display the predicted labels. Results show that the most confusion within the antivaccine data was with the pro-vaccine posts, which can be explained by the posts with jokes or irony. Also, pro-vaccine posts are confused with neutral posts, and the same is true for the neutral posts mixing with the pro-vaccine class.



(a) **(b)**

Figure 6. Five-fold cross-validation results of the fully fine-tuned BERTurk model.

6. Discussion and ethical implications

Numerous Turkish misinformation text classification studies summarized in the background section found BERTurk to be the best-performing method among the compared ones [4, 6, 14]. Our results validate this finding by revealing BERTurk as the best method through a high mean and a low standard deviation AUC score from the five-fold cross-validation experiments. Further, BERTurk with a single-layer tuned performs almost as well as BERTurk with all layers tuned. Compared to the BERT Multilingual results, all these findings highlight that even the simple frozen BERTurk is powerful in capturing the Turkish context. One can argue that there is merely a difference between the LR and BERTurk results, which is 2.4% AUC. However, considering the feature engineering steps applied after extracting n-grams while BERTurk uses the raw texts without feature engineering, the results further validate the feasibility of employing the BERTurk model.

The highest performance we achieved is 75.9% AUC. Although this score is not as high as those we summarized in the background section, achieving scores above 90% in multiclass classification of real-life Turkish texts is unrealistic. Yet, we should consider human error in the labeling process. Despite the agreement score between two annotators, agreement on incorrect labels is possible. Thus, increasing the number of annotators and changing the labels by a majority vote can improve the observed performance.

The present study also considers ethical implications that are important to acknowledge. As with any social media dataset, there is a risk that the evaluated models may unintentionally learn latent biases, such as those related to race, sex, religion, or age [39]. However, because social media posts typically contain limited to no personal information, systematically identifying and mitigating these biases is out of reach. This inherent limitation is one of the challenges of working with social media data. Another ethical concern is the potential real-world application of our framework. Automatically detecting, moderating, or even removing antivaccine posts could raise concerns about freedom of speech. Moderators and policymakers should carefully consider these implications, particularly given the possibility of false positives where posts that do not express antivaccine opinions might be mistakenly blocked.

Throughout this study, we maintained an objective approach — from data collection to labeling and model evaluation — ensuring that our personal views on vaccines did not influence the process. For the purposes of this research, and in line with established scientific consensus, we assumed that vaccines are generally safe and necessary for public health. However, if that consensus were to change for a specific vaccine in the future, the current antivaccine labels might need reconsideration, discussions would need to adapt to reflect new scientific findings, and the models might require updates to avoid misclassifying fact-based concerns as misinformation. Until then, our framework and results remain valid within the existing scientific understanding.

7. Conclusion

Throughout this study, we demonstrated that a BERT model, pretrained in Turkish, can correctly classify provaccine, antivaccine, and neutral Turkish social media posts, with results reaching an AUC score of 75.9%. Meanwhile, the model did not need to be fully fine-tuned to achieve 75% AUC. This finding is substantial enough to highlight the under-researched Turkish vaccine hesitancy problem. From a public health and policy perspective, we underline the growing challenge of vaccine hesitancy in Türkiye, where a lack of regulation on social media platforms allows misinformation to spread uncontrolled. By leveraging artificial intelligence for automated content moderation, social media platforms could help reduce the power of harmful misinformation, ultimately promoting a more informed and positive public perception of vaccines. Hence, we provided a foundation for future research on fighting vaccine hesitancy. Further, beyond the immediate application to Turkish social media, our findings contribute to the broader field of NLP research: The superior performance of transformers validates the importance of continued investment in developing language-specific models rather than relying on older machine learning techniques. These insights can guide NLP researchers to train more non-English transformers, which would improve the state-of-the-art in low-resource languages.

In future work, we will expand the dataset to incorporate other misinformation varieties to evaluate the possibility of moderating widespread detrimental posts. Additionally, we will investigate how we can utilize network science for the problem to propose a more detailed understanding of misinformation. Another promising direction is the integration of multilingual datasets to introduce a cross-cultural perspective, helping to identify a global approach to antivaccine and other harmful ideologies and providing insights applicable beyond the Turkish context.

References

[1] Conroy NK, Rubin VL, Chen Y. Automatic deception detection: Methods for finding fake news. Proceedings of the association for information science and technology. 2015;52 (1):1-4. https://doi.org/10.1002/pra2.2015.145052010082

[2] Jolley D, Douglas KM. The effects of anti-vaccine conspiracy theories on vaccination intentions. PloS one. 2014 Feb 20;9 (2):e89177. https://doi.org/10.1177/0894439318795849

[3] Sher L. COVID-19, anxiety, sleep disturbances and suicide. Sleep medicine. 2020 Jun;70:124. https://doi.org/10.1016/j.sleep.2020.04.019

[4] Bozuyla M, Özçift A. Developing a fake news identification model with advanced deep languagetransformers for Turkish COVID-19 misinformation data. Turkish Journal of Electrical Engineering and Computer Sciences. 2022;30 (3):908-26. https://doi.org/10.55730/1300-0632.3818

[5] Çılgın C, Gökçen H, Gökşen Y. Twitter'da COVID-19 aşılarına karşı kamu duyarlılığının çoğunluk oylama sınıflandırıcısı temelli makine öğrenmesi ile duygu analizi. Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi. 2023;38 (2):1093-104. https://doi.org/10.17341/gazimmfd.1030198 (in Turkish).

[6] Koru GK, Uluyol Ç. Detection of Turkish Fake News from Tweets with BERT Models. IEEE Access. 2024 Jan 15. https://doi.org/10.1109/ACCESS.2024.3354165

[7] Taskin SG, Kucuksille EU, Topal K. Detection of Turkish fake news in Twitter with machine learning algorithms. Arabian Journal for Science and Engineering. 2022;47 (2):2359-79. https://doi.org/10.1007/s13369-021-06223-0

[8] Kesgin Y, Ünlü DG. COVID-19 aşısı ve yalan haber: Aşılanma öncesinde bireylerin yalan haber içeriklerini fark etme, takip etme ve teyit etme eğilimlerinin belirlenmesi. Galatasaray Üniversitesi İletişim Dergisi. 2021;30 (35):32-55. https://doi.org/10.16878/gsuilet.914875 (in Turkish).

[9] Toğaçar M, Eşidir KA, Ergen B. Yapay Zekâ Tabanlı Doğal Dil İşleme Yaklaşımını Kullanarak İnternet Ortamında Yayınlanmış Sahte Haberlerin Tespiti. Journal of Intelligent Systems: Theory and Applications. 2021;5 (1):1-8. https://doi.org/10.38016/jista.950713 (in Turkish).

[10] Kozik R, Kula S, Choraś M, Woźniak M. Technical solution to counter potential crime: Text analysis to detect fake news and disinformation. Journal of Computational Science. 2022;60:101576. https://doi.org/10.1016/j.jocs.2022.101576

[11] Devarajan GG, Nagarajan SM, Amanullah SI, Mary SS, Bashir AK. AI-assisted deep NLP-based approach for prediction of fake news from social media users. IEEE Transactions on Computational Social Systems. 2023 Mar 29. https://doi.org/10.1109/TCSS.2023.3259480

[12] Mayank M, Sharma S, Sharma R. DEAP-FAKED: Knowledge graph based approach for fake news detection. In2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) 2022 (pp. 47-51). IEEE. https://doi.org/10.1109/ASONAM55673.2022.10068653

[13] Zovikoğlu M, Çetin U. Detecting misinformation on social networks with NLP. Transactions on Computer Science and Applications. 2024;1 (1):11-6. https://doi.org/10.1016/j.sleep.2020.04.019

[14] Bozuyla M. Advanced Turkish fake news prediction with bidirectional encoder representations from transformers. Konya Journal of Engineering Sciences. 2022;10 (3):750-61. https://doi.org/10.36306/konjes.995060

[15] Ruchansky N, Seo S, Liu Y. Csi: A hybrid deep model for fake news detection. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management 2017 Nov 6 (pp. 797-806). https://doi.org/10.1145/3132847.3132877

[16] Pestian J, Santel D, Sorter M, Bayram U, Connolly B et al. A machine learning approach to identifying changes in suicidal language. Suicide and Life-Threatening Behavior. 2020;50 (5):939-47. https://doi.org/10.1111/sltb.12642

[17] Köksal A, Özgür A. Twitter dataset and evaluation of transformers for Turkish sentiment analysis. In2021 29th Signal Processing and Communications Applications Conference (SIU) 2021 Jun 9 (pp. 1-4). IEEE. https://doi.org/10.1109/SIU53274.2021.9477814

[18] Warrens MJ. Five ways to look at Cohen's kappa. Journal of Psychology & Psychotherapy. 2015; 28:5. https://doi.org/10.4172/2161-0487.1000197

[19] Bayram U, Pestian J, Santel D, Minai AA. What's in a word? Detecting partisan affiliation from word use in congressional speeches. In2019 International Joint Conference on Neural Networks (IJCNN) 2019 Jul 14 (pp. 1-8). IEEE. https://doi.org/10.1109/IJCNN.2019.8851739

[20] Zhang Y, Yang Y. Cross-validation for selecting a model selection procedure. Journal of Econometrics. 2015;187 (1):95-112. https://doi.org/10.1016/j.jeconom.2015.02.006

[21] Bates S, Hastie T, Tibshirani R. Cross-validation: what does it estimate and how well does it do it?. Journal of the American Statistical Association. 2024;119 (546):1434-45. https://doi.org/10.1080/01621459.2023.2197686

[22] Domingos P. A few useful things to know about machine learning. Communications of the ACM. 2012;55 (10):78-87. https://doi.org/10.1145/2347736.2347755

[23] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B et al. Scikit-learn: Machine learning in Python. the Journal of machine Learning research. 2011; 12:2825-30. https://doi.org/10.5555/1953048.2078195

[24] Bayram U, Benhiba L. Determining a person's suicide risk by voting on the short-term history of tweets for the clpsych 2021 shared task. In Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access 2021 Jun (pp. 81-86). https://doi.org/10.18653/v1/2021.clpsych-1.8

[25] Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A next-generation hyperparameter optimization framework. In Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining 2019 Jul 25 (pp. 2623-2631). https://doi.org/10.1145/3292500.3330701

[26] Kim HY. Statistical notes for clinical researchers: logistic regression. Restorative Dentistry & Endodontics. 2017;42 (4):342-8. https://doi.org/10.5395/rde.2017.42.4.342

[27] Hearst MA, Dumais ST, Osuna E, Platt J, Scholkopf B. Support vector machines. IEEE Intelligent Systems and their applications. 1998;13 (4):18-28. https://doi.org/10.1109/5254.708428

[28] Bentéjac C, Csörgő A, Martínez-Muñoz G. A comparative analysis of gradient boosting algorithms. Artificial Intelligence Review. 2021;54:1937-67. https://doi.org/10.1007/s10462-020-09896-5

[29] Drin S, Kriuchkova A, Toloknova V. Predictive model for a product without history using LightGBM. pricing model for a new product. https://doi.org/10.1109/VDAT53777.2021.9600900

[30] Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining 2016;785-794. https://doi.org/10.1145/2939672.2939785

[31] Buche A, Chandak MB. Bert for opinion mining and sentiment farming. Helix. 2020;10 (04):374-8. https://doi.org/10.21786/BBRC/13.14/9

[32] Tokuma W. Business Application of BERT, a General-purpose Natural-language-processing Model. NTT Technical Review. 2021;19 (5):68-73. https://doi.org/10.53829/ntr202105fa11

[33] Jawahar G, Sagot B, Seddah D. What does BERT learn about the structure of language?. In ACL 2019-57th Annual Meeting of the Association for Computational Linguistics 2019 Jul 28. https://doi.org/10.18653/v1/P19-1356

[34] Aftan S, Shah H. A survey on bert and its applications. In2023 20th Learning and Technology Conference (L&T) 2023 Jan 26 (pp. 161-166). IEEE. https://doi.org/10.1109/LT58159.2023.10092289

[35] Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. 2018 Oct 11. https://doi.org/10.48550/arXiv.1810.04805

[36] Schweter S. BERTurk: BERT models for Turkish. Version 1.0.0. Zenodo. 2020 Apr. Available from: https://doi.org/10.5281/zenodo.3770924

[37] Hajian-Tilaki K. Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. Caspian journal of internal medicine. 2013;4 (2):627. https://doi.org/10.1016/j.sleep.2020.04.019

[38] Amin F, Mahmoud M. Confusion matrix in binary classification problems: A step-by-step tutorial. Journal of Engineering Research. 2022;6 (5). https://doi.org/10.21608/erjeng.2022.274526

[39] Bayram U. Applying Machine Learning to Online Data?: Beware! Computational Social Science Requires Care. In Opportunities and Challenges for Computational Social Science Methods 2022 (pp. 100-125). IGI Global. https://doi.org/10.4018/978-1-7998-8553-5.ch005