Black Sea Journal of Engineering and Science
doi: 10.34248/bsengineering.1527551

Open Access Journal
e-ISSN: 2619 – 8991

Research Article
Volume 8 - Issue 1: xx-xx / January 2025

FAKE NEWS DETECTION ON MAINSTREAM MEDIA USING NATURAL LANGUAGE PROCESSING

İsa KULAKSIZ[1]*, Ahmet COŞKUNÇAY[2]

[1]*Atatürk University, Faculty of Engineering, Department of Computer Engineering, 25000, Erzurum, Türkiye*
[2]*Atatürk University, Faculty of Engineering, Department of Computer Engineering, 25000, Erzurum, Türkiye*

Abstract: In light of recent advances in online journalism, the diversity, abundance, and accessibility of news have increased exponentially. However, the growth of online journalism also brings issues, especially regarding the reliability of the news. Notably, news widely shared on social media during the US presidential election campaign and the UK Brexit referendum caused millions of reactions from the public. This concerning scenario prompted industry and academia to address the pressing issue of fake news. Detecting fake news is a meticulous, time-consuming, and labor-intensive task that requires expert judgment. To mitigate this challenge, this study proposes a linguistic based model for Turkish fake news detection. In this dataset was collected from TRT's RSS service and through web scraping from the Teyit.org platform. It contains news titles and summaries related to significant events in Türkiye between 2015 and 2023. The research compares classical machine learning classifiers including SVM, Logistic Regression, Random Forest, k-NN, Decision Tree, and Naive Bayes, against a neural based sequential learning model such as LSTM using real world datasets. Furthermore, the research investigates the impacts of different word representation techniques, including TF-IDF and CountVectorizer, and also hyperparameter optimization on the classification results. The findings revealed that using hyperparameter tuning, the TF-IDF method yielded the highest accuracy rate of 93.12% on the SVM model and that TF-IDF is more effective.

Keywords: Fake News Detection, Machine Learning, Classification, LSTM, NLP

***Corresponding author:** Ataturk University, Faculty of Engineering, Department of Computer Engineering, 25000, Erzurum, Türkiye
E mail: isakulaksiz@outlook.com (İ. KULAKSIZ)
İsa KULAKSIZ https://orcid.org/0009-0000-1138-7130
Ahmet COŞKUNÇAY https://orcid.org/0000-0002-7411-310X

Received: August 07, 2024
Accepted: December 19, 2024
Published: January 15, 2025

Cite as: Kulaksız İ, Coşkunçay A. 2025. Fake news detection on mainstream media using natural language processing. BSJ Eng Sci, 8(1): xx-xx.

1. Introduction

In the initial phase of the World Wide Web or Web 1.0 (1995 -onwards), its content primarily consisted of monologues meant to convey ideas mainly generated by mainstream media agencies (i.e. Haberturk, TRT). However, Web 2.0 (2000 -onwards) facilitated the sharing of news by ordinary users by empowering them to create large-scale interactive content on social media platforms (Choudhury, 2014; Lazer et al., 2018) such as Facebook and X. The consequential impact of news production is significant; it spreads rapidly throughout society and generates both positive and negative information regarding journalism, democracy, and political agendas. In addition, certain entities with opposing viewpoints have used similar platforms to make money (Kucharski, 2016; Garcia et al., 2020) or to create biased opinions on social media platforms. The events surrounding the 2016 United Kingdom's withdrawal from the European Union (i.e. Brexit) and the results of the US presidential election notably heightened the prevalence of fake news in society (Kucharski and Adam, 2016). While numerous studies in Indo-European languages focus on detecting fake news, the analysis of Turkish presents a significant challenge due to its distinctive morphological structure. This is primarily because Turkish's classification within the Ural-Altaic language family allows for the addition of multiple suffixes to a word root, resulting in the formation of intricate "beads-on-a-string" structures (Oflazer, 2014; Yamanan, 2016).

- Çekoslovakyalılaştıramadıklarımızdan mısınız?
- "Are you one of those who we could not make czechoslovakian?"

This example highlights the complexity, but native speakers can readily understand many words of similar size (Çöltekin, 2014). Distinguishing the fakeness of a news story requires an expert's meticulous attention. Fake news spreads misleading claims in its content, so the best way to detect it is to inspect news articles for hints. In (Shu et al., 2017), the topic of fake and real news articles is discussed. TRT dataset includes real news has a stronger relationship with the agenda. The articles collected from TRT news sources were cross-validated. This necessitates using Natural Language Processing (NLP) techniques to automate it. This technique eliminates stopwords (prepositions, conjunctions), stemming, and word representation. In order to understand how news can influence society, it is essential to examine the relationship between journalistic

language and media tools. Although television and newspaper news presentations use formal and professional language, ordinary users share news via social media using casual language. These studies contributed to the depth and comprehensiveness of the study. Figure 1 shows a screenshot of the TRT and Teyit websites.

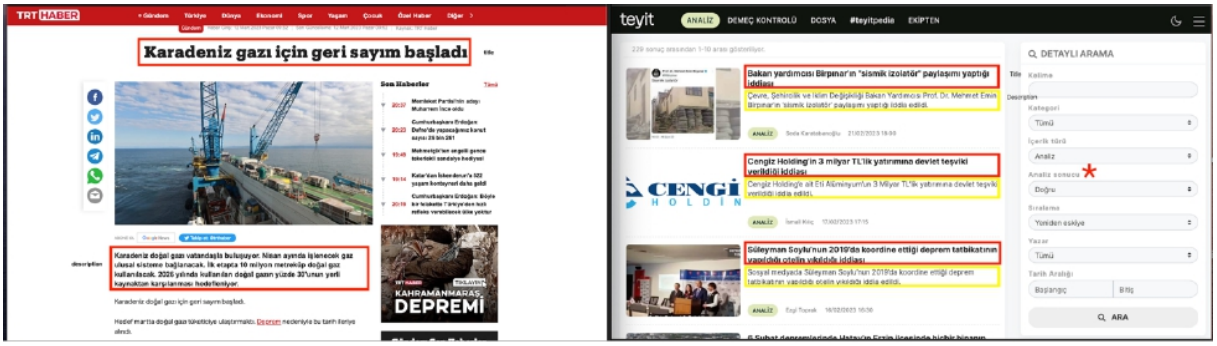

Figure 1. An example of news shared by TRT and Teyit.org platform.

Several communities are involved in this effort, including Teyit.org, which utilizes an expert-oriented fact-checking method. This platform verifies whether the news item is real or fake and that a reputable news source issues it. Notably, Teyit.org is part of IFCN (Ünver, 2023). In this study, the open source library Zemberek is used for stemming and lemmatization in Turkish NLP (Akın and Akın, 2007).

This research investigates the classification performance of the classical machine learning (i.e., SVM, Logistic Regression, Random Forest, k-NN, Decision Tree, and Naive Bayes) and Long Short-Term Memory (LSTM) algorithms using real world dataset including TRT and Teyit.org. It also uses n-gram, bag of words, and word representation methods (TF-IDF and CountVectorizer) for feature extraction. Furthermore, pre-processing steps are performed using the Zemberek library, and structural similarity is calculated to illustrate this difference. For classification tasks, classical machine learning algorithms were preferred over transformer-based models (e.g., BerTurk, Roberta) due to their lower resource requirements.

In section 3. The collection and preprocessing stages of datasets were explained. The rest of the paper contains experimental results, a discussion of this research and future natural language processing studies in fake news in section 4, and the conclusion in section 5.

2. Related Works

In the field of NLP, research focuses on facilitating communication between people and machines by providing translation, summarization, and the creation of chatbots and other tools to enhance interaction. The studies in this topic area include those that analyze and compare the use of machine learning, LSTM, and transformer algorithms to detect fake news in mainstream and social media. Reis et al, (2019) have achieved the best classification with XGB 86% according to the BuzzFeed set of tagged posts by journalists during the 2016 US presidential campaign. In contrast, Kaliyar et al. (2021a, 2021b) proposed an EchoFakeD model based on ANN, which reached 92.30% accuracy and the CNN and Bert models reached 98.90% accuracy. In addition to studies using XGB and machine learning algorithms (Khanam et al., 2021), the BI-LSTM deep learning algorithm was also used (Aslam et al., 2021) in the public LIAR dataset collected from news sites such as Politifact and FactCheck.org. The BI-LSTM model achieved the highest accuracy of 89%, followed by XGBoost at 75% and Random Forest at 73%. In a study on detecting fake news from English news shared on X, Monti et al. (2019) CNN achieved a remarkable 92.7% accuracy, significantly higher than Random Forest 87% and Geometric Deep Learning 73% by Meyers et al. (2020). Another deep learning method is LSTM 82% by Ajao et al. (2018). These results indicate that deep learning methods provide higher accuracy. Ahmad et al. (2020) presented a study on the effects of machine learning algorithms compared on DS4 created by combining DS2, DS3 and ISOT datasets taken by the Kaggle platform. In terms of results, Random Forest and Perez-LSVM achieved 99% on the ISOT dataset, Decision Tree and XGBoost achieved 94% on DS2, Perez-LSVM achieved 96% on DS3 and Random Forest achieved 91% on DS4. In 2020, Mertoğlu and Genç (2020) proposed an innovative method of detecting fake news in digital libraries by automating the classification of news to be included therein. On the TRFN dataset created with GDELT, Teyit.org and MVN, the model trained on classification algorithms (K-Nearest Neighbor, Decision Trees, Gaussian Naive Bayes, Random Forest, Support Vector Machines, ExtraTrees Classifier, Logistic Regression) found that the ExtraTrees algorithm provided the best performance with an accuracy of 96.81%. Taskin et al. (2022) presented work on the two most shared topics on social media in Türkiye (i.e., Galatasaray's Falcao transfer). They collected data using Teyit.org and X API and compared the performance of supervised and unsupervised machine learning and deep learning (DL) algorithms in detecting fake news. This study focusing on social media contributed to developing

our model for mainstream media. Based on the model's classification results, the SVM algorithm achieved the highest accuracy at 90%. Bozuyla and Özçift (2022) presented a model analysis of COVID-19 fake news during the pandemic, which analyzed data collected through X API and several Turkish expert-oriented fact-checking systems (i.e. Teyit, Malumatfurus, and Dogrulugune), using machine and deep learning algorithms and transformers to improve accuracy. In the study, BerTURK Turkish transformer detected fake COVID-19 news with 98.5% accuracy. Güler and Gündüz (2023) investigated the comparison of CNN and RNN-LSTM deep learning algorithms with Word2vec word representation approaches using datasets collected from social media using English BuzzFeed, ISOT, and SOSYalan via X APIs. The results showed that the RNN-LSTM algorithm achieved 93.41% and 85.66% CNN accuracy. Koru and Uluyol (2024) achieved 90% to 94% accuracy using BERT and BERTurk+CNN models on the TR_FaRe_News dataset created from tweets of Turkish fake news and BuzzFeedNews, LIAR, GossipCop, ISOT, Twitter15 and Twitter16 datasets pre-processed. Güler and Gündüz (2023) compared a set of Turkish tweets called SOSYalan using CNN and RNN-LSTM algorithms in comparison with BuzzFeed and ISOT. The results showed that the English-based model achieved an accuracy ranging from 85.16% to 99.9% while the Turkish-based model achieved an accuracy ranging from 87.14% to 92.48%. Kaliyar et al. (2020c). proposed a convolutional neural network (CNN) based FDNet model for fake news spread during the controversial US presidential election in 2016. The proposed model achieved a performance of 98.36% on the test dataset Natural Language Processing has proved successful in data mining, summarization and sentiment analysis thanks to classification algorithms and transformer models. NLP has been revolutionized by the development of transformer models such as BERT and GPT which provide advanced capabilities for analyzing and generating human language. These technologies play a critical role in the news verification process by analyzing text and evaluating the accuracy of the information presented.

3. Materials and Methods

The flowchart in Figure 2 illustrates the process of detecting Turkish fake news using machine learning algorithms and LSTM, a modified RNN architecture developed by Hochreiter and Schmidhuber (1993). It involves collecting data, extracting features, training the model with linguistic and similarity analysis, and using the trained model for classification.

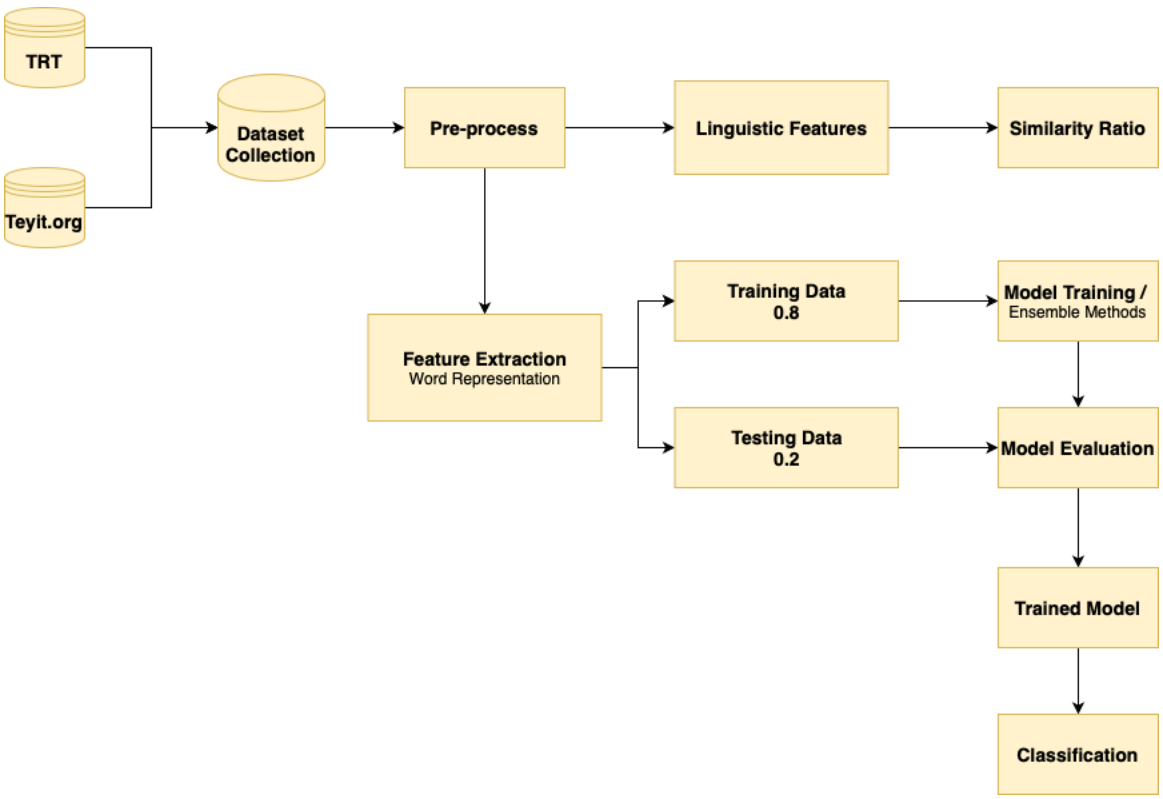

Figure 2. Flowchart for fake news detection.

3.1. Dataset Collection

This study employed supervised learning, utilizing a model trained on the titles and summaries of news articles sourced from TRT Agenda's RSS service and Teyit.org, an expert-oriented fact-checking platform. The news articles from TRT, the official broadcasting organ of the Republic of Türkiye, were collected a few days after the incident. Therefore, if errors or misleading information are detected after publication, necessary corrections will be made to the news. Consequently, it

will be possible to include only the real news instead of those removed from the official Turkish broadcasting channel TRT. As a result of these factors, TRT, the official state broadcaster, was assumed to provide real news. Teyit.org is a platform in Türkiye that verifies the accuracy of claims made in the news. As a member of the International Fact-Checking Network (IFCN), Teyit.org maintains its neutralism by relying on transparent sources and sharing its methodology with the public. Fact-checking is performed by experts and volunteers, as well as machine and deep learning algorithms (Ünver, 2023). The news dataset established and preprocessed in this study contains 5325 records. The amount of fake news received from Teyit.org is significantly higher than that of real news with a ratio of 211 to 2845. Overfitting might happen in cases where some possible outcomes are not represented with enough number samples in the dataset. This is because the overfitted model memorizes the training set without generalizing on the test set, which may negatively affect the classification result. Several methods are available for reducing the risk of overfitting (e.g., early stopping, network reduction, data expansion and regularization) (Ying, 2019). The distribution of resources, rates, and dates within the dataset is illustrated in Figure 3.

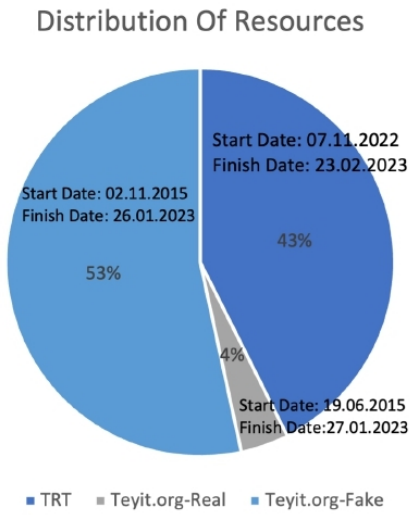

Figure 3. Information about the dataset.

In order to reduce overfitting, the dataset has been expanded with TRT news to ensure better classification results. Additionally, stopwords have been removed from the dataset to prevent the model from memorizing frequently used words during training. Determining whether different data sources have structural similarities will explain why the data collected in the study were combined. Several models for the distance-based measure of similarities between documents are frequently used in natural language processing, including cosine similarity and the Levenshtein distance metrics (Siahaan et al., 2018). The Cosine Similarity uses vector modeling to determine the structural similarity between two text documents. It can take values ranging from 0 to 1 with 0 indicating that the documents are unrelated and 1 indicating identical (Rahutomo et al., 2012; Falah and Suryawan, 2022). The internal similarities between TRT and Teyit.org were calculated by dividing this dataset into two parts and comparing them. TRT's internal similarity was calculated as 0.53 and Teyit.org's internal similarity was calculated as 0.37. This shows that the similarity of news texts within news sources is low. Then, TRT and Teyit.org news texts were compared and their similarities were calculated. The news text similarity of these two news sources is 0.23 as shown in Figure 4.

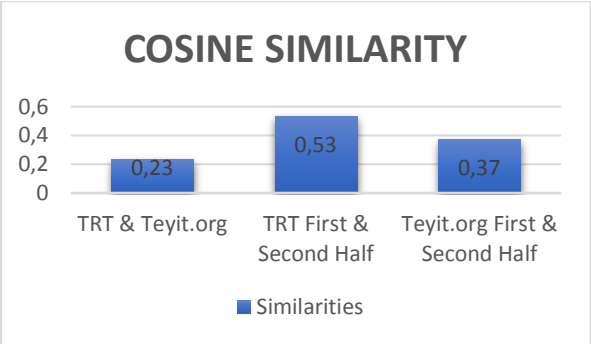

Figure 4. The Cosine similarity ratio.

3.2. Data Preprocessing

Data preprocessing refers to a series of processes that are performed before training machine learning models. In order to achieve these processes, Turkish stopwords were removed, data was transformed, and stemming was applied to ensure consistency and high quality. Noise reduction is one of the most critical steps in data preprocessing. Noise in a dataset refers to inconsistent and inaccurate data. For example, there may be inconsistencies caused by incorrect or missing data. The data must be cleaned of these problems (Han et al., 2011). The steps used are shown in Figure 5.

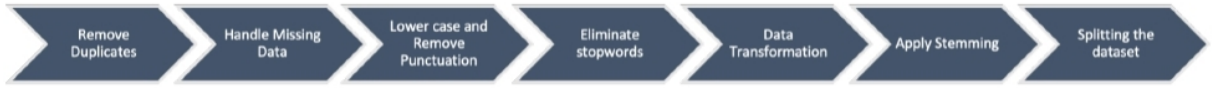

Figure 5. Data preprocessing steps.

One of the most frequent issues of the dataset was regarding the use of the wording "claimed" (i.e., in Turkish, "iddia edildi"). There might be a bias in that many of the news containing this expression are fake news. Since this would cause memorization on the training set, the places where this expression was used were replaced with more neutral and unpretentious expressions. Therefore, the places where this expression was used were replaced with more neutral and unpretentious expressions.

Figure 6. Most frequently used words with WordCloud.

During preprocessing, stopwords are often removed from texts in studies of NLP. By removing words that do not add much valuable information to the language (prepositions, conjunctions, etc.), dimension reduction brings attention to the critical information within the texts. The Turkish stopwords were removed from our dataset. Furthermore, in order not to possibly harm the dignity of individuals and to avoid conflict of interest, proper names have been removed from Wordcloud. In Figure 5, the final preprocessing step is stemming. Stemming is the most common morphological technique in NLP to determine a word's root. Stemming uses a method that allows words with similar meanings to be represented as the same, with the correct cutoff point and hence the size is reduced (Korenius et al., 2004). Many open source NLP libraries work with Indo-European texts. NLP researchers face a challenge due to the excessive use of affixes in Turkish, an agglutinative language. Therefore, linguistic feature based approaches were used to extract the texts' characters, terms and sentences based on the Zemberek library (Akın and Akın, 2007). The dataset is divided into training and testing sets with an 80 to 20 split.

3.3. Feature Extraction

When dealing with very high dimensional datasets, text grouping poses a challenge. The classification results can be adversely affected by unimportant or unconnected words. Therefore, the frequency and importance of words in a document can be measured by several word representation approaches that convert words into numerical vectors (Kaur et al., 2020). In this study, it was transformed into numerical vectors using both TF-IDF and CountVectorizer methods.

TF-IDF (Term Frequency-Inverted Document Frequency)

is a metric used to determine the importance of each word in a text. In this method, the importance of a word in a document increases as its frequency increases. The TF-IDF metric consists of two metrics: term frequency (TF) and inverse document frequency (IDF). The equations 1 and 2 formulas used to compute the tf-idf for a term t within a document set are as follows:

$$tf - idf = tf(t,d) * idf(t) \quad (1)$$

N is the number of documents in the document set and df(t) is the document frequency at t;

$$idf(t) = log\frac{n}{df(t)} + 1 \quad (2)$$

In order to identify fake news, researchers have employed a variety of methods. Bag of Words (BoW) is one of their most essential methods for feature extraction (Ahmed et al., 2017). In one respect, it is similar to Term Frequency (TF) in that word groups (n-grams) can be extracted from the text content based on word frequencies.

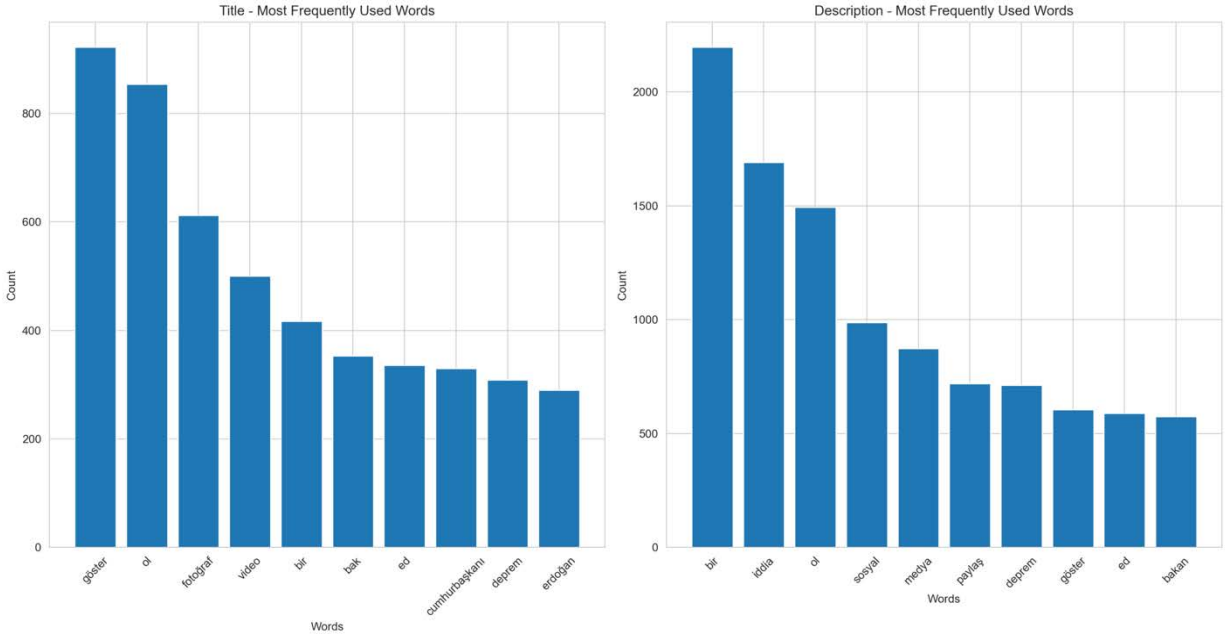

Figure 7. Title and summary: Identifying most frequent words using Bag of Words (BoW) approach.

CountVectorizer is based on word frequency in a text document. It is used for fragmentation (i.e. tokenization) in text processing. CountVectorizer provides many parameters to improve the selection of features. There are three parameters available for building features: unigrams (min d f = 1), bigrams (min d f = 2) and trigrams (min d f = 3) (Korenius et al., 2004). In this study, min d f as 3 was used. To facilitate understanding of the occurrence of each feature in a document, each vector (term) represents the feature name shown in Figure 8.

3.4. Evaluation Metrics

Several evaluation metrics are used to identify fake news, real news, misclassified real news, and undetected fake news. These metrics include True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN). They provide valuable insights into models' performance and effectiveness in distinguishing between different types of news. Significantly, the Accuracy value is the percentage of accurate predictions to be made based on the algorithm over the total number of predictions. Each evaluation metrics: accuracy is equation 3, precision is equation 4, recall (or sensitivity) is equation 5, and f1 is equation 6 is represented below by a mathematical formula.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

$$F1 = 2 \times \frac{Precision \text{ x } Recall}{Precision + Recall} \quad (6)$$

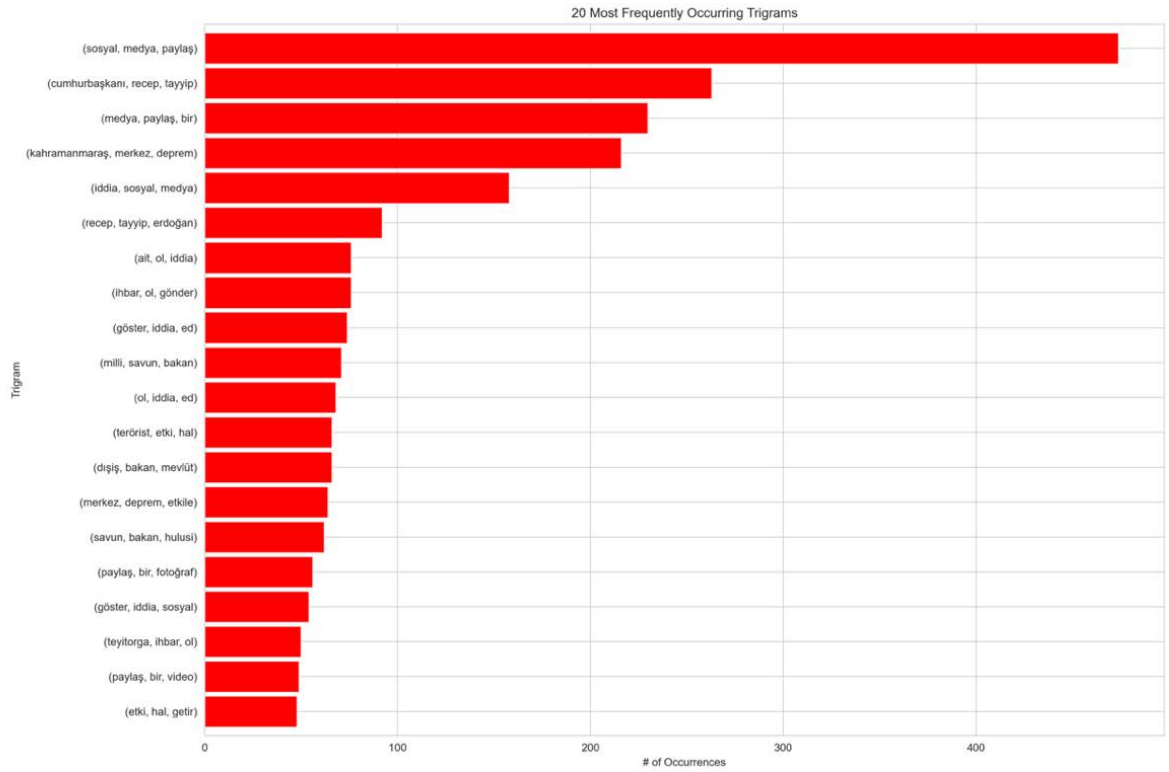

Figure 8. Twenty most frequently used trigram.

4. Results and Discussion

In mainstream media with widespread of fake news the need for artificial intelligence systems to fact-check news has become inevitable. TF-IDF and CountVectorizer word representation methods are evaluated to improve classification accuracy, and hyperparameters are optimized in this study. Additionally, many methods were applied to ensure that the model was free from the effects of extinction gradients and overfitting. This is prevented by randomization, data transformation, removing Turkish stopwords, and stemming. For LSTM, regularization techniques (L2), dropout, and early stopping were applied. In Table 1, illustrates a detailed of the hyperparameters used, which yielded the best-tuned parameters across various algorithms.

In Table 2, illustrates the classifiers with the highest hyperparameter tuning and accuracy using bold characters. Figure 9 and Figure 10 show charts of all evaluation metrics obtained using the tf-idf and countvectorizer methods.

The findings revealed that among the methods evaluated in the research, SVM achieved the highest accuracy rate of SVM (tf-idf 93.12%, countvectorizer 92.86%), followed by LSTM (tf-idf 92.21%, 92.40%) and Logistic Regression (tf-idf 92.72%, countvectorizer 92.75%). Decision Tree and Naive Bayes exhibited lower accuracy, with Decision Tree achieving 87.35% (tf-idf) and 88.40% (CountVectorizer) and Naive Bayes reaching 74.39% (tf-idf) and 75.33% (CountVectorizer). Mainly, LSTM, Logistic Regression, and Random Forest, one of the classification learning algorithms used in the study, were very close competitors to SVM in terms of performance. The findings revealed that TF-IDF had a relatively better effect on classification results than CountVectorizer. Consequently, the best classification algorithm, SVM, has detected 93.12% of fake news. In addition, the model trained with the best-fine-tuned parameters obtained through hyperparameter optimization demonstrated relatively better accuracy. Consequently, SVM and Logistic Regression have been effective for classification problems, producing binary outputs, while LSTM has shown effectiveness through its forget gates.

Table 1. Model hyperparameters and best parameters

Model Name	Default Parameters	Best Parameters
SVM	C: 1.0 Kernel: RBF Gamma: Scale Degree: 3	C: 1 Kernel: RBF
k-NN	Number of Neighbors: 5 Weight: Uniform Algorithm: Auto	Number of Neighbors: 7 Weight: Distance Algorithm: Auto
Logistic Regression	C: 1.0 Penalty: l2 Solver: lbfgs Max Iterations: 100	C: 1 Penalty: l2
Decision Tree	Criterion: Gini Max Depth: None Min Samples Split: 2 Min Samples Leaf: 1	Criterion: Gini Max Depth: 20 Min Samples Split: 10 Min Samples Leaf: 4
Random Forest	Number of Estimators: 100 Max Depth: None Min Samples Split: 2 Min Samples Leaf: 1 Criterion: Gini Bootstrap: True	Number of Estimators: 100 Max Depth: None Min Samples Split: 2 Min Samples Leaf: 1
Naive Bayes	-	-
LSTM	Optimizer: Adam Activation: Relu & Sigmoid Loss: Categorical Crossentropy Epochs: 10/7 (Early Stopping) Dropout: 0.2 Batch Size: 32	

Table 2. Model evaluation metrics with hyperparameter tuning and test accuracy

Model	Hyperparameter Tuning (%)		Test Accuracy (%)	
	TF-IDF	CountVectorizer	TF-IDF	CountVectorizer
SVM	93.12	92.86	92.40	91.65
Logistic Regression	92.72	92.75	91.74	91.46
Random Forest	92.61	92.72	92.12	91.56
k-NN	87.86	84.23	84.62	82.93
Decision Tree	87.74	88.15	85.74	86.30
Naive Bayes	74.39	75.33	74.39	75.33
LSTM	92.21	92.40	92.21	92.40

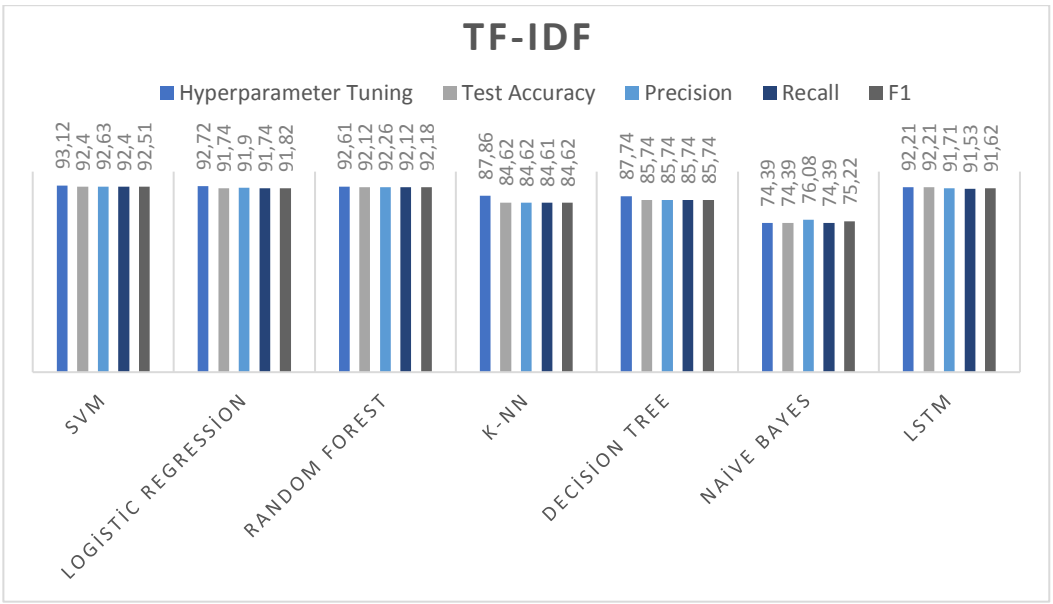

Figure 9. Evaluation metrics obtained by TF-IDF vectorization method.

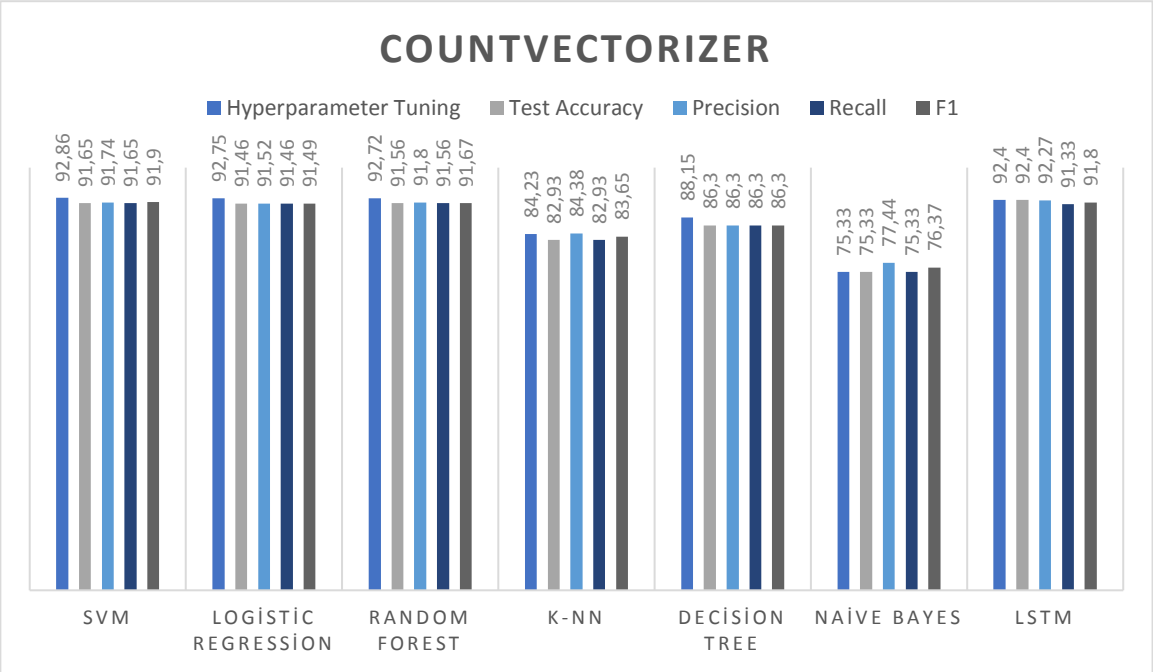

Figure 10. Evaluation metrics obtained by Countvectorizer vectorization method.

This study evaluated the performance of previous research conducted in the Turkish language and the algorithms used. Taskin et al. (2022) achieved 90% accuracy with the SVM algorithm. Bozuyla and Özçift (2022) achieved an accuracy of 98.5% using the BerTURK model. Güler and Gündüz (2023) obtained accuracies ranging from 85.66% to 93.41% with CNN and RNN-LSTM. Koru and Uluyol (2024) demonstrated that The_News dataset showed performance between 90% and 94% with BERT and BERTurk+CNN algorithms. Using the SOSYalan dataset, Koru and Uluyol (2024) achieved 87.14% and 92.48% accuracy with CNN and RNN-LSTM algorithms, respectively. This study has demonstrated that the SVM algorithm achieved an accuracy of 93.12%, which is better than many other models. This model outperformed other approaches in several studies, except for the BerTURK approach by Bozuyla and Özçift (2022). Therefore, SVM algorithms have indicated effectiveness in detecting Turkish fake news.

In addition, future studies are planned to increase the quality and quantity of the datasets used for model training, which will likely result in continuous improvements in the performance of supervised learning classification algorithms. These studies are purposed to improve the accuracy and reliability of fake news detection on mainstream and social media by using their results. Furthermore, this work may be extended to other ensemble methods.

5. Conclusion

Traditional and social media contains substantial news that damages society and the economy psychologically and materialistically. In recent years, it has been seen that the need for expert-based control systems to distinguish the accuracy of news has decreased with the developments in the field of artificial intelligence. In this study, the data collected from TRT and Teyit.org news sources were classified as fake and real news using hyperparameter tuning with machine learning and LSTM algorithms. Furthermore, Zemberek, one of the best libraries on Turkish texts, was used in the data pre-processing stage for stemming and lemmatization. When the results were investigated, it was seen that the highest classification accuracy was achieved with tf-idf 93.12% SVM. In addition to feature extraction using tf-idf and counvectorizer, n-gram and BoW methods are also applied. In future studies, transformer models (GPT or BERT) or boosting algorithms can be trained and compared on the dataset.

Author Contributions

The percentages of the authors contributions are presented below. All authors reviewed and approved the final version of the manuscript.

	İ.K.	A.C.
C	70	30
D	70	30
S	40	60
DCP	70	30
DAI	70	30
L	80	20
W	70	30
CR	70	30
SR	80	20
PM	70	30

C=Concept, D= design, S= supervision, DCP= data collection and/or processing, DAI= data analysis and/or interpretation, L= literature search, W= writing, CR= critical review, SR= submission and revision, PM= project management, FA= funding acquisition.

Conflict of Interest

The authors declared that there is no conflict of interest.

Ethical Consideration

Ethics committee approval was not required for this study because of there was no study on animals or humans.

Acknowledgments

This research is based on a master's thesis.

References

Ahmad I, Yousaf M, Yousaf S, Ahmad M. 2020. Fake news detection using machine learning ensemble methods. Complexity, 2020: 8885861. https://doi.org/10.1155/2020/8885861

Ahmed H, Traore I, Saad S. 2017. Detection of online fake news using n-gram analysis and machine learning techniques. International Conference On Intelligent, Secure, And Dependable Systems In Distributed And Cloud Environments, 28-30 November; Vancouver, Canada, pp: 127-138.

Akın A, Akın M. 2007. Zemberek, an open source NLP framework for Turkic languages. Structure, 10(2007): 1-5.

Ajao O, Bhowmik D, Zargari S. 2018. Fake news identification on twitter with hybrid cnn and rnn models. SMSociety '18: International Conference on Social Media and Society, July 18-20, New York USA, pp: 226-230.

Aslam N, Khan I, Alotaibi F, Aldaej L, Abdulbaikil A. 2021. Fake detect: A deep learning ensemble model for fake news detection, Complexity, 2021(4): 1-8.

Bozuyla M, Özçift A. 2022. Developing a fake news identification model with advanced deep language transformers for Turkish COVID-19 misinformation data. Turk J Electr Eng Comput Sci, 30(3): 908–926.

Choudhury N. 2014. World wide web and its journey from web 1.0 to web 4.0. Int J Comput Sci Inf Technol, 5(6): 8096–8100.

Çöltekin Ç. 2014. A set of open source tools for Turkish natural language processing. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), 26-31 May, Reykjavik Iceland, pp: 1079–1086.

Falah Z, Suryawan F. 2022. Recommendation system to propose final project supervisors using cosine similarity matrix. Khazanah Informatika: Jurnal Ilmu Komputer dan Informatika, 8(2).

García S, Garcia G, Prieto M, Guerrero A, Jimenez, C. 2020. The impact of term fake news on the scientific community. Scientific performance and mapping in web of science. Soc Sci, 9(5): 73.

Güler G, Gündüz S. 2023. Deep learning based fake news detection on social media. Int J Inf Secur, 12(2): 1-21.

Han J, Kamber M. 2011. Data mining: concepts and techniques. Morgan Kaufmann, Elsevier, Waltham, MA 02451, USA, pp: 47-67.

Hochreiter S, Schmidhuber J. 1997. Long short-term memory. Neural Comput, 9(8), 1735–1780.

Kaliyar R, Goswami A, Narang P. 2021a. FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. Multimed Tools Appl, 80(8): 11765-11788.

Kaliyar R, Goswami A, Narang P. 2021b. EchoFakeD: improving fake news detection in social media with an efficient deep neural network. Neural Comput Appl, 33(14): 8597-8613.

Kaliyar R, Goswami A, Narang P. 2020c. FNDNet – A deep convolutional neural network for fake news detection, Cogn Syst Res, 61: 32-44.

Kaur S, Kumar P, Kumaraguru P. 2020. Automating fake news detection system using multi-level voting model. Soft Comput, 24(12): 9049-9069.

Khanam Z, Alwasel B, Sirafi H, Rashid M. 2021. Fake news detection using machine learning approaches. International Conference on Applied Scientific Computational Intelligence using Data Science (ASCI 2020), 22-23 December, Jaipur India.

Korenius T, Laurikkala J, Jarvalin K, Juhola M. 2004. Stemming and lemmatization in the clustering of Finnish text documents. Proceedings of the thirteenth ACM international conference on Information and Knowledge Management, 13 November, New York USA, pp: 625–633.

Koru G, Uluyol Ç. 2024. Detection of Turkish fake news from tweets with BERT models. IEEE, 12: 14918-14931.

Kucharski, A. 2016. Study epidemiology of fake news. Nature, 540(7634): 525-525.

Lazer D, Baum M, Benkler Y, Berinsky A, Greenhill K, Menczer F, Metzger M, Nyhan B, Pennycook G, Rothschild D, Schudson M, Sloman S, Sunstein C. Thorson E, Watts D, Zittrain J. 2018. The science of fake news. Sci, 359(6380): 1094-1096.

Mertoğlu U, Genç B. 2020. Automated fake news detection in the age of digital libraries, Inf Technol Libr, 39(4).

Meyers M, Weiss G, Spanakis G. 2020. Fake news detection on twitter using propagation structures. Disinformation in Open Online Media, 26-27 October, Leiden Netherlands, pp: 138-158.

Monti F, Frasca F, Eynard D, Mannion D, Bronstein M. 2019. Fake news detection on social media using geometric deep learning. URL: https://arxiv.org/abs/1902.06673 (access date: December 17, 2023).

Oflazer K. 2014. Turkish and its challenges for language processing. Lang Resour Eval, 48(4): 639-653.

Rahutomo F, Kitasuka T, Aritsugi M. 2012. Semantic cosine similarity. The 7th International Student Conference on Advanced Science and Technology ICAST 2012, October 29-30, Seoul, South Korea, pp: 54.

Reis J, Correia A, Murai F, Veloso A, Benevenuto F. 2019. Supervised learning for fake news detection. IEEE Intell Syst, 34(2): 76-81.

Shu K, Amy S, Wang S, Tang J, Liu H. 2017. Fake news detection on social media: a data mining perspective. ACM SIGKDD Explor Newsl, 19(1): 22-36.

Siahaan A, Aryza S, Hariyanto E, R, Lubis A. 2018. Combination of Levenshtein distance and Rabin-Karp to improve the accuracy of document equivalence level. Int J Eng Technol, 7(27): 17-21.

Taskin S, Kucuksille E, Topal K. 2022. Detection of Turkish fake news in twitter with machine learning algorithms. Arab J Sci Eng, 47(2): 2359-2379.

Ünver A. 2023. Emerging technologies and automated fact-checking: tools, techniques and algorithms. URL: https://ssrn.com/abstract=4555022 (accessed date: August 29, 2023).

Yamanan E. 2016. Türkçenin güncel söz varliği. Mill Eğ Derg, 45(210): 85-91.

Ying X. 2019. An overview of overfitting and its solutions. J Phy Conf Ser, 1168: 022022.