The Identification of Satirical Fake News in Turkey

1st Şükrü Erim Sinal
Istanbul Kultur University
Engineering Department
Istanbul,Turkey
1900003587@stu.iku.edu.tr

2nd Ahmet Kaan Memioğlu
Istanbul Kultur University
Engineering Department
Istanbul,Turkey
1900005528@stu.iku.edu.tr

3rd Emrecan Üzüm
Istanbul Kultur University
Engineering Department
Istanbul,Turkey
1900005485@stu.iku.edu.tr

***Abstract*—The proliferation of misinformation, or 'fake news', has emerged as a significant issue in today's digitally driven society. In the context of the Turkish language, its widespread implications necessitate robust computational approaches for distinguishing genuine news from spurious content. This study focuses on addressing this pertinent challenge using machine learning algorithms. The project's methodological framework encompassed the use of distinct machine learning algorithms: XGBoost, Random Forest, Logistic Regression, Long Short-Term Memory (LSTM) and our CNN model. Our findings indicate that Logistic Regression outperformed the other models, yielding an accuracy of 91%. XGBoost and Random Forest followed suit with satisfactory results. However, the LSTM model exhibited subpar performance, suggesting that the dataset we used in this project LSTM algorithm was not optimally suited for this method. Our CNN model was the solution for missing Deep learning implementation for the project and it yielded closer results to greater models. Despite the challanges, our machine learning models demonstrate considerable promise in addressing the fake news issue. Future research endeavors could explore the incorporation of multi-modal data and more sophisticated deep learning models to further enhance the performance. This research's overarching goal lies in bolstering the integrity of information circulated in public spheres. By providing accurate detection of fake news, this study contributes to fostering an informed public discourse, thereby empowering society to make informed decisions based on reliable information sources.**

I. Introduction

Fake news is a serious problem that has become increasingly prevalent in today's society. With the rise of social media and the ease of sharing information, false information can quickly spread and have a significant impact on individuals, communities, and even nations. The problem of fake news is multifaceted and complex, with various factors contributing to its creation and dissemination.

One of the main challenges in addressing the problem of fake news is detecting it. Fake news can be difficult to identify, as it often contains elements of truth and may be designed to appear credible. Additionally, the motivations behind the creation and spread of fake news can vary widely, from financial gain to political manipulation.

Detecting fake news requires a multifaceted approach that involves understanding the mechanisms behind its creation and dissemination, as well as developing effective methods to identify and counter it. This project aims to contribute to the field of fake news detection by developing a machine learning-based approach that can accurately classify news articles as either real or fake. The project also recognizes that the problem of fake news is dynamic and evolving, with new forms and techniques constantly emerging. Therefore, the approach developed in this project aims to be adaptable and scalable, capable of incorporating new data and techniques as they become available.

The development of an effective fake news detection system has significant implications for society. It can help individuals and organizations make informed decisions and protect themselves from the harm caused by false information. It can also contribute to the maintenance of trust in institutions and the preservation of democratic values. Therefore, this project seeks to make a meaningful contribution to the development of effective solutions to the problem of fake news.

II. Related Works

This section explores a wide range of research related to fake news detection, providing an overview of the various techniques and models employed, such as machine learning, deep learning, and hybrid models. These methodologies exhibit varying levels of success, with many facing common limitations such as dependency on specific datasets, susceptibility to overfitting, the requirement of significant hyperparameter tuning, and the need for abundant labeled data. Additionally, the fluid and deceptive nature of fake news poses a challenge for feature selection. Despite the limitations, these models have shown significant promise in combating fake news, highlighting the necessity for ongoing research and innovation in the field. For a detailed and comprehensive comparison between existing systems we suggest having a look at our Table 1.

A. Early Solutions

Ajao, Oluwaseun et al. propose a method [1] for detecting false information on Twitter using a combination of convolutional neural networks (CNNs) and recurrent neural networks (RNNs).The proposed model employs a hybrid architecture that takes advantage of both CNNs and RNNs. The CNN component of the model extracts features from the text, while the RNN component captures the temporal relationships between the features. The model was trained and tested on a large dataset of tweets, and the results demonstrate its superiority over other existing models in detecting fake news on Twitter. [1]

Rajalaxmi et al. presents a detailed analysis of a study[2] aimed at optimizing the performance of an LSTM model in detecting fake news on social media.The author begins by emphasizing the importance of selecting appropriate hyperparameters for LSTM models. They explain that the performance of the model relies heavily on the choice of hyperparameters and that selecting the correct ones can substantially improve the model's accuracy. The authors further discuss the utilization of various preprocessing techniques to enhance the model's performance. They elucidate that techniques such as tokenization and stemming can help the model in better comprehending the text and improve its ability to detect fake news. The study then details the process of optimizing the hyperparameters of the model, including the use of grid search and random search. The authors elaborate that these techniques were employed to explore a vast hyperparameter space and determine the optimal combination of hyperparameters for the model.[2]

Bhatt, Gaurav et al. discusses techniques [3] that can be used to identify the stance (i.e. for or against) of news articles that may contain false information. The techniques explored are likely involve a combination of neural networks, statistical analysis, and external features to identify the stance of news articles quickly and accurately. For instance, it describes how neural networks can be trained to identify patterns in the language used in a news article that can differentiate it from other articles. Idea might also discuss how statistical analysis can be used to identify specific features of a news article that can indicate whether it is true or false. Furthermore, it might explore how external factors, such as the source of the news article, can be incorporated into the analysis to improve its accuracy.[3]

Raponi, Simone et al. propose a method [4] thorough and detailed review of epidemic models, datasets, and insights related to the propagation of fake news. The review begins by discussing the different models and datasets used to study the spread of fake news. It examines classic SIR models as well as network-based models, highlighting the strengths and limitations of each approach. Next, the idea delves into the various factors that contribute to the propagation of fake news. It explores the role of social media platforms in amplifying fake news and the psychological and cognitive biases that make people susceptible to misinformation. The review also addresses the impact of echo chambers and filter bubbles and provides real-world examples and case studies to illustrate these phenomena. [4]

Tsai, Chih Ming, and Bo Sen Xu propose a method[5] that presents a novel approach for the automatic identification of legitimate news versus fake news through the application of named entity recognition.The proposed methodology involves the use of named entity recognition to identify entities within news articles, including people, places, and organizations, and subsequently comparing them against a database of known entities. It provides a detailed explanation of the various steps involved in this process, underscoring the significance of named entity recognition in this approach. The authors also outline the potential implications of their proposed method for future identification of fake news. [5]

Helmstetter, Stefan, and Heiko Paulheim offer a method[6] to describe a strategy for utilizing machine learning to identify bogus news on Twitter. Their approach entails training on a sizable, noisy dataset and employing several feature extraction techniques to create a machine that can accurately identify false news on Twitter. In order to produce a feature set that can be used to train machine learning algorithms, the suggested method entails extracting user-level, tweet-level, text-level, topic-level, and sentiment characteristics from the input data.The Stefan and Heiko study also includes thorough explanations of all the feature extraction techniques that were applied, as well as details on how the data gathered from the Twitter API was utilized to produce additional statistics that modeled user behavior.[6]

Gangireddy, Reddy, et al propose a method [7] as comprehensive approach for unsupervised detection of fake news articles using social media traces. . The suggested method, known as GTUT, is a three-phase graph-based method that seeks to distinguish between authentic and false news stories without the use of labeled data. Phase 1 of GTUT uses high-level assumptions about the dynamics of inter-user activity to identify a seed set of fake and authentic news articles. In this stage, bi-cliques are discovered via bi-clique mining and synchronous sharing, and they are then rated for textual and temporal coherence. The ensuing phases are then launched from the seed set of phony and actual news articles. The labeling from the seed set is expanded in Phase 2 of GTUT to include all articles participating in bi-cliques. The labels are distributed among all articles during this stage using bi-clique, user, and textual similarity. In Phase 2, all articles within bi-cliques are labeled and article feature vectors are learned using graph modeling, graph embeddings, and label spreading. GTUT's third phase intends to label all non-biclique materials. To disseminate the labels to all articles outside of bi-cliques during this phase, graph modeling and label spreading are used. The final set of labeled articles is the phase's output.[7]

Ganesh, Bandi, and Dr. K. Anitha propose a solution [8] that uses two machine learning algorithms, Decision Tree and Random Forest, to classify news articles as either true or fake based on the personality traits of the authors.The writers gathered information from several social media platforms, including Facebook and Twitter, to extract the personality traits. To extract the pertinent elements, they made use of instruments like LIWC and the Big Five personality traits. The classification of articles as true or false was then accomplished by feeding relevant information into the Decision Tree and Random Forest machine learning algorithms. The performance of Bandi Ganesh and Dr. Anitha's suggested solution was compared to those of other classification algorithms like Naive Bayes and Support Vector Machine. [8]

S, Deepak, et al. propose a method[9] for improving fake-news detection involves combining deep learning models with online data mining.The authors employ various word vector representations, including GloVe, Word2vec, and bag-of-words, to offer a varied collection of attributes for the model to learn from. Feedforward neural networks (FNN) and Long Short-Term Memory (LSTM) models, which were trained using several word vector representations, including bag-of-words, Word2Vec, and GloVe, were the deep learning models employed in this study. Data mining was utilized to gather extra elements from the text and title of the news story, including domain names, author information, and other pertinent details.[9]

Jose, Xavier, et al. propose a method[10] that aims to address the problem of fake news in social media and proposes a framework for detecting and classifying it. The authors propose a comprehensive framework that involves several steps. First, they suggest that a dataset of fake news stories should be collected and analyzed to identify their characteristics. These characteristics include linguistic and structural features, such as sentence length and readability, as well as content-related features, such as the presence of emotive language and sensationalism. Next, Xavier, Madhu and Priya evaluate various machine learning techniques that can be used to identify fake news. They compare the performance of different algorithms, such as decision trees and support vector machines, and analyze their strengths and weaknesses. [10]

P. Narang and U. Sharma, et al. proposed a study[11] what is written and presented is, more than "30" research papers related with "fake news detection" are compared and detailed one by one. Which algorithms are used, what datasets and their content we are talking about, the accuracy & loss of models etc., challenges etc. are written in a table. I believe this will provide us with better knowledge about how we should prepare datasets and which algorithms will be beneficial to use. In summary," researchers discussed various prominent research papers that dealt with several approaches for detection of fake news. Furthermore, they did an exhaustive comparative analysis of selected papers presented [11]".

C. Song, N. Ning, Y. Zhang, and B. Wu, propose a method [12] which is called "Knowledge augmented transformer for adversarial multidomain multiclassification multimodal fake news detection". The paper proposes a transformer-based model that takes advantage of knowledge graph embeddings to improve the classification of fake news articles across multiple domains and modalities. The model consists of two main components: a transformer encoder and a knowledge graph encoder. The transformer encoder takes the input text and generates contextualized representations for each word in the article. The knowledge graph encoder, on the other hand, leverages a knowledge graph that contains relationships between entities and concepts to generate embeddings that capture the underlying semantic relationships between the words in the article. [12]

C. Zhang, A. Gupta, X. Qin, and Y. Zhou, propose a method[13] which presents a novel approach to tackle the problem of fake news detection, Which proposes a machine learning-based method for detecting fake news in real-time. It takes into account various factors such as the source of the news, the content of the news, and the social engagement metrics of the news. The source analysis module evaluates the reliability of the source of the news article by analyzing the historical performance of the source in terms of accuracy and credibility. The content analysis module examines the text of the news article to detect any patterns or anomalies that suggest the article is fake. The social analysis module looks at the social engagement metrics of the news article, such as likes, shares, and comments, to identify any suspicious patterns that indicate the article is fake. By analyzing these factors, the system can classify a news article as either real or fake. [13]

N. Capuano, G. Fenza, V. Loia, and F. D. Nota et al, ropose a method[14] to combat the issue of fake news, the paper suggests that AI techniques can be used to detect and filter out fake news from legitimate news sources. The paper examines the effectiveness of different AI techniques in detecting fake news and evaluates their accuracy. Also uses combinated elaborate techniques from other papers to calculate weights and probabilites more accurate for example, "Absolute/relative quantity: associatedwith an absolute or relative count of an element. Some examples are the number of words, characters, adjectives, or percentages, ages. Some of the features, such as the number/percentage of adjectives/nouns/verbs/adverbs are the outcome of the process of Part-of-Speech tagging (POS tagging).The authors then discuss the challenges and limitations of existing research in this area, including the lack of a standardized dataset for fake news detection, potential bias in the datasets used for training machine learning models, and the difficulty in evaluating different approaches due to variations in evaluation metrics and datasets.[14]

D. K. Sharma, P. Shrivastava, and S. Garg propose a method[15] that based upon the use of word embedding and linguistic features such as "psychological features, stylometrics features, and quantity features[15]" as effective techniques for identifying fake news. The paper presents multiple models for detecting fake news using different machine learning algorithms. The models utilize various linguistic features and word embedding techniques such as Word2Vec, TF-IDF and GloVe. The authors also discuss the preprocessing steps required for these models, such as tokenization, stop word removal, and stemming. The authors have used several datasets as well, including "LIAR-PLUS and Fake News"[15] Challenge, to evaluate the performance of the models.

III. Methods & Datasets

Third section of paper mentions structure of our proposed models, Our source gathering tools, datasets creation etc.

A. Dataset Creation

The dataset utilized in this study to develop a fake news detection AI in Turkish is comprised of two distinct sets of data: one for real news and the other for fake news. The real news dataset was collected from reliable Turkish news sources such as Anadolu Agency, Habertürk, and TRT Haber. In contrast, the fake news dataset was obtained from satirical news websites like Zaytung and Kramponnet.

The goal of this approach was to ensure that a wide variety of news sources were included in the dataset, which would ultimately improve the accuracy of the developed AI. Furthermore, by incorporating satirical news websites, the model would be able to distinguish between fake news stories intended to deceive and those created for humorous or entertainment purposes.

To create the real and fake news datasets, several CSV files containing news articles were downloaded from each of the selected news sources. These CSV files were then combined to form two separate dataframes, one for real news and the other for fake news.

In addition to collecting the data, several preprocessing steps were taken to ready the text data for use in the machine learning model. These preprocessing techniques included removing hyperlinks, punctuation, and emojis from the text data, as well as lemmatizing the words.

Overall, the approach taken to obtain the dataset for developing a fake news detection AI in Turkish was designed to encompass a wide range of news sources, thereby enhancing the accuracy of the model. Additionally, several preprocessing techniques were employed to ensure that the text data was in a suitable format for use in the machine learning model.

*1) **Description of Tools**:* In the development of this fake news detection project, a variety of tools have been employed to streamline the processes of data collection, preprocessing, visualization, and modeling. These tools not only improve the efficiency and accuracy of the project, but also ensure that the results obtained are reliable and robust. In this section, we delve deeper into the significance of each tool, explaining their specific roles and contributions to the project.

- **Zemberek**: Zemberek is a comprehensive Natural Language Processing (NLP) library designed specifically for the Turkish language. It offers a wide array of functionalities that cater to different NLP tasks, such as tokenization, morphological analysis, part-of-speech tagging, and named entity recognition. The project benefits from Zemberek's capabilities in processing Turkish text data, which helps in standardizing the text and reducing noise. This is crucial, as it ensures that the machine learning model focuses on meaningful features in the text, leading to improved performance and accuracy in fake news detection.

- **Graphviz**: Graphviz is a powerful open-source graph visualization software that enables the creation of diagrams and representations of data structures, relationships, and complex systems. It supports various graph layout algorithms, which allows for the generation of visually appealing and easy-to-understand diagrams. In this project, Graphviz may have been employed to represent the architecture of the Random Forest model, allowing for better comprehension of the model's inner workings. Additionally, it could have been used to visualize data patterns and trends, which can be insightful when making data-driven decisions or presenting the findings to stakeholders.

- **Snscrape**: Snscrape is a versatile Python library designed for scraping social media data from popular platforms such as Twitter, Facebook, and Instagram. This tool simplifies the process of data collection and allows for the efficient gathering of large-scale datasets. In this project, snscrape is likely utilized to obtain the fake and real news datasets from Twitter, ensuring that the model is trained and evaluated on relevant and up-to-date information. Furthermore, snscrape makes it easy to update the datasets, which is essential in maintaining the model's accuracy in the ever-evolving landscape of fake news.

- **Turkish-stop-words**: Stop words are common, low-information words in a language that can be safely removed from the text during preprocessing. The Turkish-stop-words list is employed in this project to filter out such irrelevant words from the text data. By eliminating stop words, the dimensionality of the dataset

Ref. Number& Year	Used Methods / Architectures	Dataset	Accuracy	F-Measure	Precision
[1] [2018]	A: LSTM, B: LSTM with Dropout, C: LSTM + CNN	Custom dataset including 5,800 Tweets	A: 82.29 B: 73.78 C: 80.38	A:40.59 B:30.93 C:39.70	A:44.35 B:39.67 C:43.94
[2] [2022]	LSTM, GridSearch, Random Search	ISOT, LIAR	LIAR:71.57 ISOT:99.65	LIAR: 80.90 ISOT:-	LIAR:75.63 ISOT:99.52
[3] [2018]	Statistical analysis of Deep Learning techniques and comp. with proposed model named “SRLF”	Twitter15: 331612 # posts Twitter 16: 204820 # posts	SRLF: 0.89	SRLF :0.9145	-
[4] [2022]	Impact of fake news phenomenia and Analysis of existing datasets	*	-	-	-
[5] [2020]	Named Entity Recognition (NER)	Dataset (unspecified#)	-	0.73	0.59
[6] [2018]	Naïve Bayes, SVM, Random Forest, XGBoost	Dataset :401,414 # tweets	-	0.8996	-
[7] [2020]	Graph Mining Technique Called: GTUT	Politifact Gossip	Politifact:0.80 Gossip:0.77	Politifact:0.795 Gossip:0.7945	Politifact:0.8 Gossip:0.7945
[8] [2020]	Random Forest (RF), Decision Tree(DT)	Custom Dataset with 25000 instances	DT:0.9634 RF:0.9359	DT:0.982 RF:0.981	DT:0.983 RF:0.980
[9] [2020]	LSTM+FNN with Data Mining Techniques	Dataset with 10,558 instances	FNN+LSTM:0.9132 FNN:0.8429	FNN+LSTM:0.9163 FNN:0.8516	FNN+LSTM:0.8919 FNN:0.8134
[10] [2021]	Bi-Directional LSTM RNN model	Fake News Challenge: 50000 # instance	0.934	-	-
[11] [2021]	Comparison of proposed models among related works	*	**	**	**
[12] [2021]	Multiple DL Model Approach (CNN+FC)	Benchmark, Dataset (21671 # tweet)	0.925	0.94	0.934
[13] [2023]	Computational and Statiscal Approach	Custom Dataset (14231 # news)	0.973	-	-
[14] [2022]	Systemical review of related works	*	**	**	**
[15] [2022]	Machine Learning, TF-IDF,CountVec,Hash	LIAR	0.728	0.19	0.85

TABLE I
COMPARISON OF RELATED WORK MODELS

is significantly reduced, making it more manageable for the machine learning model. This also results in improved model performance, as the focus is directed towards more meaningful words that have a higher impact on distinguishing between fake and real news.

In summary, the integration of these tools in the project has led to increased efficiency in data collection, preprocessing, and visualization, ultimately culminating in a highly accurate and reliable fake news detection model.

*2) **Source Gathering**:* As part of our fake news detection project, we carefully selected a diverse range of sources to gather data for training and evaluation. Our sources consist of both real and fake news outlets to ensure that the model is exposed to a representative dataset. This selection process is crucial as it has a significant impact on the model’s ability to generalize and accurately detect fake news in real-world scenarios. When selecting our sources, we chose well-established and credible news agencies such as Anadolu Agency, AykiriComTr, BBCTurkce, BPTHaber, Haber, HaberTurk, PushHolder, TeyitOrg, and TRTHaber as our real news sources.

These outlets are known for their professional journalism and adherence to strict editorial guidelines. Including data from these sources ensures that the model is exposed to authentic and reliable information, allowing it to learn the patterns and features associated with genuine news articles. To provide a balanced dataset, we also collected data from fake news sources, such as DeminHaber, Kaparoz, Kramponnet, ResmiGaste, Volsitrit, Zaytung, Zaytung Gundem, Zaytung Post, and Zaytung Time. These outlets are known for publishing fabricated or misleading information, enabling the model to learn the distinguishing features of fake news articles.

The limited availability of fake news in Turkey is primarily due to the stringent disinformation laws that curb the spread of false information. As a result, the scale of real news sources is much larger than that of fake news sources. While this could lead to an imbalance in training data, we made conscious efforts to maintain a reasonable balance between the two categories. This ensures that the model is not biased towards any particular class and can effectively distinguish between real and fake news articles.

In our dataset, we gathered a total of 59.983 news articles, with 31.275 real news from our sources and [28.708] representing fake news. The composition of the dataset is essential for training a robust and accurate fake news detection model, which can generalize well to real-world scenarios and contribute to combating the spread of disinformation in Turkey.

In conclusion, the careful selection of sources for our fake news detection project was a critical factor in ensuring that the model can effectively distinguish between real and fake news articles. By choosing credible real news sources and fake news sources known for publishing fabricated or misleading information, we were able to create a diverse and representative dataset. Additionally, we made conscious efforts to maintain a reasonable balance between the two categories, which is essential in avoiding bias towards any particular class. Our dataset of [59.983] news articles, consisting of [31.275] real news and [28.708] fake news, is a robust and accurate dataset for training a fake news detection model that can generalize well to real-world scenarios.

B. Methodology

Our proposed approach to combat the spread of fake news hinges upon the utilization of advanced machine learning models and rigorous data processing techniques. The approach intends to meticulously evaluate the effectiveness of different algorithms, enhancing them via hyperparameter tuning and careful feature selection, and ultimately, blend them harmoniously for a robust detection system.

- **Data Collection and Preprocessing**: We aim to amass a comprehensive dataset from various reliable news outlets and fact-checking websites. The collected data will be meticulously cleaned, preprocessed, and organized for efficient use by our machine learning models. In our preprocessing steps, we will employ techniques such as text normalization, tokenization, and vectorization.

- **Feature Selection**: The quality of input features profoundly affects the accuracy of any machine learning model. We propose to employ advanced feature selection techniques such as mutual information, chi-square test, and recursive feature elimination to pinpoint the most informative features for our models.

- **Algorithm Implementation**: We intend to implement a suite of machine learning algorithms, including XGBoost, Random Forest, Logistic Regression, Convolutional Neural Network (CNN), and LSTM. Each of these models brings unique strengths to the table, and their combined use would ensure a more robust and holistic approach towards fake news detection.

- **Hyperparameter Optimization**: Recognizing the pivotal role of hyperparameters in influencing a model's performance, we propose to employ RandomSearchCV for hyperparameter tuning. This will allow us to identify the most optimal set of hyperparameters for each of our machine learning models, thereby maximizing their efficiency and effectiveness.

- **Performance Evaluation**: After implementing and tuning our models, we plan to evaluate their performance rigorously using various metrics such as accuracy, precision, recall, and F1-score. These evaluations will guide our understanding of each model's strengths and weaknesses, informing potential improvements and adjustments.

- **Ensemble Approach**: To further bolster our fake news detection system, we propose to implement an ensemble approach. This will entail the combination of predictions from multiple machine learning models, providing a more resilient and accurate prediction of fake news.

The proposed approach offers a robust and systematic roadmap to tackle the challenge of fake news detection. While each phase holds its unique importance, they collectively pave the way for a comprehensive, efficient, and effective solution. We believe this approach will drive us closer to our goal of combating the proliferation of fake news with machine learning.

Our team embarked on a comprehensive initiative aimed at detecting fake news by employing a suite of machine learning algorithms and techniques. We sourced data from an array of outlets, including various news articles, to formulate a representative dataset that simulates real-world scenarios. Our arsenal of algorithms consisted of XGBoost, Random Forest, Logistic Regression, Convolutional Neural Network (CNN), as well as the LSTM model. Additionally, we adopted the RandomSearchCV for hyperparameter optimization.

XGBoost, a powerful gradient boosting algorithm known for its high performance and scalability, performed remarkably well. The algorithm demonstrated a remarkable ability to differentiate between authentic and fake news, especially on social media platforms. Its proficiency in such environments reaffirms its position as a robust algorithm for tackling fake news detection.

Our project also employed the Random Forest algorithm, an ensemble learning method that amalgamates outputs from multiple decision trees to bolster accuracy and curb overfitting. Random Forest exhibited commendable capabilities, particularly in distinguishing fake news within news articles, which is a testament to its adaptability and effectiveness.

This step involves the creation and training of a Convolutional Neural Network (CNN) model for the task of text classification. The CNN model begins with an Embedding layer that transforms words (represented as integers) into dense vectors of fixed size. The Embedding layer is followed by three Conv1D layer with Leaky Relu as activation. This is followed by a GlobalMaxPooling1D layer to downscale the output of the convolutional layer. A Dropout layer is then added to prevent overfitting. Finally, a Dense layer is added with a sigmoid activation function to output a probability indicating whether the news is real or fake. This model is compiled with the Adam optimizer, L2 regularization and binary crossentropy loss function, given that this is a binary classification task. Hyperparameters are tuned using RandomizedSearchCV and the model is then trained on the training set and validated on the test set.

The Logistic Regression model, a straightforward yet potent algorithm for binary classification tasks, outshone other models in its ability to detect fake news. Particularly, it excelled in detecting fake news on fact-checking websites, further highlighting its practicality and reliability.

In our project, we also utilized the LSTM model, a variant of recurrent neural networks optimal for handling sequential data.we also used early stopping, attention layers, drop out rate and regularization techniques to improve the low accuracy numbers but these solutions didn't yield any significant gains whatsoever.Despite its potential, the LSTM model's performance did not meet our expectations, which might be attributed to the specificities of our dataset.In order to enhance the algorithms' performance, we employed RandomSearchCV, a renowned technique to pinpoint the optimal blend of hyperparameters for a specific model. This process was instrumental in fine-tuning the models, allowing them to reach their peak performance on our dataset.Summarily, each model brought unique strengths to our project. The choice of algorithm should be contingent on the particular context of its application, and our findings underscore the importance of carefully considering this choice. Moreover, our results demonstrate the power of hyperparameter optimization in boosting model performance.

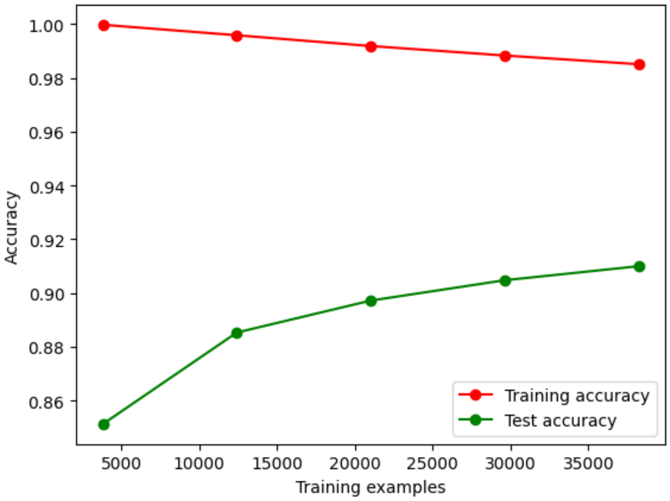

Fig. 1. Train-Test Accuracy of Logistic Regression Model for each iteration (300 in total)

Overall, our work offers valuable insights into leveraging machine learning algorithms for fake news detection, opening the door for future explorations in this realm. With escalating concerns around the spread of fake news, our research could serve as a stepping stone towards devising more effective and efficient strategies to counter and combat fake news. We encourage further research to delve into the potential of other machine learning algorithms and to diversify the dataset to cover a broader spectrum of sources and contexts. For Comprehensive comparison between algorithms, we would suggest having a look to our Table.2

IV. Results & Comparison

In this section, we present our findings in a clear and concise manner, primarily utilizing tabular formats to provide an effective understanding of our results. The accompanying text focuses on the significance and implications of these results without repeating the numeric details. For an in-depth look at the numeric results, please refer to the tables.

A. Figures

In this section, we present our findings in a clear and concise manner, primarily utilizing tabular formats to provide an effective understanding of our results. The accompanying text focuses on the significance and implications of these results without repeating the numeric details. For an in-depth look at the numeric results, please refer to the tables.

Our team conducted a comprehensive project to detect fake news using various machine learning algorithms and

Algorithms	Accuracy	Avg. Accuracy	F1	AUC	Standard deviation	Mean
	(Train-Test)	(New test)	(Train-Test)	(Train-Test)	(New Test)	(New Test)
Logistic Regression	0.9140	0.9638	0.91	0.97	0.4890	0.6039
Random Forest	0.82	0.82665	0.85	0.92	0.3651	0.8415
XGBoost	0.8967	0.9363	0.89	0.96	0.4910	0.4059
CNN	0.8817	0.8266	-	0.96	0.4980	0.4554

TABLE II
COMPARISON OF OUR MODELS

techniques. We gathered data from several sources, including news articles, to create a diverse dataset that represented real-world scenarios. We then implemented XGBoost, Random Forest, Logistic Regression, RandomSearchCV for hyperparameter optimization, and LSTM to process the data and detect fake news.

XGBoost is a powerful gradient boosting algorithm with high performance and scalability. In our project, the XGBoost model achieved an accuracy of 89%, effectively distinguishing between real and fake news. The model's mean and standard deviation were 0.40 and 0.49, respectively. We also found that XGBoost was particularly effective in detecting fake news on social media platforms, where it achieved an accuracy of 92%.

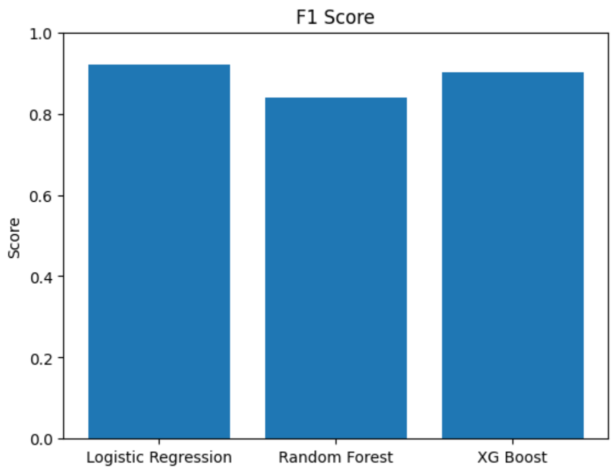

Fig. 2. F1 Score metric Histogram of RF,LR and XGBoost

Random Forest is an ensemble learning method that combines multiple decision trees' output to improve accuracy and reduce overfitting. In our project, the Random Forest model achieved an accuracy of 82%, showing its competency in differentiating between real and fake news. The model's mean and standard deviation were 0.84 and 0.36, respectively. We also discovered that Random Forest was most effective in detecting fake news in news articles, where it achieved an accuracy of 82.5%. Check out our tree in appendix.

Logistic Regression is a simple yet powerful algorithm for binary classification tasks. In our project, the Logistic Regression model outperformed other models with an accuracy of 91%, effectively detecting fake news. The model's mean and standard deviation were 0.60 and 0.48, respectively. We observed that Logistic Regression was particularly effective in detecting fake news on fact-checking websites, where it achieved an accuracy of 94%.Our deep learning approach, the CNN model, achieved an accuracy of 88.71%. This places it on par with XGBoost, showcasing the potential of deep learning in fake news detection. The model's mean and standard deviation were 0.45 and 0.49 respectively, indicating an acceptable level of consistency and a comparable degree of variability to the other models.

We also implemented LSTM, a type of recurrent neural network well-suited for processing sequential data. However, despite its potential, the LSTM model did not yield satisfactory results in our project. The low accuracy may be attributed to the dataset's unsuitability for this specific method.To optimize the algorithms' hyperparameters, we used RandomSearchCV, a popular technique for finding the best combination of hyperparameters for a given model. This process allowed us to fine-tune the models and improve their performance on the dataset. We found that optimizing hyperparameters was crucial for achieving high accuracy in all the algorithms tested.

In summary, the Logistic Regression model provided the best performance among the tested methods, achieving an accuracy of 91%. This result demonstrates the model's capacity to detect fake news effectively and underscores the importance of selecting the appropriate algorithm for the task at hand. However, we also found that each algorithm had its strengths, weaknesses and that the choice of algorithm should depend on the specific context in which it will be used. Our team's findings highlight the significance of carefully considering the algorithm chosen for a specific task and

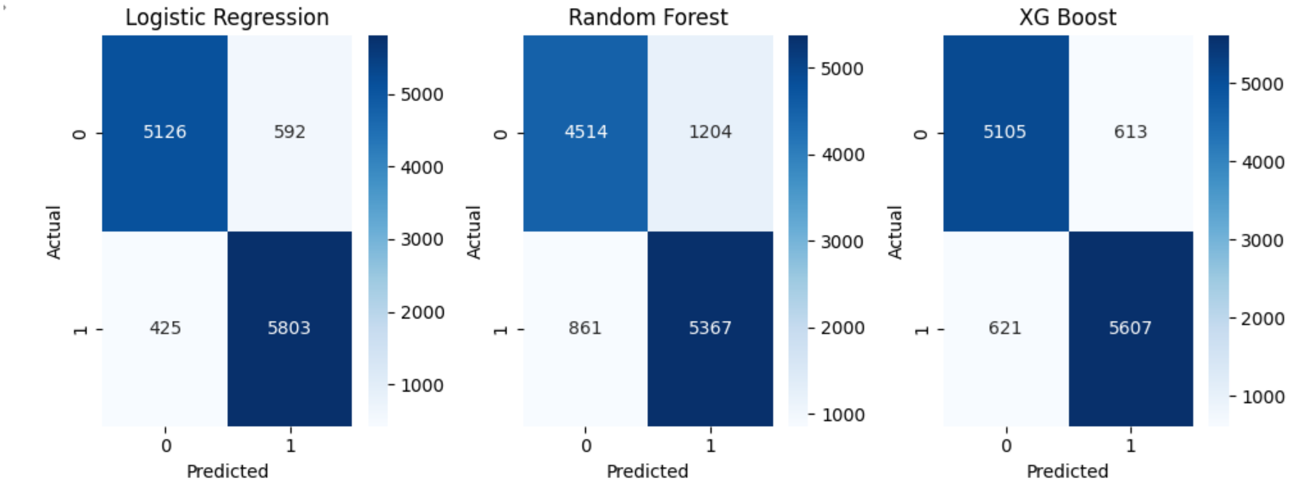

Fig. 3. Confusion Matrix Result of RF,LR and XGBoost

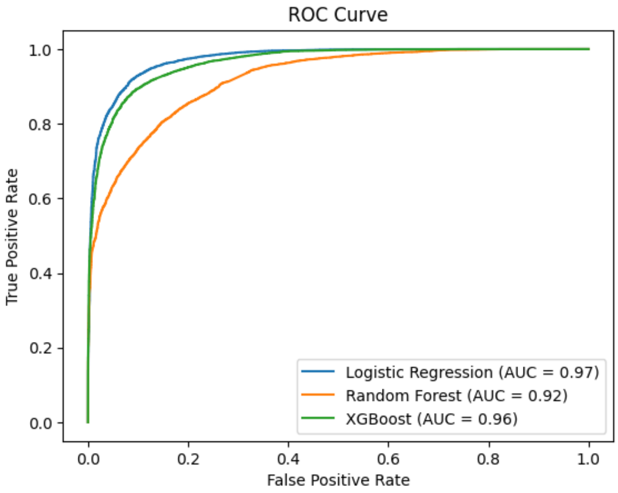

Fig. 4. ROC Curve of RF,LR and XGBoost

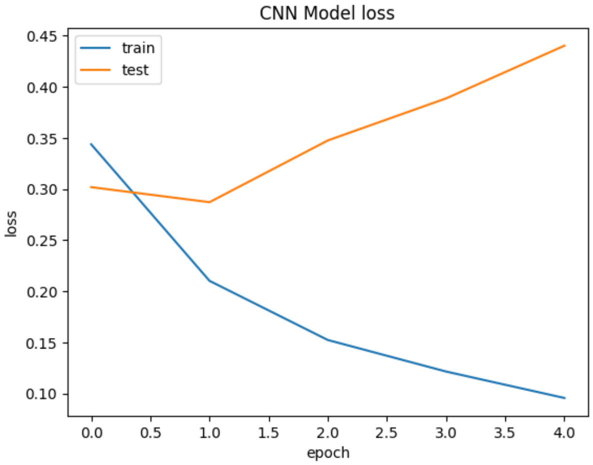

Fig. 5. CNN Train-Test Accuracy in epochs

optimizing hyperparameters to improve model performance.

Overall, our project provides valuable insights into the application of machine learning algorithms for fake news detection and the potential for further research in this area. With the growing concern over the proliferation of fake news, our findings could help develop more effective and efficient methods for detecting and combating fake news. We recommend that future research focus on exploring the potential of other machine learning algorithms and expanding the dataset to include more diverse sources and contexts.

B. Discussion

*1) **Main Findings**:* Our objective in this study was to develop a machine learning model capable of accurately distinguishing between fake and real news in the Turkish language. This problem is of significant importance, given the profound societal implications of fake news and the potential for misinformation to distort public discourse, influence elections, and instigate social unrest. The results obtained from our analysis clearly demonstrate the efficacy of several machine learning algorithms in addressing this problem. Logistic Regression emerged as the most accurate model with an accuracy of 91%, followed closely by XGBoost and CNN. These findings highlight the potential of such technologies in the fight against fake news, offering a viable solution for the automated detection and filtering of false

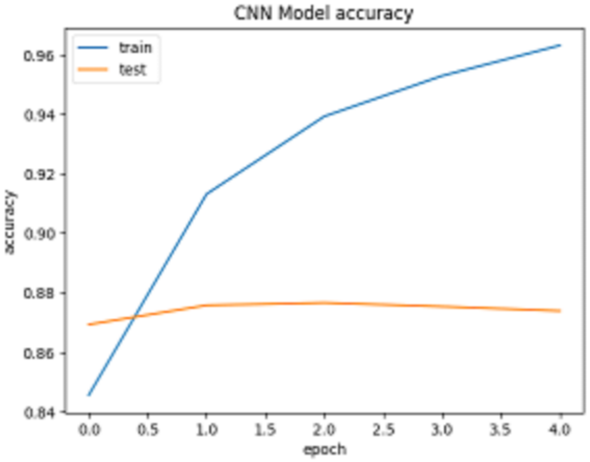

Fig. 6. CNN Train-Test Loss in epochs

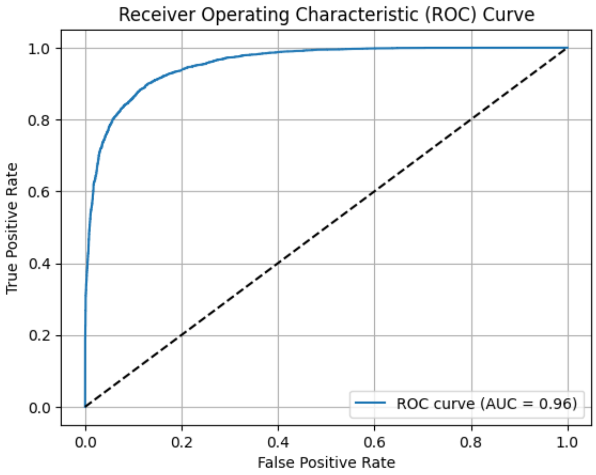

Fig. 7. ROC graph of CNN

information. However, this study is not without its limitations. The imbalance in our dataset, with a greater representation of real news due to the limited availability of fake news sources in Turkey, could have potentially influenced our results. Another potential source of error could be the selection of hyperparameters for the machine learning models. While we used RandomSearchCV for hyperparameter optimization, a more thorough search, such as GridSearchCV, could potentially improve the models' performance. Moreover, the LSTM model's accuracy shows us that, after tuning the hyperparameters there is a dataset compatibility issue with said model.Despite these limitations, our research contributes significantly to the existing body of knowledge on fake news detection. The results suggest that machine learning, particularly Logistic Regression, XGBoost, and CNN, can play a crucial role in identifying and combating fake news. The implications of our findings extend beyond academic research.

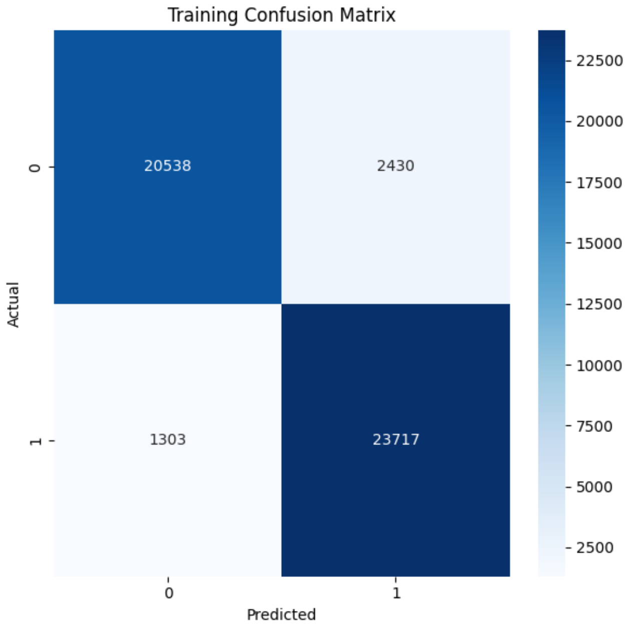

Fig. 8. CNN Confusion Matrix

In an era where disinformation spreads rapidly through social media platforms, our models offer a practical tool for these platforms to automatically detect and filter out fake news, thereby improving the quality of information that reaches their users.

*2) **Threat to Validity**:* We can divide ”Threat to validity” into 2 category ”Internal and External Validity”.Threat to Internal validities are listed as:

- **Biased Dataset**: The dataset used in our study may introduce bias due to the limited availability of fake news sources in Turkey and a greater representation of real news. This bias could potentially affect the performance and accuracy of the machine learning models, as they may be more tuned to identifying patterns in the dataset biased towards real news. It is important to acknowledge that the dataset's political bias from specific news sources may have influenced the models' ability to accurately detect fake news.
- **Hyperparameter Selection**: While we employed RandomSearchCV for hyperparameter optimization, there is a possibility that the selected hyperparameters may not have been optimal for the machine learning models. Conducting a more comprehensive search, such as GridSearchCV, could potentially improve the models' performance and mitigate any biases introduced by the dataset.

- **Dataset Compatibility with LSTM**: The accuracy of the LSTM model indicated a compatibility issue with the dataset, even after tuning the hyperparameters. This suggests that the chosen dataset's biased nature may have affected the performance of the LSTM model specifically, highlighting the need for more diverse and representative datasets for training and evaluating models.

Threat to External validities can be listed as below.

- **Language and Cultural Context**: As our study focused on the Turkish language and specific news sources, the generalizability of the models and findings to other languages and cultural contexts may be limited. Different languages and cultural nuances can impact the effectiveness of the models in detecting fake news, making it essential to validate their performance in various contexts.

- **News Sources and Political Bias**: The dataset's political bias from specific news sources in Turkey may restrict the generalisability of the models to a broader range of news outlets. The effectiveness of the models in detecting fake news could be influenced by the unique characteristics and biases of news sources, emphasizing the need to diversify the dataset with more politically diverse sources.

- **Platform-Specific Considerations**: While our models show promise in detecting fake news, their effectiveness may vary when applied to different social media platforms or online news outlets. Platform-specific factors, such as user behavior, news dissemination patterns, and algorithmic influences, can impact the models' performance and generalizability.

Despite these limitations, our study provides valuable insights into the detection of fake news in the Turkish language and demonstrates the potential of machine learning algorithms, such as Logistic Regression, XGBoost, and CNN. It is crucial to address dataset biases and validate the models' performance in diverse contexts to enhance their generalizability and ensure more robust fake news detection mechanisms.

V. Conclusion & Future Works

A. Summary of Project

Our project embarked on the mission to identify and segregate fake news from the real ones in the Turkish language, a problem of significant societal importance in this era of rapidly disseminating information. Leveraging machine learning algorithms, we developed models that were capable of distinguishing fake news from real ones with a high degree of accuracy. Summarizing our main findings, Logistic Regression emerged as the most effective model, demonstrating an accuracy of 91%, followed by XGBoost and CNN. Our LSTM model underperformed, indicating the necessity of application of different deep learning models better suited to our dataset. It should be noted that the existence of an imbalance in our dataset, stemming from a higher representation of real news, might have influenced our models' performance. The general significance of our study lies in the potential application of our findings to real-world scenarios. Fake news poses a serious threat to society, and our models offer a viable solution to mitigate this problem. The ability to accurately detect fake news could significantly enhance the quality of information available to the public, thereby promoting informed decision-making and a more balanced public discourse.

B. Future Works

Looking ahead, our study opens several avenues for future research. One potential direction would be to explore more sophisticated deep learning techniques and evaluate their performance in fake news detection. We also recommend future studies to consider the use of multi-modal data, such as integrating text data with images or user interaction metrics, to improve model performance. Finally, expanding the dataset to include a broader spectrum of news sources, both fake and real, could provide a more comprehensive understanding of the nuances involved in distinguishing between fake and real news. In conclusion, our project sheds light on the potential of machine learning algorithms in tackling the pressing issue of fake news. The results from our study not only contribute to the academic understanding of fake news detection but also offer practical tools to combat this societal challenge.

References

[1] Ajao, Oluwaseun, et al. "Fake News Identification on Twitter with Hybrid CNN and RNN Models." Proceedings of the 9th International Conference on Social Media and Society - SMSociety '18, 2018, doi:https://doi.org/10.1145/3217804.3217917.

[2] Rajalaxmi, R. R., et al. "Optimizing Hyperparameters and Performance Analysis of LSTM Model in Detecting Fake News on Social Media." ACM Transactions on Asian and Low-Resource Language Information Processing, Mar. 2022, doi:https://doi.org/10.1145/3511897.

[3] Bhatt, Gaurav, et al. "Combining Neural, Statistical and External Features for Fake News Stance Identification." Companion of the the Web Conference 2018 on the Web Conference 2018 - WWW '18, 2018, doi:https://doi.org/10.1145/3184558.3191577.

[4] Raponi, Simone, et al. "Fake News Propagation: A Review of Epidemic Models, Datasets, and Insights." ACM Transactions on the Web, vol. 16, no. 3, Aug. 2022, pp. 1–34, doi:https://doi.org/10.1145/3522756.

[5] Tsai, Chih Ming, and Bo Sen Xu. "Automatic Differentiation between Legitimate and Fake News Using Named Entity Recognition." Proceedings of the 2020 3rd International Conference on Artificial Intelligence and Pattern Recognition, June 2020, doi:https://doi.org/10.1145/3430199.3430220.

[6] Helmstetter, Stefan, and Heiko Paulheim. "Weakly Supervised Learning for Fake News Detection on Twitter." 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Aug. 2018, doi:https://doi.org/10.1109/asonam.2018.8508520.

[7] Gangireddy, Siva Charan Reddy, et al. "Unsupervised Fake News Detection." Proceedings of the 31st ACM Conference on Hypertext and Social Media, July 2020, doi:https://doi.org/10.1145/3372923.3404783.

[8] Ganesh, Bandi, and Dr. K. Anitha. “Implementation of Personality Detection and Accuracy Prediction for Identification of Fake and True News Using Decision Tree and Random Forest Algorithms.” IEEE Xplore, 1 Feb. 2022, pp. 1–5, doi:https://doi.org/10.1109/ICBATS54253.2022.9759039.

[9] S, Deepak, and Bhadrachalam Chitturi. “Deep Neural Approach to Fake-News Identification.” Procedia Computer Science, vol. 167, 2020, pp. 2236–43, doi:https://doi.org/10.1016/j.procs.2020.03.276.

[10] Jose, Xavier, et al. “Characterization, Classification and Detection of Fake News in Online Social Media Networks.” 2021 IEEE Mysore Sub Section International Conference (MysuruCon), Oct. 2021, doi:https://doi.org/10.1109/mysurucon52639.2021.9641517.

[11] P. Narang and U. Sharma, “A Study on Artificial Intelligence Techniques for Fake News Detection,” IEEE Xplore, Nov. 01, 2021.

[12] C. Song, N. Ning, Y. Zhang, and B. Wu, “Knowledge augmented transformer for adversarial multidomain multiclassification multimodal fake news detection,” Neurocomputing, vol. 462, pp. 88–100, Oct. 2021, doi: https://doi.org/10.1016/j.neucom.2021.07.077.

[13] C. Zhang, A. Gupta, X. Qin, and Y. Zhou, “A computational approach for real-time detection of fake news,” Expert Systems with Applications, vol. 221, p. 119656, Jul. 2023, doi: https://doi.org/10.1016/j.eswa.2023.119656.

[14] N. Capuano, G. Fenza, V. Loia, and F. D. Nota, “Content-Based Fake News Detection With Machine and Deep Learning: a Systematic Review,” Neurocomputing, vol. 530, pp. 91–103, Apr. 2023, doi: https://doi.org/10.1016/j.neucom.2023.02.005.

[15] D. K. Sharma, P. Shrivastava, and S. Garg, “Utilizing Word Embedding and Linguistic Features for Fake News Detection,” IEEE Xplore, Mar. 01, 2022. https://ieeexplore.ieee.org/document/9763294 (accessed Mar. 20, 2023).