EPiC Series in Computing

Volume 91, 2023, Pages 124–137

Proceedings of 38th International Conference on Computers and Their Applications

Fake News Detection System: An implementation of BERT and Boosting Algorithm

Raquiba Sultana[1] and Tetsuro Nishino[2]

Graduate School of Informatics and Engineering,
The University of Electro-Communications, Tokyo, Japan
sultana.ra@uec.ac.jp, nishino@uec.ac.jp

Abstract

On social media, false information can proliferate quickly and cause big issues. To minimize the harm caused by false information, it is essential to comprehend its sensitive nature and content. To achieve this, it is necessary to first identify the characteristics of information. To identify false information on the internet, we suggest an ensemble model based on transformers in this paper. First, various text classification tasks were carried out to understand the content of false and true news on Covid-19. The proposed hybrid ensemble learning model used the results. The results of our analysis were encouraging, demonstrating that the suggested system can identify false information on social media. All the classification tasks were validated and shows outstanding results. The final model showed excellent accuracy (0.99) and F1 score (0.99). The Receiver Operating Characteristics (ROC) curve showed that the true-positive rate of the data in this model was close to one, and the AUC (Area Under The Curve) score was also very high at 0.99. Thus, it was shown that the suggested model was effective at identifying false information online.

1 Introduction

The use of social media has been steadily growing in recent years. Most Internet users are frequently active on websites like Facebook, Instagram, Twitter, and others. Social media users were forecast to number 3.6 billion in 2020, and by 2025, that number is projected to rise to 4.41 billion [1]. People frequently rely on social media for daily news. As a result, this has turned social media into the center for spreading false information. A global issue has been caused by the proliferation of fake news, which has been especially noticeable during COVID-19. Because of the fear of COVID-19, people were more likely to believe false information.
News that is false and disseminated through social media or news outlets are called fake news. In mass media, information accuracy is occasionally compromised in order to boost revenue. As a result, readers might be misled, and false information might be disseminated regarding subjects like politics, religious affiliation, branding, and financial services [3]. False information is also propagated to attract public attention, making people more vulnerable to security attacks and harming social and political factors. Maybe that's why the current era is defined as the "post-truth" era [4].

The concept of fake news came into the limelight during the 2016 United States presidential election and the subsequent social, political, and economic damage caused by the online transmission of misinformation has been well discussed. The prevalence of social media, where it is so simple to spread false information, has made this issue worse. In reality, this is frequently done in order to deceive those who believe those news and accomplish economic and political milestones. In addition, the mainstream media has gotten more and more biased, and yellow journalism has become more common. Political news like Election, Democracy, war, and conflict are the main topics of news. In traditional media, politically biased reporting and pulling a predetermined line are frequently used to win over the public. Even though such reporting does not spread factually incorrect information, it frequently presents incomplete information to deceive the public in order to further complicit political interests. Many misleading and inappropriate claims concerning the SARS-CoV-2 novel coronavirus (COVID-19) have indeed been made in conjunction with the virus's outbreak, notably on social media [5]. In fact, the World Health Organization (WHO) issued a warning about an ongoing "infodemic," or an excess of information—especially false information—during the epidemic, as a result of the propagation of false information about the virus [6]. Ever since the corona outbreak, there have been numerous claims that the illness may be cured, including that consuming methanol, ethanol, and bleach can protect one against covid-19 [7]. The WHO (World Health Organization) had to issue a warning to people not to consume these poisonous substances as a consequence [8]. Political leaders like President Donald Trump endorsing this assertion sparked controversy over it. He frequently described this disease as the Wuhan virus or the China virus. Asians were targeted for their race in America as a response. The spread of racial hate crimes was a direct result of this misleading information.

Another well-liked hoax involved the 5G network. A rumor that 5G spreads the coronavirus or impairs human immunity systems first appeared at the start of the lockdown. There are worries that people ignited communication masts on fire across the UK as a result of the false reports. According to a spokesman for the industry group Mobile UK, "more than 50" of these arson attacks have occurred[9]. Rumors were spread across the globe regarding the coronavirus vaccine also. Researchers conducted numerous studies on the relationship between coronavirus vaccine hesitancy and fake news. Anti-vaccine groups tried to demotivate mass people with their far-fetched conspiracy theories [10]. A famous conspiracy theory claimed that vaccines will permanently damage DNA or alter genes [11]. This myth was about only mRNA (messenger-RNA) vaccines as they implement the genetic approach. These are some of the examples among many fake rumors spread in recent years. It's increasing at an alarming rate and needs to take immediate action to prevent the spreading of fake information online. To detect and prevent the spread of disinformation, the first step is to understand the information contained therein. For example, writing patterns, emotions, expression styles, and grammatical accuracy must be analyzed. In other words, it is necessary to identify the standard patterns throughout the story. The purpose of this research is to analyze the characteristics of fake and real news. Based on these characteristics, we try to find out the similarities and differences between the two types of news. Many research has been done on this topic in recent years. For example, a Naive Baise classifier has been proposed and implemented for spam filtering via email [12]. The authors used a Buzzfeed dataset and collected data from three major Facebook pages and three political news pages (Politico, CNN, and ABCNews). The model showed a classification accuracy of 75.40%. In another study, we proposed a hybrid fake news detection system focusing on BERT and Ensemble Learning models [38]. The goal of that study was to analyze the characteristics of fake news by implementing text classification tasks and detect fake news by implementing an ensemble learning model. The result was pretty impressive. The

accuracy score was 0.97, f1-score was 0.98. Our proposed model in this study is a modified and extended version of the aforementioned hybrid fake news detection system [38].

2 Related Works

Several deep learning-based methods have been proposed to diminish the online spread of fake news, which have performed well on a variety of datasets. A recent study proposed a hybrid CNN model that integrates metadata with the text [13]. The authors sought to demonstrate that a hybrid approach could enhance a text-only deep learning model. The results of the hybrid CNN were compared with those of the support vector machines (SVM), logistic regression, Bi-LSTM, and also CNN.

In recent years, Transformers have become the most widely used deep learning model. It was first introduced in a seminar paper, published by several researchers from Google and The University of Toronto [15]. It is a self-attention-based deep-learning language model. The authors suggested a brand new, straightforward network architecture that is solely based on attention mechanisms by rejecting the concept of recurrence and convolutions entirely. These models exhibit superior quality while being more parallelizable and taking a significant reduction in training time, according to experiments on two machine translation tasks. Since then, many more new transformer models have been introduced in recent years. These are the modified version of the base model. Transformer models have become extremely popular in recent years for fake news detection. Several studies have been published based on this topic.

Another research proposed the utilization of a transformer-based ensemble of COVID-Twitter-BERT (CT-BERT) models [17]. The authors described the models that were utilized, the methods for text preprocessing, and how to add more data. The best-performing model demonstrated a weighted F1 score of 98.69 on the test set. Transformer-based models were used to perform text classification tasks. BERT, RoBERTa and CT-BERT have been used successfully. The authors also empirically evaluated the effectiveness of a linear support vector baseline (linear SVC) and various text preprocessing techniques and added additional data. Finally, an ensemble learning technique was used to obtain the average of the above models.

Models built on transformers have had great success identifying the features of social media news. The TweetEval framework, which evaluates tweet classification for various tasks, was recently proposed. The benchmark for tweet classification known as TweetEval consists of seven fundamental heterogeneous tasks in social media NLP research. The authors compared various language modeling pre-training strategies and proposed a strong set of baselines as the starting point. The effectiveness of starting with pre-trained generic language models and continuing their training on Twitter corpora was first demonstrated by these experimental results [18].

In a different study, news articles are analyzed to determine whether they are accurate, partially true, false, or something else altogether [19]. The dataset comprises of news articles, titles, and article ratings. The data was preprocessed using TF-IDF vectorization, and several machine-learning techniques were employed to select the most effective classification models. The Gradient Boosting technique outperformed all other models. With the best classification accuracy of 0.57 and the highest f1-macro score of 0.54 on the provided dataset, the techniques were quite interpretive. Different findings are shown by other classification models, such as Passive Aggressive Classifiers, Logistic Regression Classifiers, and Random Forest Classifiers.

Another study examines the rapid expansion of online news content and establishes whether the news is true or false [20]. For this reason, the research suggests a mechanism to identify rumors and claims that need to be fact-checked, particularly those that receive thousands of views and likes before being refuted and debunked by reliable sources. To identify and cate-

gorize fake news, several machine-learning algorithms have been used. However, the accuracy of these methods is constrained. To distinguish between fake and real news, this study used a random forest (RF) classifier. The chosen News Dataset is used to extract twenty-three (23) textual features for this purpose. Out of twenty-three features, fourteen are chosen as the best using four techniques, including chi2, univariate, information gain, and feature importance. On a benchmark dataset with the best features, the proposed model and other benchmark techniques were assessed. According to experimental results, the proposed model performed better in terms of classification accuracy than other machine learning methods like GBM (Gradient Boosting Machine), XGBoost (eXtreme Gradient Boosting)), and Ada Boost Regression Model.

3 Dataset Description

Our society has been impacted by COVID-19 for more than two years. The quality of life suffered as a result of the disruption of supply chains and the impact on the economies of several nations. The disease, infection rates, preventative measures, vaccinations, etc., have all received daily, top-priority news coverage during this time. Many people believed information shared online to be true without checking the source because of the widespread panic; the spread of false information was almost as bad as the pandemic itself. This problem has been referred to as an "infodemic". Social media sites like Facebook and Twitter served as the focal points of this "infodemic." The Co-Aid (Covid-19 Healthcare Misinformation) dataset was chosen for analysis due to this issue [22]. It consists of a variety of healthcare-related Covid-19 data that was obtained from social media. The information was gathered from December 1, 2019, to September 1, 2020. Total three versions were released during the period. In this research, data was collected from all of the versions and combined together. The information includes news reports, facts, and false information about Covid-19. Covid-19, coronavirus, pneumonia, flu9, lockdown, staying at home, quarantine, and ventilators were among the main topics. The dataset contained 4,251 news articles, 296,000 user interactions, 926 posts on social media platforms using Covid-19, and ground truth labels. This dataset included information about user engagement on social media as well as information about true and false claims. These were purposefully put into separate files.
Only true and false data were taken into account in this study. This information included posts on social media and news articles. The majority of the posts were gathered from Tiktok, Facebook, Twitter, Instagram, and YouTube. In this study, the data on real and fake news from the entire period of data collection were combined separately. In total, 4532 real data and 925 fake data were utilized in this study. For ease of analysis, fake and real data were combined. Various fact-checking websites were used to validate all news articles and blog posts. Both true and fake data comprised a statement of the news type (articles/post, etc.), fact-checking URL, news URL, title, news title, content, abstract, publishing date, and meta keywords. Considerable information was gathered from the news URL, title, content, and abstract columns. The title refers to the news or the title of the article, and the content refers to the content of the news. The abstract refers to a brief description of the news. Fig-2 shows a representation of all the analysis performed on the title, content, and abstract. It illustrates the pattern of information dissemination via social media during Covid-19.

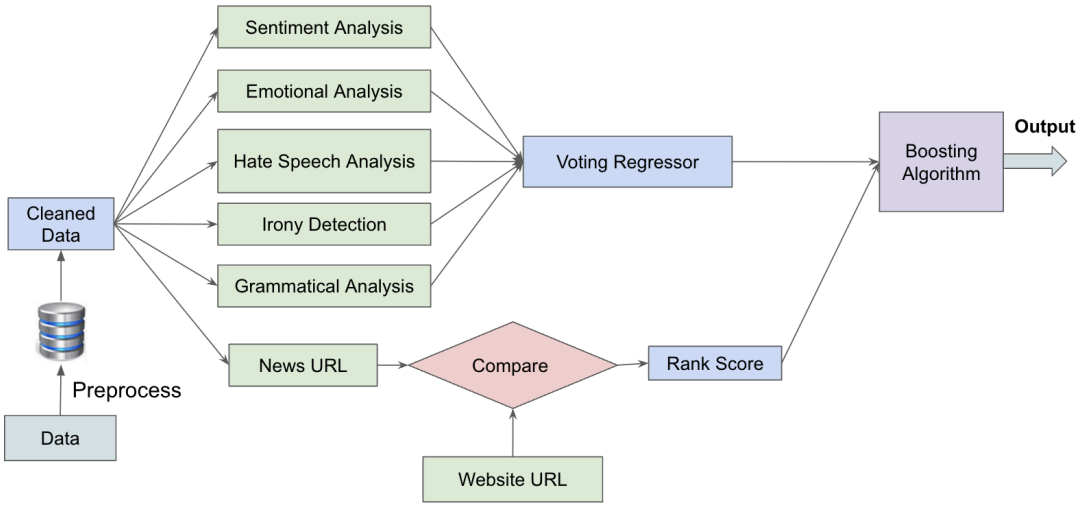

Figure 1: Architecture of proposed model

4 Methodology

The primary objective of this research was to develop a model that could accurately identify false news on social media. In order to achieve this, we took into account information gathered from Twitter and examined the characteristics of news articles and social media posts to build a hybrid system to identify false information in social media. The understanding of the characteristics of tweets was aided by a variety of text classification tasks. The TweetEval framework influenced our research [23]. When training the data for multiple text-classification tasks, such as sentiment analysis, emotional analysis, hate speech detection, irony detection, and grammatical analysis, we first attempted to investigate the characteristics and patterns of tweets related to COVID-19. Utilizing pre-trained transformer models from Hugging Face, all classification tasks were carried out. All news items were then rated according to how reliable their sources were. The ensemble learning model was then updated with all of the results. In this part, the Voting Regressor model received the prediction scores from each classification task as an input. The boosting ensemble model then received the output score of voting regressor and rank scores, which predict whether the news is true or false. Fig - 1 depicts the overall process' architecture.

4.1 Data Pre-processing

The first step was to pre-process the entire data. Data pre-processing was the most important step, as raw data was difficult to train. Unprocessed data often outputs bad results. Especially, when there were enormous amount of missing data. Missing value was a crucial issue for this dataset, as lots of content and abstract data were missing. Missing data of content columns was handled by replacing the value with the value of title. On the other hand, for missing value of abstract was replaced by title. Punctuation was also removed to clean the data.

4.2 Information Analysis

The data were trained based on all five classification tasks. After training, the prediction scores were transmitted to the ensemble model section. Fig - 2 depicts the prediction scores of the trained data on all five tasks.

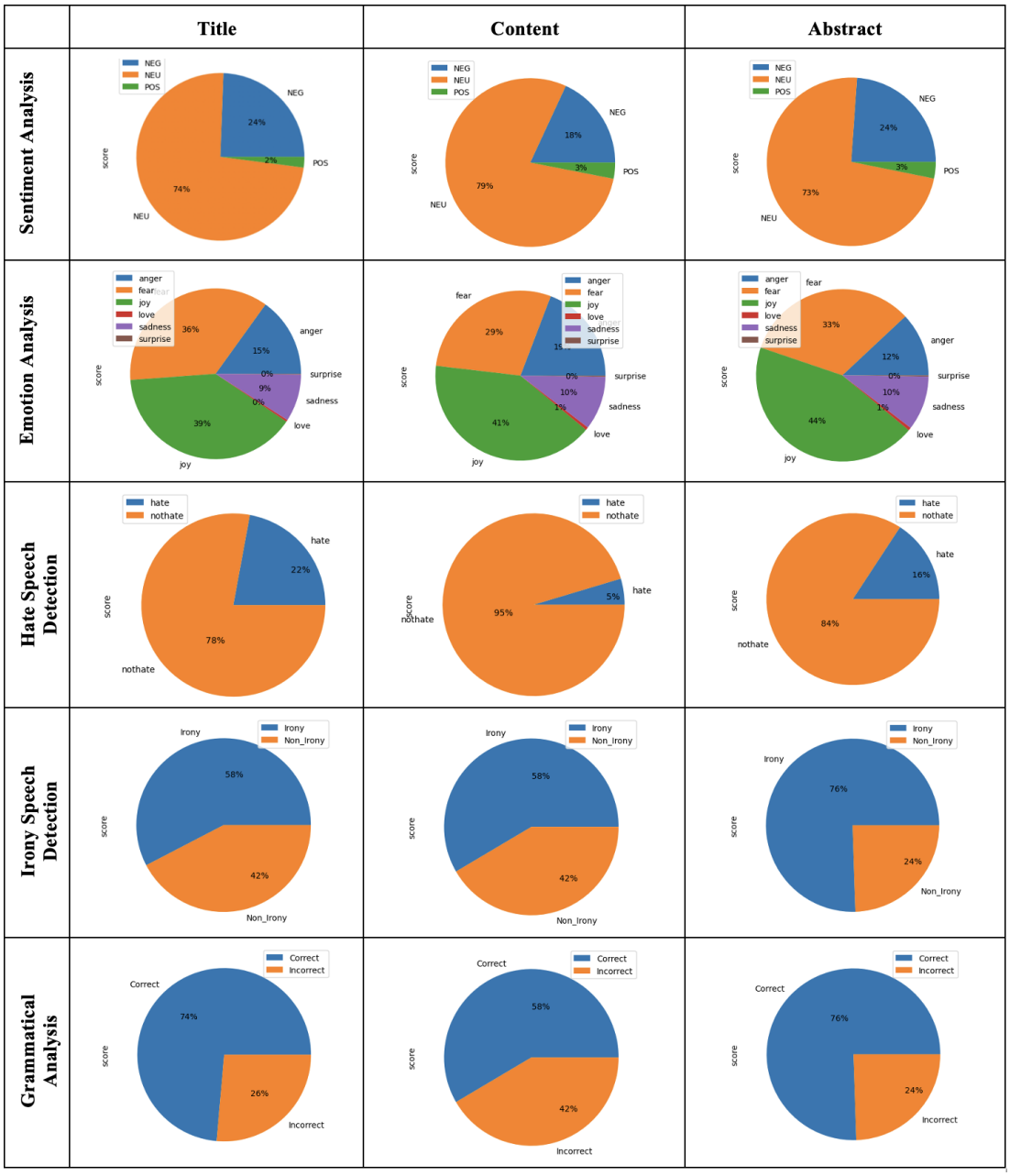

Figure 2: Comparative Representation of All the Analysis Tasks

- **Sentiment analysis:** The goal of sentiment analysis is to determine whether tweets were positive, negative, or neutral in nature. The pre-trained transformer model ardiffNLP's twitter-roBERTa-base-sentiment-latest [35] was employed to analyze the sentiment [18]. Using this model, the title, content, and abstract were trained. This specific model was pre-trained on around 124M tweets. The tweets were collected from 2018 to 2021 and then fine-tuned for Sentiment Analysis with TweetEval benchmark. This pre-trained model was applied to the Co-Aid dataset to analyze sentiment of the data. The sentiment analysis findings from Covid-19 are clearly displayed in Fig - 2. According to the analysis, neutral news was the most prevalent kind of news, making up a significant portion of the title, content, and abstract. Neutral news is resulting more than 70% in all three cases. Even so, the prevalence of negative emotions was much lower than that of neutral emotions. Negative emotions are varying from 18% to 24% in these cases.

- **Emotion analysis:** Another text-classification task is emotion analysis, which divides data into six categories: anger, fear, joy, love, sadness, and surprise. The purpose of this assignment is to identify various emotional states in tweets [24]. A pre-trained Distilbert model obtained from Hugging Face was employed to train the data. "bhadresh-savani/distilbert-base-uncased-emotion" [25] was employed in this research. Originally, the developer fine tuned Distilbert-base-uncased model on the emotion dataset [27] using HuggingFace Trainer with specific Hyperparameters. The patterns that posts follow can be explained by emotion analysis. As illustrated in the Fig-2, angry, happy, and fearful feelings were all frequently expressed in news articles. The data were gathered at the start of the Covid-19 pandemic, which was characterized by anxiety about the illness and resentment toward the government over measures like the lockdown. On the other hand, when news about vaccines was reported, people felt relieved.

- **Hate speech detection:** Hateful content is frequently found in fake news. Even though this is also true in the case of true news, it is much less likely to happen in latter case. People occasionally make conscious attempts to spread divisive propaganda. Bots have been employed in recent years to spread false propaganda on social media. Therefore, it is crucial to confirm whether information in news articles is true or false. To train the data and spot offensive or hateful content in news data, Hugging Face's BERT base transformer model was used. During analysis, we designated offensive information as "hate" and neutral information as "not hate". This model was pre-trained on the HateXplain dataset [25]. From the comparative analysis depicted in Fig-2, it is clear that most of the covid data are normal. However, the abusive/hateful news percentage is also too high to be ignored.

- **Irony detection:** Sarcasm is a common way for people to convey their emotions. Sarcastic posts might include both accurate and inaccurate information. This factual ambiguity aids in the online dissemination of fake content. Ironic language on social media needs to be examined to stop this. This study used the RoBERTa-based transformer model to examine ironic content in social media. Data were divided into "ironic" and "non-ironic" categories. The outcomes are shown in Fig-2. Even though there were more ironic posts and news stories, there was still a significant amount of non-ironic posts about titles, content, and abstracts.

- **Grammatical analysis:** The number of people using social media is growing rapidly along with the number of internet users. The number of online newspapers has also grown concurrently. In place of traditional newspapers, people now rely on online news

portals and social media for their news. Online news portals' content quality, however, is not sufficiently standardized. These tabloids occasionally circulate false information to boost their audience. They frequently lack an appropriate editorial board and speak in grammatically incorrect ways. Therefore, it is important to consider the grammar of any news article. In order to achieve this, a BERT-based model was used to train the study's data. The Corpus of Linguistic Acceptability (CoLA), which concentrates on linguistic aspects of text, was used to pre-train the model. Label 0 (grammatically incorrect) and Label 1 (grammatically acceptable) were used to categorize the data [26]. Surprisingly, Fig-2 shows that, aside from the title, most news content and abstracts on social media were grammatically correct. This held true for both social media posts and news articles. The amount of data with label one was very high in abstract, content, and abstract. This is very alarming because, in numerous nations, newspapers are considered an excellent resource for young people learning foreign languages.

After training the data using the aforementioned BERT models, some post-processing tasks were performed. The first step was to determine the performance of these models. Due to that reason, it's crucial to validate all the aforementioned models. As a part of the evaluation process, accuracy score, precision, recall, and f1 scores were calculated. The final prediction scores of these models consisted of a label and a score, e.g., sentiment analysis yields positive/negative/neutral labels and their corresponding scores. These two pieces of data were subsequently combined to yield a single final score: Final Score = Prediction Score + Label Score
On a scale from 0 to 1, the label score represents the frequency of the label among all data. In sentiment analysis title, negative data comprised 24% of the total data, giving it a label score of 0.24, positive data comprised 2% of the total data, giving it a label score of 0.02 and neutral data comprised 74% of the total data, giving it a label score of 0.74. According to the aforementioned formula, the final score would be 0.75 + 0.74 = 1.49 if neutral news had a prediction score of 0.75. Similar to this, if a piece of positive news had a prediction score of 0.5, its final score would be 0.5 + 0.02 = 0.52. All five of the tasks perform this processing. Apart from calculating the final score, its crucial to validate all the classification tasks. All of these tasks are validated to verify if those models are working as per our expectations.

4.3 Rank Score

News websites may be biased or poorly ranked. The ranking of various news websites serves as the foundation for the rank score. The credibility of a website affects the quality of the news. For instance, traditional newspapers like the New York Times rank higher than satirical news websites like The Onion. To rank news websites according to various criteria, researchers from Stony Brook University developed the website Media Rank [28, 29]. Six different rankings were employed by the authors.
1. Reputation Rank
2. Popularity Rank
3. Breadth Rank
4. Ads Indicator
5. Spammer Indicator
6. Political Bias

As the ranking process was incomplete during the composition of this study, only the breadth rank is considered here. The reporting of trustworthy news organizations aims to be politically unbiased. Unlike narrow domains with few and repeating entity occurrences, reliable news sources work hard to cover the full spectrum of important news [29]. As a result, the depth of

insight, scope, relevance, clarity, and accuracy of reporting are all reflected in the breadth of coverage, which is a key sign of news quality [30]. Based on the quantity of distinct entities that appear in news reports, breadth rank quantifies the breadth of coverage. In this study, the rank score for each news source was determined using the breadth rank:

$$RankScore = 1/BreadthRank \tag{1}$$

It was not possible to obtain the breadth rank of all news data taken into account in this study because Media Rank does not cover all news websites. The breadth rank was estimated in cases where it wasn't available. The rank score was then used in the ensemble learning model after being normalized between a range of 0 and 1.

4.4 Ensemble Learning Model

The second half of the experiment was dedicated to ensemble learning. Our objective was to learn a stable model that performs well all around using a supervised machine learning algorithm. However, this requirement was met by multiple models in some circumstances. An ensemble learning model was used to lessen over-fitting and increase the generalizability of the model in order to address this problem. To create a stronger, more complete supervised model, ensemble learning involves combining a number of weak supervised models. The fundamental tenet of ensemble learning is that other weak classifiers will correct the error even if one weak classifier makes an incorrect prediction. Because of this, ensemble learning models are frequently used to combine various fine-tuned models [31]. In this study, two different types of ensemble models were used.
i) Voting Regressor
ii) Boosting Ensemble
i) Voting Regressor: An ensemble machine learning model called a voting ensemble (or "majority voting ensemble") combines the predictions from various other models. It is a method that can be applied to enhance model performance, ideally producing results that are superior to those of any individual model used in the ensemble. By combining the results from various models' predictions, a voting ensemble operates. It can be applied to regression or classification. Calculating the average of the model predictions is necessary in the case of regression [32]. When classifying data, each label's predictions are added up, and the label with the most votes is predicted. This research uses Voting Ensemble for regression. This implies that the average of all the input models must be calculated. The final score will be transmitted to Boosting Ensemble Model. Boosting Model is the last model applied on our data.
i) Boosting Ensemble: Another type of Ensemble Model is boosting. By developing a series of weak models, prediction power is generally increased[33]. Each model makes up for the shortcomings of its predecessors. It employs a gradual learning process, an iterative method that aims to reduce the errors of previous estimators. The entire process is sequential, and in order to make better predictions, each estimator relies on the one before it[34]. Extreme Gradient Boosting, also known as the XGBoost algorithm, is one of the most widely used boosting techniques. In order to increase the voting regressor's prediction score and determine the study's final output, the XGBoost algorithm was used. Here, the prediction score obtained from the voting regressor and rank score serve as the model's inputs. This entails rank score and the prediction score of title, content and abstract. The result is a binary score that can either be 0 (false) or 1. (true). After the completion of this study, the model was validated to determine how well the suggested model would perform. The previous version of our suggested model uses the aforementioned classification tasks. These tasks were implemented using the

identical pre-trained huggingface BERT models. The outcome of these classification tasks was the prediction scores. The final scores (achieved from label score and prediction score) were transmitted to weighted average ensemble model as input. On the other hand, rank score was calculated of the given news. Output of weighted ensemble score and rank score were fed into Stacking Ensemble classifier. The output of Stacked model successfully distinguishes between true and false news. Output 0 denotes fake news and output 1 denotes true news. In our previous system, the classification tasks were not validated. However, in this study, those tasks are validated in Co-Aid dataset. We are implementing voting regressor in the proposed model. In earlier studies, we implemented Weighted Average Ensemble model. Previous research used Stacking Ensemble model and in this study we replaced stacked model with XGboost model.

4.5 Results

The project was implemented using Python version 3.9 and the NVIDIA environment. The proposed solution was employed using PyTorch. The data was cleaned in the beginning. Handling missing value was crucial, as abstract column Hugging Face Transformer models were used to analyze the title, content, and abstract columns. The following transformer models, which are available on the Hugging Face website, were used to calculate the prediction scores:
1) Sentiment Analysis: CardiffNLP's twitter-roBERTa-base-sentiment-latest [35]
2) Emotion Analysis: Bhadresh Savani's distilbert-base-uncased-emotion [25]
3) Hate Speech Detection: Hate speech CNERG bert-base-uncased-hatexplain-rationale-two[36]
4) Irony Detection: CardiffNLP's twitter-roberta-base-irony [37]
5) Grammatical Analysis: textattack's bert-base-uncased-CoLA [26]
The dataset was trained using the aforementioned transformer models. The maximum length of the input data was set to 512 for all models. Default tokenizers from pretrained models were applied in this study. The prediction scores, collected from classification tasks were applied in second part of proposed model. All the classification tasks were validated and accuracy, precision, recall and f1 scores were calculated. For validation purpose, 4500 data were used as training and 957 data were used for testing the whole data set. The number of epochs = 3 and batch size = 8. The result was surprisingly well. The prediction and rank scores were normalized using minimum-maximum feature scaling. Subsequently, a voting regressor ensemble model was applied to the title, content, and abstract columns. As output, continuous prediction scores for each column were generated. The performance of classification tasks were measured. Due to this purpose, Accuracy, Precision, Recall and F-1 Scores were calculated. Table 1 clearly explains the performance measurement of all the classification tasks. The table successfully represented the accuracy, precision, recall and f1-score. The scores were amazing almost every where. That means, the models provide perfect prediction in majority cases. This is the end of text classification part. In the next step, prediction scores will be transmitted to Ensemble Learning Part. In this part, the first step was to apply Voting Regressor on Title, Content and Abstract. The prediction output needed to be boosted as the result was not satisfactory. Title, content, abstract, and rank scores were taken as the inputs of the XGBoost model. The goal of implementing XGBoost model was to achieve a final score for the whole news including rank score and evaluate the final model. The output column represented the output; output = 0 if the news was false and output = 1 if it was true. Scikit Learn was employed on XGBoost model. 80% of the entire data were used for training and 20% for testing. Boosting model performed surprisingly well. It successfully boosted the input score, which was the output of Voting Regressor. According to fig - 3 a), the confusion matrix elaborated more about the prediction employed on the test data set. Out of 1092 test samples, our model accurately predicted 893

Table 1: Evaluation of Text Classification Models

	Sentiment Analysis			
	Accuracy	***Precision***	***Recall***	***f1-Score***
Title	0.999	0.999	1.0	0.993
Content	0.997	1.0	0.996	0.998
Abstract	0.996	0.997	0.997	0.997
	Emotion Analysis			
	Accuracy	***Precision***	***Recall***	***f1-Score***
Title	0.979	0.992	0.982	0.987
Content	0.994	1.0	0.993	0.997
Abstract	0.987	0.992	0.992	0.992
	Hate Speech Analysis			
	Accuracy	***Precision***	***Recall***	***f1-Score***
Title	0.994	0.997	0.995	0.996
Content	0.994	0.999	0.994	0.996
Abstract	0.817	0.817	1.0	0.89
	Irony Speech Analysis			
	Accuracy	***Precision***	***Recall***	***f1-Score***
Title	0.969	0.997	0.965	0.981
Content	0.994	0.997	0.995	0.996
Abstract	0.993	0.995	0.996	0.996
	Grammatical Speech Analysis			
	Accuracy	***Precision***	***Recall***	***f1-Score***
Title	0.991	0.999	0.989	0.994
Content	0.989	0.998	0.987	0.993
Abstract	0.972	0.987	0.978	0.983

true and 186 fake data. On the other hand, 10 true data were predicted as fake and 3 fake data were predicted as true. This matrix proves that the model was accurately predicting majority of the time. As a result, the Accuracy, Precision, Recall and F1-Scores came out extremely well. The accuracy score = 0.98, precision = 1.0, recall = 0.99, f1-score = 0.98 and AUC score = 0.99 On the other hand, ROC AUC curve also provided an excellent result. Fig 3 b) depicts the ROC Curve of XGBoost model.

5 Discussion and Conclusion

The study presented a fantastic model that is capable of accurately identifying fake news. However, it only addressed the two categories of news—fake news and legitimate news. Implementing the proposed model on a dataset that is divided into more than two categories, such as true, partially true, fake, partially fake, etc., will be beneficial to obtain a better understanding. Another issue was the Accuracy score and F1-Score of Hate Speech Analysis. Other than Hate Speech Analysis, all these models have higher accuracy rate and also F1-score. These problems can be solved during future research. Another shortcoming is, this research is implemented only in Co-Aid dataset. Applying this model on a different dataset can be more helpful to verify the efficacy of this model. This model is only capable of detecting fake news online. We

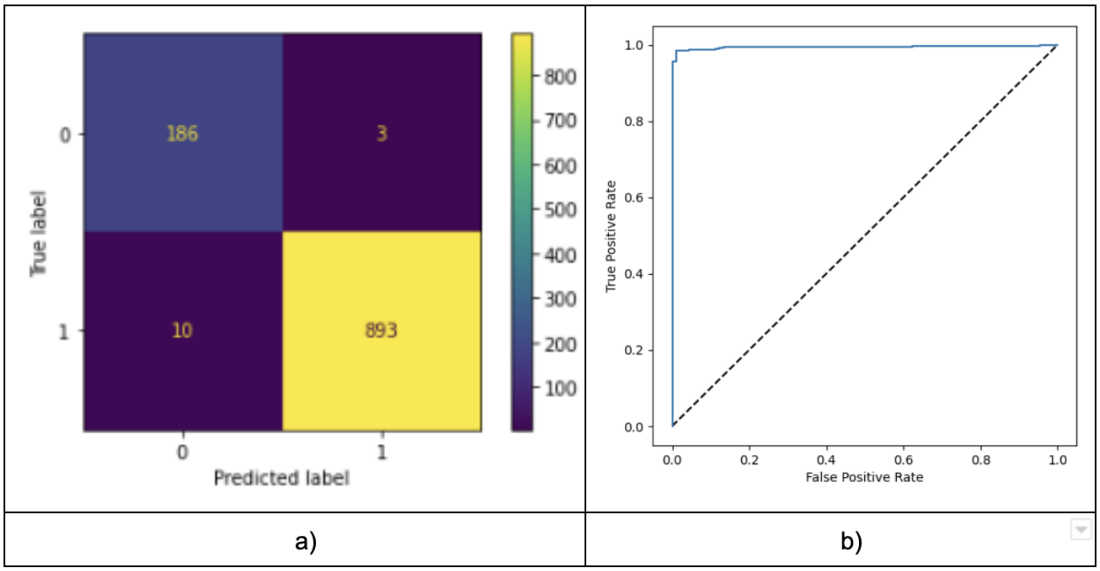

Figure 3: a) Confusion Matrix and b) ROC Curve of proposed model

didn't consider to track the news propagation and verify the source authenticity. Monitoring the propagation of fake news can be more helpful to identify the source of the news. This part will be covered in our future research.
The comparison between original model [38] and our suggested model is displayed in table 2. The proposed model is performing more effectively than the existing model. Accuracy, precision, recall, f1-score and AUC scores, all exhibit improved performance in this new model. The accuracy score was 0.97 in original model and 0.99 in XGboost model. f1-score and AUC score is also 0.99 in proposed model, wheres, those were 0.98 in the original model. This indicates that the proposed model outperforms the original model overall. Fake News has become a ma-

Table 2: Model Comparison

	Accuracy	***Precision***	***Recall***	***f1-Score***	***AUC***
Proposed Model	0.99	1.0	0.99	0.99	0.99
Original Model	0.97	0.98	0.98	0.98	0.98

jor issue due to overwhelming amount of news floating around mankind. Spreading fake news, caused enormous amount of harm in our society. Proposed model is a small initiative to control false and misleading information around us. The model was a two step process, where initial step was to understand the insight of given information based on different perspective of human behavior. The prediction scores were calculated successfully employing pre-trained BERT text classification models, e.g., sentiment analysis, emotion analysis, hate speech detection, irony detection, and grammatical analysis. The model was used to identify fake information in the second step by employing Voting Regressor followed by Boosting algorithms. The model performed admirably, displaying high accuracy and F1 score of (0.99) in both cases. The final outcome showed the highest AUC rating (0.99). The TPR rate in this model was close to one, according to the ROC curve, which supports the proposed model's quality of performance.

Before carefully selecting the final model, several experiments were run; the selected combination produced the best outcomes for spotting false information on social media. Calculating the variables for each threshold and plotting them on a plane is required in order to draw the curve. The model's performance is shown by the curve. Here, the true-positive rate is represented by the blue line, while the false-positive rate is represented by the black line. The ROC curve's close proximity to the axis in the figure shows how well the stacking ensemble performs.

References

[1] Cement, J. (2020, October). Number of social media users 2025. statista [online]. Available: https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

[2] Karimi, N. & Gambrell, J. (2022, October). Hundreds die of poisoning in Iran as fake news suggests methanol cure for virus. Times of Israel. [online] Available: https://www.timesofisrael.com/hundreds-die-of-poisoning-in-iran-as-fake-news-suggests-methanol-cure-for-virus/

[3] Wardle, C. & Derakhshan, H. (2017). "Information disorder: Toward an interdisciplinary framework for research and policymaking."

[4] Qayyum, A., Qadir, J., Janjua, M. U. & Sher, F. (2019). Using blockchain to rein in the new post-truth world and check the spread of fake news. IT Professional, 21(4), 16–24.

[5] van Der Linden, S., Roozenbeek, J., & Compton, J. (2020). Inoculating against fake news about COVID-19. Frontiers in psychology, 11, 566790.

[6] World Health Organization (2020b). Novel Coronavirus (2019-nCoV) Situation Report - 13. Available online at: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200202-sitrep-13-ncov-v3.pdf (accessedTo the Editor — Since the first reports of novel pneumonia (COVID-19) in Wuhan, Hubei province, China1,2, there has been considerable discussion on the origin of the causative virus, SARS-CoV-23 (also referred to as HCoV-19)4. Infections with SARS-CoV-2 are now widespread, and as of 11 March 2020, 121,564 cases have been confirmed in more than 110 countries, with 4,373 deaths5. November 8, 2022).

[7] World Health Organization (2020a). Coronavirus disease (COVID-19) advice for the public: Mythbusters. Available online at: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public/myth-busters (accessed November 8, 2022).

[8] Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C., and Garry, R. F. (2020). The proximal origin of SARS-CoV-2. Nat. Med. 26, 450–452. doi: 10.1038/s41591-020-0820-9

[9] BBC News (2020). Ofcom: Covid-19 5G Theories are "Most Common" Misinformation. www.Bbc.Co.Uk. Available online at: https://www.bbc.co.uk/news/technology-52370616 (accessed November 8, 2022).

[10] Khan, Y. H., Mallhi, T. H., Alotaibi, N. H., Alzarea, A. I., Alanazi, A. S., Tanveer, N., & Hashmi, F. K. (2020). Threat of COVID-19 vaccine hesitancy in Pakistan: the need for measures to neutralize misleading narratives. The American journal of tropical medicine and hygiene, 103(2), 603.

[11] Ahuja, A., & Bhaskar, S. (Eds.). (2021, February 19). COVID-19 vaccines myth vs fact: No vaccines do not alter DNA. NDTV. https://swachhindia.ndtv.com/covid-19-vaccinesmyth-vs-fact-novaccines-do-not-alter-dna-56612/

[12] Granik, M. & Mesyura, V. (2017, May). Fake news detection using naive Bayes classifier. In 2017 IEEE first Ukraine conference on electrical and computer engineering (UKRCON) (pp. 900–903). IEEE.

[13] Wang, W. Y. (2017). "Liar, liar pants on fire:" A new benchmark dataset for fake news detection. arXiv preprint arXiv: 1705.00648.

[15] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., & Polosukhin, I.(2017). Attention is all you need. Advances in neural information processing systems, 30.

[17] Glazkova, A., Glazkov, M. & Trifonov, T. (2021, February). g2tmn at constraint@ aaai2021: exploiting CT-BERT and ensembling learning for COVID-19 fake news detection. In International Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation (pp. 116–127). Springer, Cham.

[18] Barbieri, F., Camacho-Collados, J., Neves, L. & Espinosa-Anke, L. (2020). Tweeteval: Unified benchmark and comparative evaluation for tweet classification. arXiv preprint arXiv: 2010.12421.

[19] Utsha, R. S., Keya, M., Md. Arid Hasan, & Islam, M. S. (2021). Qword at CheckThat! 2021: An Extreme Gradient Boosting Approach for Multiclass Fake News Detection. In CLEF (Working Notes) (pp. 619-627).

[20] Fayaz, M., Khan, A., Bilal, M., & Khan, S. U. (2022). Machine learning for fake news classification with optimal feature selection. Soft Computing, 1-9.

[22] Cui, L. & Lee, D. (2020). Coaid: Covid-19 healthcare misinformation dataset. arXiv preprint arXiv: 2006.00885.

[23] (2022, October) Huggingface Website [Online]. Available: https://huggingface.co/models

[24] Tunstall, L., von Werra, L. & Wolf, T. (2022). Natural language processing with transformers. "O'Reilly Media, Inc."

[25] (2022, October) Huggingface Website [Online]. Available: https://huggingface.co/bhadresh-savani/distilbert-base-uncased-emotion

[26] (2022, October) Huggingface Website [Online]. Available: https://huggingface.co/textattack/bert-base-uncased-CoLA

[27] (2022, October) Huggingface Website [Online]. Available: https://huggingface.co/datasets/viewer/?dataset=emotion

[28] (2022, October) MediaRank Website [online]. Available: https://media-rank.com/

[29] Ye, J. & Skiena, S. (2019, July). MediaRank: Computational ranking of online news sources. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 2469–2477).

[30] Plasser, F. (2005). From hard to soft news standards? How political journalists in different media systems evaluate the shifting quality of news. Harvard International Journal of Press/Politics 10, 2 (2005), 47–68

[31] Zhou, S., Li, J. & Ding, H. (2021, February). Fake News and Hostile Posts Detection Using an Ensemble Learning Model. In International Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation (pp. 74–82). Springer, Cham.

[32] Machine Learning Mastery Website [online]. Available: https://machinelearningmastery.com/voting-ensembles-with-python/ Accessed on 17th November, 2022

[33] Machine Learning Mastery Website [online]. Available: https://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/ Accessed on 17th November, 2022

[34] TOPBOTS Website [online]. Available: https://www.topbots.com/practical-guide-to-ensemble-learning/ Accessed on 17th November, 2022

[35] (2022, October) Huggingface Website [Online]. Available: https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest

[36] (2022, October) Huggingface Website [Online]. Available: https://huggingface.co/Hate-speech-CNERG/bert-base-uncased-hatexplain-rationale-two

[37] (2022, October) Huggingface Website [Online]. Available: https://huggingface.co/cardiffnlp/twitter-roberta-base-irony

[38] Raquiba S., Tetsuro N. (2022). (Waiting for publication) Fake News Detection Using Transformer and Ensemble Learning Models