Fake News Detection System using BERT and Boosting Algorithm

Raquiba Sultana* and Tetsuro Nishino*
The University of Electro-Communications, Tokyo, JAPAN

Abstract

False information can proliferate and cause significant issues on social media platforms. To minimize the harm caused by false information, understanding its sensitivity and content is essential. This research analyzes the characteristics of human expression and, based on the results, successfully detects fake news by implementing different machine learning models. To identify false information on the Internet, we propose an ensemble model based on transformers. First, various text classification tasks were conducted to understand the contents of false and true news about COVID-19. The proposed hybrid ensemble learning model utilizes these results. The results of our analysis were encouraging, demonstrating that the proposed system can identify false information on social media platforms. All the classification tasks were validated and displayed outstanding results. The final model exhibited 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 Area Under the Curve (AUC) score was very high at 0.99. Thus, it was demonstrated that the proposed model effectively identified false information online.

Key Words: NLP, deep learning, text classification, BERT, boosting algorithm.

1 Introduction

The use of social media has steadily increased in recent years. Most Internet users are frequently active on websites such as Facebook, Instagram, and Twitter. Social media users were forecast to number 3.6 billion in 2020; by 2025, that number is projected to rise to 4.41 billion [8]. People frequently rely on social media for daily news. As a result, social media has become the center for spreading false information. The proliferation of fake news has become a global issue, especially during the COVID-19 pandemic. Due to fear of COVID-19, people are more likely to believe false information.

News that is false and disseminated through social media or news outlets is called fake news. In mass media, information accuracy is occasionally compromised to boost revenue. As a result, readers might be misled, and false information might be disseminated regarding politics, religious affiliations, branding, and financial services [35]. False information is propagated to attract public attention, making people more vulnerable to security attacks and harmful social and political issues. This may explain why the current era is defined as the "post-truth" era [24].

Daily news consumption alters how we see the world. The proliferation of false news has jeopardized the integrity of journalism and media. Governments and businesses have traditionally taken measures to define, recognize, and halt the spread of fake news as key goals. Nevertheless, millions receive falsified information daily, which is pervasive on social media. By fostering prejudice and intolerance, misinformation prepares the path for enduring issues.

Many aspects of society have suffered significant damage owing to fake news. For instance, in the stock market, a false story about the parent company of United Airlines declaring bankruptcy in 2008 resulted in a decrease in the stock price by 76% in a matter of minutes, a closing price that was 11% below the previous day, and the negative effect lasted for more than six days [7].

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 spreading false information can easily be done, has worsened this issue. This is frequently carried out to deceive those who believe the news and accomplish economic and political milestones. In addition, the mainstream media has become increasingly biased, and yellow journalism has become more common. Elections, democracy, war, and conflict are the main topics of political news.

In traditional media, politically biased reporting and pulling a predetermined line are frequently used to win over the public. Although such reporting does not spread factually incorrect information, it frequently presents incomplete information to deceive the public and further complicit political interests. Many misleading and inappropriate claims concerning the SARS-CoV-2 novel coronavirus (COVID-19) have been made in conjunction with the virus's outbreak, notably on social media [19]. The World Health Organization (WHO) warned about an ongoing "infodemic", or an excess of information, especially false information, during the pandemic due to the propagation of false information [16].

Since the outbreak of COVID-19, there have been numerous claims that the illness may be cured, including that consuming methanol, ethanol, and bleach can protect one against COVID-19 [38]. The WHO (World Health Organization) had to issue a warning to people not to consume these poisonous substances as a consequence [2]. Political leaders such as President Donald Trump endorsed this assertion, sparking controversy. He frequently described this disease as the Wuhan or China

*Graduate School of Informatics and Engineering, Email: sultana.ra@uec.ac.jp, nishino@uec.ac.jp

virus. In response, Asians had been targeted for their race in America. The spread of racial hate crimes is a direct result of misleading information.

Another well-linked hoax involves the 5G network. A rumor that the network was spreading the coronavirus or disrupting human immunity systems first appeared at the start of the lockdown. There were concerns that people ignited communication masts on fire across the UK as a result of the false reports. According to a spokesperson for the Mobile UK industry group, "more than 50" of these arson attacks occurred [4]. Rumors regarding the coronavirus vaccine also spread globally. Numerous studies have examined the relationship between coronavirus vaccine hesitancy and fake news. Anti-vaccine groups tried to demotivate mass people with their far-fetched conspiracy theories [17]. One well-known conspiracy theory claims vaccines permanently damage DNA or alter genes [1]. This myth was only about messenger RNA (messenger-RNA) vaccines, as they implement a genetic approach. These are examples of the spread of fake rumors in recent years. It is increasing alarmingly and requires immediate action to prevent the spread of fake information online.

Fake news creators frequently combine facts from reliable news sources with false materials to purposefully or inadvertently mislead readers. It is increasingly viewed as dangerous to democracy, public peace, and free speech and can confuse people and spark unrest. Many websites have taken on the responsibility of debunking and dismissing rumors and claims, especially those that receive thousands of views and likes before being proven false. A near real-time response is essential to prevent fake information from spreading among online users. Fact-checking websites frequently cannot verify the accuracy of all the latest information fast enough. Identifying fake news aims to save time and effort when examining news veracity[33].

The US 2016 election was questionable for many people as it caused a lot of fake news to spread online. Many researchers have attempted to determine fake news patterns during election periods. Recently, we have proposed a broad framework [27] which might be used in future elections worldwide to help people make better decisions in recognizing news deception and identifying an author's hidden bias. To conduct this study, the researchers built a dataset of 200 tweets about "Hilary Clinton" and conducted a truthfulness assessment. They started by "text normalizing" tweets, examined feature extraction techniques to categorize news, conducted a thorough linguistic analysis of tweets, and extracted the bag-of-words to find observable patterns, and then used the k-nearest neighbor algorithm to distinguish between polarized and credible news. They then discussed the outcomes of implementing the KNN algorithm, interconnected research domains, and future research directions for building an ideal model for a fake news detection system around social media before quantifying the success rate of the proposed framework.

The first step in detecting and preventing the spread of disinformation is understanding the information contained in it. For example, writing patterns, emotions, expression styles, and grammatical accuracy must be analyzed. In other words, it is necessary to identify standard patterns throughout the story. This current study aims to analyze the characteristics of fake and real news. Based on these characteristics, we determine the similarities and differences between the two news types. Recently, several studies have been conducted on this topic. For example, a Naive Bayes classifier was proposed and implemented for spam filtering via emails [13]. This research used Buzzfeed dataset and collected data from three major Facebook pages and three political news pages (Politico, CNN, and ABC News). The model exhibits a classification accuracy of 75.40%.

In another study, we proposed a hybrid fake news detection system focusing on BERT and Ensemble Learning models [28]. This study aimed to analyze the characteristics of fake news by implementing text classification tasks and detecting fake news using an ensemble learning model. These results were impressive. The accuracy score was 0.97, and the f1-score was 0.98.

2 Related Works

Several deep-learning-based methods that perform well on various datasets have been proposed to diminish the online spread of fake news. A recent study proposed a hybrid CNN model that integrated metadata with the text [34]. The authors sought to demonstrate that a hybrid approach could enhance text-only deep learning models. The results of the hybrid CNN were compared with those of support vector machines (SVM), logistic regression, Bi-LSTM, and CNN. Another study suggested an automatic fake news detection system based on a multi-perspective speaker profile [20].

The authors proposed a novel approach for integrating speaker profiles into an attention-based LSTM model to detect falsified news. The profile information served as an attention factor and additional input data. The system performance was assessed using a dataset from [34], and it was shown that adding speaker profiles significantly enhanced the output. The accuracy of this model on the benchmark dataset was 0.415, which is approximately 14.5% greater than that of the most advanced hybrid CNN model.

Public and academic communities have expressed interest in fake news [25]. Such false information has the potential to affect public perception, giving malicious groups a chance to influence the results of public events such as elections. Because of the high stakes involved, automatically identifying fake news is a vital but complex problem that remains poorly understood. However, three features of false news have been universally acknowledged: the language of the article, user responses it receives, and source users endorsing it. The existing research has mainly concentrated on developing solutions specific to a single attribute, which limits their applicability and success. To mitigate this issue, researchers have proposed a CSI model that consists of three modules: capture, score, and integrate

[10]. A recurrent neural network (RNN) is used in the first module, which is based on responses and text, to record the temporal pattern of user activity in a particular article. To evaluate whether an item is fake, the third module is paired with the second module, which learns the source characteristics based on user behavior. CSI outperforms current models in accuracy tests using real-world data and recovers valuable latent representations of users and articles. In recent years, transformers have become the most widely used deep learning model. It was first introduced in a seminar published by several researchers from Google and the University of Toronto [31].

This is a self-attention-based deep-learning language model. The authors suggested a new, straightforward network architecture based solely on attention mechanisms by rejecting the concepts of recurrence and convolutions. According to experiments on two machine translation tasks, these models exhibited superior quality while being more parallelizable and requiring a significant reduction in training time. Since then, several new transformer models have been proposed. These are the modified versions of the base model. In recent years, transformer models have become extremely popular for fake news detection. Several studies have been published based on this topic.

Another study proposed using a transformer-based ensemble of COVID-Twitter-BERT (CT-BERT) models [12]. The authors described the models utilized, methods used 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 have been 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 successfully identified features of social media news. The TweetEval framework, which evaluates tweet classifications for various tasks, was recently proposed. The benchmark for tweet classification, TweetEval,consists of seven Fundamental Heterogeneous Tasks in Social Media NLP Research. The authors compared various pretraining strategies for language modeling and proposed a strong set of baselines as the starting point.The effectiveness of starting with pretrained generic language models and continuing their training on Twitter corpora was demonstrated in these experimental results [3].

In another study, news articles were analyzed to determine whether they were accurate, partially true, false, or something else [30]. The dataset comprised news articles, titles, and article ratings. The data were 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 (0.54 on the provided dataset, the techniques were interpretive. Other classification models, such as Passive Aggressive Classifiers, Logistic Regression Classifiers, and Random Forest Classifiers, have shown different findings.

Another study demonstrated a straightforward method for detecting false information using a Naïve Bayes classifier [13]. The strategy was implemented as a software system and evaluated using data from Facebook news posts. Given the relative simplicity of the model, the classification accuracy of the test set was approximately 74%, which is reasonable. Several methods, which are also explained in this article, can be used to improve the outcomes. According to the results, artificial intelligence techniques can be used to address the challenge of detecting fake news.

Another study examined the rapid expansion of online news content and established whether the news was true or false [11]. Therefore, this 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. Several machine-learning algorithms have been used to identify and categorize fake news. However, the accuracies of these methods are limited. This current study uses a random forest classifier to distinguish between fake and real news. The selected News Dataset was used to extract twenty-three (23) textual features. Out of twenty-three features, 14 were chosen as the best using four techniques, including chi2, univariate, information gain, and feature importance. The proposed model and other benchmark techniques were assessed using the benchmark dataset with the best features. According to the 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 the Ada Boost Regression Model.

Social media and news media spread false information to increase the number of viewers or as part of the psychological competition. To mitigate this issue, another study determines a classification of the ensemble using a set of marked as true and false news articles [15]. This study develops a text-based classification approach using an SVM, Random Forest, and Naïve Bayes, and Decision Tree are used as base learners in Bagging and AdaBoost. The goal is to find an answer that allows the user to classify and filter fake material. Consequently, the authors determine that the best-performing classifiers were AdaBoost–Linear SVM and AdaBoost-Random Forest with an accuracy of 90.70% and 80.17%, respectively.

Fact-checking websites play crucial roles in identifying fake news. The difficult process of identifying fake news aims to save time and effort when examining news veracity. For this reason, another study proposed an approach that could identify possible fake news spreaders on social media as the first step towards preventing fake news from being propagated among online users. Therefore, they conducted different learning experiments from multilingual perspectives: English and Spanish. They evaluated different textual features primarily not tied to a specific language and compared different machine-learning algorithms. The results indicated that language-independent features could be used to distinguish between possible fake news spreaders and users who share credible information, with an average detection accuracy of 78% for English and 87% for the Spanish corpus [33].

3 Dataset Description

The fake news dataset aims to develop useful features that can distinguish fake news from legitimate news more precisely. Several methods have been developed to acquire news and determine its accuracy. Linguistic traits of news are present in many benchmark datasets for detecting fake information.

The novel coronavirus known as SARS-CoV-2, which was first identified in Wuhan, China, in December 2019, is thought to be the source of COVID-19. SARS-CoV-2 has rapidly spread around the globe. On January 30, 2020, the WHO labeled the outbreak a Public Health Emergency of International Concern [39]. Coughing, shortness of breath, fever, sore throat, and loss of taste or smell are typical COVID-19 symptoms. According to estimates, the incubation period lasts up to 14 days, with a median duration of 5.1 days[18].

Our society has been affected by COVID-19 for more than two years. The quality of life suffers due to the disruption of supply chains and the impact on the economies of several nations. The disease, infection rates, preventative measures, and vaccinations received daily top-priority news coverage during this time. Because of widespread panic, many people believed that the information shared online was true without checking the source; the spread of false information was almost as bad as the pandemic. This problem is referred to as an "infodemic". Social media sites such as Facebook and Twitter have served as the focal points of this "infodemic". The Co-Aid (COVID-19 Healthcare Misinformation) dataset was chosen for analysis because of this issue [9]. It consists of a variety of healthcare-related COVID-19 data that was obtained from social media.

Information was gathered from December 1, 2019, to September 1, 2020. Three versions were released during this period. In this study, the data were collected from all versions and combined. This information includes news reports, facts, and false information regarding COVID-19. COVID-19, coronavirus, pneumonia, flu9, lockdown, staying at home, quarantine, and ventilators are among the main topics. Most of the posts were gathered from Tiktok, Facebook, Twitter, Instagram, and YouTube. To collect news articles, the author retrieved URLs from several fact-checking websites, including LeadStories, PolitiFact, FactCheck.org, CheckYourFact, AFP Fact Check, and Health Feedback. After obtaining all the URLs of true and fake news related to COVID-19, the authors used newspapers to fetch their corresponding titles, contents, abstracts, and keywords. The original dataset contained 4,251 news articles, 296,000 user interactions, and 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 placed in separate files. Only the true and false data were considered in this study. Figure 1 and Figure 2 represent few examples of fake and real news used in the Co-Aid dataset.

Fake News		
Title	**Content**	**Abstract**
Regarding the risks of coronavirus transmission on an airplane "It's as safe as an environment as you're going to find."	on this face the nation broadcast moderated by margaret brennan click here to browse full transcripts of face the nation. margaret brennan i 'm margaret brennan in washington. and this week on face the nation moving on to may is proving to be even more challenging as the emotional dilemma between personal and economic well being intensifies. with restrictions on americans and businesses across the country easing by the day the trump administration says there are positive signs in th.	on this quot;face the nation&quot broadcast we sat down with illinois governor jb pritzker gilead sciences ceo daniel o'day and dr. scott gottlieb.
"The (corona)virus just isn't nearly as deadly as we thought it was."		many politicians couldn't seem less interested in asking. foxnews fox news operates the fox news channel fnc fox business network fbn fox news radio ....
"Children don't seem to be getting this virus."	but every judge mayor sheriff clerk and trustee was on the ballot. wisconsin would have been without elected officials from around the state during covid. this election was necessary as it was not just a democratic primary.	glad the governor is opening parks but what is the science of keeping restrooms and playgrounds closed. if social distancing works in the bathroom at ....

Figure 1: Example of fake news

Real News		
Title	**Content**	**Abstract**
Here's Exactly Where We Are with Vaccines and Treatments for COVID-19	scientists around the world are working on a number of vaccines and treatments for covid-19. Xinhua Zhang Yuwei via getty images scientists around the world are working on potential treatments and vaccines for the new coronavirus disease known as covid-19. several companies are working on antiviral drugs some of which are already in use against other illnesses to treat people who already have covid-19. other companies are working on vaccines that could be used as a preventive measure aga.	scientists around the world are working on a number of vaccines and treatments for covid-19 ..
Screen Time Doesn't Hurt Kids' Social Skills, Study Finds	a new study finds an increase in screen time does nt hurt kids social skills. getty images a new study found that despite the time spent on smartphones todays young people are as socially skilled as those of the previous generation. researchers compared teacher and parent evaluations of kids who entered kindergarten in 1998 years before facebook with children who did so in 2010. even children within both groups who experienced the heaviest exposure to screens showed similar developme.	new research found that school age children in 2010 despite the time spent on smartphones and social media are as socially skilled as those at the same age in 1998 ..
1 in 5 Cancer Survivors Stays at Their Job Due to Fears of Losing Health Insurance	experts say cancer survivors as well as their spouses and partners will experience job lock where they continue at their workplace to maintain their current health insurance. getty images researchers say 20 percent of cancer survivors have job lock where they stay in jobs mainly to keep their health insurance. experts say cancer survivors should take the time to fully understand their health coverage at work. there are alternatives for health insurance under the affordable care act.	experts say cancer survivors as well as their spouses and partners will experience job lock where they continue at their workplace to maintain their current health insurance ..

Figure 2: Example of real news

This information included social media posts and news articles. This study separately combined real and fake news from the entire data collection. A total of 4532 real data and 925 fake data points were utilized in this study. Fake and real data were combined for ease of analysis. Various fact-checking websites validated all the news articles and blog posts. Both true and fake data comprise a statement of the news type (articles/posts, etc.), fact-checking URL, news URL, title, news title, content, abstract, publishing date, and meta-keywords. Information was gathered from news URLs, titles, contents, and abstract columns. The title refers to the news or title of the article and content refers to the content of the news. The abstract refers to a brief description of the article. This research used the title, content, abstract, and URLs among all the information provided. Figure 4 shows a representation of all the analyses performed on the title, content, and abstract. This illustrates the patterns of information dissemination through social media during COVID-19.

4 Methodology

The primary objective of this study was to develop a model that could accurately identify false news on social media. To achieve this, we considered the information gathered from Twitter and examined the characteristics of news articles and social media posts to build a hybrid system for identifying false information on social media. Various text-classification tasks have aided in understanding the characteristics of tweets. This study was influenced by the TweetEval framework [3].

To detect fake news, it is crucial to analyze the data and determine their patterns. This section focuses on the analysis of the data patterns. 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 investigated the characteristics and patterns of tweets and news articles. All classification tasks were performed using pretrained transformer models from Hugging Face website. Huggingface website offers various pretrained transformer models for different purposes. Available pretrained models can be used for different tasks; such as text classification, image classification, feature extraction, question answering etc. All news items were then rated according to the reliability of their sources. The ensemble learning model was updated based on all the results. The Voting Regressor model received the prediction scores from each classification task as input. The Boosting Ensemble model then received the output score of the voting regressor and the rank score, which predicted whether the news was true or false. Figure 3 depicts the overall process architecture.

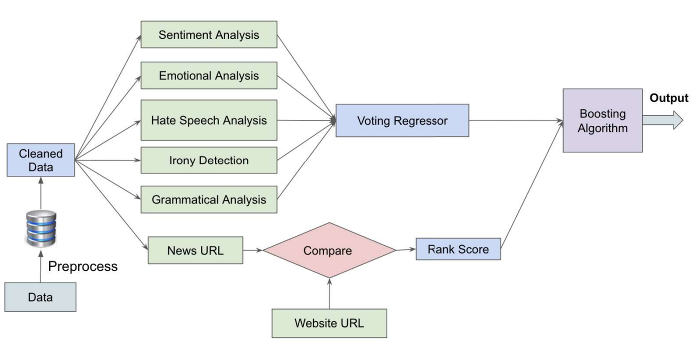


Figure 3: Architecture of proposed model

4.1 Data Pre-processing

The first step was to preprocess the entire dataset. Data preprocessing was the most important step because the raw data were difficult to train. Unprocessed data often yield poor results. This is particularly true when there are large amounts of missing data. The missing values was a crucial issue for the Co-Aid dataset, as many content and abstract data were missing. Those missing from the content columns were handled by inputting the value with the title. However, the missing value in the abstract was replaced with the title. The punctuation was also removed to clean the data. Consequently, the rank scores were normalized using minimum-maximum feature scaling.

4.2 Information Analysis

The data were trained on the basis of all five classification tasks. After training, the prediction scores were transmitted to the ensemble model section. Figure 4 presents the prediction scores of the trained data for all five tasks in the Co-Aid dataset.

The information analysis aimed to determine how people behaved during the COVID-19 pandemic. The “infodemic” era began during the COVID-19 period. They were anxious and believed in anything that could stop the outbreak. Some people made an effort to use this circumstance by spreading false information regarding diseases, prevention, governmental policies, etc. This makes it necessary to examine the patterns of fake news during the pandemic.

- **Sentiment Analysis:** The goal of sentiment analysis is to determine whether tweets are positive, negative, or neutral. The pretrained transformer model CardiffNLP’s twitter-roBERTa-base-sentiment-latest [21] was employed to analyze the sentiments [3]. Using this model, the titles, contents, and abstracts were trained. This specific model was pretrained on approximately 124M tweets. The tweets were collected from 2018 to 2021 and fine-tuned for sentiment analysis using TweetEval.

  This pretrained model was applied to the Co-Aid dataset to analyze the sentiments of the data. The findings of the sentiment analysis for COVID-19 are displayed in Figure 4. According to sentiment analysis in Figure 4, neutral news was the most prevalent type, comprising a significant portion of titles, contents, and abstracts. Neutral news accounted for more than 70% of the three cases. However, the prevalence of negative emotions was much lower than that of neutral emotions. Negative emotions ranged from 18% to 24%. Surprisingly, the percentage of positive sentiments is 3%, which is negligible compared to the other cases.
- **Emotion Analysis:** Another text classification task is emotion analysis, which divides data into six categories: anger, fear, joy, love, sadness, and surprise. This assignment aimed to identify the various emotional states in tweets [29]. A pretrained DistilBERT model obtained from the Hugging Face was employed to train the

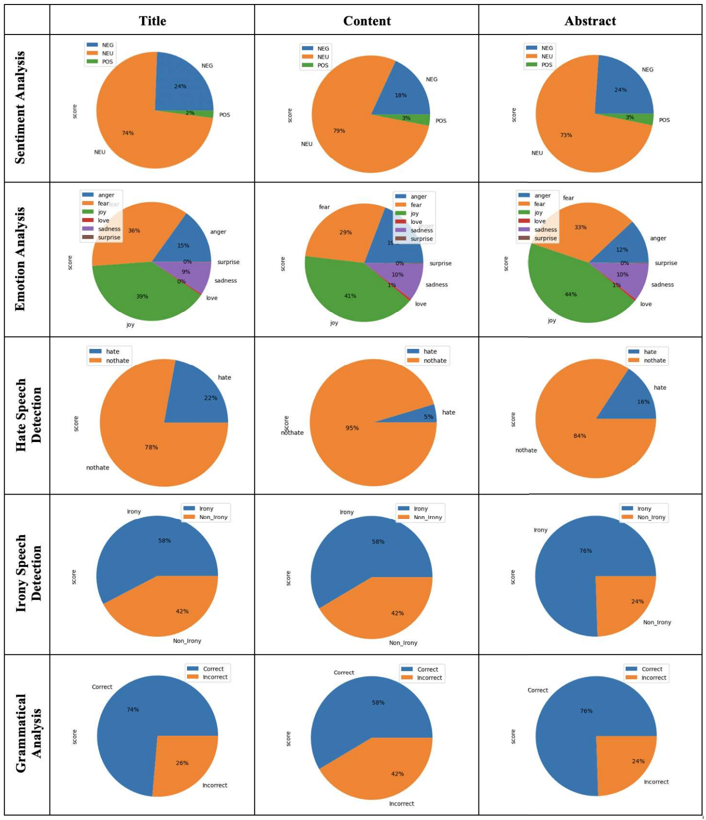


Figure 4: Comparative representation of all the analysis tasks of CoAid dataset

data. "bhadresh-savani/distilbert-base-uncased-emotion" [26] was employed in this study. Originally, the developer fine-tuned the distilbert-base-uncased model on the emotion dataset [32] using HuggingFace Trainer with specific hyperparameters. The patterns that the posts follow can be explained by emotion analysis.

As illustrated in Figure 4, angry, happy (joyful), and fearful feelings were frequently expressed in news articles. The amount of joyful news was the highest in all three cases. In contrast, the frequency of sad posts was approximately 10%. In contrast, romantic (love) posts were negligible (approximately 1%). The data were gathered at the start of the COVID-19 pandemic, which was characterized by anxiety about the illness and resentment towards the government over measures such as the lockdown. However, when news about vaccines was reported, people felt relieved.

- **Hate Speech Detection:** Hateful content is frequently found in fake news. Although this is true in the case of true news, it is much less likely. Occasionally, people make conscious attempts to spread divisive propaganda. In recent years, bots have been used to spread false propaganda on social media platforms. Therefore, confirming whether the information in news articles is true or false is crucial. HuggingFace's BERT-based transformer model was used to train the data and spot offensive or hateful content in the news data. During the analysis, we designated offensive information as "hate" and neutral information as "not hate." The model was pretrained using the HateXplain dataset [22].

  The comparative analysis depicted in Figure 4, shows that most of the cloud data are normal. However, the percentage of abusive or hateful news was too high to be ignored, especially in the title (22%) and abstract (16%).

- **Irony Detection:** Sarcasm is a common way in which people convey emotions. Sarcastic posts contain both accurate and inaccurate information. This ambiguity aids the online dissemination of fake content. Ironic language on social media must be examined to prevent this. This study used the RoBERTa-based transformer model to examine the ironic content in social media. The data were divided into "ironic" and "non-ironic" categories. The results are shown in Figure 4. Surprisingly, most of the posts contained ironic data. The title and content both consist of 58% of the ironic data. This amount was the highest in the abstract (76%). Although there were more ironic posts and news stories, significant percentages of non-ironic posts regarding titles (42%), content (42%), and abstracts (24%) remain.

- **Grammatical Analysis:** The number of people using social media and internet users is proliferating. The number of online newspapers has increased concurrently. Instead of traditional newspapers, people rely on online news portals and social media for news. However, the content quality of online news portals is not sufficiently standardized. These tabloids occasionally circulate false information to boost their audiences. They frequently lack an appropriate editorial board and speak grammatically incorrectly. Therefore, it is important to consider the grammar of any news article.

  To achieve this, a BERT-based model was used to train the data. The Corpus of Linguistic Acceptability (CoLA), which concentrates on the linguistic aspects of texts, was used to pretrain the model. The labels 0 (grammatically incorrect) and 1 (grammatically acceptable) were used to categorize the data [40]. Surprisingly, Fig-4 shows that, aside from the title, most news content and abstracts on social media were grammatically correct. This is true for both social media posts and news articles. The amount of grammatically acceptable data was very high for the title (74%), content (58%), and abstract (76%). This is alarming because newspapers are considered excellent resources for young people learning foreign languages in numerous nations.

After training the data using the aforementioned BERT models, postprocessing tasks were performed on both datasets. The first step was to determine the performance of the models. Therefore, it is crucial to validate all the aforementioned models. As part of the evaluation process, accuracy, precision, recall, and f1 scores were calculated. The final prediction

scores of these models consisted of a label and a score; for example, sentiment analysis yielded positive/negative/neutral labels and their corresponding scores. These two data sets were subsequently combined to yield a final score:
Final Score = Prediction Score + Label Score

On a scale of 0 to 1, the label score represents the frequency of the label among all the data. For example, in the sentiment analysis title of the Co-Aid dataset, negative data comprised 24% of the total data with a label score of 0.24. In contrast, positive data comprised 2% of the total data, giving them a label score of 0.02, and neutral data comprised 74% of the total data, giving them a label score of 0.74. According to the aforementioned formula, if the neutral news had a prediction score of 0.75, the final scores would be 0.68, 0.31, and 0.75. Similarly, if a piece of positive news had a prediction score of 0.5 and a label score of 0.02, its final score would be 0.5 + 0.02 = 0.52. All five participants performed the task. In addition to calculating the final score, it is crucial to validate all classification tasks. All these tasks were validated to verify whether these models functioned per our expectations.

4.3 Rank Score

News websites can be biased or poorly ranked. The rankings of various news websites served as the foundation for the ranking scores. The credibility of a website affects the news quality. For instance, traditional newspapers such as the New York Times rank higher than satirical news websites such as The Onion. Researchers from Stony Brook University developed the Media Rank website to Rank [36]. 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

Because the ranking process was incomplete during the composition of this study, only the breadth rank was considered. The reporting of trustworthy news organizations aims to be politically unbiased. Unlike narrow domains with a few repeating entity occurrences, reliable news sources work hard to cover the full spectrum of important news [36]. Consequently, the depths of insight, scope, relevance, clarity, and reporting accuracy are reflected in the breadth of coverage, which is a key indicator of news quality [23]. Based on the number of distinct entities appearing in news reports, breadth rank quantifies the breadth of coverage. This study determined the rank score for each news source using breadth rank.

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

It was not possible to obtain the breadth rank of all the news data considered in this study because it did not cover all news websites. The breadth rank was estimated for cases in which it was not available. In particular, the breadth and rank scores of all government websites were estimated to be 1, as we assumed that government websites provide correct information. The rank score was then used in the ensemble learning model after normalization between 0 and 1.

4.4 Ensemble Learning Model

The second half of the experiment was dedicated to ensemble learning. We aimed to develop a stable model that performs well using a supervised machine learning algorithm. However, under certain circumstances, this requirement can be satisfied by multiple models. To address this problem, an ensemble learning model was used to reduce overfitting and increase the model's generalizability. Ensemble learning involves combining several weakly supervised models to create a stronger and more complete supervised model. The fundamental tenet of ensemble learning is that the other weak classifiers correct errors even if one weak classifier makes an incorrect prediction. Therefore, ensemble-learning models are frequently used to combine various fine-tuned models [37]. Two different types of ensemble models were used in the study.

i) Voting Regressor
ii) Boosting Ensemble

i) Voting Regressor: An ensemble machine-learning model called a voting ensemble (or "majority voting ensemble") combines predictions from various other models. This method can be applied to enhance the model performance, ideally producing results superior to those of any individual model used in the ensemble. By combining the results from various models' predictions, a voting ensemble operates. This method can be applied to regression or classification. Calculating the average of the model predictions is necessary for regression [5]. When classifying the data, the predictions of each label were added, and the label with the most votes was predicted. This study used a Voting Ensemble for the regression because the average of all input models must be calculated. The final score is transmitted to the Boosting Ensemble Model. The Boosting model was the last one applied to our data.

ii) Boosting Ensemble: Boosting is another type of Ensemble Model. Developing a series of weak models generally increases the prediction power [6]. Each model compensates 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 to make better predictions, each estimator relies on the one before it [14]. Extreme Gradient Boosting, also known as the XGBoost algorithm, is one of the most widely used boosting techniques. The XGBoost algorithm was used to increase the voting regressor's prediction score and determine the final output of the study. The prediction score obtained from the voting regressor and the rank score served as the model's inputs. This entails the ranking and prediction scores of the title, content, and abstract. The result was a binary score of either zero (false) or one. (true). After the completion of this study, the model was validated to determine how well the suggested model would perform.

The previous version of the proposed model used the aforementioned classification tasks. These tasks were implemented using identical pre-trained hugging-face BERT models. The outcome of these classification tasks was the prediction score. The final scores (obtained from the label and prediction scores) were transmitted to the weighted average ensemble model as the input. In contrast, the rank score was calculated for the given news item. The outputs of the weighted ensemble and rank scores are fed into a Stacking Ensemble classifier. The output of the stacked model successfully distinguishes between true and false news items. Output 0 denotes fake news, and output 1 denotes true news. In our previous system, classification tasks were not validated. However, in the present study, these tasks were validated using the Co-Aid dataset. We implement a voting regressor in the proposed model. In previous studies, we implemented a Weighted Average Ensemble model. Previous research used the Stacking Ensemble model; in this study, we replaced the stacked model with the 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 values is crucial, because the abstract column contains many missing data points. HuggingFace 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-twitter-roBERTa-based-sentiment-latest [21]
2) Emotion Analysis: Bhadresh-Savani-distilbert-based-uncased-emotion [26]
3) Hate Speech Detection: Hate-speech-CNERG-bert-base-uncased-hatexplain-rationale-two [22]
4) Irony Detection: CardiffNLP-twitter-roberta-base-irony [3]
5) Grammatical Analysis: textattack-bert-base-uncased-CoLA [40]

The dataset was trained using the transformer models. The maximum length of the input data was set to 512 for all the models. The default tokenizers from the pretrained models were used in this study. The prediction scores collected from the classification tasks were applied in the second part of the proposed model. All classification tasks were validated, and the accuracy, precision, recall, and f1 scores were calculated. For validation purposes, 4500 data points were used for training, and 957 data points were used for testing the entire dataset. There were three epochs=3, and the batch size was eight. Surprisingly, the results are satisfactory.

The prediction and rank scores were normalized using minimum–maximum feature scaling. Subsequently, a voting regressor ensemble model is applied to the title, content, and abstract columns. The continuous prediction scores for each column were generated as outputs. Subsequently, the performance of the classification task was measured. Due to this purpose, accuracy, precision, recall, and f1 scores were calculated. Table 1 clearly explains the performance measurements of all classification tasks applied to the Co-Aid dataset. The table successfully presents the accuracy, precision, recall and f1-score. The scores were impressive almost everywhere. This implies that the models provide perfect

Table 1: Evaluation of text classification models

Co-Aid	**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

prediction in the majority cases for Co-Aid dataset. The model performed well in Co-Aid dataset. In most cases, the accuracy, precision, recall, and f1-score of the text-classification n tasks were approximately d 99%. In some cases. precision and recall achieved 100% scores. However, there were some exceptions in which the scores were much lower. This is the end of text classification. In the next step, the prediction scores were transmitted to the ensemble learning module.

The first step is to apply a Voting Regressor on Title, Content and Abstract. The prediction output needs to be boosted because the results are unsatisfactory. The title, content, abstract, and rank scores were used as inputs for the XGBoost model. The goal of implementing the XGBoost model was to achieve a final score for all news items, including the rank score, and to evaluate the final model. The output column represents the output: Output = 0 if the news is false and output = 1 if it is true. SciKit Learn is employed in the XGBoost model. Approximately 80% of the entire Co-Aid dataset was used for training and 20% for testing. Surprisingly, the Boosting model performed well on the Co-Aid dataset. This successfully boosts the input score, which is the output of the Voting Regressor.

According to figure 5 a), the confusion matrix elaborates more on the prediction employed for the test data of the Co-Aid dataset. Out of the 1092 test samples, our model accurately predicted 893 true and 186 fake data. In contrast, ten true data points were predicted as fake, and three fake data points were predicted as true. This matrix proves that the model accurately predicted most of the time. Consequently, the accuracy, precision, Recall, and f1-Scores were extremely good (accuracy score = 0.98, precision = 1.0, recall = 0.99, f1-score = 0.98, AUC (Area Under the Curve) score = 0.99). By contrast, the ROC AUC curve showed excellent results. Figure 5 b) shows the receiver operating characteristic (ROC) curve of the XGBoost model.

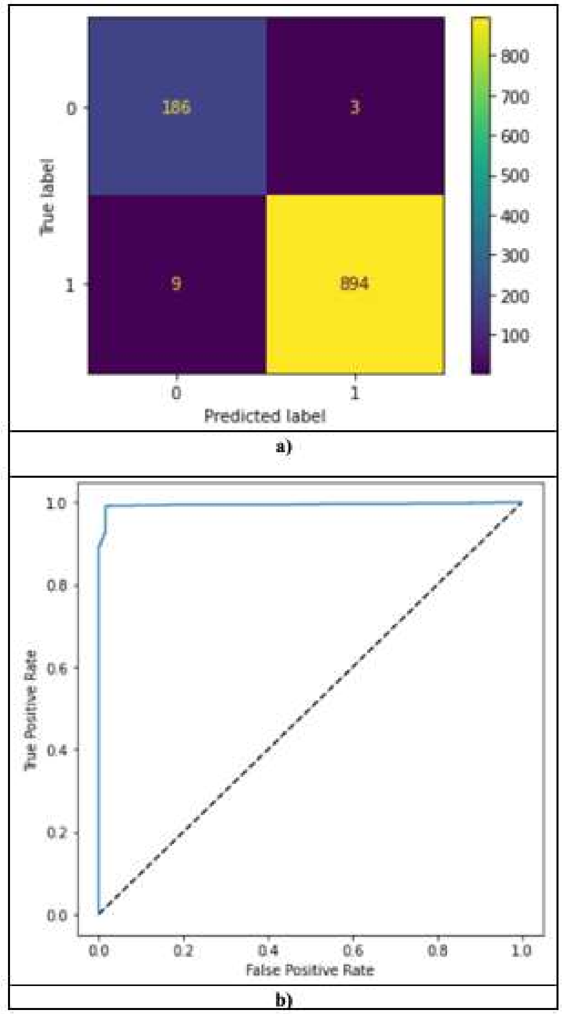


Figure 5: a) Confusion matrix and b) ROC Curve of the proposed model for CoAid dataset

5 Discussion and Conclusion

This study presents an excellent model capable of accurately identifying fake news. However, it only addresses two categories of news: fake and legitimate. Implementing the proposed model on a dataset divided into more than two categories–true, partially true, fake, and partially fake–will be beneficial for obtaining a better understanding. Another issue was the accuracy score and f1-Score of the Hate Speech analysis. In addition to Hate speech analysis, all these models had higher accuracy rates and also f1-score. These issues should be addressed in future studies. Another shortcoming is that this study was implemented only on the Co-Aid dataset. Applying this model to a different dataset can help verify its efficacy. This model can only detect fake online news. We did not consider tracking news propagation or verifying source authenticity. Monitoring the propagation of fake news can help identify the source of the news. This issue will be addressed in future studies.

The comparison between the original model [28] and our suggested models are presented in Table 2. The proposed model performed better than the existing models. The accuracy, precision, recall, f1-score and AUC scores all exhibited improved performance in this new model. The accuracy score was 0.97 in the original model and 0.99 in the XGboost model. f1-score and AUC score is also 0.99 in the proposed model, whereas those were 0.98 in the proposed and original models, respectively. This indicates that the proposed model outperformed the original one. Fake News has become a major issue due to the overwhelming amount of news floating around humans. The spread of fake news has caused enormous harm to society. The proposed model is a small initiative to control false and misleading information. The model is a two-step process in which the initial step is to understand the given information based on different perspectives of human behavior. Prediction scores were successfully calculated by employing pretrained BERT text classification models, such as 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 a Voting Regressor, followed by Boosting algorithms. The model performed admirably, displaying high accuracy and an f1 score

Table 2: Comparison between the proposed and original model

Model	***accuracy***	***precision***	***recall***	***f1-Score***	***AUC***
Proposed	0.99	1.0	0.99	0.99	0.99
Original	0.97	0.98	0.98	0.98	0.98

of (0.99) in both cases. The final outcome exhibits the highest AUC rating of 0.99 for the Co-Aid dataset. The TPR rate in this model was close to one, according to the ROC curve, which supports the performance of the proposed model.

Before carefully selecting the final model, several experiments were conducted. 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 are required to draw the curve. The performance of the model is illustrated by a curve. Here, the true-positive rate is represented by the blue line, whereas the false-positive rate is represented by the black line. The close proximity of the ROC curve to the axis in the figure indicates the performance of this Boosting model.

References

[1] A. Ahuja, "COVID-19 Vaccines Myth vs Fact: No Vaccines do not Alter DNA," https://swachhindia.ndtv.com/covid-19-vaccinesmyth-vs-fact-novaccines-do-not-alter-dna-56612/, NDTV, Feb. 2021.

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

[3] F. Barbieri, J. Camacho-Collados, L. Neves, and L. Espinosa-Anke, "Tweeteval: Unified Benchmark and Comparative Evaluation for Tweet Classification," arXiv preprint arXiv: 2010.12421, 2020.

[4] BBC News, "Ofcom: Covid-19 5G Theories are "Most Common" Misinformation," Available, Online: https://www.bbc.co.uk/news/technology-52370616, April 2020, [accessed: Nov. 8, 2022].

[5] J. Brownlee, "How to Develop Voting Ensembles With Python," Machine Learning Mastery Website, Online, Available: https://machinelearningmastery.com/voting-ensembles-with-python/, Nov. 2022.

[6] J. Brownlee, "A Gentle Introduction to the Gradient Boosting Algorithm for Machine Learning," Machine Learning Mastery Website, Online, Available: https://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/, Aug. 2020 [accessed: Nov. 10 2022].

[7] C. Carvalho, N. Klagge and E. Moench, "The Persistent Effects of a False News Shock," *Journal of Empirical Finance* 18(4): 597–615, 2011.

[8] J. Cement, "Number of Social Media Users 2025," Statista, Online, Available: https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/, Oct. 2020.

[9] L. Cui, and D. Lee, "Coaid: Covid-19 Healthcare Misinformation Dataset," arXiv preprint arXiv: 2006.00885, 2020.

[10] A. Dey, R. Z. Rafi, S. H. Parash, S. K. Arko, and A. Chakrabarty, "Fake News Pattern Recognition Using Linguistic Analysis," *In 2018 Joint 7th International Conference on Informatics, Electronics & Vision (ICIEV) and 2018 2nd International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*, IEEE, pp. 305-309, June 2018.

[11] M. Fayaz, A. Khan, M. Bilal and S. U. Khan, "Machine Learning for Fake News Classification with Optimal Feature Selection," *Soft Computing*, 26(16): 7763-7771, 2022.

[12] A. Glazkova, M. Glazkov and T. Trifonov, "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*, Cham: Springer International Publishing, pp. 116–127, Feb. 2021.

[13] M. Granik, and V. Mesyura, "Fake News Detection Using Naive Bayes Classifier," *In 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)*, IEEE, pp. 900–903, May 2017.

[14] I. Ismiguzel, "Practical Guide to Ensemble Learning," TOPBOTS Website, Online, Available: https://www.topbots.com/practical-guide-to-ensemble-learning/, Sep. 2021 [accessed: Nov. 2022].

[15] M. Z. Khan, and O. H. Alhazmi, "Study and Analysis of Unreliable News Based on Content Acquired Using Ensemble Learning (Prevalence of Fake News on Social Media)," *International Journal of System Assurance Engineering and Management*, 11(2): 145-153, 2020.

[16] N. Kassam, "Disinformation and Coronavirus," The Interpreter, Online, Available: https://www.lowyinstitute.org/the-interpreter/disinformation-coronavirus, Lowy Institute, March 2020, [accessed: June 26, 2022].

[17] Y. H. Khan, T. H. Mallhi, N. H. Alotaibi, A. I. Alzarea, A. S. Alanazi, N. Tanveer, and F. K. Hashmi, "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, 2020.

[18] S. A. Lauer, K. H. Grantz, Q. Bi, F. K. Jones, Q. Zheng, H. R. Meredith, A. S. Azman, N. G. Reich, and J. Lessler, "The Incubation Period of Coronavirus Disease 2019 (covid-19) From Publicly Reported Confirmed Cases: Estimation and Application," *Annals of Internal Medicine*, 172(9):557-582, 2020.

[19] S. van Der Linden, J. Roozenbeek, and J. Compton, "Inoculating Against Fake News About COVID-19," *Frontiers in Psychology*, 2928, 2020.

[20] Y. Long, Q. Lu, R. Xiang, M. Li, and C. R. Huang, "Fake News Detection Through Multi-Perspective Speaker Profiles," *In Proceedings of the eighth international joint*

conference on natural language processing, Short papers 2:252-256, Nov. 2017.

[21] D. Loureiro, F. Barbieri, L. Neves, L. E. Anke, and J. Camacho-Collados, ”Timelms: Diachronic language models from twitter,” arXiv preprint arXiv:2202.03829, 2022.

[22] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee, ”Hatexplain: A Benchmark Dataset for Explainable Hate Speech Detection,” *In Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14867-14875, May 2021.

[23] F. Plasser, ”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): 47–68, 2005.

[24] A. Qayyum, J. Qadir, M. U. Janjua, and F. Sher, ”Using Blockchain To Rein in The New Post-Truth World and Check The Spread of Fake News,” *IT Professional*, 21(4): 16–24, 2019.

[25] N. Ruchansky, S. Seo, and Y. Liu, ”Csi: A Hybrid Deep Model for Fake News Detection,” *In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pp. 797-806, Nov. 2017.

[26] B. Savani, ”Distilbert-Base-Uncased-Emotion,” Online, Available: https://huggingface.co/bhadresh-savani/distilbert-base-uncased-emotion, Oct. 2022.

[27] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, ”Fakenewsnet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media,” *Big data* 8(3):171-188, 2020, 2020.

[28] R. Sultana and T. Nishino (2022), ”Fake News Detection Using Transformer and Ensemble Learning Models,” *2022 13th International Congress on Advanced Applied Informatics Winter (IIAI-AAI-Winter)*, IEEE, 2022.

[29] L. Tunstall, L. von Werra and T. Wolf, ”Natural Language Processing With Transformers,” O’Reilly Media, Inc., 2022.

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

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, and I. Polosukhin, ”Attention is All You Need,” *Advances in Neural Information Processing Systems*, pp. 30, 2017.

[32] A. P. B. Veyseh, F. Dernoncourt, Q. H. Tran, and T. H. Nguyen, ”What does this Acronym Mean? Introducing a New Dataset for Acronym Identification and Disambiguation,” arXiv preprint arXiv:2010.14678, 2020.

[33] I. Vogel and M. Meghana, ”Detecting Fake News Spreaders on Twitter from a Multilingual Perspective,” *In 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, IEEE, pp. 599-606, Oct. 2020.

[34] W. Y. Wang, “Liar, Liar Pants on Fire: A New benchmark Dataset for Fake News Detection,” arXiv preprint arXiv: 1705.00648, 2017.

[35] C. Wardle, and H. Derakhshan, “Information Disorder: Toward an Interdisciplinary Framework for Research and Policymaking,” 2017.

[36] J. Ye and S. Skiena, ”MediaRank: Computational Ranking of Online News Sources,” *In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2469–2477, July 2019.

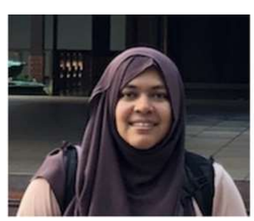
[37] S. Zhou, J. Li, and H. Ding, ”Fake News and Hostile Posts Detection Using an Ensemble Learning Model,” *In Combating Online Hostile Posts in Regional Languages during Emergency Situation: First International Workshop, CONSTRAINT 2021, Collocated with AAAI 2021, Virtual Event, Springer International Publishing, Revised Selected Papers 1*, pp. 74-82, Feb. 2021.

[38] World Health Organization, ”Coronavirus Disease (COVID-19) Advice for the Public: Mythbusters,” Online, Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public/myth-busters, Aug. 2020, [accessed November 8, 2022].

[39] World Health Organization, ”Statement on the Second Meeting of the International Health Regulations (2005) Emergency Committee Regarding the Outbreak of Novel Coronavirus (2019-ncov),” Online, Available: https://www.who.int/news-room/detail/30-01-2020-statement-on-the-second-meeting-of-the-international-health-regulations-(2005)-emergency-committee-regarding-the-outbreak-of-novel-coronavirus-(2019-ncov), Jan. 2020, [accessed: Oct. 10, 2022].

[40] ”bert-base-uncased-CoLA”, Online, Available: https://huggingface.co/textattack/bert-base-uncased-CoLA, Oct. 2022.

[41] ”Hugging Face – The AI Community Building the Future,”, Online, Available: https://huggingface.co/, [accessed: Aug. 4, 2023].



Raquiba Sultana is a Ph.D. Researcher at The University of Electro-Communications, Tokyo, Japan. She is awarded with Japanese government scholarship, MEXT (Monbukagakusho Scholarship) for pursuing her Ph.D. Her research focuses on Artificial Intelligence and Blockchain. She is enthusiastic about working on applications of artificial intelligence, especially on Natural Language Processing. She also has experience in software engineering in the healthcare industry, and project management in the EdTech field.

Tetsuro Nishino is a Professor at The University of Electro-Communications. He is also a theoretical computer scientist. His research focuses on quantum computations, neural nets, and circuit complexity theory. His primary interest is the computational complexity theory for finding new computation models. He is affiliated with the Information Processing Society of Japan; the Japanese Society for Artificial Intelligence; the Institute of Electronics, Information and Communication Engineers; the Mathematical Society of Japan; the Association for Computing Machinery (ACM).