Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 2, April 2025
https://doi.org/10.47709/cnahpc.v7i2.5717

Submitted : Mar 24, 2025
Accepted : Apr 04, 2025
Published : Apr 04, 2025

Fake News Detection Using Machine Learning Techniques

Fazal Wahab[1], Imran Khan[2], Anju Shankar[3]
[1] National University of Computer and Emerging Science, Faisalabad, Pakistan, studentpk2024@gmail.com
[2] Department of Informatics, University of Sussex, UK; imran.khan@sussex.ac.uk
[3] Department of Engineering and Computer Science, University of Hertfordshire, UK; anju.shankar@herts.ac.uk
Corresponding Author: studentpk2024@gmail.com

ABSTRACT

One of the most important initiatives has been the widespread adoption of internet use. People use the Internet for a wide range of reasons. These people are well-versed in the use of numerous social media channels. These web platforms allow any user to publish content or disseminate news. Neither the individuals nor their content is authenticated on these networks. Individuals occasionally utilise these mediums to disseminate deceptive information. The exponential spread of fabricated information has emerged as a crucial concern in the contemporary culture that heavily relies on information. The spread of deceptive and inaccurate information presents significant concerns to public perception, democratic procedures, and the overall welfare of society. In response, researchers have progressively relied on machine learning (ML) algorithms to create automated systems for identifying fraudulent news. Motivated by this, we propose to create a robust and efficient system that can accurately detect and counteract the spread of misinformation. In this study, we employed eight different ML models to detect fake news, and their performance is assessed on two real-world datasets obtained from Kaggle. For feature extraction, these models utilize the Term Frequency-Inverted Document Frequency (TF-IDF) technique. Among the other eight models, the Passive Aggressive Classifier (PAC) demonstrated the highest performance, achieving an average accuracy of 97.26%.

Keywords: Decision Tree, Fake News Detection, SVM, Linear Regression, PAC

INTRODUCTION

In the current digital age, the conflict between truth and deception has grown increasingly complex, fueled by the vast and often overwhelming volume of information available [1]. The rise of social network services within today's cyber-physical environment has played a pivotal role in the aggregation and dissemination of data [2]. However, this has also led to the rampant spread of misinformation, disinformation, and fabricated content, commonly referred to as fake news, which has emerged as a critical issue in recent years. While the digital era offers undeniable advantages, it is not without its challenges. Among the many concerns associated with digitalization is the proliferation of deceptive or false news. The internet serves as a valuable tool for individuals seeking information [3], but the rapid growth of online platforms like Facebook and Twitter, combined with the effects of globalization, has created an unprecedented medium for information exchange [4]. These platforms are favored by users due to their accessibility and ease of use, making them a popular choice for news consumption. Additionally, they provide opportunities for public engagement through comments, reactions, and other interactive features. Despite these benefits, such platforms have also become a breeding ground for malicious actors. Hackers and unethical individuals exploit these networks to distribute false information, often leveraging the ability to share posts or news articles to amplify the reach of misinformation. The deliberate circulation of counterfeit or misleading content disguised as legitimate news has become a widespread and concerning trend, aimed at confusing and manipulating audiences. This phenomenon poses a significant threat to media integrity, public discourse, and the foundations of democratic processes.
The rapid spread of misinformation has made it difficult to tell the difference between real and fake information. The ease of information sharing on social networking platforms has made it challenging for individuals to find trustworthy sources. This proliferation of easily shared false information has bad consequences for individuals, organizations, and society [5]. With the shift towards online platforms, including social media, search engines, and digital news outlets, as the main places people get information, the possible effect of made-up stories is bigger. Sharing these stories can change what people think, affect election results, and cause social problems. Also, it can hurt reputations and add to confusion and anger, especially during important events [6]. False information can also weaken public trust in news sources and the whole news system, which are important for people to know what's going on in a democracy [7]. So, finding fake news is very important to lessen its bad effects [8], as the information shown greatly affects how people see things and make decisions [9]. Recent events have shown how people can react irrationally to news that was later shown to be wrong. For example, during the COVID-19

* Corresponding author

Submitted : Mar 24, 2025
Accepted : Apr 04, 2025
Published : Apr 04, 2025

pandemic, lots of wrong information about the virus's beginnings, features, and biology spread online [10]. The 2016 US election also showed how much fake news can affect what people think and their general ideas [11]. Machine learning (ML) has emerged as a critical field, developing mathematical algorithms and statistical models that enable computers to learn from data without explicit programming. The efficacy of ML hinges on these algorithms, which provide the computational framework for automated learning, adaptation, and progressive improvement. Researchers have extensively explored the application of ML in detecting misinformation. For instance, [8] employed data mining techniques to identify instances of fabricated information disseminated through social media platforms. Their work emphasized the importance of analyzing network topology, user activity patterns, and content attributes for the identification of deceptive data. In another study, [12] performed an analysis of fabricated news articles, demonstrating the effectiveness of examining syntactic and lexical features to differentiate reliable from unreliable sources. Furthermore, numerous ML models have been constructed for fake news classification, exhibiting a wide range of potential applications [13].

Machine learning has emerged as a highly effective approach for detecting fake news. By analyzing large datasets from social media, news articles, and other information sources, ML algorithms are trained to identify patterns and distinguishing features that help differentiate between authentic and deceptive content. The task of fake news detection has gained significant attention, with ML playing a crucial role due to its ability to learn from data and enhance detection accuracy [14]. ML-based techniques facilitate the development of advanced systems capable of identifying false information by leveraging sophisticated methods to assess both textual and contextual elements. Through computational analysis, ML algorithms examine textual content, recognize patterns, and detect subtle linguistic cues that contribute to authenticity assessment. By training on extensive datasets containing both real and fabricated news articles, ML models can discern linguistic structures and writing styles commonly associated with misleading content [15]. Natural Language Processing (NLP) techniques are widely employed to process and analyze text, incorporating aspects such as sentiment analysis, phrase structure, and word usage. ML algorithms can automatically extract textual features, including word frequency, syntactic structures, and grammatical patterns, which are then used to train models capable of identifying deceptive news. Furthermore, ML models can assess the coherence of information against verified facts and real-world events, allowing them to detect contradictions or inconsistencies. Supervised learning approaches enable training on labeled datasets that classify news stories as either true or false [16]. In scenarios where labeled data is unavailable, unsupervised learning techniques, such as clustering and anomaly detection, can be utilized to uncover suspicious patterns indicative of misinformation [17].

False information, whether deliberately created to mislead or unintentionally propagated, can lead to severe societal consequences, such as eroding public trust, manipulating public sentiment, and interfering with democratic processes. To effectively address the challenge of fake news detection, it is essential to develop automated systems capable of accurately distinguishing between credible and deceptive sources of information. Numerous researchers have explored the application of ML techniques in mitigating the spread of misinformation by analyzing both textual and contextual factors. The work of [18] proposed a fake news detection model that utilized the Term Frequency-Inverse Document Frequency (TF-IDF) method for feature extraction, employing a Support Vector Machine (SVM) classifier as the primary learning model. Similarly, TF-IDF was also implemented in the study conducted by [5], but instead of relying on a single classifier, the researchers evaluated multiple classification algorithms, including the Passive Aggressive Classifier (PAC), Naïve Bayes (NB), and SVM, to determine the most effective approach. In another study, [19] trained and tested a model using four different classifiers: Random Forest (RF), NB, Logistic Regression (LR), and PAC. A more recent study [20] compared six ML algorithms such as Linear Regression (LR), Linear Support Vector Machine (LSVM), Decision Tree (DT), SVM, Stochastic Gradient Descent (SGD), and K-Nearest Neighbors (KNN) to assess their performance in detecting fake news.

This study employs several ML techniques such as NB, DT, LR, RF, PA, KNN, GB, and SVM to determine which model obtains the most reliable and efficient results when recognizing fake news posts and misleading material circulating online on social media. To determine which algorithm is most effective at detecting false news, this work compares their performance in terms of standard evaluation metrics such as accuracy, precision, recall and F1-score. The models are trained and evaluated with two distinct datasets containing both fake and true news. This research also compares the average accuracy of the model with these two datasets. Furthermore, we compare the results with current cutting-edge techniques. The following are the research objectives.

- To perform a thorough comparison of different ML algorithms in terms of their effectiveness in detecting fake news. This objective aims to systematically assess the accuracies of different algorithms in differentiating between true and false news content in order to highlight their strengths and limitations.

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 2, April 2025
https://doi.org/10.47709/cnahpc.v7i2.5717

Submitted : Mar 24, 2025
Accepted : Apr 04, 2025
Published : Apr 04, 2025

- To identify the most suitable and efficient ML algorithm for identifying falsified news by conducting a comprehensive assessment of precision, recall, accuracy, and F1-score. This objective aims to find an algorithm that achieves overall good performance by balancing precision and recall.
- To examine and compare the efficacy of different ML algorithms on two separate datasets for identifying fake news. This objective seeks to determine if algorithms consistently outperform others across different data contexts or if their performance depends on particular data attributes. The process involves running the same set of algorithms on multiple datasets and comparing the outcomes.

This article proceeds as follows: Section 2 evaluates prior research and comparative frameworks. Section 3 outlines the methodological approach. Section 4 interprets experimental outcomes and performance metrics. Section 5 contextualizes the study's contributions and limitations. Finally, Section 6 synthesizes key conclusions and proposes future directions.

LITERATURE REVIEW

The widespread dissemination of misleading content and fabricated narratives has become a critical societal challenge in today's digitally driven landscape. Academic investigations have rigorously explored the ramifications of misinformation on contemporary communities, with studies demonstrating that repeated exposure to deceptive content can skew cognitive interpretations of factual events and undermine rational decision-making capabilities [21]. This phenomenon has intensified concerns among policymakers, researchers, and citizens regarding its capacity to manipulate collective beliefs, electoral outcomes, and behavioral norms. ML has emerged as a pivotal tool for countering disinformation, owing to its scalability in analyzing large-scale data patterns. Significant research efforts have focused on deploying ML techniques for automated deception detection in media. The following discussion critically evaluates existing literature on ML-driven fake news identification frameworks. It synthesizes recent advancements, including comparative assessments of algorithmic performance, optimization of models via feature engineering, and innovations in hybrid ensemble architectures and deep learning methodologies. This review further highlights specialized approaches for enhancing detection accuracy through multimodal data integration and adaptive learning paradigms.

The study by [22] introduces a framework for disinformation detection, integrating linguistic analysis NLP with ML classifiers such as Naive Bayes (NB) and SVM. Experimental results were benchmarked against existing solutions using a curated dataset of RSS feeds compiled from aggregated news sources. The SVM-driven model outperformed comparative systems, achieving 93.6% classification accuracy, and was supplemented by a corrective mechanism that suggests verified alternative articles when misinformation is detected. This work underscores the complexity of disinformation mitigation, advocating for hybrid methodologies that combine feature engineering with algorithmic synergy. Complementing this, [5] conducted a systematic evaluation of classifier efficacy in fake news identification. Their analysis identified ensemble architectures combining linguistic features and behavioral metadata as particularly effective for improving detection robustness. The study emphasizes iterative model refinement and data diversity as critical factors for optimizing accuracy in dynamic information ecosystems. This study included three distinct classification techniques, namely SVM, PA, and NB. This research utilized a news dataset with a size of (63354). Among the methods tested, SVM demonstrated the highest level of accuracy, with a precision of 95.05%. However, it is worth noting that SVM requires a longer processing time compared to PA or NB. Other ML algorithms can be employed to enhance the outcomes as a progression in their task. Specifically for the Bengali language, [23] investigates how well various ML systems can spot fabricated news stories. The study accomplishes this by developing a unique dataset specifically for the Bengali language. They used the Gaussian Naïve Bayes model and obtained an accuracy rate of 87.4%. Through the utilization of more refined data, their objective is to construct a model that exhibits enhanced precision and efficiency in the identification of features.

Several types of ML techniques and strategies for spotting false news have been explored in a recent work [24]. The study also addressed the difficulties linked to identifying false news and emphasized the significance of recognizing fake news. Acquired the dataset on fake news from Kaggle for the research. The DT algorithm outperforms all other ML techniques with an accuracy of 99.36%. Future research directions include integrating additional classifiers and refining feature engineering to improve model robustness in identifying unreliable content. In [25], the efficacy of five ML classifiers such as LR, RF, GB, NB, and SVM were evaluated for binary news classification (authentic vs. deceptive). Performance analysis, based on classification accuracy metrics, revealed RF and GB as superior in handling textual and contextual features. The article [26] proposed a disinformation detection framework that benchmarks ML approaches (RF, LR, NB, SVM, deep neural networks/DNN) across computational efficiency, accuracy, and resource utilization. Using a news dataset derived from web-based news platforms, their DNN model achieved 91% accuracy with optimized inference times,

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 2, April 2025
https://doi.org/10.47709/cnahpc.v7i2.5717

Submitted : Mar 24, 2025
Accepted : Apr 04, 2025
Published : Apr 04, 2025

outperforming traditional methods. Future extensions of this work could explore hybrid architectures or adaptive training protocols. These studies collectively highlight the importance of algorithmic adaptability and computational efficiency in scalable fake news detection systems.

Improvement of ML strategies with feature extraction methods

This section emphasizes research that employed particular feature extraction techniques to improve the efficiency of ML algorithms.

The study published by [18] developed a framework that uses n-gram analysis and term frequency-inverse document frequency (TF-IDF) to identify fake news. They employ a bag of words as a technique for extracting features and employ SVM as the classifier. In their strategy, a software tool known as WEKA and SMO library is used to enhance the decision model and achieve improved precision. The researchers determined that the most influential parameters for the SVM were Cost C, gamma γ, and epsilon ϵ. The findings collected demonstrated the efficacy of the proposed framework. There is potential for conducting the same study using a larger dataset in the future. In a similar direction, [27] concentrated on constructing an ML model that can detect false information. They employed PA and NLP approaches, such as TF-IDF feature extraction. The ML club gathered the UTK dataset from Kaggle for this purpose. The study further demonstrated enhanced model performance through TF-IDF-based classification, achieving 96% classification accuracy. [28] proposed a multimodal framework integrating TF-IDF vectorization, count vectorization, and word embedding techniques to enable granular linguistic analysis. Using the Kaggle fake news corpus, they implemented an SVM classifier with TF-IDF feature engineering, attaining 94% accuracy. Recognizing the limitations of conventional ML approaches, the authors augmented their architecture with DNNs to exploit hierarchical feature learning, further elevating detection precision. This hybrid methodology underscores the value of combining feature diversity with computational linguistics for robust disinformation analysis.

Ensemble and advanced ML models

The research by [29] explored the viability of KNN and quantum KNN algorithms for disinformation detection, leveraging the Buzzface dataset for empirical validation. To optimize model efficacy, the team integrated Genetic and Evolutionary Feature Selection (GEFeS), a metaheuristic technique for dimensionality reduction, which enhanced KNN's discriminative power to achieve 91.3% accuracy. While classical KNN demonstrated robust performance, the study highlighted unexplored opportunities in quantum-inspired machine learning for parallelized pattern recognition in high-dimensional data spaces. These findings underscore the necessity for expanded experimentation with quantum-classical hybrid architectures to advance computational efficiency in large-scale misinformation analysis.

Research by [30] evaluated DT and KNN alongside established ensemble classifiers such as RF and GB. To enhance detection capabilities, they designed novel hybrid architectures, including stacked generalization and majority voting classifiers, for binary news classification. Using temporally stratified datasets (2016–2017) containing both factual and deceptive articles, their composite framework integrating NB, SVM, and LR achieved 91.5% accuracy. In [34], the authors introduced a credibility assessment framework that evaluates semantic coherence and contextual plausibility to flag disinformation. Comparative analysis of classifiers (NB, RNN, SVM, KNN, LSTM) revealed long short-term memory (LSTM) networks as superior, attaining 97% accuracy through sequential pattern recognition. [31] proposed an NLP-driven pipeline for high-stakes news verification, leveraging the synthesized WEL Fake dataset. Through eight experimental configurations, SVM achieved peak performance with 98% accuracy, outperforming RF and baseline models. This work underscores the scalability of DL in addressing disinformation when paired with curated, multimodal datasets, emphasizing the need for adaptive architectures in evolving information landscapes.

The work by [32] introduces a multimodal framework for detecting medical disinformation, integrating ten machine learning algorithms with seven feature extraction methods spanning lexical, structural, and domain-specific attributes. Training and evaluation were performed on a verified ground-truth dataset of health-related articles, with reliability ensured through stratified 5-fold cross-validation. This approach demonstrated consistent performance metrics such as precision of 92%, and recall of 89% across validation splits, emphasizing the role of multi-algorithm benchmarking and domain-aware feature engineering in combating health misinformation. The study advocates for adaptive validation protocols and computational epidemiology techniques to minimize algorithmic bias in high-stakes applications. LR demonstrated superior performance compared to the other methods, achieving an accuracy rate of 99.87%. As part of their plans, they contemplated developing a framework that encompasses data in languages other than English. The work authored by [33] conducted a comprehensive analysis of ML models, including LSVM, SVM, RNN, CNN, KNN, Naïve Bayes, DT, Stochastic Gradient Descent, and LR. The study

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 2, April 2025
https://doi.org/10.47709/cnahpc.v7i2.5717

Submitted : Mar 24, 2025
Accepted : Apr 04, 2025
Published : Apr 04, 2025

found that CNN achieved a superior accuracy rate of 99.8%. Their primary objective was to detect and classify misinformation utilizing ML and AI methodologies.

Comparative analysis

Ongoing research on detecting false news involves a thorough examination of ML techniques that are specially developed to improve accuracy. The authors [5] utilized a simple methodology and performed a comparative analysis of three classification techniques. The researchers found that SVM provided the highest level of accuracy, although it necessitated a longer computing time. In contrast, [22] employed a comprehensive methodology that integrated many attributes and ML techniques, achieving an accuracy of 93.6%, comparable to that of SVM. The linguistic particularities of the Bengali language were explored by [23], which resulted in a somewhat lower accuracy of 87% when using GNB.

The studies conducted by [24, 29] achieved accuracy rates over the 90% threshold, although employing distinct methodologies. [29] Combined quantum algorithms with Genetic and Evolutionary Feature Selection to reach an accuracy of 91.3% utilizing the KNN method. In contrast, [24]utilized a DT strategy and achieved an impressive accuracy of 99.36%. This exemplifies the diverse range of methodologies at one's disposal, each offering distinct benefits. [31] Highlighted the significance of having a diverse dataset and achieved significantly better results than Gupta et al. by an impressive proportion of 98% utilizing SVM. Nevertheless, the meticulous approach employed by [18], which focuses on optimizing parameters in SVM, highlights that the choice of algorithm does not solely determine the effectiveness of the outcomes but also the extent of algorithm refinement. The study conducted by [30] examined ensemble methodologies that combine the strengths of many algorithms. The results, with an accuracy rate of 91.5%, demonstrate the promise of these hybrid models. This aligns with the emphasis of [27] on feature extraction approaches, in which their model highlights the importance of obtaining high-quality data representation to achieve successful outcomes.

The studies conducted by [25, 33] explored various ML techniques. However, they obtained contrasting results, with [33] using CNN and [25] employing GB and RF. This suggests that the efficacy of the algorithm may be dependent upon the context. [34] Examined the LSTM and demonstrated its dynamic nature in the field by employing advanced DL techniques. The significance of domain specificity and data representation is emphasized by the health-specific approach of [32] and the focus on feature extraction approaches by [28]. Both had impressive accuracies exceeding 90%, indicating that employing specialized tactics in specific contexts has advantageous outcomes. To summarize, the comparison study demonstrates that the different strategies employed in the research all strive for reasonable accuracy in detecting false news. However, it is crucial to strike the right balance between selecting the appropriate algorithm, extracting relevant features, and ensuring the uniqueness of the dataset. The accuracy range, spanning from 87% to over 99%, underscores the intricacies and challenges inherent in this ever-evolving field.

Critical assessment

The investigator's collective attempts in the domain of ML to detect false news exhibit an impressive dedication to tackling a critical issue in the present information ecosystem. Their research employs a diverse array of algorithms, datasets, and methodologies, all aimed at enhancing the precision of fake news identification. While these attempts have shown promising results, some crucial issues demand consideration. The utilization of several ML approaches, such as NB, DT, and SVM, has exhibited exceptional levels of accuracy, indicating the possibility of practical implementation. The research often focuses on specific algorithms, perhaps overlooking the benefits of incorporating alternative ways to enhance resilience. Moreover, the model's capacity to be applied to real-world occurrences may be limited due to its dependence on specific datasets, which could introduce biases. To improve the effectiveness and dependability of machine learning models in detecting fake news. Future research might explore hybrid methodology, diverse datasets, and robust verification techniques on a broader scale.

Notable scholars have conducted in-depth research that has provided valuable insights closely related to the purpose of our proposed technique; examples are [35]. We are even more dedicated to developing precise ML models after reading their in-depth analysis and evaluation of numerous ML techniques, such as SVM and DT, for the detection of false information. The focal point of our research revolves around achieving enhanced accuracy rates, which is notably prominent in all articles. These experiments showcase the feasibility and importance of our desired objective by employing a range of ML approaches to classify news stories as either genuine or fraudulent. Their research establishes a robust structure for our attempts to enhance the precision of ML-based false news detection.

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 2, April 2025
https://doi.org/10.47709/cnahpc.v7i2.5717

Submitted : Mar 24, 2025
Accepted : Apr 04, 2025
Published : Apr 04, 2025

METHODOLOGY

In this section, we explain the procedures that were used to put the proposed approach into practice. It includes both the methods for gathering and preparing the data, as well as the explanations of the models used for analysis and prediction.

Data Collection

As information spreads rapidly through digital channels, the ability to identify and challenge fake news has assumed heightened significance, necessitating a comprehensive exploration within the realm of data-driven solutions; we set out to choose an appropriate dataset from the Kaggle community site. We made our pick based on numerous crucial factors, each of which had a substantial impact on our decision-making process.

The size and quality of a dataset

The quantity and excellence of the dataset are crucial elements in the selection process. We were seeking a dataset including a substantial quantity of diverse news stories encompassing a wide range of themes and originating from numerous sources. Due to the wide variety of writing styles, tonalities, and subject areas, the model will have endured training, allowing it to distinguish between genuine and fabricated news more effectively.

Balanced distribution

A proper balance between authentic and fake news examples is essential. If the distribution is skewed, the model may perform well for the majority group while ignoring essential details for the minority [16]. A balanced dataset will ensure that the model is not biased and can accurately classify both forms of news.

Review and Community Feedback

Our methodology incorporated systematic reviews of peer assessments and experiential data from prior studies using the dataset. These insights into the dataset's capabilities, constraints, and operational risks informed our feasibility analysis for achieving project goals. Collaborative knowledge-sharing with domain specialists enhanced our evaluation framework for dataset alignment with investigative requirements. Following rigorous evaluation of these parameters, we developed two Kaggle-hosted repositories optimized for disinformation detection systems. These datasets balance technical robustness for misinformation analysis with practical utility to support operational objectives in countering deceptive content, validated through iterative testing and cross-disciplinary peer validation. We selected the Fake and true news dataset [20] as our first dataset, and the second dataset is the fake News and misinformation text data sets. The initial dataset comprises 44,898 records, which are then categorized into two separate datasets: true and fake. The dataset of fake news has 23,481 entries, whereas the dataset of true news consists of 21,417 records. Both examples possess four attributes: Subject, Title, Date, and Text. Table 1 displays the data pertaining to the first dataset. Figure 1 displays the features of the false news dataset in DS1, whereas Figure 2 presents the details of the true news dataset in DS1.

Table 1: First dataset

News type	Dataset entries
True news	21417
Fake news	23481
Total news	44898

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 2, April 2025
https://doi.org/10.47709/cnahpc.v7i2.5717

Submitted : Mar 24, 2025
Accepted : Apr 04, 2025
Published : Apr 04, 2025

title	text	subject	date
Donald Tr	Donald Tru	News	December 31, 2017
Drunk Bra	House Inte	News	December 31, 2017
Sheriff Da	On Friday,	News	December 30, 2017
Trump Is S	On Christm	News	December 29, 2017
Pope Fran	Pope Franc	News	December 25, 2017
Racist Ala	The numbe	News	December 25, 2017
Fresh Off	Donald Tru	News	December 23, 2017
Trump Sai	In the wak	News	December 23, 2017
Former CI	Many peo	News	December 22, 2017
WATCH: E	Just when	News	December 21, 2017
Papa John	A centerpi	News	December 21, 2017
WATCH: P	Republicar	News	December 21, 2017
Bad News	Republicar	News	December 21, 2017
WATCH: L	The media	News	December 20, 2017
Heiress Tc	Abigail Dis	News	December 20, 2017
Tone Dea	Donald Tru	News	December 20, 2017
The Interr	A new anir	News	December 19, 2017
Mueller Sp	Trump sup	News	December 17, 2017
SNL Hilari	Right now,	News	December 17, 2017
Republica	Senate Ma	News	December 16, 2017
In A Heart	It almost s	News	December 16, 2017
KY GOP St	In this #ME	News	December 13, 2017

Figure 1: Fake news data of dataset 1

title	text	subject	date
As U.S. bu	WASHING	politicsNe	December 31, 2017
U.S. milita	WASHING	politicsNe	December 29, 2017
Senior U.S	WASHING	politicsNe	December 31, 2017
FBI Russia	WASHING	politicsNe	December 30, 2017
Trump wa	SEATTLE/V	politicsNe	December 29, 2017
White Hou	WEST PALI	politicsNe	December 29, 2017
Trump say	WEST PALI	politicsNe	December 29, 2017
Factbox: T	The follow	politicsNe	December 29, 2017
Trump on	The follow	politicsNe	December 29, 2017
Alabama o	WASHING	politicsNe	December 28, 2017
Jones certi	(Reuters) -	politicsNe	December 28, 2017
New York	NEW YORI	politicsNe	December 28, 2017
Factbox: T	The follow	politicsNe	December 28, 2017
Trump on	The follow	politicsNe	December 28, 2017
Man says I	(In Dec. 2!	politicsNe	December 25, 2017
Virginia of	(Reuters) -	politicsNe	December 27, 2017
U.S. lawm	WASHING	politicsNe	December 27, 2017
Trump on	The follow	politicsNe	December 26, 2017
U.S. appea	(Reuters) -	politicsNe	December 26, 2017
Treasury S	(Reuters) -	politicsNe	December 24, 2017
Federal juc	WASHING	politicsNe	December 24, 2017
Exclusive:	NEW YORI	politicsNe	December 23, 2017
Trump tra	(Reuters) -	politicsNe	December 23, 2017
Second co	WASHING	politicsNe	December 23, 2017

Figure 2: True news data of dataset 1

There are total of 79k records in the second dataset, and they are split between two groups: DatasetMisinfoFAKE, which has 43642 records, and DatasetMisinfoTRUE, which has 34975 records. There are two characteristics present in both examples: an index and a body of text. The data from the second set are listed in Table 2. Figure 3 and Figure 4 depict specifics of the fabricated and authentic news datasets used in DS2.

Table 2: Second dataset

News Type	Entries
True News	34975
Fake News	43642
Total News	78617

	text
0	Donald Trump just couldn t wish all Americans a Happy New Year and leave it at that. Instead, he had to give a shout out to his enemies, haters and the very dishonest fake news media. The former reality show star had just one job to do and he
1	House Intelligence Committee Chairman Devin Nunes is going to have a bad day. He s been under the assumption, like many of us, that the Christopher Steele-dossier was what prompted the Russia investigation so he s been lashing out at the De
2	On Friday, it was revealed that former Milwaukee Sheriff David Clarke, who was being considered for Homeland Security Secretary in Donald Trump s administration, has an email scandal of his own.In January, there was a brief run-in on a plane
3	On Christmas day, Donald Trump announced that he would be back to work the following day, but he is golfing for the fourth day in a row. The former reality show star blasted former President Barack Obama for playing golf and now Trump is o
4	Pope Francis used his annual Christmas Day message to rebuke Donald Trump without even mentioning his name. The Pope delivered his message just days after members of the United Nations condemned Trump s move to recognize Jerusalem a
5	The number of cases of cops brutalizing and killing people of color seems to see no end. Now, we have another case that needs to be shared far and wide. An Alabama woman by the name of Angela Williams shared a graphic photo of her son, lyi
6	Donald Trump spent a good portion of his day at his golf club, marking the 84th day he s done so since taking the oath of office. It must have been a bad game because just after that, Trump lashed out at FBI Deputy Director Andrew McCabe on T
7	In the wake of yet another court decision that derailed Donald Trump s plan to bar Muslims from entering the United States, the New York Times published a report on Saturday morning detailing the president s frustration at not getting his way a
8	Many people have raised the alarm regarding the fact that Donald Trump is dangerously close to becoming an autocrat. The thing is, democracies become autocracies right under the people s noses, because they can often look like democracies i
9	Just when you might have thought we d get a break from watching people kiss Donald Trump s ass and stroke his ego ad nauseam, a pro-Trump group creates an ad that s nothing but people doing even more of those exact things. America First Po
10	A centerpiece of Donald Trump s campaign, and now his presidency, has been his white supremacist ways. That is why so many of the public feuds he gets into involve people of color. One of his favorite targets, is, of course, the players in the Nat
11	Republicans are working overtime trying to sell their scam of a tax bill to the public as something that directly targets middle-class and working-class families with financial relief. Nothing could be further from the truth, and they re getting hamme
12	Republicans have had seven years to come up with a viable replacement for Obamacare but they failed miserably. After taking a victory lap for gifting the wealthy with a tax break on Wednesday, Donald Trump looked at the cameras and said, W
13	The media has been talking all day about Trump and the Republican Party s scam of a tax bill; as well as the sheer obsequiousness of Trump s cabinet, and then members of Congress, after their tax scam was all but passed. But the media isn t quit
14	Abigail Disney is an heiress with brass ovaries who will profit from the GOP tax scam bill but isn t into f-cking poor people over. Ms. Disney penned an op-ed for USA Today in which she rips the GOP a new one because she has always been cogniz
15	Donald Trump just signed the GOP tax scam into law. Of course, that meant that he invited all of his craven, cruel GOP sycophants down from their perches on Capitol Hill to celebrate in the Rose Garden at the White House. Now, that part is bad
16	A new animatronic figure in the Hall of Presidents at Walt Disney World was added, where every former leader of the republic is depicted in an audio-animatronics show. The figure which supposedly resembles Jon Voight Donald Trump was add
17	Trump supporters and the so-called president s favorite network are lashing out at special counsel Robert Mueller and the FBI. The White House is in panic-mode after Mueller obtained tens of thousands of transition team emails as part of the R
18	Right now, the whole world is looking at the shocking fact that Democrat Doug Jones beat Republican Roy Moore in the special election to replace Attorney General Jeff Sessions in the United States Senate. Of course, Moore s candidacy was roc
19	Senate Majority Whip John Cornyn (R-TX) thought it would be a good idea to attack Special Counsel Robert Mueller over the Russia probe. As Mueller s noose tightens, Republicans are losing their sh-t and attacking Mueller and the FBI in order to
20	It almost seems like Donald Trump is trolling America at this point. In the beginning, when he tried to gaslight the country by insisting that the crowd at his inauguration was the biggest ever or that it was even close to the last couple of inaugurat
21	In this #METOO moment, many powerful men are being toppled. It spans many industries, from entertainment, to journalism, to politics and beyond. Any man that ever dared to abuse his power to sexually harass, molest, or assault women bette

Figure 3: Fake news data of dataset 2

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 2, April 2025
https://doi.org/10.47709/cnahpc.v7i2.5717

Submitted : Mar 24, 2025
Accepted : Apr 04, 2025
Published : Apr 04, 2025

	text
0	The head of a conservative Republican faction in the U.S. Congress, who voted this month for a huge expansion of the national debt to pay for tax cuts, called himself a â€œfiscal conservativeâ€ on Sunday and urged budget restraint in 2018. In ke
1	Transgender people will be allowed for the first time to enlist in the U.S. military starting on Monday as ordered by federal courts, the Pentagon said on Friday, after President Donald Trumpâ€™s administration decided not to appeal rulings that b
2	The special counsel investigation of links between Russia and President Trumpâ€™s 2016 election campaign should continue without interference in 2018, despite calls from some Trump administration allies and Republican lawmakers to shut it d
3	Trump campaign adviser George Papadopoulos told an Australian diplomat in May 2016 that Russia had political dirt on Democratic presidential candidate Hillary Clinton, the New York Times reported on Saturday. The conversation between Papa
4	President Donald Trump called on the U.S. Postal Service on Friday to charge â€œmuch moreâ€ to ship packages for Amazon (AMZN.O), picking another fight with an online retail giant he has criticized in the past. â€œWhy is the United States P
5	The White House said on Friday it was set to kick off talks next week with Republican and Democratic congressional leaders on immigration policy, government spending and other issues that need to be wrapped up early in the new year. The expe
6	President Donald Trump said on Thursday he believes he will be fairly treated in a special counsel investigation into Russian meddling in the U.S. presidential election, but said he did not know how long the probe would last. The federal investigatio
7	While the Fake News loves to talk about my so-called low approval rating, @foxandfriends just showed that my rating on Dec. 28, 2017, was approximately the same as President Obama on Dec. 28, 2009, which was 47%...and this despite massiv
8	Together, we are MAKING AMERICA GREAT AGAIN! bit.ly/2lnpKaq [1814 EST]
9	Alabama Secretary of State John Merrill said he will certify Democratic Senator-elect Doug Jones as winner on Thursday despite opponent Roy Mooreâ€™s challenge, in a phone call on CNN. Moore, a conservative who had faced allegations of gr
10	Alabama officials on Thursday certified Democrat Doug Jones the winner of the stateâ€™s U.S. Senate race, after a state judge denied a challenge by Republican Roy Moore, whose campaign was derailed by accusations of sexual misconduct with
11	The new U.S. tax code targets high-tax states and may be unconstitutional, New York Governor Andrew Cuomo said on Thursday, saying that the bill may violate New York residentsâ€™ rights to due process and equal protection. The sweeping R
12	Vanity Fair, which looks like it is on its last legs, is bending over backwards in apologizing for the minor hit they took at Crooked H. Anna Wintour, who was all set to be Amb to Court of St Jamesâ€™s & a big fundraiser for CH, is beside herself in gr
13	â€œOn 1/20
14	A man claiming to be the person who delivered a gift-wrapped package of horse manure at the Los Angeles home of U.S. Treasury Secretary Steven Mnuchin said on Monday he did it to protest the federal tax overhaul signed into law last week by
15	A lottery drawing to settle a tied Virginia legislative race that could shift the statehouse balance of power has been indefinitely postponed, state election officials said on Tuesday, after the Democratic candidate mounted a legal fight. The decisio
16	A Georgian-American businessman who met then-Miss Universe pageant owner Donald Trump in 2013, has been questioned by congressional investigators about whether he helped organize a meeting between Russians and Trumpâ€™s eldest so
17	Based on the fact that the very unfair and unpopular Individual Mandate has been terminated as part of our Tax Cut Bill, which essentially Repeals (over time) ObamaCare, the Democrats & Republicans will eventually come together and develop a

Figure 4: True news data of dataset 2

Dataset Pre-processing
Due to the quick dissemination of false information in the internet age, reliable methods of identifying fake news are more important than ever. Only high-quality data must be fed into these systems. The Pre-processing of the dataset consisted of the following procedures.

Merging and labelling of the data
We added a column labelled "class," where "class 0" indicates false news and "class 1" indicates true news in both of our datasets. Samples of fake and accurate news from both sets were combined to create a third set. To get more valuable insights from the predictive model, the data must first be cleansed. The groundwork for our essential pre-processing operations is our newly combined dataset.

The process of data cleaning
A dataset may contain information that is wholly or partially structured, semi-structured, or unstructured. Therefore, data cleaning is crucial for ensuring the reliability, precision, and applicability of test and training data in the ML model. This will improve the quality of the data and get rid of any noise that could prevent the model from correctly identifying bogus news. After we combined the dataset, we eliminated columns like subject, title, and date because we just needed text and class columns for our training. All punctuation, URLs, HTML tags, and newline characters were eliminated from the dataset. The text was converted to lowercase for consistent processing, any text enclosed in square brackets that might have contained metadata or citations was eliminated, alphanumeric strings containing numbers that might have been identifiers or other irrelevant information was eliminated, and any non-word character that separated words was replaced with a space. The cleaned-up text was then stored for later use in the research. The second dataset also underwent these meticulous cleaning procedures to ensure standardization.

The Removal of Stopwords
A "stopword" is a word or phrase that has been deemed unnecessary for a given text analysis activity. The NLTK library allows us to filter out these terms by downloading a list of stopwords. Common words like "the," "is," "and," "in," etc. are called "stopwords," and they are taken out of our text data because they are not beneficial to our text analysis. In our situation, we have processed each text to eliminate all instances of stopwords in English and replaced them with the original content.

Data splitting
It is a method which is used to generate distinct subsets from a given dataset, which are then used for testing and training purposes. Seventy-five per cent of each dataset is used for training, while the remaining 25 per cent is used for evaluation in both datasets. The model's learning and generalization capabilities improve significantly when it is trained on a larger dataset. As opposed to being used during model training, the test dataset is only used to evaluate how well the model performs on entirely new data. The quantity of the dataset, the complexity of the model, and the existence of class imbalances are only a few of the factors that influence the decision of the data-splitting ratio. The 75:25 split was chosen for this research because it strikes an appropriate balance between training the model thoroughly and keeping the evaluation reliable.

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 2, April 2025
https://doi.org/10.47709/cnahpc.v7i2.5717

Submitted : Mar 24, 2025
Accepted : Apr 04, 2025
Published : Apr 04, 2025

Feature extraction
We used the TF-IDF vectorizer technique for feature extraction. TF-IDF enables the conversion of text data into a numerical format, allowing for the identification of unique characteristics in each document. This numerical representation can then be utilized as input for a variety of ML methods. TF is a statistical measure of how often a certain word appears in a given text. A higher TF value indicates that the document places greater emphasis on that word. The IDF evaluates how distinctive a term is across the entire corpus. The ratio of total papers to documents containing the word in question is used to get this conclusion.

$$\mathrm{TF} - \mathrm{IDF} = \mathrm{TF} \times \mathrm{IDF} \quad (1)$$

The final score indicates how significant a term is within a specific text in comparison to how often it occurs across the entire corpus. The ML model is trained using the TF-IDF vectors to identify associations between words and, fake news and real news labels.

Machine learning algorithms
Online misinformation can be identified using various ML algorithms. We use the algorithms NB, DT, RF, GB, KNN, LR, PA, and SVM for additional analysis and predictions.
Decision tree: The Decision Tree (DT) classifier is a widely used ML technique known for its high accuracy and precision in classification tasks [36]. Our research findings indicate that DT effectively identifies key factors influencing classification, such as distinctive word patterns or combinations that differentiate real from fake news. Its adaptability to both numerical and categorical data, along with its ability to handle missing values, makes it well-suited for processing the diverse information found in news articles
Logistic Regression: Logistic Regression (LR) is a widely used approach in both statistical analysis and ML for binary classification problems [37]. In this study, LR is employed due to its ability to learn optimal feature weights, enabling it to determine whether a given piece of content is authentic or not. The model estimates the probability of an instance belonging to a particular class, assigning a score between 0 and 1 using the logistic function.
K-Nearest Neighbours: Among the ML toolkits, the KNN is the most basic and least intricate classifier. In this method, training is done using a labelled dataset of news articles that have been manually labelled as fraudulent or genuine. Article features, such as text, are extracted from the dataset during pre-processing. In the prediction stage, KNN finds the unseen news articles KNN using a distance metric of choice and then labels the article according to the distribution of its neighbour classes.
Random Forest: The RF classifier, similar to the DT classifier, is a popular tree-based method extensively employed for classification purposes. The ensemble approach is a technique that enhances accuracy and robustness by combining numerous decision trees [38]. The RF model exhibits a lower error rate compared to other models due to its low tree correlation, as demonstrated by [39]. The problems it has solved in the past with complicated data structures and nonlinear relationships are a good fit for the ones we are working on.
Support Vector Machine: The SVM classifier achieves excellent performance in binary classification by utilizing support vectors as anchor points to identify an optimal hyperplane that maximizes the separation between classes [40]. The SVM algorithm is highly proficient at discerning between various categories of news stories, making it a valuable tool for our research. Moreover, SVM exhibits the capability to effectively process intricate textual patterns by utilizing kernel functions, enabling it to stand out in scrutinizing linguistic subtleties and detecting deceptive information.
Passive Aggressive: The PA is an online learning algorithm that performs binary classification. It can adjust its parameters based on wrong predictions, which makes it particularly suitable for scenarios involving sequential data with concept drift. The computational efficiency and little memory consumption of our study make it well-suited for processing huge, streaming datasets such as Internet news.
Naive Bayes: The NB algorithm is a probabilistic model that computes probabilities by considering the independence of features given class labels. This property allows it to be efficient in spaces with a large number of dimensions [39, 40]. NB employs observed characteristics to compute the probability of an article's category and has demonstrated significant outcomes, mainly when dealing with a scarcity of training data or when processing in real-time.
Gradient Boosting: It iteratively merges multiple weak decision tree models sequentially to generate a robust predictive model. Through the process of iteratively rectifying errors in the prior ensemble and modifying the weights of data points, it acquires knowledge from incorrectly classified examples, resulting in precise forecasts. The study demonstrates that the system's capacity to adjust to complex patterns and manage datasets with noise improves its ability to detect deceptive information subtly.

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 2, April 2025
https://doi.org/10.47709/cnahpc.v7i2.5717

Submitted : Mar 24, 2025
Accepted : Apr 04, 2025
Published : Apr 04, 2025

Evaluation metrics

Metrics are used to evaluate the performance and quality of ML algorithms. These indicators provide an essential understanding of the model's strengths and weaknesses. Most of them are based on the confusion matrix. A classification model's effectiveness on a test set is displayed in a confusion matrix. False negative (FN), true negative (TN), true positive (TP), and false positive (FP) are the four variables that make up this matrix are listed in Table 3.

Table 3: Confusion metrics

	Predicted Positive	Predicted Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

We used the following matrices to evaluate all the classifiers on both datasets:

Accuracy: Accuracy is the most straightforward metric because it simply indicates whether or not the prediction was correct. Correctly predicted samples are divided by the total number of samples in the dataset to determine the accuracy rate.

$$Accuacy = \frac{(\text{Number of Correct Predictions})}{(\text{Total Number of Samples})} \tag{2}$$

A better model is represented by one with a more excellent accuracy score. However, there is still a potential that the forecast will be wrong. Therefore, three extra metrics were used to account for the misclassified observation: recall, precision, and F1-score.

Precision: The accuracy of a model is measured by how many correct predictions it makes relative to how many overall correct predictions it makes. It indicates the ratio of actual events to predicted events.

Recall: The recall measures how many correct predictions were made out of a total number of events. This exemplifies the model's ability to identify favourable outcomes accurately.

F1-Score: It is the average of the two measures: recall and precision. It's an excellent all-around metric because it takes recall and precision into account.

Tools and software used in Our Experiment

Our system uses Jupyter Notebook as the development environment and Python 3.9 for all coding, which reduces the time it takes to complete the project. We're using Python because, unlike many other programming languages, it can be deployed on multiple platforms without any additional modification. It made our code easier to read and more concise. Numerous libraries have been utilized, such as Pandas, NLTK, Scikit-Learn, Seaborn, NumPy, and Matplot.

Proposed Method

The following steps (see Figure 5) will be employed in this work to achieve the goal: In the first stage, we will perform the pre-processing on the dataset and extract the features. The pre-processing is explained earlier. Separate training and test sets will be created from the entire dataset. The following phase involves training the model with training data. Standard assessment criteria, including Recall, Precision, Accuracy, and F1-Score, will be used to assess the models after their performance has been recorded. We will evaluate the model's output and choose the most effective approach based on how well it performs and how accurately it can recognize the target. We will consider the computation time and accuracy level of all the models.

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 2, April 2025
https://doi.org/10.47709/cnahpc.v7i2.5717

Submitted : Mar 24, 2025
Accepted : Apr 04, 2025
Published : Apr 04, 2025

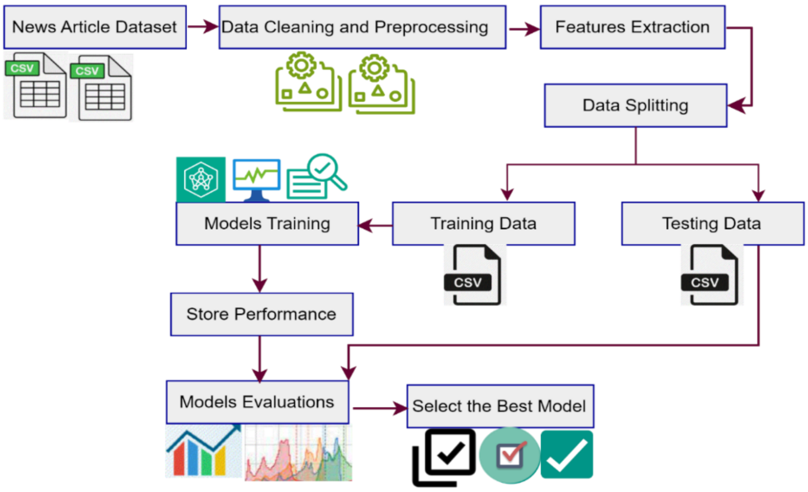


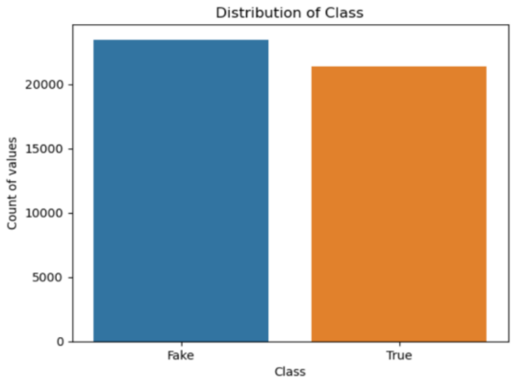
Figure 5: Working architecture of the proposed methodology

RESULT AND ANALYSIS

We have thoroughly explained the analysis of the collected results in this part. It also includes the subsections listed below.

Classes distribution

We used a count plot to show how the datasets were split between true and false labels. The count plot is a vital tool for evaluating the balance or imbalance of the two classes in the dataset shown in Figure 6 and Figure 7.



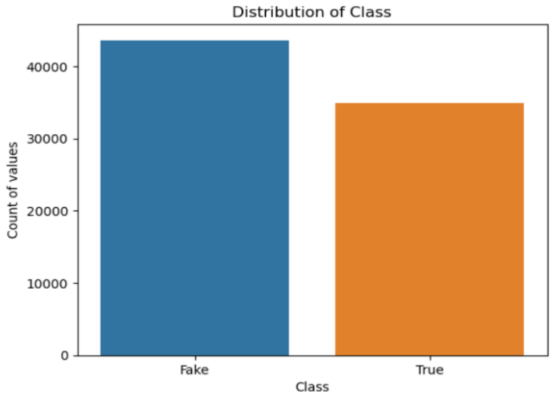


Figure 6: True and False label class distribution for DS1. Figure 7: True and False label class distribution for DS2.

Text distribution

Using a bar chart, we compared the number of texts representing different topics, as shown in Figure 8. This is useful for seeing how the DS1 text data is dispersed throughout various subjects, which in turn aided in spotting trends and inequities among those subjects.

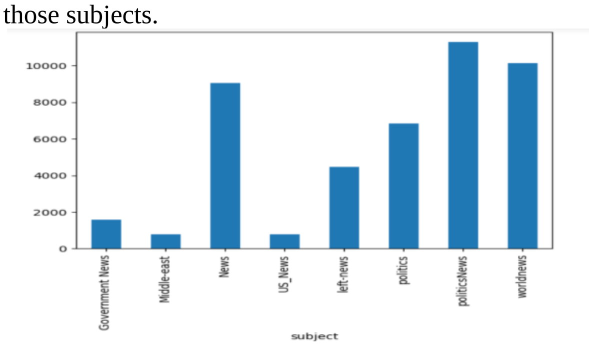


Figure 8: Text data subject classification for DS.

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 2, April 2025
https://doi.org/10.47709/cnahpc.v7i2.5717

Submitted : Mar 24, 2025
Accepted : Apr 04, 2025
Published : Apr 04, 2025

Top bigrams in the text

A bigram is a pair of consecutive components in a text or data set, usually words or letters. Bigrams are frequently utilized for text analysis and processing in the fields of NLP and computational linguistics. A bigram is a particular case of an n-gram, a series of n-consecutive elements. For example, n = 2 for bigrams. We looked at the most common pairs of words (bigrams) in the combined dataset. This aids in the identification of critical phrases, pattern recognition, etc., all of which improve the accuracy of our model, as shown in Figure 9 and Figure 10.

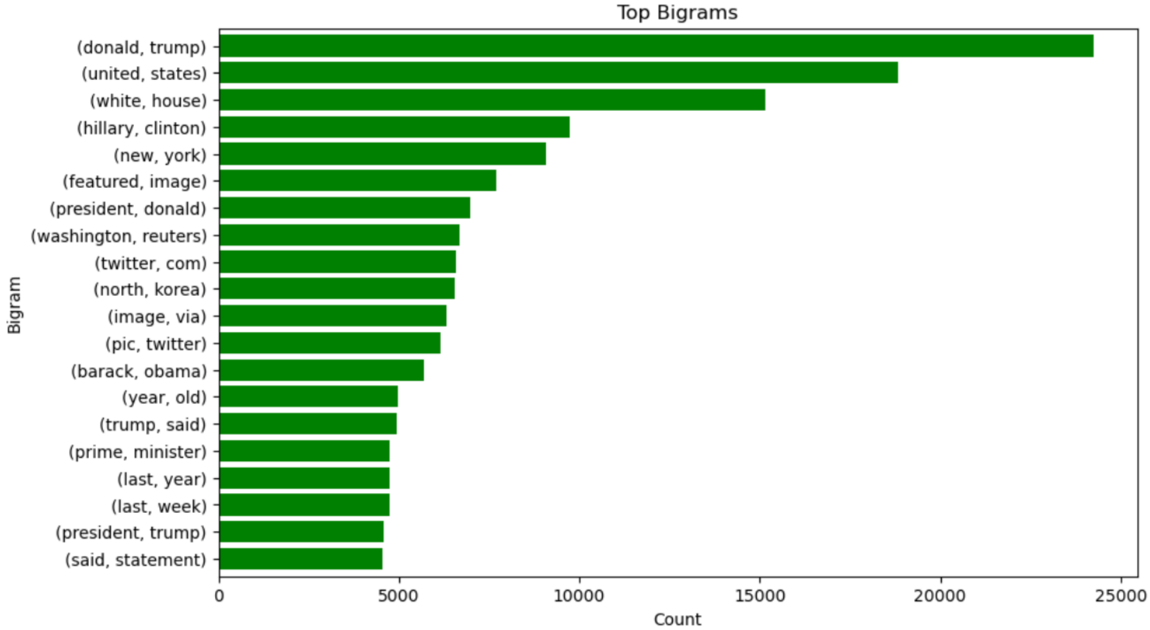


Figure 9: DS1 words data set for top bigrams

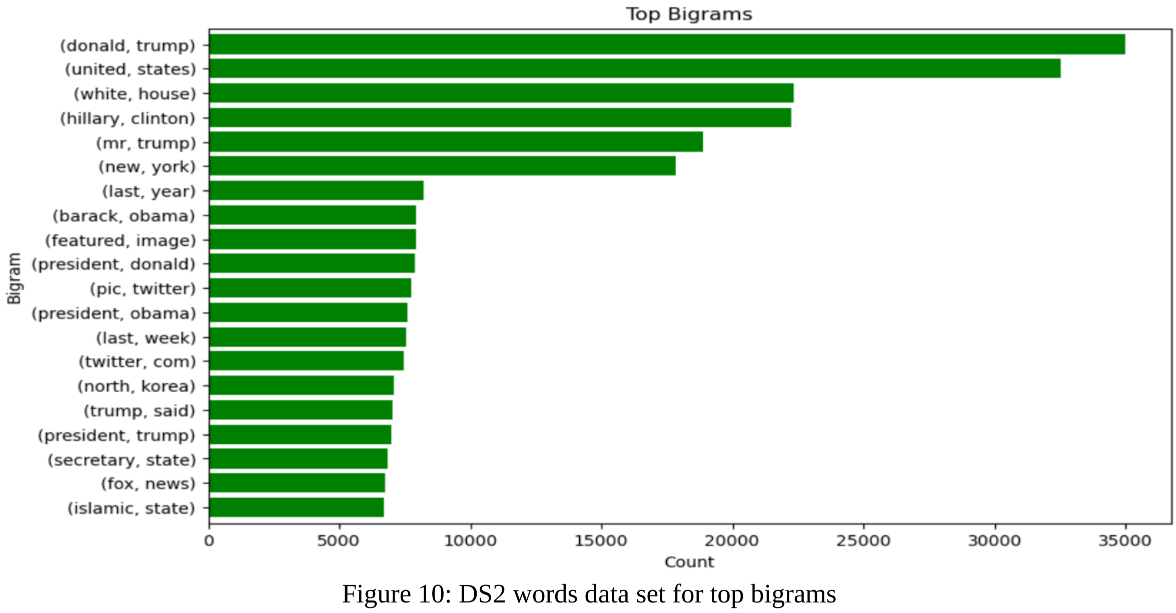


Figure 10: DS2 words data set for top bigrams

The discussion of results we obtained using eight ML models

Eight distinct ML algorithms, including GB, DT, RF, LR, NB, PA, SVM, and KNN, have been applied to the separated training and testing data. Figures 11 and 12 display the accuracy of eight methods on two separate datasets. The DT algorithm outperforms other algorithms in DS1 with an impressive 99.591% accuracy. The GB classifier, with its accuracy of 99.537%, is quite close to this performance. The PA achieved a maximum accuracy of 99.511%. The 99.341% precision rate achieved by the SVM is also awe-inspiring. When applied to DS2, the SVM method once again demonstrates its durability by achieving an accuracy of 95.179%, solidifying its position as the top performer. The PAC is just as reliable after SVM, with an accuracy of 95.027%, while the LR is just as

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 2, April 2025
https://doi.org/10.47709/cnahpc.v7i2.5717

Submitted : Mar 24, 2025
Accepted : Apr 04, 2025
Published : Apr 04, 2025

reliable, with an accuracy of 93.882%. Notable is the RF algorithm's constant performance, which, while lower at 93.434% in Dataset 2, is still sizeable. KNN performs poorly, with accuracy levels of 64.134% and 67.399% in Dataset 1 and Dataset 2, as shown in Figure 11 and Figure 12, respectively, indicating a lower level of efficacy. Due to its poor ability to capture complicated decision boundaries in high-dimensional environments, KNN is highly susceptible to background noise.

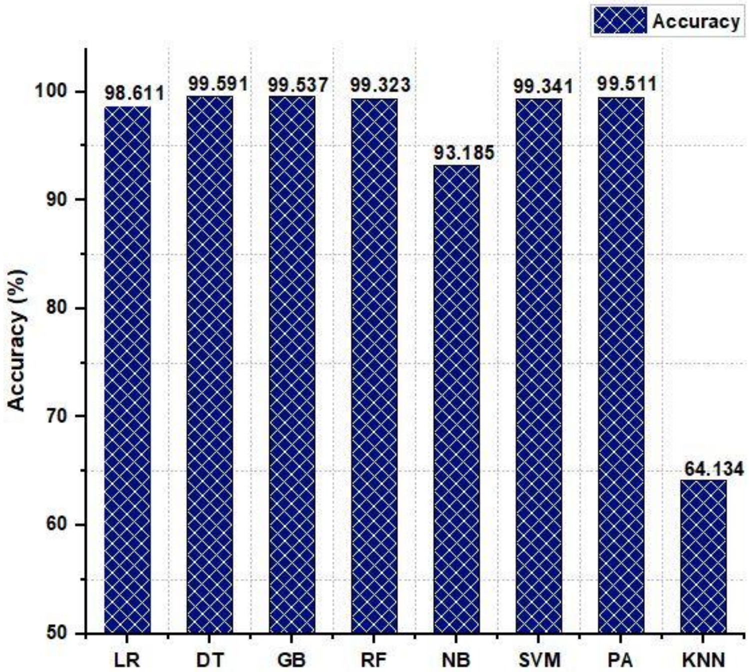


Figure 11: Accuracy of Dataset 1

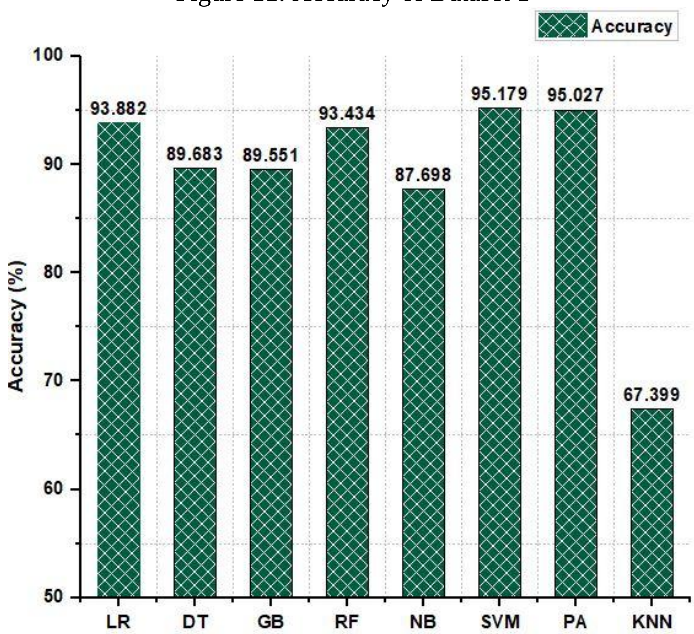


Figure 12: Accuracy of Dataset 2

Figure 13 displays the average accuracy of both of the models on both datasets. The average accuracy of PA and SVM is 97.269% and 97.26%, respectively, while the average accuracy of KNN is just 65.766%. In addition to the accuracy score, we employ recall, precision, and F1-score to evaluate the effectiveness of learning models.

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 2, April 2025
https://doi.org/10.47709/cnahpc.v7i2.5717

Submitted : Mar 24, 2025
Accepted : Apr 04, 2025
Published : Apr 04, 2025

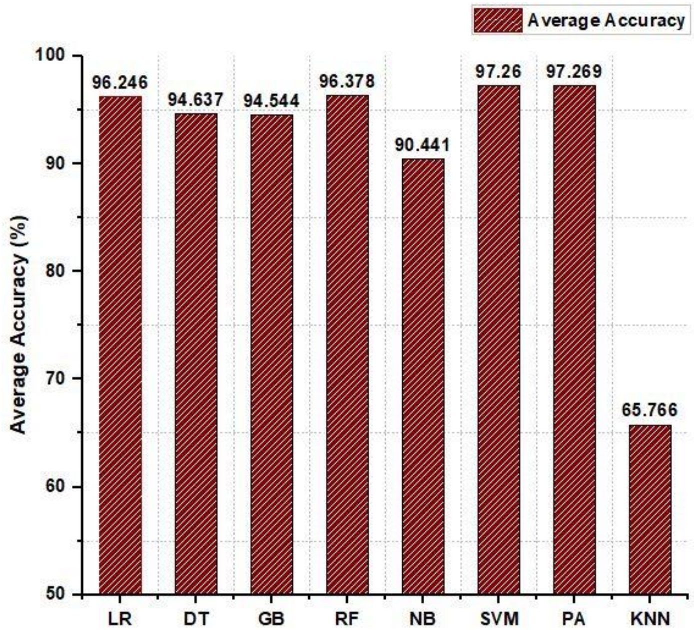


Figure 13: Average accuracy of models on both datasets

Figures 14 and Figure 15 for data set 1 and Figure 16 and Figure 17 for data set 2 display precision, recall, and F1 scores for each model across two datasets. In DS1, the Recall, Precision, and F1-score metrics remain stable across both classes for all models. The high accuracy and recall of DT, LR, and RF confirm their effectiveness in distinguishing between real and fake news. Meanwhile, GB's strong recall, precision, and F1-score highlight its strength in detecting authentic news. SVM stands out as a top performer, achieving consistently high recall and precision across both classes, demonstrating its robustness in classification. PA maintains a balanced trade-off between recall and precision, ensuring reliable differentiation between genuine and fabricated content. However, KNN struggles with precision for the real news class, indicating challenges in accurately identifying authentic news articles

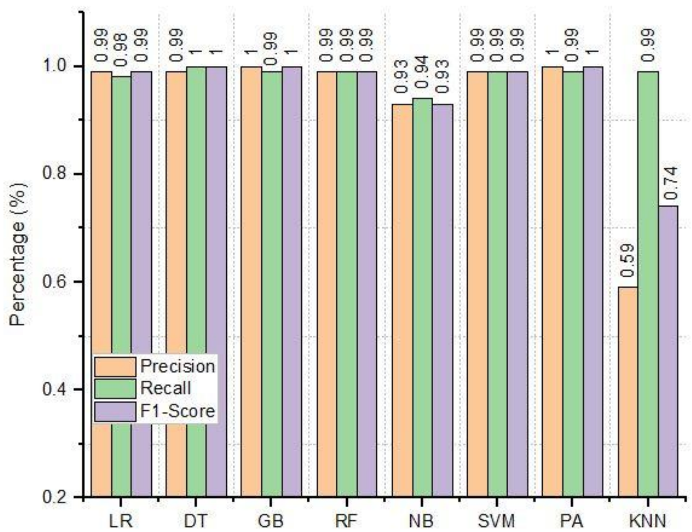


Figure 14: Evaluation Metrics of Fake News of Dataset 1

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 2, April 2025
https://doi.org/10.47709/cnahpc.v7i2.5717

Submitted : Mar 24, 2025
Accepted : Apr 04, 2025
Published : Apr 04, 2025

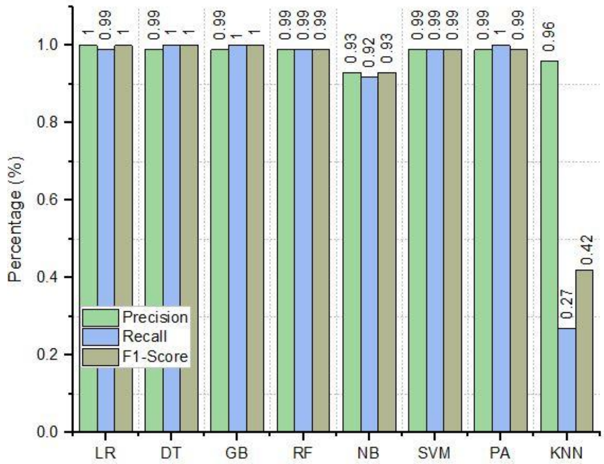


Figure 15: Evaluation Metrics of True News of Dataset 1

When evaluating classifiers based on several performance parameters, the PAC and SVM models demonstrate remarkable performance. Through comprehensive research and rigorous testing, they have demonstrated an exceptional capacity to outperform their rivals in several tasks. Consistently, it has surpassed other classifiers in terms of recall, precision, and F1-score metrics, establishing itself as the optimal choice for a wide range of applications.

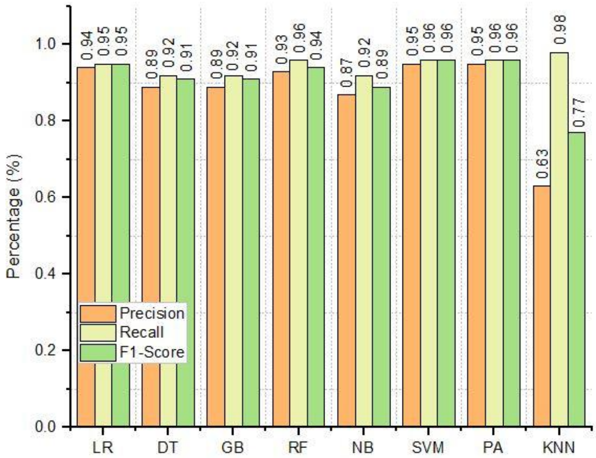


Figure 16: Evaluation Metrics of Fake News of Dataset 2

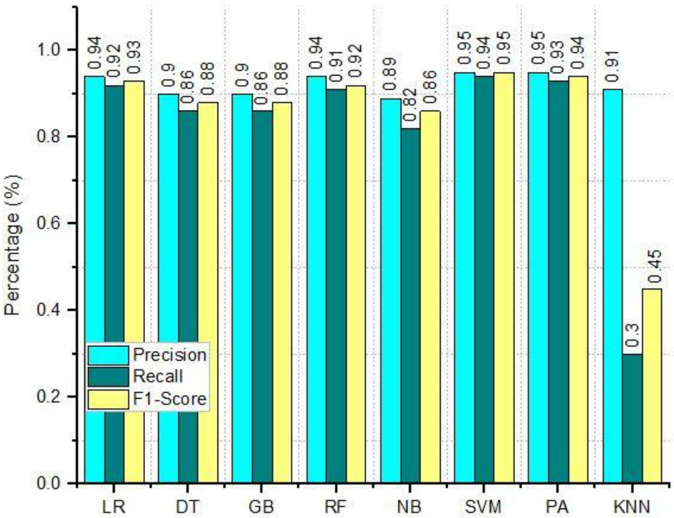


Figure 17: Evaluation Metrics of Fake News of Dataset 2

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 2, April 2025
https://doi.org/10.47709/cnahpc.v7i2.5717

Submitted : Mar 24, 2025
Accepted : Apr 04, 2025
Published : Apr 04, 2025

Confusion matrix

Evaluating the performance of a classification model can be accomplished by employing a confusion matrix, as shown in Figure 18. Given the problems with classification, a system with two or more classes may be required. Confusion matrices determine the precise count of accurate, inaccurate, and uncertain classifications by comparing the predicted label from the classification technique with the actual classes from the original dataset.

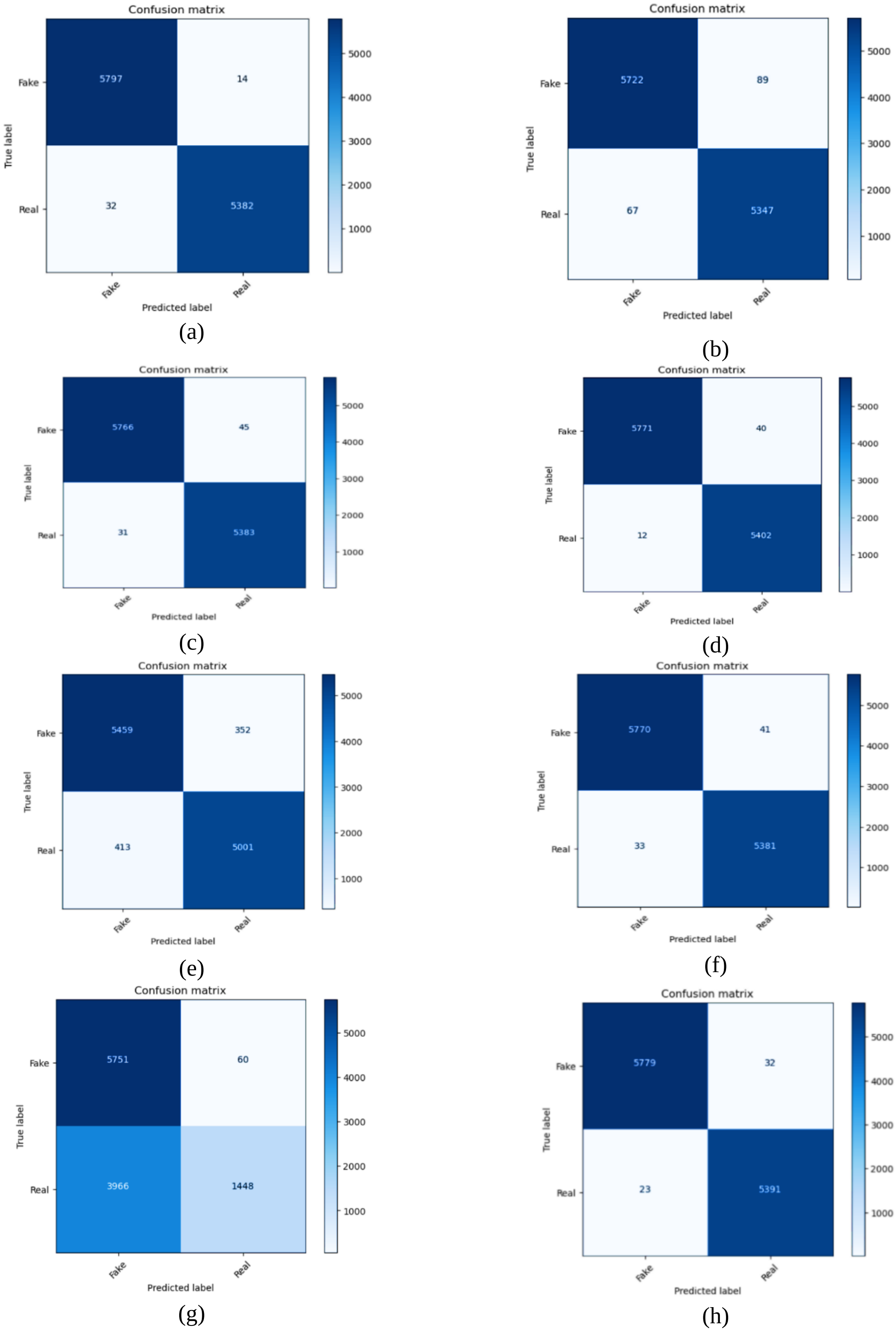


Figure 18: Confusion matrix of Dataset 1: (a)DT (b)LR (c) RF (d) GB (e)NB (f)SVM (g) KNN (h) PA

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 2, April 2025
https://doi.org/10.47709/cnahpc.v7i2.5717

Submitted : Mar 24, 2025
Accepted : Apr 04, 2025
Published : Apr 04, 2025

The confusion matrix tree of various techniques in two datasets is illustrated in Figures 18 and Figure 19. Based on Figure 18, the GB classifier and the DT classifier demonstrated superior performance, whereas KNN did not meet expectations and yielded the lowest results. Figure 19 demonstrates that SVM and PA exhibited the most impressive achievements. Similarly, in the DS1 scenario, KNN displayed the lowest results.

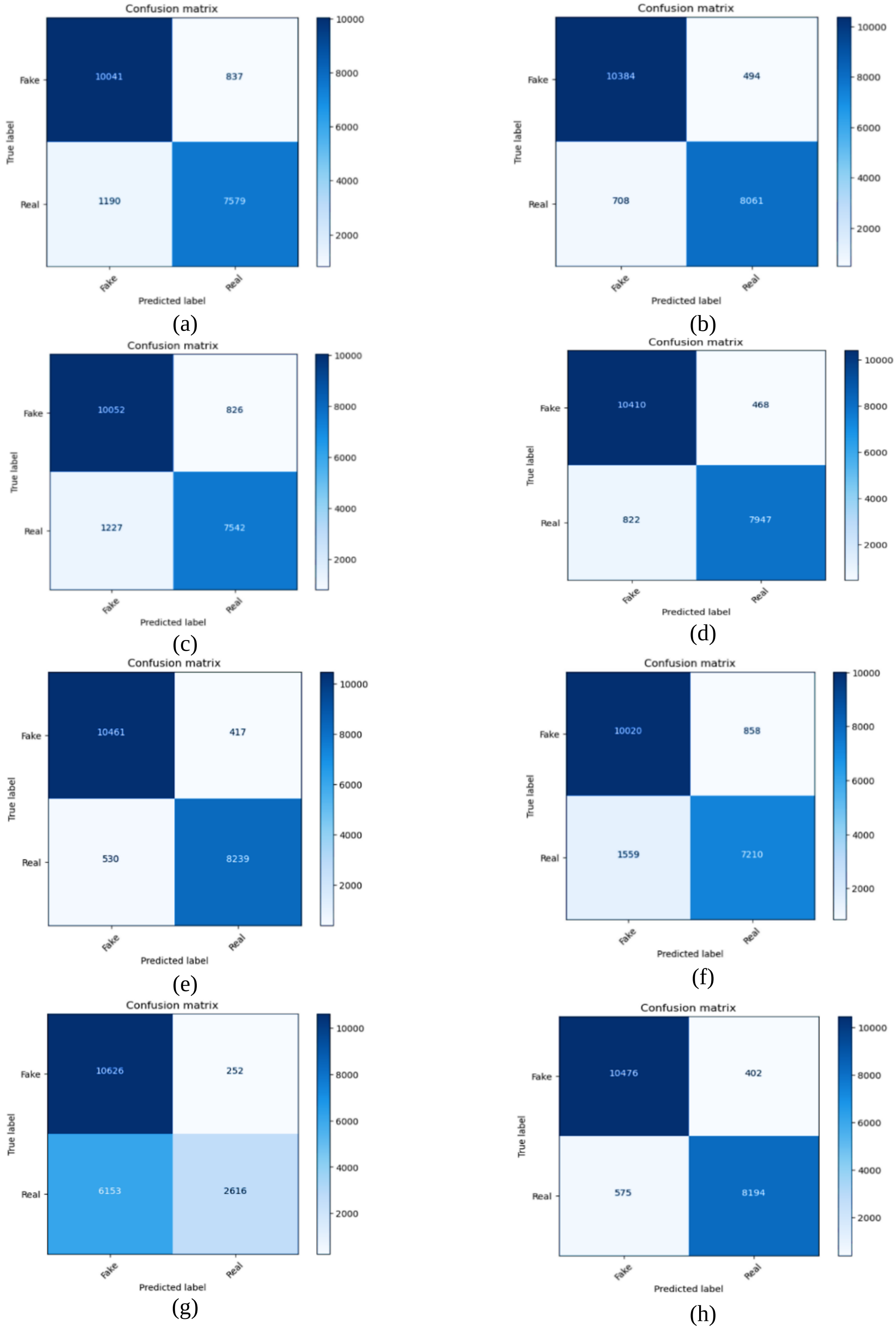


Figure 19: Confusion matrix of Dataset 2: (a)DT (b)LR (c)GB (d)RF (e)SVM (f)NB (g)KNN (h)PA

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 2, April 2025
https://doi.org/10.47709/cnahpc.v7i2.5717

Submitted : Mar 24, 2025
Accepted : Apr 04, 2025
Published : Apr 04, 2025

DISCUSSIONS

Finding and performance evaluation

We have presented the results of our study on identifying false news by employing various ML algorithms on two separate datasets, DS1 and DS2. Our goals were determining the most suitable ML algorithm and investigating how the size and complexity of the dataset impact the efficacy of different ML techniques. All models exhibited comparable accuracy and performance across metrics on the smaller DS1 dataset. However, as the DS2 dataset expanded to 79,000 instances, most models experienced a notable decline in performance. This highlights the difficulty of maintaining detection precision when handling larger and more diverse datasets.

Automated news classifier

In an era where news credibility is increasingly contested, automated verification systems are critical for systematically evaluating content legitimacy. Our research advances this objective through a computational framework designed to classify news articles as authentic or deceptive with high precision. By integrating eight machine learning classifiers, the system employs a structured framework for evaluating textual patterns, semantic coherence, and contextual anomalies. Users can input articles to receive a consensus-driven classification, leveraging ensemble predictions to minimize individual model biases. For validation, we applied the system to textual datasets derived from verified news repositories. Each classifier generates probabilistic outputs, enabling granular analysis of algorithmic confidence. Empirical results indicate consistent alignment across all models (excluding KNN) in classifying test articles, with robust correspondence to ground-truth labels. This underscores the framework's reliability in supporting evidence-based assessments of information integrity, equipping users to navigate complex media ecosystems with data-driven insights.

Research's novelty

This study's unique contribution stems from its broad evaluation of multiple ML algorithms, including passive-aggressive classifiers a less common choice in prior fake news detection research. By testing these models on two distinct datasets, the work demonstrates their adaptability to varied data structures and sources, reinforcing their applicability across different misinformation scenarios. The analysis acknowledges the contextual complexity of disinformation, where deceptive patterns shift across sociopolitical or topical domains. Detailed accuracy metrics (ranging from 84% to 96%) provide actionable guidance for researchers and developers to select context-appropriate techniques. For instance, the underperformance of KNN in cross-dataset validation underscores challenges in handling high-dimensional text features or sparse data common in misinformation datasets. This research presents a ground-breaking framework that broadens the scope of methodological approaches in the ongoing struggle against disinformation.

Comparison with state-of-the-art methods

Previous studies on ML-based false news detection have systematically evaluated various algorithms to measure their effectiveness in tackling key challenges. These comparisons typically involve multiple classifiers, including DT, RF, SVM, NB, and others. Researchers assess these models using metrics such as recall, F1-score, precision, and accuracy to enhance the distinction between fake and authentic news articles.

Decision Tree

The study by [24] evaluated six different ML algorithms for detecting and classifying fake news. After extensive testing and analysis, DT emerged as the most effective approach. It achieved an impressive accuracy of 99.36%, providing clear and interpretable classification rules Our research using a DT yielded an accuracy of 99.591% on DS1 and 89.683% on DS2. While DS1's accuracy improved slightly, DS2's accuracy dropped by 10% in comparison to this current model. Both experiments utilize datasets sourced from Kaggle. The study conducted by [24] utilized a dataset consisting of 20,000 records. In comparison, our DS1 dataset contains 49,000 entries, while DS2 contains 79,000 entries. There has been a notable enhancement in utilizing the DS1 dataset for the purpose of identifying and detecting fake news.

Logistic Regression

The article authored by [32] examined ten ML algorithms in conjunction with seven feature extraction approaches to analyse fake news pertaining to healthcare. The researchers reached an impressive accuracy rate of 99.87% utilizing the LR algorithm. Based on our research, DS1 accuracy is 98.611%, and DS2 accuracy is 93.882%, respectively, by the utilization of LR. By utilizing both datasets, we attained a lower level of accuracy in comparison to the previous research.

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 2, April 2025
https://doi.org/10.47709/cnahpc.v7i2.5717

Submitted : Mar 24, 2025
Accepted : Apr 04, 2025
Published : Apr 04, 2025

Gradient Boosting

The approach developed by [28] utilizes seven ML frameworks and three distinct feature extraction techniques. The GB technique, along with the TF-IDF feature extraction method, yielded a 90% accuracy rate. On the other hand, DS1 attained a precision level of 99.537 per cent, and DS2 achieved a precision level of 89.551 per cent when utilizing GB.

Random Forest

The research conducted by [24] achieved a high accuracy of 99.29% using the RF algorithm, which was somewhat lower than the accuracy achieved by the DT algorithm. In contrast, the research conducted by [19] reported a minimum accuracy of 59% when employing the RF algorithm. However, in our given scenario, DS1 and DS2 achieved better levels of accuracy utilizing the RF model, with individual scores of 99.323% and 93.434%, respectively.

Naïve Bayes

The study conducted by [23] evaluates the effectiveness of different ML algorithms in detecting fake news in the Bengali language. This is achieved by developing a unique dataset specifically for the Bengali language. They achieved their objective by employing the NB method and obtained an accuracy of 87.4%. Our accuracy rates for DS1 and DS2 were 93.185% and 87.698% respectively. The research conducted was superior to previous studies, and NB outperformed other methods when applied to our English Language datasets.

Support Vector Machine

SVM consistently demonstrated superior performance in the majority of prior studies. The study conducted by [5] utilized a reduced dataset obtained from Kaggle. They employed SVM as an ML method and achieved an impressive accuracy of 95.05%, surpassing the performance of other ML techniques. The authors [22] only employed NB and SVM as the ML algorithms. SVM exhibited superior performance compared to NB, achieving an accuracy of 93.6%. In contrast to RF, the SVM model was used in the study conducted by [31]. Achieved superior performance. The inclusion of several supplementary fake news datasets allowed for this, resulting in a 98% accuracy rate. Seven ML models and three different feature extraction algorithms were used in another investigation [28]. With the use of SVM and a variety of TF-IDF extraction algorithms, they were able to achieve a remarkable 94% accuracy. Our research outperformed the vast majority of prior studies, with an accuracy of 99.341% for DS1 and 95.179% for DS2. Our analysis confirms that SVM outperformed all other methods, with the exception of the study conducted by [31].

Passive Aggressive Classifier

In a study conducted by [27], just the PAC method was employed for false news identification, achieving an accuracy of 96% with the utilization of TF-IDF as the feature extraction technique. The accuracy of our DS1 model was 99.511%, while DS2 obtained an accuracy of 95.027%. DS1 achieved a remarkable accuracy compared to DS2 and the previous research.

K-Nearest Neighbour

The authors of [29] explored using ML models, especially KNN and QKNN, to identify instances of fake news. They observed that the KNN algorithm had the highest accuracy, reaching 91.3%. On the other hand, when we applied KNN to our scenario, neither of the datasets performed exceptionally well. DS1's accuracy was measured at 64.134%, while DS2's accuracy was measured at 67.399%. Regarding previous research, the authors employed the genetic engineering feature selection method, which outperforms TF-IDF when applied to KNN. The methods selected for this study include DT, LR, NB, RF, GB, PA, KNN, and SVM, based on previous model evaluations. The news articles were successfully classified upon implementing these classifiers on the dataset. The PA classifier has superior average accuracy in this classification, with a rate of 97.269%, surpassing the performance of previous models. This research has enhanced the process of identifying bogus news in online media by utilizing the passive-aggressive classifier. The study has resulted in the development of a sophisticated automated system that utilizes eight ML classifiers to accurately differentiate between authentic and falsified news stories. This system promotes well-informed decision-making in the current intricate information environment.

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 2, April 2025
https://doi.org/10.47709/cnahpc.v7i2.5717

Submitted : Mar 24, 2025
Accepted : Apr 04, 2025
Published : Apr 04, 2025

Findings Explanation

Dataset's robustness
In our results, the PA and SVM demonstrated excellent performance on both datasets, consistently achieving high accuracy. The proficiency in efficiently overseeing varied data characteristics underscores its potential for practical application in real-life applications.

Accuracy complexity
In DS1, the GB and DT showed excellent accuracy, indicating their ability to understand subtle relationships within the data. However, RF's reduced accuracy suggests a trade-off between complexity and precision.

KNN's sensitivity
Due to its reduced accuracy in both datasets, KNN's sensitivity to noise and difficulties in high-dimensional spaces are further reinforced. This result is consistent with its poor ability to record intricate patterns.

CONCLUSION

Manual news classification requires subject knowledge and text inconsistencies. This study uses ML algorithms to detect fake news. We classified news using eight ML algorithms. We chose two Kaggle datasets with lots of fake and real news. Learning models were trained and parameterized to improve accuracy. Only text and class features from the datasets were used for detection. This developed and evaluated the ML model, which enhances the accuracy of identifying false news. The study successfully achieved its objective by comprehensively analysing the strengths and drawbacks of different ML algorithms for detecting false news. In this work, we have employed eight ML algorithms to identify the most effective technique for detecting false news. Standard evaluation metrics were utilized to compare the performance of all the models. We utilized two separate datasets from Kaggle for this purpose. The ML algorithms utilized for the study include NB, DT, KNN, LR, PA, RF, GB, and SVM. PAC has a superior average accuracy of 97.269% compared to other models. By employing this research, we gained a deeper understanding of how effectively each algorithm distinguished truth from falsehood in news content. Identifying fake news presents numerous unresolved matters that necessitate additional investigation. A critical initial measure in curbing the dissemination of false information is acknowledging the fundamental components implicated in the transmission of news. Graph theory and transformer techniques, integrating multimodal analysis, can be utilized to identify the primary sources responsible for spreading fake news. Another probable future trend is the real-time detection of fake news in videos. The insufficient availability of labelled data can be addressed by constructing unsupervised or semi-supervised algorithms, which may become beneficial. These strategies aim to identify patterns or clusters in the data without relying heavily on labelled samples.

Funding: This work has not received any funding.
Conflict of interest: The authors declare that the research was conducted without any commercial or financial relationships that could be construed as a potential conflict of interest.
Data Availability: Misinformation & Fake News text dataset 79k. [online] Available at: https://www.kaggle.com/datasets/stevenpeutz/misinformation-fake-news-text-dataset-79k
Ethical approval: We confirm that relevant guidelines and regulations are carried out in all methods.
Competing interests: The authors declare no competing interests.

REFERENCES

[1] Q. Zhang, Z. Guo, Y. Zhu, P. Vijayakumar, A. Castiglione, and B. B. Gupta, "A deep learning-based fast fake news detection model for cyber-physical social services," *Pattern Recognition Letters*, vol. 168, pp. 31-38, 2023.
[2] Z. Guo, K. Yu, A. K. Bashir, D. Zhang, Y. D. Al-Otaibi, and M. Guizani, "Deep information fusion-driven POI scheduling for mobile social networks," *IEEE Network*, vol. 36, pp. 210-216, 2022.
[3] P. K. Donepudi, "Crowdsourced Software Testing: A Timely Opportunity," *Engineering International*, vol. 8, pp. 25-30, 2020.
[4] M. Aldwairi and A. Alwahedi, "Detecting fake news in social media networks," *Procedia Computer Science*, vol. 141, pp. 215-222, 2018.
[5] J. Shaikh and R. Patil, "Fake news detection using machine learning," in *2020 IEEE International Symposium on Sustainable Energy, Signal Processing and Cyber Security (iSSSC)*, 2020, pp. 1-5.

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 2, April 2025
https://doi.org/10.47709/cnahpc.v7i2.5717

Submitted : Mar 24, 2025
Accepted : Apr 04, 2025
Published : Apr 04, 2025

[6] S. Lewandowsky, U. K. Ecker, and J. Cook, "Beyond misinformation: Understanding and coping with the “post-truth” era," *Journal of applied research in memory and cognition,* vol. 6, pp. 353-369, 2017.

[7] L. Graves, *Deciding what's true: The rise of political fact-checking in American journalism*: Columbia University Press, 2016.

[8] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD explorations newsletter,* vol. 19, pp. 22-36, 2017.

[9] I. Ahmad, M. Yousaf, S. Yousaf, and M. O. Ahmad, "Fake news detection using machine learning ensemble methods," *Complexity,* vol. 2020, pp. 1-11, 2020.

[10] J. Hua and R. Shaw, "Corona virus (Covid-19)“infodemic” and emerging issues through a data lens: The case of china," *International journal of environmental research and public health,* vol. 17, p. 2309, 2020.

[11] E. Dwoskin, C. Dewey, and C. Timberg, "Why Facebook And Google Are Struggling to Purge Fake News," *Washington Post,* vol. 1, 2016.

[12] M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff, and B. Stein, "A stylometric inquiry into hyperpartisan and fake news," *arXiv preprint arXiv:1702.05638,* 2017.

[13] N. K. Conroy, V. L. Rubin, and Y. Chen, "Automatic deception detection: Methods for finding fake news," *Proceedings of the association for information science and technology,* vol. 52, pp. 1-4, 2015.

[14] H. U. Khan, M. Sohail, S. Nazir, T. Hussain, B. Shah, and F. Ali, "Role of authentication factors in Fin-tech mobile transaction security," *Journal of Big Data,* vol. 10, p. 138, 2023/09/08 2023.

[15] F. Wahab, I. Ullah, A. Shah, R. A. Khan, A. Choi, and M. S. Anwar, "Design and implementation of real-time object detection system based on single-shoot detector and OpenCV," *Frontiers in Psychology,* vol. 13, p. 1039645, 2022.

[16] A. Shah, B. Ali, F. Wahab, I. Ullah, F. Alqahtani and A. Tolba, "A Three-Way Clustering Mechanism to Handle Overlapping Regions," in *IEEE Access*, vol. 12, pp. 6546-6559, 2024, doi: 10.1109/ACCESS.2024.3349620

[17] H. Sadia, M. Ullah, T. Hussain, N. Gul, M. F. Hussain, N. ul Haq, *et al.*, "An efficient way of finding polarity of roman urdu reviews by using Boolean rules," *Scalable Computing: Practice and Experience,* vol. 21, pp. 277-289, 2020.

[18] N. F. Baarir and A. Djeffal, "Fake news detection using machine learning," in *2020 2nd International workshop on human-centric smart environments for health and well-being (IHSH)*, 2021, pp. 125-130.

[19] U. Sharma, S. Saran, and S. M. Patil, "Fake news detection using machine learning algorithms," *International Journal of Creative Research Thoughts (IJCRT),* vol. 8, pp. 509-518, 2020.

[20] H. Ahmed, I. Traore, and S. Saad, "Detecting opinion spams and fake news using text classification," *Security and Privacy,* vol. 1, p. e9, 2018.

[21] N. F. Johnson, N. Velásquez, N. J. Restrepo, R. Leahy, N. Gabriel, S. El Oud, *et al.*, "The online competition between pro-and anti-vaccination views," *Nature,* vol. 582, pp. 230-233, 2020.

[22] A. Jain, A. Shakya, H. Khatter, and A. K. Gupta, "A smart system for fake news detection using machine learning," in *2019 International conference on issues and challenges in intelligent computing techniques (ICICT)*, 2019, pp. 1-4.

[23] S. B. S. Mugdha, S. M. Ferdous, and A. Fahmin, "Evaluating machine learning algorithms for bengali fake news detection," in *2020 23rd International Conference on Computer and Information Technology (ICCIT)*, 2020, pp. 1-6.

[24] V. Gupta, R. S. Mathur, T. Bansal, and A. Goyal, "Fake News Detection using Machine Learning," in *2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON)*, 2022, pp. 84-89.

[25] S. Elyassami, S. Alseiari, M. ALZaabi, A. Hashem, and N. Aljahoori, "Fake news detection using ensemble learning and machine learning algorithms," *Combating Fake News with Computational Intelligence Techniques,* pp. 149-162, 2022.

[26] C. K. Hiramath and G. Deshpande, "Fake news detection using deep learning techniques," in *2019 1st International Conference on Advances in Information Technology (ICAIT)*, 2019, pp. 411-415.

[27] S. Deepak and P. Ameer, "Retrieval of brain MRI with tumor using contrastive loss based similarity on GoogLeNet encodings," *Computers in Biology and Medicine,* vol. 125, p. 103993, 2020.

[28] N. Smitha and R. Bharath, "Performance comparison of machine learning classifiers for fake news detection," in *2020 Second international conference on inventive research in computing applications (ICIRCA)*, 2020, pp. 696-700.

[29] Z. Tian and S. Baskiyar, "Fake News Detection using Machine Learning with Feature Selection," in *2021 6th International Conference on Computing, Communication and Security (ICCCS)*, 2021, pp. 1-6.

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 2, April 2025
https://doi.org/10.47709/cnahpc.v7i2.5717

Submitted : Mar 24, 2025
Accepted : Apr 04, 2025
Published : Apr 04, 2025

[30] S. Neeraj, L. Singh, S. Tripathi, and N. Malik, "Detection of Fake News using Machine Learning," in *2023 13th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, 2023, pp. 20-24.

[31] M. Aljabri, D. M. Alomari, and M. Aboulnour, "Fake News Detection Using Machine Learning Models," in *2022 14th International Conference on Computational Intelligence and Communication Networks (CICN)*, 2022, pp. 473-477.

[32] M. K. Elhadad, K. F. Li, and F. Gebali, "Detecting misleading information on COVID-19," *Ieee Access,* vol. 8, pp. 165201-165215, 2020.

[33] K. Patel and I. U. Hassan, "Identifying fake news with various machine learning model," in *2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)*, 2021, pp. 1-5.

[34] A. Agarwal and A. Dixit, "Fake news detection: an ensemble learning approach," in *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, 2020, pp. 1178-1183.

[35] Y.-Y. Song and L. Ying, "Decision tree methods: applications for classification and prediction," *Shanghai archives of psychiatry,* vol. 27, p. 130, 2015.

[36] Y. Yin, Y. Shi, Y. Zhao, and F. Wahab, "Multi-graph learning-based software defect location," *Journal of Software: Evolution and Process,* p. e2552, 2023.

[37] J. L. Speiser, M. E. Miller, J. Tooze, and E. Ip, "A comparison of random forest variable selection methods for classification prediction modeling," *Expert systems with applications,* vol. 134, pp. 93-101, 2019.

[38] B. Gregorutti, B. Michel, and P. Saint-Pierre, "Correlation and variable importance in random forests," *Statistics and Computing,* vol. 27, pp. 659-678, 2017.

[39] X. Liu, Y. Zhao, T. Xu, F. Wahab, Y. Sun, and C. Chen, "Efficient False Positive Control Algorithms in Big Data Mining," *Applied Sciences,* vol. 13, p. 5006, 2023.

[40] Y. Yalman, T. Uyanık, İ. Atlı, A. Tan, K. Ç. Bayındır, Ö. Karal*, et al.*, "Prediction of Voltage Sag Relative Location with Data-Driven Algorithms in Distribution Grid," *Energies,* vol. 15, p. 6641, 2022.

* Corresponding author