International Journal of Applied Sciences: Current and Future Research Trends
(IJASCFRT)
ISSN: 2790-3990
https://ijascfrtjournal.isrra.org/index.php/Applied_Sciences_Journal

Detection of the Fake News Using Machine Learning Algorithms and Data Analysis Techniques

Mirza Nikšić[a*], Dželila Mehanović[b]

[a,b]*International Burch University, Francuske revolucije bb, Sarajevo and 71000, Bosnia and Herzegovina*
[a]*Email: mirza.niksic@stu.ibu.edu.ba*
[b]*Email: dzelila.mehanovic@ibu.edu.ba*

Abstract

Due to the rapid advancement of online social networks in recent years, the prevalence of fake news has increased significantly. Fake news is deliberately created to deceive users by imitating real news, making it challenging to identify early on. So, we need to explore the accompanying information to improve its disclosure such as the publisher. This study focuses on analyzing and investigating various traditional machine learning models to determine the most effective one. The goal is to develop a supervised machine learning algorithm that can classify news articles as either true or fake, utilizing tools like Python's scikit-learn and NLP for text analysis. The proposed approach involves feature extraction and vectorization. To accomplish this, the scikit-learn library in Python is utilized, which offers helpful tools like CountVectorizer and TfidfVectorizer. The experiment involved implementing well-known algorithms: Logistic regression, Neural networks and SVM, and comparing their performance to determine the most suitable one. Each of the three algorithms performed well, but SVM demonstrated superior outcomes across nearly all categories.

Keywords: Fake news; Detection; Machine learning; Algorithm; Text Classification.

1. Introduction

The proliferation of social media and online news platforms has led to an unprecedented increase in the dissemination of information. While this has undoubtedly enhanced the accessibility and speed of news consumption, it has also created fertile ground for the spread of misinformation and fake news. Fake news, defined as deliberately false or misleading information presented as news, has become a significant concern, as its impact can be far-reaching and detrimental to individuals, societies, and even democratic processes. The intentional spread of false information can have serious consequences, such as undermining trust in institutions, influencing public opinion, and even posing a threat to national security.

Received: 7/24/2023
Accepted: 8/29/2023
Published: 9/11/2023

---

* Corresponding author.

The issue of fake news has become increasingly urgent, and detecting and combating it has become a critical challenge for researchers, policymakers, and technology companies. Social media is dramatically changing the way news is produced, disseminated, and consumed, opening unforeseen opportunities but also creating complex challenges. A key problem today is that social media has become a place for disinformation campaigns which affect the credibility of the entire news ecosystem. A unique feature of social media news is that anyone can register as a news publisher at no upfront cost. Approximately 56.8% of people worldwide use social media. With a global population of 8 billion people, this percentage corresponds to around 4.48 billion individuals actively using social media platforms [1]. In 2020, most American adults get their news from digital platforms, and for young Americans, social media is the new primary source of news. In 2020, 90% of American millennials received the latest news from popular social platforms [2]. In this modern world, data is very important and in 2022, 2.5 quintillion bytes of data are created every day. The average social media user engages with an average of 6.6 different social media platforms, so there are plenty of opportunities to catch some fake news [3]. According to the Tandoc, Lim, and Ling [4], the term "fake news" has evolved into a popular buzzword, and its current usage appears to deviate from earlier definitions. In the past, the term encompassed various forms of content that were related but distinct from each other, including news parodies, political satires, and news propaganda. However, presently, it is commonly used to describe false stories that circulate on social media platforms. Additionally, there have been instances where the term "fake news" has been invoked to undermine critical reporting by certain news organizations, thereby adding further complexity to discussions surrounding fake news. The objective of this study is to make a meaningful contribution to the expanding realm of fake news detection. Our focus involves an examination of diverse strategies and methods aimed at recognizing and curbing the dissemination of false information. Through an exploration of the fundamental drivers, incentives, and attributes associated with fake news, our aim is to cultivate a thorough comprehension of this phenomenon. In our research, we have used three different machine learning models—logistic regression, neural networks, and support vector machines—assessing and contrasting their effectiveness and outcomes.

1.1. Structure

The remaining sections of this thesis are structured as follows. We are getting familiar with "fake news" term and the various consequences it can bring about in Chapter Two. In Chapter Two, an in-depth review of relevant literature related to our research topic is conducted. Furthermore, it provides a detailed explanation of the methodology employed in our study. This chapter includes a clear elucidation of each machine learning algorithm used, accompanied by visual aids such as images and figures to enhance comprehension. Additionally, the steps taken to complete this research paper are outlined. This chapter also provides comprehensive information about the dataset utilized and the data manipulation techniques employed is presented. This chapter also encompasses data analysis, where the findings obtained from the dataset are presented. Chapter Three is dedicated to presenting the obtained results, engaging in a thorough discussion of these outcomes, and comparing results of used algorithms. Lastly, in Chapter Four, we conclude our work by examining its implications and potential limitations. Furthermore, we propose avenues for future research in this domain.

1.2. Fake News Definition and Effects

Fake news refers to a misleading or sensational report that is intentionally created to attract attention, deceive, or harm someone's reputation. It differs from misinformation, which arises from journalists mistakenly confusing facts. Fake news is deliberately fabricated with the aim of manipulating individuals or situations. The dissemination of false information is rapid, as people can easily download articles, share them with others, who in turn share them further. By the end of the day, the fake news has traveled so far from its original source that it becomes indistinguishable from genuine news [5]. Fake news can indeed have detrimental impacts on individuals and society.

- It can mislead people and lead them to adopt false beliefs based on inaccurate information. This can have far-reaching consequences, as individuals make decisions and form opinions based on these false premises.
- The presence of fake news can influence how people respond to genuine news. When individuals are repeatedly exposed to false information, it can distort their perception of reality and erode their trust in reliable news sources.
- The widespread dissemination of fake news can undermine the overall credibility of the news ecosystem. If people become increasingly skeptical of news in general, it becomes more challenging to distinguish between accurate reporting and deliberately misleading content.

2. Materials and Methods

2.1. Literature Review

Utilizing social media as a means of staying updated with news is a practice fraught with both advantages and disadvantages. On the positive side, social media platforms offer a convenient and readily accessible avenue for accessing information, often at minimal or no cost. They excel at facilitating the swift dissemination of news to a broad and diverse audience. In this digital age, news has the potential to spread like wildfire on social networks, obliterating traditional constraints of time and geography. However, this immense convenience and speed come at a price. Social media platforms have evolved into ideal breeding grounds for the creation and propagation of fake news, a significant drawback. The virality of information on social networks, coupled with the absence of rigorous fact-checking mechanisms, creates fertile ground for the rapid and extensive circulation of misinformation and deceptive content [6]. The consequences of misinformation should not be underestimated, as they often extend well beyond the initial dissemination of false information. Exposure to inaccurate or false information, whether intentionally or inadvertently, can profoundly influence individuals' reasoning and decision-making processes. This is particularly concerning in an era where information flows swiftly and widely through various online platforms. An illustrative example highlighting the potential consequences of misinformation occurred in October 2008. During that period, a journalist published a false report claiming that Steve Jobs, the esteemed co-founder of Apple Inc., had suffered a heart attack. The speed at which this fabricated news permeated social media platforms and other online channels was astonishing. As a result, the stock market reacted instantaneously, witnessing significant fluctuations in Apple Inc.'s shares throughout that fateful day [7]. This incident vividly underscores the profound impact misinformation can have

on financial markets, investor confidence, and the reputations of individuals or organizations. The rapid dissemination of false information can incite panic, trigger market volatility, and inflict financial losses. Moreover, its consequences often endure over time, tainting reputations and molding public perceptions. Addressing the challenges posed by misinformation and rectifying its effects presents its own set of formidable obstacles. Once false information gains traction and spreads widely, reversing its impact can be a daunting task. Even when the truth ultimately emerges, the initial false narrative can linger in people's minds, shaping their beliefs and attitudes. The persistence of misinformation underscores the critical importance of proactive measures to prevent its dissemination and counter its effects. In the domain of fake content detection, previous research has explored a multitude of domains to identify deceptive information. These domains encompass forums, consumer review websites, online advertising, online dating platforms, and crowdsourcing platforms [8] [9, 10]. Researchers have harnessed linguistic cues to distinguish between genuine communicators and deceivers. For instance, self-references and the use of positive or negative language have been leveraged to profile deceptive individuals [11]. Other studies have delved into the analysis of deceptive content characteristics, including word and sentence counts, the presence of self-references, affective language, as well as the spatial and temporal dimensions of deceptive behavior [12]. Furthermore, researchers have delved into additional features related to deceptive behaviors. Expressiveness, informality, variety, and immediacy have all been explored as potential indicators of deceptive content [13]. These features aim to capture the linguistic and behavioral patterns that differentiate between authentic and deceptive communication. By examining diverse domains and employing a variety of linguistic and behavioral cues, researchers have made significant strides in automating the detection of fake content. These investigations have shed light on the underlying characteristics of deceptive behavior and offered valuable insights into the development of effective detection models. As we embark on our own research journey, it is paramount that we build upon these prior findings and explore the application of traditional machine learning models. Through the utilization of techniques such as text analysis, feature extraction, and vectorization with tools like Python's scikit-learn library, we have the opportunity to compare different algorithms and identify the most effective approach for classifying fake news as either true or false. This iterative process contributes to the ongoing efforts to combat the dissemination of deceptive information and reinforce the credibility of news sources in the digital age.

2.2. Methodology

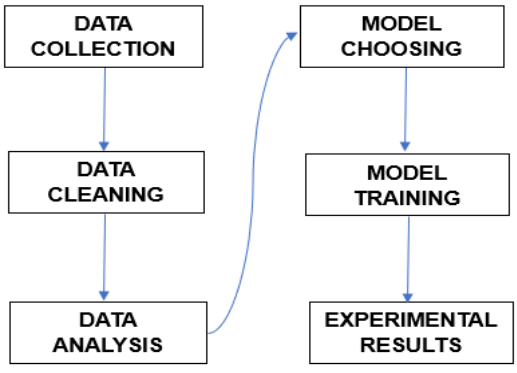

Figure 1: Fake news section methodology.

Figure 1 illustrates the steps followed to develop this article. The initial stage involved data collection, followed by data cleaning and analysis. Subsequently, the models were selected, trained, and finally evaluated to assess their performance.

2.2.1. Data collection

The methodological process began with the crucial step of data collection. However, during this initial phase, we encountered a challenging decision: whether to invest time and effort into crafting a custom script for data scraping or to leverage an existing dataset. Data scraping, also commonly known as web scraping, involves the automated extraction of information from websites, which is then organized and stored in formats such as spreadsheets or local files on a computer [14]. To address this dilemma, we carefully evaluated the pros and cons of each approach. Creating a custom script for data scraping would have allowed us to tailor the data collection process precisely to our specific requirements. It would have granted us more control over the data extraction and allowed us to target specific sources or websites relevant to our research. However, we also acknowledged that developing a robust and reliable data scraping script would require a considerable amount of time and technical expertise, potentially diverting our focus from other critical aspects of the project. On the other hand, opting for an already existing dataset offered the advantage of immediate access to a substantial pool of relevant data. By using a pre-existing dataset, we could save valuable time and resources and promptly proceed to subsequent stages of the research. Moreover, depending on the dataset's source and size, it could provide a diverse and comprehensive collection of data points, enriching the analytical potential of our study. After careful consideration and consultation with our research team, we ultimately decided to forego the custom script development and leverage an existing dataset. This decision aligned with our project's timelines and allowed us to dedicate more attention to refining our research questions, designing the analytical framework, and delving into the subsequent phases of the study. By choosing this approach, we gained quick access to a well curated dataset, meticulously compiled by other researchers or organizations, which aligned closely with the specific subject matter of our study. The use of a pre-existing dataset provided us with a strong foundation to initiate data analysis promptly, leading to more efficient progress in our research journey.

2.2.2. Data cleaning

The second step in our methodology involved the critical process of data checking and cleaning. This crucial stage ensured the quality and reliability of the dataset, preparing it for further analysis. During this phase, we have used various data cleaning techniques to enhance the integrity of the information. The first aspect of data cleaning involved the identification and removal of duplicates. Duplicate entries within the dataset can distort analysis results and lead to biased conclusions. By eliminating duplicate records, we ensured that each data point contributed uniquely to our analysis, enhancing the accuracy of our findings. Additionally, we addressed the issue of empty fields or missing values in the dataset. Missing data can pose challenges in statistical analysis and machine learning modeling. To mitigate this problem, we systematically checked for empty fields and adopted appropriate strategies to handle them. Depending on the specific scenario, we used methods such as imputation or exclusion, ensuring that missing values did not undermine the overall integrity of the dataset. To simplify the subsequent analysis process and ensure consistent comparisons, we standardized the text data.

Punctuation marks were removed from the text, thereby eliminating any potential discrepancies arising from punctuation usage. Additionally, to avoid the complications of case sensitivity, we converted all text into lowercase. This step ensured that our analysis treated similar words with different cases as identical, making it easier to identify patterns and trends within the text data. With the data now refined and ready for analysis, we are better equipped to delve into the core of our research questions. The cleaned dataset, free from duplicates, empty fields, and punctuation marks, enables us to conduct more precise and reliable analyses, drawing meaningful insights from the text data. This meticulous data preparation ensures that our research outcomes are based on robust and accurate information, facilitating more informed conclusions, and enhancing the overall integrity of our study.

2.2.3. Data analysis

After completing the data cleaning process, the next step in our methodology involved conducting data analysis to gain a deeper understanding of the dataset. Data analysis is the process of inspecting, cleaning, transforming, and interpreting raw data with the goal of extracting useful information, patterns, and insights to support decision-making and problem-solving. It involves various techniques, methodologies, and tools to examine data from different perspectives, identify trends, correlations, and outliers, and make informed conclusions [15]. To facilitate the data analysis process, we utilized Python libraries such as Pandas [16] and Seaborn [17]. Seaborn is a graphic visualization library that is built on the primary configurations of Matplotlib. It provides accessibility to the users with some of the most commonly provides data visualizations processes with certain data visualizations necessities such as mapping color to a variable or using faceting requirements across the globe [18]. Pandas is an open-source library used in Python that provides enhanced performance metrics, easy to use data structures and data analysis packages, tools, and libraries for Python Programming Language. D. Choosing models After completing the data analysis phase, the subsequent step in our methodology was to determine the machine learning models we would employ for our classification problem. Machine learning (ML) serves to teach machines how to handle data more efficiently. There are instances where the data itself contains valuable information that may not be immediately apparent or interpretable to humans. In such cases, machine learning techniques come into play to extract meaningful insights and patterns from the data [19]. Given that our objective involves predicting class labels (either fake or real news) based on input data, it falls under the realm of classification. In this case, we are dealing with binary classification, as there are only two possible outputs. Classification refers to a predictive modeling problem where a class label is predicted for a given example of input data [20]. To ensure a comprehensive evaluation, we decided to utilize some of the most popular and widely used classification-based machine learning models. Each of these models brings distinct strengths and characteristics to the table, allowing us to compare their performance and identify the most suitable one for our specific research problem.

- Support Vector Machines: Support Vector Machines (SVM) function by skillfully mapping data into a higher dimensional feature space, strategically enabling the categorization of data points, even when they exhibit non-linear separability in their original form. Through a meticulous process, SVM identifies a separator that distinguishes between different categories, and subsequently transforms the data in such a way that the separator can be visualized as a hyperplane in higher-dimensional space. This transformation empowers SVM to

effectively handle complex data distributions and find meaningful patterns within them [21]. By establishing this hyperplane, SVM is equipped to efficiently classify new data points based on their proximity to either side of the separator. Such a classification process allows SVM to make predictions regarding the group to which a new record should be assigned, utilizing the distinctive characteristics exhibited by the data [21]. Due to its ability to deal with both linear and nonlinear classification problems, SVM has gained widespread application in diverse domains, including image recognition, text classification, biological data analysis, financial modeling, and more. Its capability to handle high-dimensional data and generalize well to unseen samples makes SVM a valuable tool for various machine learning tasks, contributing significantly to the advancement of data analysis and decision-making processes across industries [21].

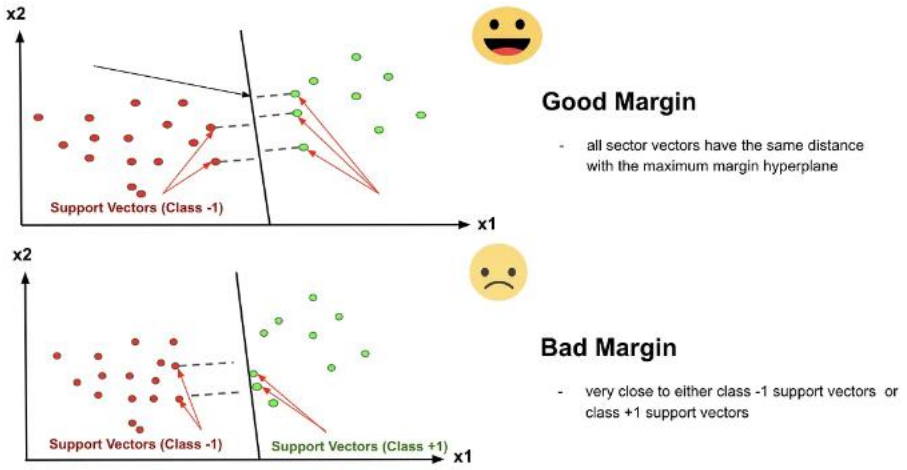

Figure 2: Example of good and bad margin [22].

In Figure 2, you can see examples of good and bad margin positions. In the first example, the margin is evenly positioned and equally distanced from the support vectors of both “Class +1” and “Class -1”. In contrast, the second example illustrates a bad margin position where the margin is too close to the support vectors of “Class +1” and too far from the support vectors of “Class -1”. In the context of fake news detection, SVMs can be trained on a set of labeled news articles to identify patterns and features that distinguish between real and fake news. SVMs can use various types of features, such as the content of the article, the source of the article, and the metadata associated with the article, to learn a decision boundary that separates real and fake news. Once trained, the SVM can classify new news articles as real or fake based on the learned decision boundary.

- Logistic regression: Logistic regression is used to calculate the odds ratio when there are multiple explanatory variables. It follows a procedure akin to multiple linear regression, except that the response variable is binomial. The outcome of logistic regression reveals the influence of each variable on the odds ratio of the observed event of interest. One of its primary advantages lies in its ability to analyze the association of all variables simultaneously, thereby circumventing confounding effects. In this article, we elucidate the logistic regression process using illustrative examples to simplify comprehension. After defining the technique, we emphasize the fundamental interpretation of the results and subsequently delve into certain special considerations [23].

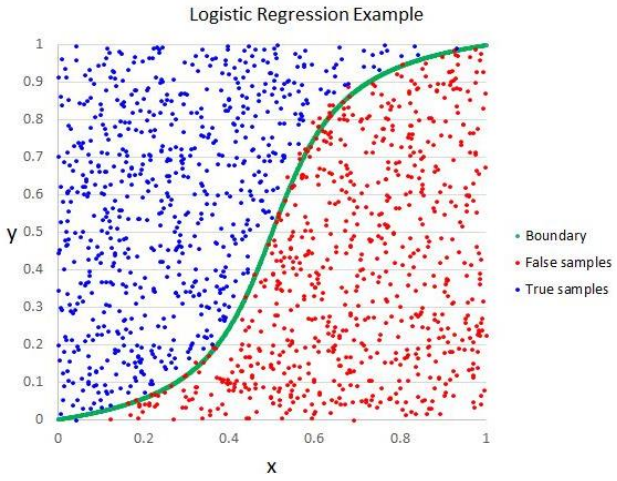

Figure 3: Logistic regression example [24].

In Figure 3, you can observe a visualization of the logistic regression model with a well-positioned boundary line that effectively separates false and true samples. In the context of fake news detection, logistic regression can be used to classify news articles as either real or fake based on a set of input features. For example, the presence of specific keywords or phrases, the number of spelling or grammatical errors, and the sentiment of the article can all be used as input features. Logistic regression works by modeling the probability of a news article being real or fake given its input features. The algorithm estimates the parameters of a logistic function that maps the input features to a probability value between 0 and 1. This probability value can be interpreted as the likelihood of the news article being real or fake, with a value closer to 0 indicating a higher probability of the article being real and a value closer to 1 indicating a higher probability of the article being fake. To train the logistic regression model, a set of labeled data is required, where each news article is labeled as either real or fake. The model is then trained to minimize the difference between the predicted probabilities and the actual labels using a loss function such as cross-entropy loss. One of the advantages of logistic regression is that it is a simple and interpretable algorithm that can be easily implemented and understood.

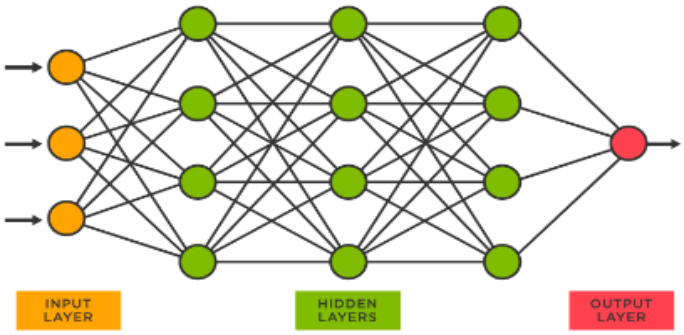

Figure 4: Neural networks architecture [26].

- Neural networks: Neural network is a method in artificial intelligence that teaches computers to process data in a way that is inspired by the human brain. It consists of layers of interconnected nodes, or neurons, that perform calculations on input data to produce output values. Deep neural networks can have many layers,

allowing them to learn complex patterns and relationships in data. It creates an adaptive system that computers use to learn from their mistakes and improve continuously [25].

The architecture of the neural networks is visually represented in figure 4. Neural networks adopt a distinct problem-solving approach compared to conventional computers. Traditional computers rely on algorithms, following predefined instructions to tackle problems. This limitation restricts their problem-solving capabilities to tasks we already comprehend and can explicitly instruct the computer to solve. However, the true potential of computers lies in their ability to handle problems beyond our current knowledge. In contrast, neural networks emulate the information processing of the human brain. These networks consist of numerous interconnected processing elements called neurons, working in parallel to address specific problems. A key characteristic of neural networks is their ability to learn from examples. Unlike conventional computers, neural networks cannot be programmed for a specific task; instead, they learn through carefully selected examples. This learning process is crucial, as inappropriate examples could lead to wasted time or incorrect functionality of the network. In the context of fake news detection, neural networks can be trained on a large corpus of news articles to identify patterns and features that distinguish between real and fake news. Neural networks can use various types of features, such as the content of the article, the source of the article, and the metadata associated with the article, to learn a decision boundary that separates real and fake news.

2.3. Data and Findings

For this project, the dataset used was obtained from Kaggle, an online community platform for data scientists and machine learning enthusiasts. The dataset consists of two files: “true” and “fake news.” The “fake news” dataset contains 17,903 unique values, while the “true” dataset contains 20,826 unique values. To create a comprehensive dataset for training machine learning models, these two files were combined, resulting in a larger dataset with 38,729 entries. The combined dataset includes four columns or features: title, text, subject, and date. These columns provide valuable information for analyzing and categorizing news articles. The “title” column contains the headline or title of the article, while the “text” column contains the main body of the news content. The “subject” column indicates the subject or topic of the article, and the “date” column represents the publication date of the news article. In total, there are eight subjects covered in the dataset, including politics news, world news, news, politics, government news, left-news, US News, and middle-east. These subjects encompass a broad range of news topics, allowing for diverse analysis and classification of the articles. In Table 1 and Figure 5, you can find a visual representation of the subjects along with their corresponding article counts.

Table 1: Subjects of the News.

Subject	**Count**
politicsNews	11272
Worldnews	10145
News	9050
Politics	6841
Left-news	4459
Government News	1570
US_News	783
Middle-east	778

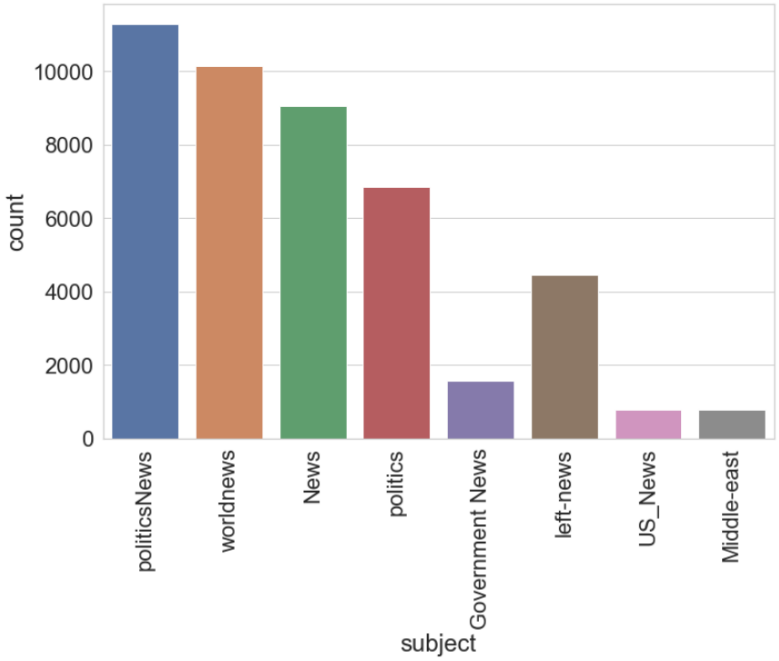

Figure 5: Visual presentation of subjects.

In Figure 6, you can find a chart that visually represents the subjects along with their corresponding article counts. This chart provides an overview of the number of articles or news pieces associated with each subject. The data indicates that most real news articles contain fewer than 2000 characters, whereas most fake news articles have character counts below 4000. This observation suggests that there is a difference in the length distribution between the two categories.

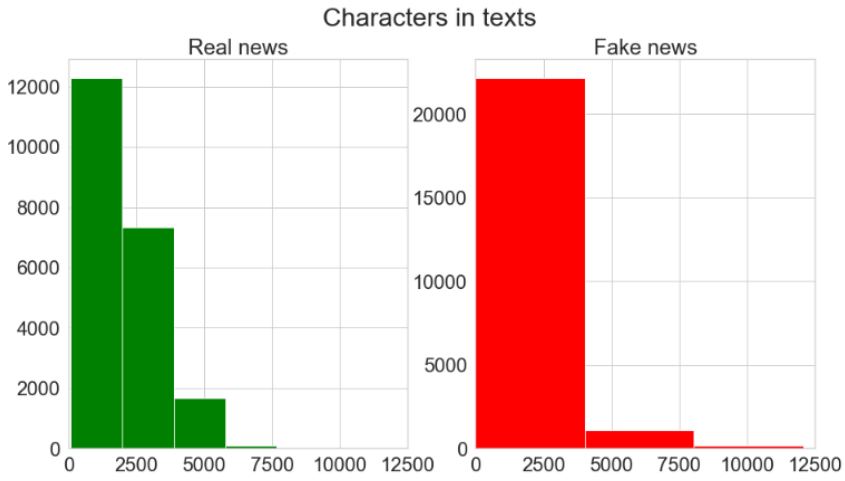

Figure 6: Number of Characters in Real and Fake News.

Table 2 shows examples of Fake and Real news in the dataset.

Table 2: Fake and Real News Examples.

Label	**Text**
Fake	A judge has refused to toss out a federal lawsuit against the city of San Jose, California, in which the city police are accused of allowing an angry mob of left-wing agitators and violent protesters to target peaceful pro-Trump fans. Several people were injured during the June 2 rally: The city was sued by attorney Harmeet Dhillon, in a pro bono case, representing some of the victims. Dhillon, the national committeewoman of the California Republican Party and a contender to lead the Civil Rights Division of the Department of Justice, also attended the rally. Citizens ranging from their teens to their 70s were assaulted, abused, chased, hunted, and terrorized in a situation for which the city is responsible, and must now answer, Dhillon told LifeZette. This lawsuit seeks to vindicate the principle that every American, regardless of his or her political beliefs, is entitled to equal protection of the laws, and to the rights of free speech and free assembly, particularly in the support of their candidate of choice. The unfortunate series of events happened just weeks after Republican businessman Donald Trump had sewn up the GOP presidential nomination in May 2016. The Trump supporters were leaving the San Jose rally, exiting the convention center on June 2.
Real	An earthquake of magnitude 6.2 struck southeast of the town of Mat as Romero in Oaxaca, Mexico, on Saturday, the U.S. Geological Survey (USGS) said. Slight quake tremors were felt, and seismic alarms sounded on Saturday in Mexico City, which earlier this week was hit by the country's most deadly earthquake in decades. That 7.1 magnitude earthquake destroyed more than 50 buildings in the sprawling Mexican capital on Tuesday, leaving thousands homeless and close to 300 people dead nationwide.

Figure 7 provides further insights into the word counts of real and fake news articles. The data indicates that the majority articles have word counts below 300, while most fake news articles have word counts below 500.

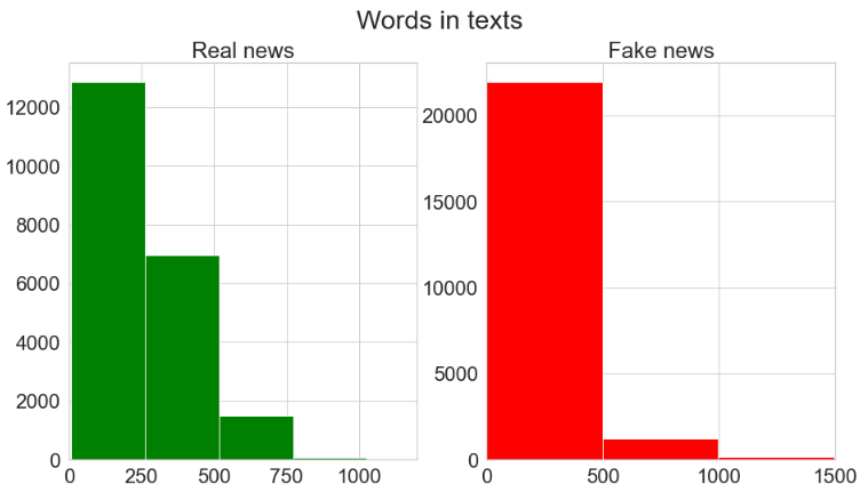

Figure 7: Number of Words in Real and Fake News.

Figure 8 provides insights into the distribution of word lengths in both real and fake news texts. The figure shows that most words in both categories have lengths ranging from 6 to 8 characters. This finding suggests a similarity in the distribution of word lengths between real and fake news. However, an interesting distinction emerges when examining words with lengths exceeding 15 characters. In the case of fake news, there are examples of words with lengths greater than 15 characters, whereas such occurrences are absent or rare in real news texts.

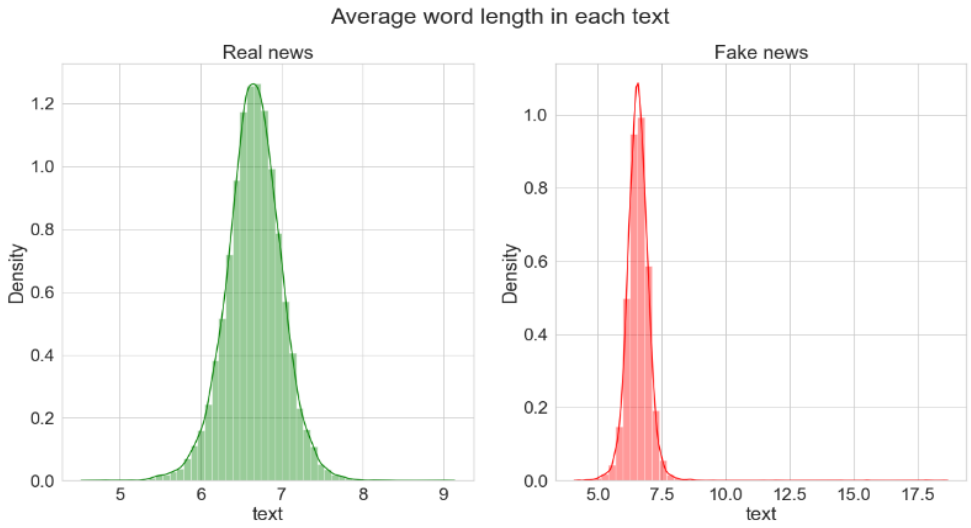

Figure 8: Number of Words in Real and Fake News.

These results suggest that true articles are often shorter, which may indicate shorter reporting or compliance with journalistic standards. Fake news, on the other hand, often uses longer text with more characters and word lengths, perhaps to present a more complex or compelling story. There are several reasons why fake news texts tend to use longer words. Intentionally using complex language or techniques to convey an air of authority or knowledge can be seen as a deliberate tactic. Alternatively, it may be a deliberate attempt to mislead or confuse the reader by using obscure or difficult to verify terms.

2.3.1. Most common words

N-Gram is a n-character chunk taken from a string. N-Gram is used in the process of making a model by dividing a sentence into parts of words. In N-Gram, 'N' shows the number of words that will be grouped into one section. In this paper, we have divided it into three types, namely:

- Unigram: token consisting of only one word.
- Bigram: a token consisting of two words.
- Trigram: a token consisting of three words.

Table 3 provides an analysis of the most frequently used words in the given text. The table presents the top 10 words along with the corresponding count of their occurrences. These words offer insights into the prominent themes and subjects discussed in the text.

Table 3: Top 10 Most Common Words.

Word	**Count**
Trump	113 014
said	93 480
would	55 354
U. S.	50 440
president	47 216
people	35 411
one	34 023
new	34 909
also	30 689
Donald	27 942

In Figure 9, you can find a representation of the top 10 bigrams in the text along with their corresponding number of occurrences.

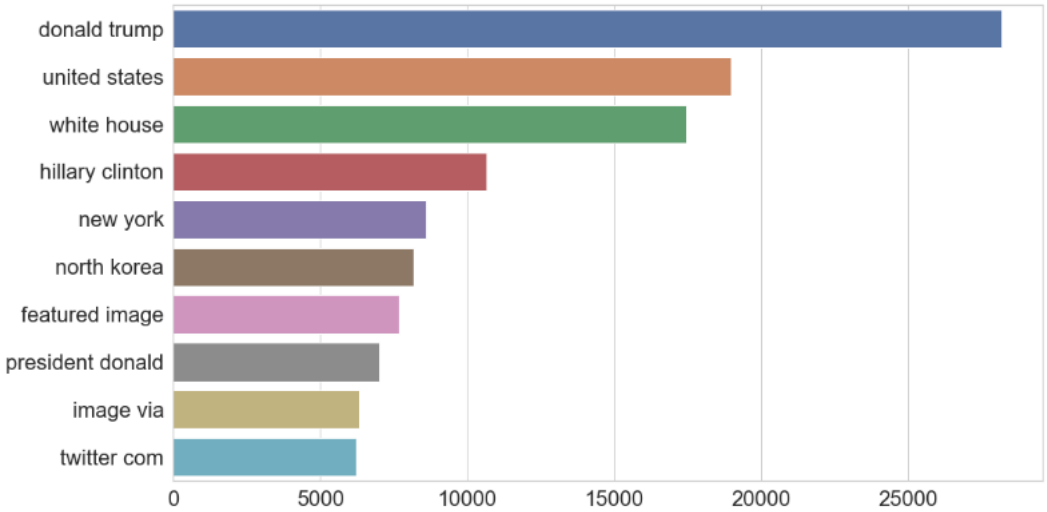

Figure 9: Top 10 Most Common Bigrams in Text.

Figure 10 presents a visual representation of the top 10 trigrams found in the text, accompanied by their respective number of occurrences.

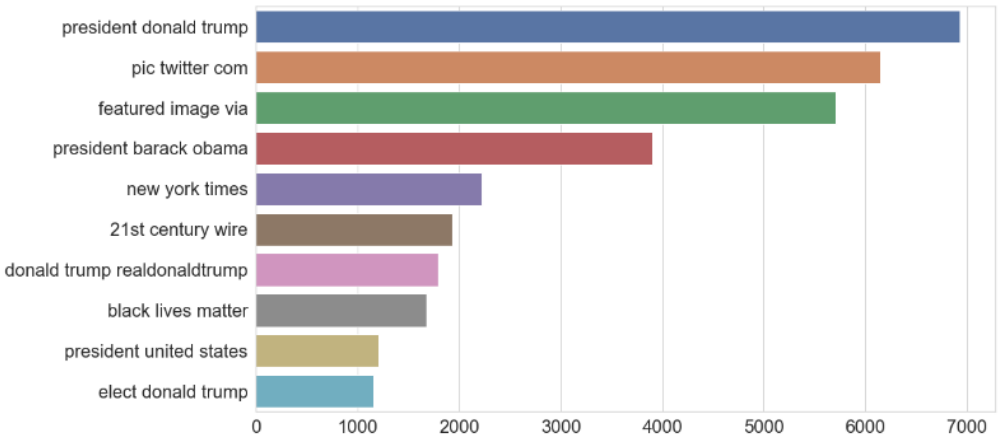

Figure 10: Top 10 most common trigrams in text.

Word clouds are a straightforward and easy-to-understand method of visually representing text documents. They serve as an initial glimpse of the most common words within the text, displayed as a list of weighted words arranged in a particular spatial layout (e.g., sequential, circular, random) [27]. The resulting image typically consists of words displayed in different sizes, colors, and orientations. The larger the size of a word, the more frequently it occurs in the text. Figures 11 and 12 display the WordCloud plot generated from the real and fake news text data.

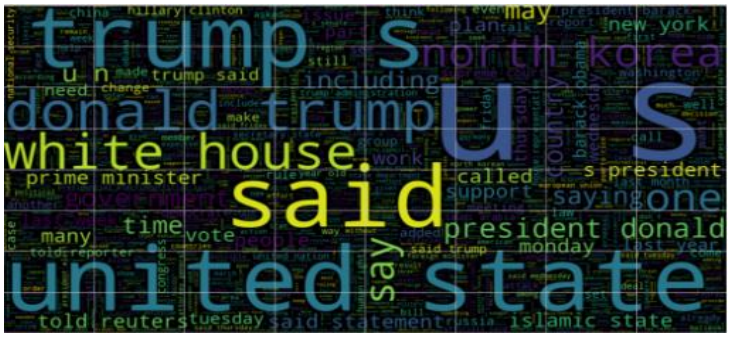

Figure 11: WordCloud Representation of Real News Text.

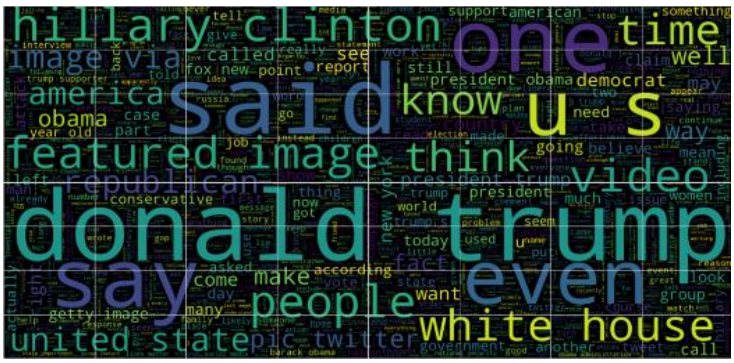

Figure 12: WordCloud Representation of Fake News Text.

Upon analyzing these two figures, it becomes evident that both fake and real news texts share some similar most common words. Notably, words like “donald”, “said” and “trump” appear prominently in both WordCloud images. However, specific to real news text, additional frequent words include “united states”, “white house” and “u s.” In contrast, the fake news text also contains distinctive common words like “hillary clinton”, “one”, “featured image” and “people.”

3. Results

Results will be evaluated using confusion matrix, accuracy, F1 score, sensitivity (recall), specificity and precision. These are metrics used for classification problems.

- **Confusion matrix** for a binary classifier is very common in evaluating results for classification problems. Actual values are marked True (1) and False (0) and are predicted as Positive (1) and Negative (0). Estimates of the possibilities of classification models are derived from the expressions TP, TN, FP, FN, which exist in the confusion matrix [28]. How the confusion matrix looks like, we can see in Table 4.

Table 4: Top 10 Most Common Words.

Class designation		**Actual Class**	
		True (1)	**False (0)**
Predicted Class	**Positive (1)**	**TP**	**FN**
	Negative (0)	**FP**	**TN**

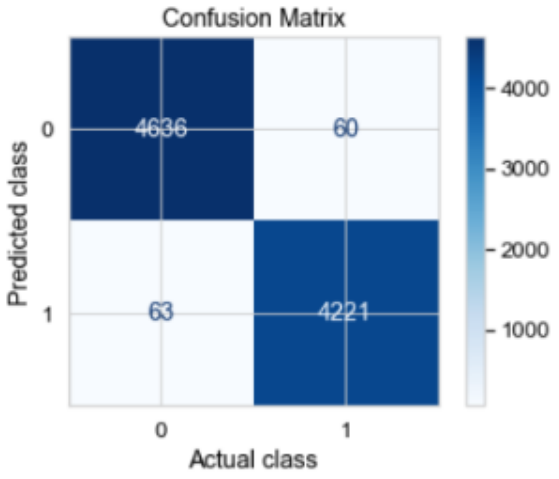

Figure 13: Confusion matrix for the logistic regression model.

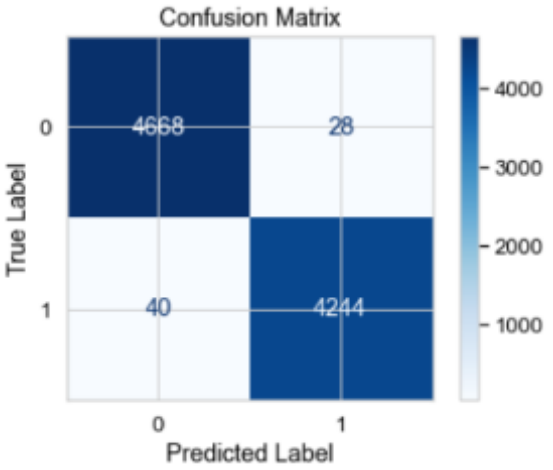

Figure 14: Confusion matrix for the neural network model.

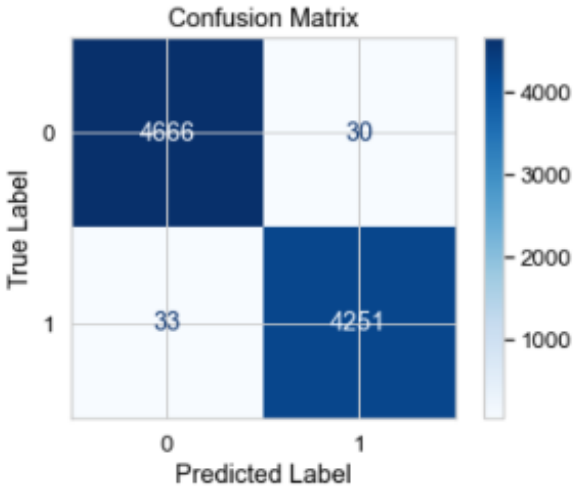

Figure 15: Confusion matrix for the support vector machine model.

In Figures 13, 14 and 15 we can observe that the logistic regression model predicted 4636 true positives, 60 false negatives, 63 false positives, and 4221 true negatives; the neural network model predicted 4668 true positives, 28 false negatives, 40 false positives, and 4244 true negatives; and the SVM model predicted 4666 true positives, 30 false negatives, 33 false positives, and 4251 true negatives.

- **Accuracy** is calculated as the sum of two accurate predictions (TP + TN) divided by the total number of data sets (P + N). The best accuracy is 1.0, and the worst is 0.00 [28].
- **F1 score** is a measure of the accuracy of the test. It is calculated, based on precision and reminders, by the formula: [29] 2 × Precision × Recall / (Precision + Recall)
- **Sensitivity (recall)** is calculated as the number of accurate positive predictions (TP) divided by the total number of positive (P). Also called Sensitivity or Recall (REC). The best TP Rate is 1.0 and the worst 0.0 [29, 30].
- **Specificity** is calculated as the number of correct negative predictions (TN) divided by the total number of negatives (N). The best specificity is 1.0 and the worst 0.0 [29].

- **Precision** is calculated as the number of correct positive predictions (TP), divided by the total number of positive predictions (TP + FP). The best precision is 1.0 and the worst 0.0 [30].

3.1. Models Comparison

In this section, we will compare the results of three different machine learning algorithms: logistic regression, support vector machine (SVM), and neural network.

Table 5: Top 10 Most Common Words.

SCORE	**VALUE**		
	Logistic regression	**SVM**	**Neural network**
Accuracy	0. 986	0.993	0.992
Sensitivity	0.985	0.992	0.991
Specificity	0.986	0.993	0.993
F1	0.986	0.996	0.992
Precision	0.987	0.994	0.994

All three algorithms exhibited impressive performance in our evaluation, as demonstrated by their high accuracy scores, as summarized in Table 5, and visualized in Figure 16. Specifically, Logistic Regression achieved an accuracy score of 0.986, SVM surpassed it with an accuracy of 0.993, and the Neural Network closely followed with an accuracy of 0.992. Notably, SVM emerged as the top-performing algorithm in terms of accuracy. When assessing sensitivity, which measures the algorithm's ability to correctly identify positive cases, SVM once again excelled, achieving the highest score of 0.992. The Neural Network closely trailed behind with a sensitivity score of 0.991, and Logistic Regression exhibited respectable performance with a sensitivity score of 0.985. Turning our attention to specificity, which gauges the algorithm's proficiency in accurately identifying negative cases, both SVM and the Neural Network achieved outstanding scores of 0.993, highlighting their capacity to discern genuine instances effectively. Meanwhile, Logistic Regression, while still commendable, displayed a slightly lower specificity score of 0.986. The F1 score, which balances precision and recall, further substantiates the superiority of SVM and the Neural Network, as both achieved exceptional scores of 0.993. Logistic Regression, while still performing well, recorded a slightly lower F1 score of 0.986. Lastly, in terms of precision, which indicates the algorithm's ability to correctly classify positive cases, SVM and the Neural Network maintained their dominance with precision scores of 0.994, while Logistic Regression achieved a slightly lower precision score of 0.987. In summary, our comprehensive evaluation revealed that SVM and the Neural Network consistently outperformed Logistic Regression across all critical evaluation metrics. SVM emerged as the top-performing algorithm, excelling in accuracy, sensitivity, specificity, F1 score, and precision.

The Neural Network, while slightly below SVM, still exhibited superior performance compared to Logistic Regression. These findings underscore the suitability of SVM and the Neural Network for the specific classification task at hand, affirming their efficacy in achieving robust and accurate results.

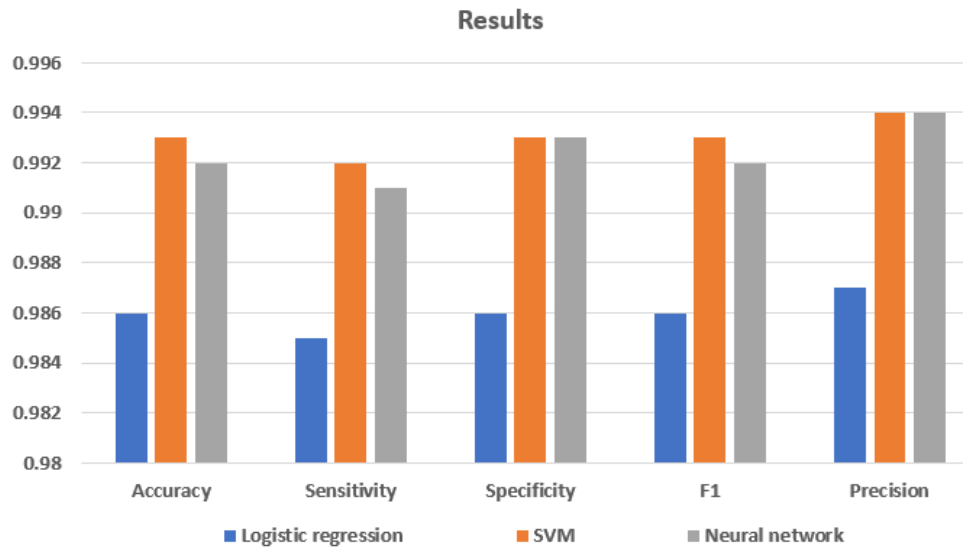

Figure 16: Results comparison visualization.

4. Conclusion

Fake news has become a significant issue in today's society, and detecting it is a challenging task. Machine learning algorithms, such as Support Vector Machines, Naive Bayes, and Neural Networks, have shown promising results in identifying fake news from real news. These algorithms use different approaches to learn patterns and features that distinguish between real and fake news, ranging from simple rule-based models to complex deep learning models. Overall, the performance of machine learning algorithms for fake news detection depends on various factors, such as the quality and size of the training data, the choice of features, the model architecture and hyperparameters, and the evaluation metrics used. Therefore, it is important to carefully evaluate and compare different algorithms and approaches to develop robust and effective fake news detection systems. In this study, multiple machine learning models, including SVM, logistic regression, and neural network, were used to address the fake news detection problem. The results obtained from each model demonstrated satisfactory performance in accurately distinguishing between fake and genuine news articles. While all models showed promising results, one model stood out as the best performer based on the evaluation metrics. Among the three models, Support Vector Machine (SVM) emerged as the most effective in terms of performance metrics. It achieved an impressive accuracy of 0.993, indicating a high proportion of correct predictions overall. The sensitivity of 0.992 highlights its ability to correctly identify a vast majority of true positive cases (correctly detecting fake news). The specificity of 0.993 signifies its capacity to accurately identify true negative cases (correctly recognizing genuine news). Furthermore, the f1-score of 0.993, which is a harmonic mean of precision and recall, demonstrates a balanced performance in capturing both true positives and true negatives. Lastly, the precision of 0.994 indicates the model's accuracy in correctly classifying fake news instances. While SVM stood out as the best model, it is essential to acknowledge that logistic regression and neural networks also yielded favorable results. Although slightly lower than SVM, the performance metrics of these models were still very good, indicating their competence in fake news detection. These positive outcomes are encouraging and validate the effectiveness of machine learning approaches in tackling the fake

news problem. The strong performance of the models demonstrates the potential of employing advanced computational methods in identifying false information and promoting more trustworthy information dissemination.

4.1. Pros and Cons

This study on fake news detection holds significant potential to benefit various key groups and society. The positive impacts include:

- Individuals and Social Media Users: More accurate fake news detection mechanisms can protect individuals from the harmful consequences of false information, promoting a safer online environment and fostering trust in social media platforms.
- Society and Democratic Processes: Improving fake news detection contributes to upholding the integrity of democratic discourse by providing reliable and credible information for informed decision-making.
- Economy and Business: By reducing the impact of unreal fake stories, the study can foster a more stable and transparent business environment, enhancing investor confidence and market efficiency.
- Media and Journalism: Effective fake news detection aids in distinguishing reputable journalism from misinformation, bolstering public trust in the media.
- Technology Companies and Platforms: Incorporating robust fake news detection mechanisms can create a more trustworthy user experience on social media platforms, mitigating the spread of false information.
- Policy-Makers and Government: The study's findings can help policymakers develop regulations and policies that promote responsible information dissemination and protect citizens from misinformation.

Machine learning-based solutions undoubtedly encounter several limitations when it comes to the task of detecting fake news, and two of the most prominent challenges revolve around the acquisition of a sizable and diverse training dataset and the selection of appropriate features that effectively capture the nuances of deception.

- Large Training Dataset: Machine learning models heavily rely on data, and the availability of an extensive and varied training dataset is pivotal to their efficacy. In the realm of fake news detection, it is imperative to amass a substantial repository of labeled examples, encompassing both fake and legitimate news articles. However, this task is far from straightforward due to the labor-intensive and time-consuming nature of manually classifying news articles as either fake or authentic. Moreover, ensuring diversity within the dataset is paramount, spanning a wide spectrum of topics, sources, and writing styles. This diversity is essential for crafting a robust and adaptable model that can effectively identify fake news across various domains and contexts.
- Feature Selection: Another crucial aspect of building effective machine learning models is feature selection. Features are the measurable attributes or characteristics of the data that the machine learning model employs to make predictions. In fake news detection, the process of selecting the most pertinent

and informative features that can capture the deceptive attributes of false information is of utmost importance. However, this task is not without complexity. Determining which features are the most discriminatory and generalizable can be intricate, given that fake news can manifest in a multitude of forms and may exhibit subtle distinctions when compared to genuine news.

These limitations collectively present substantial hurdles in the development of precise and dependable fake news detection systems. Despite the formidable nature of these challenges, it is important to note that ongoing research and continuous innovation are actively addressing these limitations. Researchers and technologists are working tirelessly to devise novel approaches and methodologies to enhance the accuracy and reliability of machine learning-based fake news detection systems. In the future, there are still many challenges and opportunities in fake news detection, such as dealing with evolving and adaptive fake news, identifying deepfakes and manipulated images and videos, and addressing the ethical and social implications of automated fake news detection. Therefore, researchers and practitioners need to continue developing and improving machine learning algorithms and techniques for fake news detection, as well as considering interdisciplinary approaches that involve human experts, social and behavioral sciences, and media literacy education. In conclusion, the potential benefits of this study are far-reaching and multi-faceted. From individuals to society, businesses, media, technology companies, and policymakers, the improved detection of fake news can lead to a more informed, secure, and trustworthy information ecosystem. By countering the harmful effects of false information, this research contributes to creating a healthier and more responsible digital society.

References

[1] How Many People Use Social Media in 2023? (65+ Statistics).

[2] M. Pengue, “How Many People Get Their News From Social Media in 2023?,” Letter.ly, February 2021.

[3] B. Marr, “How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read,” July 2021.

[4] J. Tandoc, Z. W. Lim and R. Ling, “Defining ‘Fake News.’,” Digital Journalism, August 2017.

[5] Fake News, 2017.

[6] K. Shu, A. Sliva, S. Wang, J. Tang and H. Liu, “Fake News Detection on Social Media,” ACM SIGKDD Explorations Newsletter, 2017.

[7] Y. Chen, N. J. Conroy and V. L. Rubin, “Misleading Online Content,” in Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection, 2015.

[8] D. Warkentin, M. Woodworth, J. T. Hancock and N. Cormier, “Warrants and Deception in Computer Mediated Communication,” in Proceedings of the 2010 ACM Conference on Computer Supported

Cooperative Work - CSCW ’10, 2010.

[9] C. L. Toma and J. T. Hancock, “Looks and Lies: The Role of Physical Attractiveness in Online Dating Self-Presentation and Deception,” Communication Research, 2010.

[10] L. Zhang and Y. Guan, “Detecting Click Fraud in Pay-Per-Click Streams of Online Advertising Networks,” in 2008 The 28th International Conference on Distributed Computing Systems, 2008.

[11] M. L. Newman, J. W. Pennebaker, D. S. Berry and J. M. Richards, “Lying Words: Predicting Deception from Linguistic Styles,” Personality and Social Psychology Bulletin, 2003.

[12] T. Qin, J. K. Burgoon, J. P. Blair and J. F. Nunamaker, “Modality Effects in Deception Detection and Applications in Automatic-Deception-Detection,” in Proceedings of the 38th Annual Hawaii International Conference on System Sciences.

[13] W. Shafqat, S. Lee, S. Malik and H.-C. Kim, “The Language of Deceivers,” in Proceedings of the 25th International Conference Companion on World Wide Web - WWW ’16 Companion, 2016.

[14] D. Glez-Peña, A. Lourenço, H. López-Fernández, M. Reboiro-Jato and F. Fdez-Riverola, “Web scraping technologies in an API world,” Briefings in bioinformatics, vol. 15, pp. 788-797, 2014.

[15] A. Field, Discovering Statistics Using IBM SPSS Statistics, SAGE, 2013.

[16] Pandas Documentation — Pandas 1.5.1 Documentation.

[17] Seaborn: Statistical Data Visualization — Seaborn 0.12.1 Documentation.

[18] K. Rahman, Python Data Visualization Essentials Guide: Become a Data Visualization Expert by Building Strong Proficiency in Pandas, Matplotlib, Seaborn, Plotly, Numpy, and Bokeh (English Edition), BPB Publications, 2021.

[19] B. Mahesh, “Machine Learning Algorithms-a Review,” International Journal of Science and Research (IJSR), 2020.

[20] T. Oladipupo, “Types of Machine Learning Algorithms,” New Advances in Machine Learning, 2010.

[21] W. S. Noble, “What Is a Support Vector Machine?,” Nature Biotechnology, 2006.

[22] M. S. Kawsar, “Machine Learning Quiz 03: Support Vector Machine,” Medium, March 2021.

[23] S. Sperandei, “Understanding Logistic Regression Analysis,” Biochemia Medica: Casopis Hrvatskoga Drustva Medicinskih Biokemicara / HDMB, vol. 24, p. 12–18, 2014.

[24] Logistic Regression.

[25] M. M. Cruz-Cunha, Handbook of Research on ICTs and Management Systems for Improving Efficiency in Healthcare and Social Care, IGI Global, 2013.

[26] S.-C. Wang, “Artificial Neural Network,” in Interdisciplinary Computing in Java Programming, Springer US, 2003, p. 81–100.

[27] S. Lohmann, F. Heimerl, F. Bopp, M. Burch and T. Ertl, “Concentri Cloud: Word Cloud Visualization for Multiple Text Documents,” in 2015 19th International Conference on Information Visualisation, 2015.

[28] M. Hossin, M. Hossin and M. N. Sulaiman, “A Review on Evaluation Metrics for Data Classification Evaluations,” International Journal of Data Mining & Knowledge Management Process, 2015.

[29] A. Tharwat, “Classification Assessment Methods,” Applied Computing and Informatics, 2021.

[30] Ž. Đ. Vujovic, “Classification Model Evaluation Metrics,” International Journal of Advanced Computer Science and Applications.