Journal of Computer Science

Original Research Paper

Improved SVM with Hyperparameter Tuning for Fake News Detection

Atcharaporn Nachaithong and Kittipol Wisaeng

Department of Technology and Business Information System Unit, Mahasarakham Business School, Mahasarakham University, Thailand

Article history
Received: 04-02-2024
Revised: 17-03-2024
Accepted: 30-03-2024

Corresponding Author:
Kittipol Wisaeng
Department of Technology and Business Information System Unit, Mahasarakham Business School, Mahasarakham University, Thailand
Email: kittipol.w@acc.msu.ac.th

Abstract: In today's digital age, accessing information has become effortless. An abundance of resources is available online, from trustworthy news outlets providing factual information to unverified opinions shared by anonymous individuals. With the advent of modern technology, social media platforms have revolutionized interaction and staying informed, providing instant access to news and information related to a wide range of topics. They also allow us to share valuable links and content that we find interesting or informative and express our thoughts and beliefs on various issues. However, knowing if the information you see is true or fake can be challenging. This study introduces an improved SVM with hyperparameter tuning for detecting fake news on the Twitter dataset. The proposed has two phases: Check-worthiness identification and fact-checking, which include three tasks: Feature selection, fake news detection and determining whether claims within tweets are factual. The main idea for tackling complex optimization problems is to transform them into more straightforward linear or quadratic programming problems. This transformation is made possible by approximating the Gaussian kernel using Epanechnikov kernels. The process involves selecting an optimal probability distribution from a set of choices and using the minimax strategy to construct the most effective separating functions. The approach is a highly efficient and effective way of addressing optimization problems that are too complex to solve through direct methods. According to the results, the proposed method has been able to identify fake news with accuracy, precision, recall and F-measure of 99.67, 99.61, 100 and 99.81%, respectively. This framework is a game-changer in the fight against misinformation, as it allows the classification of recurring fake news and the utilization of social network users' connections to prevent the spread of false information.

Keywords: Fake News, Support Vector Machines, Optimal Feature Selection

Introduction

Nowadays, the internet has become the most significant source of information for individuals, particularly when making commercial decisions. The availability of online news provides a vast amount of information about various entities, events, opinions, and sentiments that are relevant to business activities, making it a crucial resource for consumers (Nozza *et al*., 2021; Hogenboom *et al*., 2016; Lee *et al*., 2020; Chan and Chong, 2017). However, the rise of fake news has created significant challenges for people who rely on online news to guide their judgments. The problem with fake news is that it can distort people's perception of the truth and impact their connections with real information. This, in turn, can have severe implications on public opinions, interests, and even decisions (Talwar *et al*., 2020). Consequently, it is key to exercise caution and carefully evaluate the reliability and accuracy of the information presented. Making decisions based on unverified or false news can lead to grave consequences, both for individuals and society as a whole. The impact of fake news on people's daily lives is a matter of great concern. The proliferation of misinformation and deliberate falsehoods on social media platforms is a pressing issue that poses a significant threat to individuals and society. The spread of such fake news has the potential to mislead people, create panic and mistrust, and even cause harm in some cases. Addressing this problem effectively is essential to safeguarding people's trust and well-being (Phan *et al*., 2023). Over the past few years, the spread of fake news has emerged as a significant problem, particularly with the rise of social media. Fake news refers to deliberately false information that can be easily verified as untrue. This phenomenon poses a significant threat to democratic societies, as it undermines public trust in political institutions and has far-reaching impacts on crucial aspects of our society, including but not limited to elections, the economy, and public opinion (Capuano *et al*., 2023). Fact-checking has emerged as a popular technique for investigating and analyzing erroneous information.

Many projects have arisen to prevent the negative impacts of the spread of misleading information (Zhang and Ghorbani, 2020; Zhang *et al*., 2020). Identifying bogus news's text, image and other characteristics is crucial to effectively combat it and prevent its harmful effects (Bondielli and Marcelloni, 2019). Social media platforms have become increasingly popular and ubiquitous in today's digital age. They allow individuals to share news and information with others, leading to rapid word-of-mouth dissemination. Thanks to the digital technology that powers social media platforms, information sharing has become more accessible. As a result, social media has become a prevalent and accessible platform for people to access information worldwide. It has become many individuals' primary source of news, entertainment and communication. Nonetheless, the widespread sharing of information on social media has also given rise to a significant problem spreading false information. Fake news has become a global issue and it poses a severe threat to the credibility and reliability of social media. It can cause harm, create confusion and mislead people (Monti *et al*., 2019).

False information can cause many issues, as it can lead people to believe things that are not true and subsequently spread them further, creating a cycle of misinformation. Social media's emergence has changed how people consume news. However, it has also led to the spread of fake news, making it crucial for platforms to prevent its dissemination and ensure the authenticity of shared information. Determining the authenticity of a tweet's message is crucial in identifying whether it is fake news. As social media users, we are well aware of the prevalence of false news circulating on these platforms, which makes it challenging to distinguish between trustworthy sources and those who habitually publish unverified or fraudulent content (Sansonetti *et al*., 2020). As a result, it is crucial to exercise caution and discernment while using social media to influence how others perceive and believe the news. Failing to do so can have severe repercussions, particularly in cases where the news lacks specifics. There is a significant risk that something posted on social media could harm society, especially if the audience is small. Therefore, the abundance of online information has made it increasingly difficult to distinguish credible news sources from those lacking credibility. This challenge is of significant importance and demands attention to ensure that the public has access to trustworthy and accurate information. Awareness of this issue is crucial as it can create panic and chaos in society, leading to unnecessary harm and even loss of life. False reports of disasters and emergencies can be particularly dangerous as they can cause distress and confusion among people, making it difficult for them to take appropriate action during a crisis. In addition, misinformation about politics, safety, and national security can be spread intentionally by ill-wishers to cause division and unrest in society (Aphiwongsophon and Chongstitvatana, 2018). This can create a sense of distrust and suspicion among people, making it challenging to maintain social harmony and progress. Moreover, fake news producers often use various rhetorical techniques to simulate news and justify the distortion of reality (Shu *et al*., 2017). They may also support their assertions with erroneous citations of reliable sources, making it even more challenging to recognize bogus news. Consequently, it is important to distinguish between fabricated or misleading news and factual news that accurately represents reality. The need for reliable and accurate news has never been more critical than it is today. We must carefully examine the sources from which we receive news and verify their authenticity before accepting the information as true (Gupta *et al*., 2022). Throughout history, disseminating false information has been a persistent problem and the invention of fake news in the 15^{th} century marked a turning point in spreading disinformation. Today, fake news is more pervasive than ever and can be found on various platforms, including social media, email, radio, newspapers and television. In the current digital era, rumors and fake news proliferation have emerged as significant concerns. Social bots play a pivotal role in disseminating erroneous information, making the issue even more critical. The speed with which these misleading stories spread is alarming, making it challenging to control their impact and prevent them from causing harm. These bots are designed to spread disinformation intentionally and fabricate headlines, leading to the circulation of fake news articles. Fortunately, computer power and data processing advancements have brought about a revolutionary change in Artificial Intelligence (AI).

These advancements have enabled the development of sophisticated techniques to effectively classify and combat the spread of fake news. Advanced AI-powered tools can perform in-depth analyses of vast amounts of data from diverse sources. These tools can effectively identify fake news articles by leveraging their ability to detect patterns and anomalies. These techniques have shown positive results in tackling the problems above with false news and they offer hope for a future where accurate and reliable news is the norm (Rohera *et al*., 2022). The biggest source of fake news today is social media, where phony news websites, ads and messages that people see on their social media feeds have become a common tactic for spreading false information. This has led to widespread confusion and mistrust as people struggle to separate fact from fiction. Fake news has become a pervasive problem in our society, with its impact reaching far and wide. It can manipulate and influence public opinion on crucial issues, leading to profound consequences such as shaping the outcome of elections. Therefore, it is crucial to address this issue and ensure that people have access to accurate and reliable information. The detrimental effects of fake news can be observed in various situations, exemplified by a widely shared social media post that claimed three significant cities would be subjected to a military lockdown after May 31. This fabricated news caused immense pressure and anxiety among the public, resulting in panic buying of essential goods. One such incident involved a widely circulated rumor that Bill Gates was attempting to implant microchips into humans through a potential coronavirus vaccine. This baseless claim was spread across various social media platforms in Portuguese, leading to defamatory statements about Bill Gates and painting him as an evil mastermind (Coronavirus, 2020). The impact of fake news is not limited to individuals alone. It can have severe consequences on the economy as well. For instance, a false post on social media suggesting that the government would reduce pensions by 20% due to the pandemic led to widespread panic and anxiety among pensioners. This, in turn, had a detrimental effect on their mental health, causing unnecessary stress and worry (Mumbai *et al*., 2020). Furthermore, fake news can impact company stock returns, as some investors may be unable to distinguish whether the news is real or fake. This may cause conflict among investors and the target company's stock price may respond to fake news (Arcuri *et al*., 2023). Fake news has far-reaching consequences that affect many facets of human existence (Syed *et al*., 2023). From fake political news posts, expressions of public opposition have the effect of changing people's behavior, mainly seen in issues involving large populations (Leon *et al*., 2023). Additionally, the dissemination of false information regarding the COVID-19 pandemic through social media channels has had disastrous consequences, leading to the loss of countless innocent lives. The rampant posting of misleading news articles and inaccurate statistics has created confusion, panic and mistrust among the public, making it difficult for authorities to contain the virus's spread and provide accurate information to those in need. Therefore, we must exercise caution and diligence in verifying the authenticity of any information before sharing it with others. It has caused instability and social fear due to misinformation about COVID-19, affecting consumer behavior and resulting in product hoarding (Sarraf *et al*., 2024; Dwivedi *et al*., 2020; Kar *et al*., 2023; Naeem, 2021). Therefore, it is essential to be cautious while consuming news and information and to evaluate the sources of the information we receive carefully.

Machine learning is a powerful method for predicting, tracking and combating the spread of fake news. These studies have yielded astonishingly precise results, as machine learning algorithms have accurately identified fake news. By utilizing machine learning, we can effectively monitor and analyze the dissemination of false information across various online channels, including social media platforms. These techniques have proven valuable tools for tracking the dissemination of fake news and identifying potential sources of misinformation (Sansonetti *et al*., 2021; Jing *et al*., 2023; Hiramath and Deshpande, 2019; Neeraj *et al*., 2023). As a result, AI-powered solutions are increasingly being employed to combat misinformation and disinformation in today's digital age (Shu *et al*., 2019; Thakkar *et al*., 2019; Billones *et al*., 2022; Renuka and Anithaashri, 2022; Vadlamudi *et al*., 2023; Krishna and Adimoolam, 2022). This project aims to utilize the extensive volume of data being shared on Twitter to detect and combat the spread of false information online. The main goal is to create an automated fact-checking mechanism to identify and verify the most relevant and essential claims made within tweets. The ultimate goal is to provide a reliable source of information for the public by preventing the spread of fake news. With the help of machine learning, this system can quickly and accurately analyze the credibility of claims, thus reducing the potential harm caused by misleading information.

Related Work

In the past few years, the issue of fake news has increasingly raised concerns among people around the world. Professionals in various fields have conducted numerous research studies to develop effective strategies for detecting and identifying false information. However, a relatively new concept, detecting fake news, has garnered much interest and attention worldwide. Various methods have been presented for classifying bogus news across different data types. Machine learning approaches such as multiscale feature extraction, sentiment evolution, anomaly detection and sentiment reasoning have been widely used in various domains to tackle this issue. This chapter comprehensively summarizes current and relevant research on identifying false news. It is worth noting that while some social media users may be honest and sincere, others may be cunning, deceptive and unconventional. Therefore, it is crucial to maintain a cautious and critical approach while evaluating the information available on social media (Chen *et al*., 2022; Lin *et al*., 2022; Lu *et al*., 2023). By staying informed and aware, individuals can take steps to protect themselves from the adverse effects of fake news. Fake news is a significant issue today and can originate from various sources. The three primary sources are customers, social media and the internet. Fake news on social media platforms is rampant, and verifying social media accounts is essential to minimize the spread of fraudulent news. Social media has become a breeding ground for fake news and one primary method of spreading it is through social bots and computer programs on social networking websites. These bots can interact with users on social media and automatically generate content, which can be either helpful or harmful, depending on how they are programmed. Unfortunately, some social bots are specifically designed to spread false information and propaganda on social media platforms to cause harm. These bots can become a significant factor in disseminating bogus news, as they can be programmed to use sophisticated techniques to make their content appear authentic and trustworthy. This can make it challenging for users to distinguish between what is true and not, leading to confusion and misinformation. To counter this problem, several methods have been proposed in the literature, such as those discussed in references Xiong *et al*. (2023); Yang *et al*. (2019). These techniques aim to identify and flag potentially fake news stories that can be reviewed and removed. The network analysis approach is auspicious among the various approaches to classifying and preventing the spread of false information. This approach is based on identifying deceptive language cues and requires a large corpus of collective human details to accurately determine the veracity of novel claims. In contrast, the linguistic approach relies on identifying patterns in language use. One of the most effective ways to identify false information is to verify the accuracy of significant claims in news articles and evaluate the sincerity of the broadcast. This approach is crucial for developing fact-checking techniques and advancing the field of information verification. External references play a vital role in supporting projected statements made in news articles, providing factual meaning to arguments in a way that makes sense. Utilizing these effective strategies enables us to enhance our ability to recognize and counteract the dissemination of inaccurate or misleading information (Mishra *et al*., 2022). It is important to implement such techniques to ensure that the public receives accurate information and to prevent the spread of false narratives that could harm individuals or society.

Artificial Intelligence (AI) is a rapidly growing field based on machine learning principles. Machine learning involves using a variety of algorithms and techniques to analyze data and identify patterns. By doing so, machine learning can make predictions based on these patterns. One of the most fundamental techniques in machine learning is supervised learning, which consists of training an algorithm on a set of labeled data to recognize patterns and relationships. The algorithm uses the data's features and attributes to identify the most critical factors determining an outcome or prediction. The logistic regression method is widely used to estimate the probability of an event occurring based on the occurrence of another event.

This approach is used in several fields, such as finance, marketing and healthcare, to predict customer behavior, market trends and disease outcomes. Another popular supervised learning algorithm is SVM, which is highly effective for binary classification problems. The aim is to classify whether a new data point belongs to one of two classes. SVM works by finding a hyperplane that maximizes the distance between the two groups in the dataset, also known as the margin. The support vectors, the data points closest to the hyperplane, are crucial in defining it. SVM can handle high-dimensional data, making it suitable for various real-world systems. Additionally, it's memory-efficient, which means it can handle large datasets without much memory. Logistic regression and SVM are powerful machine-learning techniques that can help us better understand data patterns and predict outcomes more accurately. To address these limitations, researchers have developed various extensions of SVM, such as kernel SVM, which is a powerful technique for handling non-linear data. The kernel of the SVM algorithm is a technique for classifying data into two separate groups using a linear hyperplane. However, the success of this approach depends on selecting the proper kernel function, which plays a critical role in the algorithm's accuracy. Therefore, careful consideration should be given when selecting the kernel function to ensure optimal performance (Jain and Kasbe, 2018). Upon conducting a thorough analysis, we discovered that two models, namely the random forest and the Gradient boosting method, demonstrated an outstanding level of accuracy by achieving a score of approximately 97%. On the other hand, the Gaussian NB and Stochastic Gradient Descent (SGD) classifiers exhibited a considerably lower level of accuracy, with scores of only 84 and 81%, respectively (Ahmed *et al*., 2022). The different machine-learning techniques that can be used to detect malware in IoT-based enterprise information systems. The survey covers various approaches and comprehensively analyzes their effectiveness in identifying and mitigating malware threats (Gaurav *et al*., 2023). The ultimate goal is to help businesses safeguard their networks from possible cyber threats with cutting-edge malware detection methods. Fortunately, machine learning holds great promise in addressing this challenge. However, several key hurdles must be overcome to develop effective solutions. The first obstacle is accurately identifying false information within a large volume of data. Accomplishing this intricate task demands the utilization of advanced algorithms and a profound comprehension of natural language processing. The capacity to differentiate between genuine and counterfeit information can be influenced by various factors, such as the language used to express it, the context in which it is presented and the presence of misleading statements or propaganda. Moreover, preprocessing the

DOI: 10.3844/jcssp.2024.1357.1375

data is another crucial step that plays a pivotal role in determining the accuracy of the final model. Ensuring the model's accuracy is essential to achieve optimal results while dealing with large-scale data. This involves cleaning and transforming raw data, dealing with missing values and selecting relevant features for analysis. Quality data is crucial for the success of a machine learning model. Inaccuracies or prejudices in the data can significantly impact the model's effectiveness and accuracy of predictions. Finally, selecting suitable algorithms for feature selection and classification is another crucial aspect of developing effective machine-learning solutions for combating fake news in emails. Eliminating false information through email can be challenging and requires careful consideration of various factors. These factors include the dataset's size and complexity, the desired accuracy level and the computational resources available. To effectively tackle these challenges, a system can integrate several advanced machine learning algorithms. These algorithms can analyze and identify false information, enabling the system to prevent its spread through email. The suggested system employs state-of-the-art technology to ensure users can trust the information they receive without manual verification. This eliminates the need for time-consuming and error-prone manual verification, enhancing productivity and efficacy across all organizational levels. The system's comprehensive approach ensures superior efficacy in identifying and combating fake news, safeguarding corporate reputation and credibility.

The study proposes an improved approach for classifying and identifying fake news on Twitter. This study aims to improve the traits' effectiveness in categorizing objects. The study utilizes the SVM algorithm, known for its high accuracy and precision, to achieve this. The study results indicate that this proposed method performs significantly better than traditional approaches in terms of accuracy, precision, recall, and F-measure. The implications of this research are essential in identifying and combating the spread of fake or real news on the Twitter dataset. The study's findings hold significant promise in minimizing the harmful effects of misinformation on society. By improving the detection of fake news, the study aims to aid in the creation of a more informed and engaged society. This research can help individuals distinguish between reliable and unreliable sources of information and promote a more positive and constructive digital space.

Materials and Methods

The study introduces a novel method for detecting fake news on social media, focusing on Twitter. This framework accurately distinguishes between factual and non-factual information, enabling users to easily identify and filter out unreliable news sources. As shown in Fig. 1, the framework is a comprehensive approach to identifying and detecting fake news claims present in tweets by leveraging machine learning methods. The framework is divided into two phases: Data preprocessing and fake news detection. In the first phase, data preprocessing involves cleaning the data and selecting the parameters contributing to the tweet content's checkworthiness (task 1). This stage aims to enhance the accuracy of identifying fake news from the tweet content. In the second phase, the fake news detection phase, the framework determines the factual status of the fake news claims within tweets (task 2). This is done by utilizing existing datasets based on machine learning algorithms (NB, SVM, DT, RF and LR). Task 3 involves collecting data from various sources, including the Twitter platform (Kaggle, data word and UCI dataset), to enhance the accuracy of detecting fake news on Twitter.

Once the two phases are completed, the framework returns experimental results. The fake news claims within tweets are labeled as factual or non-factual for model evaluation with accuracy, precision, recall and F-measure scores. The study hypothesizes that the performance of the machine learning models with various datasets can be compared to determine the SVM with the polynomial kernel that accurately detects fake news more highly than other methods (task 2). Additionally, the study hypothesizes that the SVM with polynomial kernel trained and tested on the Kaggle dataset has the best efficacy in identifying fake news on Twitter (task 3). Finally, by combining all components with preprocessing and selecting suitable algorithms, the study hypothesizes that good detection performance can be achieved, surpassing state-of-the-art techniques in an end-to-end fashion.

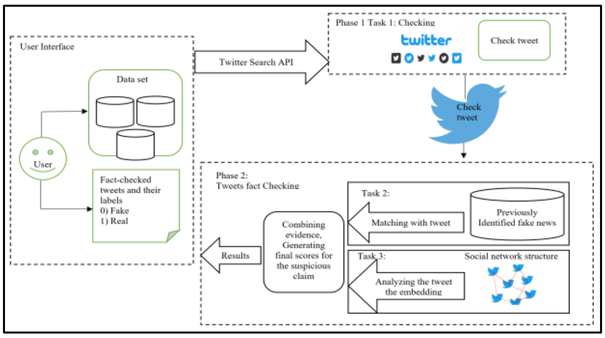

Fig. 1: The framework is a comprehensive approach to identifying and detecting fake news claims present in tweets by leveraging machine learning methods

Dataset

Numerous open-source fake news databases are available today, each with its classification systems. Some of the most popular databases include the UCI machine learning repository, Data World, subreddit/datasets R package-DS labs and Kaggle. Researchers generally use these databases to create models to help identify fake news. Researchers often scrape data from websites and other sources to create these models. The internet is flooded with fake news and it can be challenging to differentiate between real and fake news. The internet news liar, liar pants on fire dataset is a widely used fake news database, which contains a whopping 12.8 K short news items known as LIAR. This dataset comprises fake news from various sources, including Twitter, buzz feed news and weir-do (Ali *et al*., 2022; Sansonetti *et al*., 2021; Atodiresei *et al*., 2018; Nadikattu, 2023). Several benchmarked datasets have been developed to detect fake news and one such dataset is real and fake. These datasets consist of labeled news data covering a range of topics, focusing on social and political subjects. In addition, Kaggle has been a valuable source of news datasets used in various research papers to determine news specifications. Despite this, access to the first dataset specifically created for news on social context is restricted and contains only 344 labeled news retrieved from the internet. Although some datasets are available, such as those related to fake or real news and Twitter, the need for a comprehensive dataset covering a wide range of domains and scenarios still needs to be addressed. Such a dataset can significantly assist researchers and developers in creating and testing AI-powered systems that can solve real-world problems. For this study, multiple sources were utilized to gather news information on fake news, with frequently used key terms related to social context news, political news and general news, chosen with input from notable personalities, cities and lawmakers. The process of gathering data for fake news identification involves the use of various sources. One of the most widely recognized data science communities, Kaggle, is a significant source of information. Kaggle is known for its support of multiple data science objectives, including the identification of fake news. The platform offers numerous problems and modern datasets for detecting fake news, which provides a valuable resource to researchers in this field. In this regard, a developed dataset containing news entries concerning Pakistan was created. This dataset collects news information for both actual and fake news classes. This study utilized three datasets to assess the effectiveness of classifiers in classifying fake news. These datasets were carefully selected to ensure a comprehensive evaluation of the classifiers' performance. These datasets were obtained from Kaggle, Data World and the machine learning repository (UCI) library. Table 2 compiles the datasets used in the study. To evaluate the accuracy, we have split the dataset in a 70:30 ratio, where 70% of the data will be used for training and the remaining 30% will be used for testing the accuracy of the classifier. The testing data was used to adjust the model's hyperparameters. Additionally, fake news tweets from users on Twitter were collected over three years, from 2018-2020, as shown in Table 1. The datasets mentioned here contain crucial information that can help us better understand how well classifiers can differentiate between real and fake news.

Data Pre-Processing

Before conducting any data analysis, a series of crucial data preprocessing procedures were meticulously carried out to ensure the accuracy and reliability of the results. Special attention was paid to eliminating missing data points, which involved carefully examining every data record and removing any instances with incomplete or inconsistent information. Furthermore, the dataset was meticulously separated into distinct training and testing subsets, enabling the creation and verification of forecasting models on autonomous data samples. These procedures were carried out with great care and attention to detail to ensure that the resulting analyses and models were robust, reliable and trustworthy.

Table 1: Statistics of developed datasets

Datasets	Sample size	Training data (70%)	Testing data (30%)
Kaggle	20,717	14,502	6,215
Data word	19,530	13,671	5,859
UCI machine learning repository	21,248	14,875	6,373

Table 2: Methods for identifying false information

Reference	Year	The algorithms utilized in the research
Gebremariam *et al*. (2023)	2019	DT and SVM techniques
El-Azab *et al*. (2023)	2019	ML model by using SVM
Rasool *et al*. (2019)	2020	NB classifier
Jain *et al*. (2019)	2020	RF, SVM, LR, DT, gradient boosting and XG-boost
Adiba *et al*. (2020)	2020	SVM, CNN, LSTM, KNN and NB
Smitha and Bharath (2020)	2021	DT, RF and extra tree classifier
Agarwal and Dixit (2020)	2022	SVM
Hakak *et al.* (2021)	2022	SVM
Patel *et al.* (2021)	2022	NB, SVM, Neural Networks (NN) and Long Short-Term Memory (LSTM)
Rajeswari (2022)	2022	SVM
Kocherlakota *et al*. (2022)	2023	RF, DT, SVM and Hidden Markov Models (HMM)

Data Cleaning

Data cleaning is an essential stage in data mining that involves classifying and removing incomplete and inaccurate data. This process is important to ensure the data is accurate, consistent and error-free. To achieve this, several methods are carried out during data cleaning. One of the first processes in data cleaning is eliminating duplicate records (Sultana *et al*., 2023; Wang *et al*., 2023; Wang and Wang, 2020). In many cases, duplicate entries with the exact news text and label can be found in the data collected from various sources. Since the news labels from each source are the same, there is no contradiction in the duplicate entries. Eliminating duplicate records in the initial data cleaning stage is essential to ensuring data accuracy and consistency. This process involves identifying and removing identical copies of records to avoid redundancy, confusion and potential errors in data analysis. This ensures that the dataset only contains unique new data records. Another necessary process in data cleaning is the removal of any missing information. The process of detecting fake news involves several stages of data cleaning. News content is one of the most critical components of this process. Since the news content serves as the primary basis for identifying fake news, any news records that do not contain news text are usually disregarded in the next stage of the data cleaning process. This ensures that the dataset only includes complete and original text news data. In conclusion, data cleansing is a crucial step toward achieving high-quality data and ensuring that it is consistent, error-free and reliable. The dataset becomes reliable and suitable for analysis and decision-making purposes by eliminating duplicate records and removing any missing information.

Feature Selection

After preprocessing, the next phase features extraction. This involves reducing the dimension of the data by eliminating extra features, which makes it easier to classify text messages. The authors employ two well-liked feature selection techniques in this research: Correlation-Based Feature Selection (CBFS) (Singh and Singh, 2020; 2022). The CBFS method involves selecting features highly correlated with the class variable. This method is advantageous because it is computationally efficient and can handle many features. On the other hand, the other method used in this research, which should be mentioned in the original text, is the mutual information feature selection method. This method selects features with a high mutual information score with the class variable. Overall, these feature selection techniques are essential in reducing the dimensionality of the data, which improves the efficiency and accuracy of the classification methods.

Machine Learning Techniques

Machine learning techniques can simulate complex and ambiguous systems, even when non-linear relationships are not well-defined. A recent study used five distinct machine-learning methods to build a model and compare their effectiveness in handling such systems. These methods included NB, SVM, DT, RF and LR. The study also provides comprehensive information on how to improve the settings for each algorithm, which can help identify fake news (Qu *et al*., 2023; Odhiambo *et al*., 2021). Table 2 presents crucial information that can significantly improve our capacity to identify and curb the proliferation of false information. The recent studies discussed in the table have primarily concentrated on leveraging advanced machine-learning methods to detect fake news. Specifically, some researchers have proposed a highly effective Machine Learning (ML) model that employs Support Vector Machines (SVM) to identify and flag false news stories as documented in Yogendra *et al.* (2022). On the other hand, several studies as mentioned in St *et al*. (2023); Sneha and Gangil (2019); Khaled *et al*. (2020); Birzhandi *et al.* (2022) have conducted a thorough analysis of seven distinct machine learning classifiers. The primary objective of these studies was to evaluate the accuracy of these classifiers in detecting fake news.

Naïve Bayes (NB)

The Naïve Bayes is a powerful and widely used prediction technique in machine learning that helps predict class membership for unidentified data sets. This algorithm is based on Bayes probability theory, which calculates the likelihood of a particular class based on the probability of its features. Unlike other algorithms, Naïve Bayes classifies each feature independently of the others, which assumes that one feature's presence or absence does not affect another's presence or absence. To obtain the posterior probability $P(A|B)$, the algorithm uses Bayes' theorem, a fundamental concept in probability theory. Bayes' Theorem is that the probability of A given B is proportional to the probability of B given A and the prior probability of A. This formula is used to calculate the posterior probability for each class. By analyzing the data and assigning probabilities to each class, the algorithm can accurately predict the membership of unknown data sets, making it an essential tool in medical diagnosis, spam filtering and data analysis. The equation mentioned is a necessary formula for Bayes' Theorem, widely used in probability theory. It helps to calculate the probability of an event A occurring, given that another event B has already occurred. The equation is as Eq. (1):

$$P(A|B) = P(B|A) * P(A) / P(B) \tag{1}$$

The theorem can be expressed in various forms. In Bayes's theorem, the denominator is a normalization factor, while the numerator represents the probability of observing evidence *B*, given that event *A* is true. This allows us to update our belief about the probability of *A* based on new evidence of *B*. This theorem has numerous applications in fields such as machine learning, data analysis and decision-making, as Eq. (2) (Garg and Program, 2013) provides another way to define the Bayes theorem that is equivalent to Eq. (1). $P(A)$ is the prior probability of *A* and $P(B)$ is the prior probability of *B*, representing the probability of occurrence before having any information about each other. $P(A|B)$, on the other hand, represents the posterior probability of *A*, which is the probability of *A* occurring after taking into account the information provided by *B*. Lastly, $P(B|A)$ represents the conditional probability of *B* given *A*, which is the probability of *B* occurring given that *A* has occurred:

$$P\left(A|B\right)=\frac{P\left(B|A\right)P\left(A\right)}{P\left(B|A\right)\times P\left(A\right)+P\left(B|\neg A\right)\times P\left(\neg A\right)} \quad (2)$$

$P(\neg A)$ is the probability of *A* being false and $P(B|\neg A)$ represents the probability of *B* given that *A* is false. A classification algorithm is a machine learning technique used to predict the class or category of a given data sample. The algorithm relies on the Bayes theorem and the independence assumption between the predictors to make these predictions. This technique is beneficial for predicting the class of unknown data samples by computing the probability of the class in the input data. It is considered an effective classification method that works well with big datasets. To help understand the process better, Fig. 2 displays the Naïve Bayesian Flowchart (Efe and Fadipe, 2023). The flowchart outlines the various steps that need to be followed to implement the Naïve Bayesian technique. By following these steps, users can make accurate predictions about the class of unknown data samples.

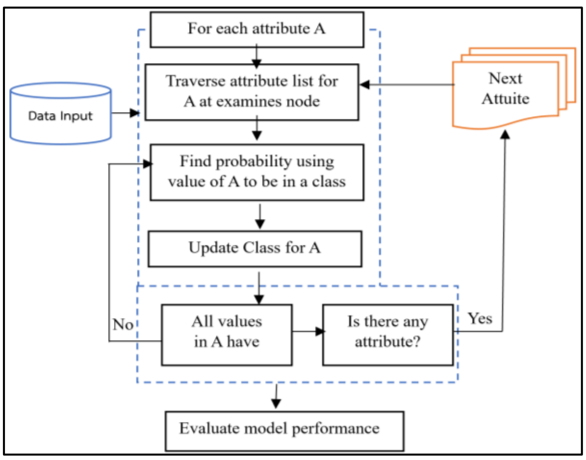

Fig. 2: Flow chat for Naïve bayes

Support Vector Machines (SVM)

SVM has emerged as a popular technique thanks to its high performance in various applications (Gereme and Zhu, 2019). SVMs are kernel-based models that can be used for identification and classification tasks. They have proven effective in multiple domains, such as text classification, statistics, pattern recognition and image processing. One of the significant advantages of SVMs is their ability to optimize the expected solution, resulting in an optimal solution that outperforms other supervised learning techniques. The data mining, pattern recognition and machine learning communities have shown great interest in SVM due to its exceptional generalization ability and discriminative capacity. SVM is particularly useful in resolving real-world binary classification problems, where it has been demonstrated to outperform other supervised learning techniques. SVM's theoretical underpinnings and strong generalization ability have made it one of today's most popular classification techniques. Its ability to learn from a limited training data set and produce accurate results on unseen data makes it a valuable tool in many applications. SVM is a robust algorithm for two-class classification problems. SVM finds an optimal hyperplane that maximizes the margin between two classes of data points. This helps generalize unseen data well and handle non-linearly separable datasets using kernel functions. SVM aims to maximize this margin, leading to better generalization performance and reducing overfitting. To illustrate the concept, let's consider a dataset $\{(x^{(i)}, y^{(i)})\}$ where *xi* is a vector of input features and $y^{(i)}$ is the corresponding binary label (+1 or -1) indicating the class membership. We assume that the dataset is linearly separable, which means a hyperplane exists that can perfectly separate the two classes. The hyperplane is defined by the equation $w^T x^{(i)} + \beta = 0$, where w is a vector of weights and β is the bias term. A vector called "*w*" determines the orientation of this hyperplane, while the bias term "*b*" controls its position in the feature space. The objective of SVM is to obtain the best possible values of *w* and β that maximize the margin between the two classes. This is achieved by solving an optimization problem represented by Eq. (3):

$$minimize\left\|w\right\|, such\ that\ y^{(i)}\left(w^T * x^{(i)} + \beta\right) ^3 1 \quad (3)$$

The norm of the weight vector w is represented by $||w||$ and the inequality constraint ensures that all data points are correctly classified and lie outside the margin. To solve this problem, Lagrange multipliers and the Krush-kuhn-tucker conditions are used, resulting in a set of support vectors that lie on the margin and determine the hyperplane's position. When dealing with a dataset that is not linearly separable, a common technique is to use the kernel trick to map the input features into a higher-dimensional space where the dataset becomes separable.

DOI: 10.3844/jcssp.2024.1357.1375

This technique is helpful in prediction and classification and fake news analysis. The kernel function is responsible for computing the inner product between the transformed feature vectors without explicitly computing the transformation, which makes the computation more efficient. The SVM algorithm can be used for linear and non-linear data and is particularly useful in various applications. Figure 3 displays the SVM process, an important tool for data analysis and is essential in multiple fields.

SVM by Epanechnikov Kernel Density Estimation

In the context of a trained Quantile Regression Neural Network model (QRNN), suppose we have generated q quantiles represented by y^1, y^2,…y^q at a specific time. Quantiles can be treated as samples from an unknown distribution using Kernel Density Estimation (KDE). This method estimates the probability density function of a random variable based on its sample by placing a kernel function at each point and adding it to produce a continuous density estimate. This function is defined mathematically and its properties depend on the choice of kernel and bandwidth (He *et al*., 2020). This function is defined as Eq. (4):

$$\hat{f}d(y) = \frac{1}{qd}\sum_{i=1}^{q} k\left(\frac{y^i - y}{d}\right) \tag{4}$$

The equation mentioned below is used in statistical analysis to estimate the density function of a given variable. The equation involves three key parameters: Bandwidth (d), number of quantiles (q) and the Kernel function (k). The kernel function is a mathematical function that transforms data points into a probability distribution. In this study, the Epanechnikov kernel is used as the kernel function, which is defined as Eq. (5). This function is a type of symmetric probability density function used to estimate a random variable's density. The Epanechnikov kernel is known for its good performance in terms of accuracy and efficiency in smoothing data:

$$K(\alpha) = \begin{cases} \frac{3}{4}\left(1-\alpha^2\right) & \alpha \in \left[-1,1\right] \\ 0 & \alpha \notin \left[-1,1\right] \end{cases} \tag{5}$$

The main objective of our study is not to evaluate the effectiveness or efficiency of a particular problem. Instead, we aim to demonstrate the application of the Epanechnikov kernel in solving the classification problem involving interval-valued data. This will help illustrate that our proposed approach, based on the Support Vector Machine (SVM) of a specific form and its dual form, can be extended to cover cases of kernels different from the traditional triangular kernel. By showcasing the versatility of our approach, we hope to provide valuable insights to researchers and practitioners looking for novel ways to tackle similar problems in their respective fields.

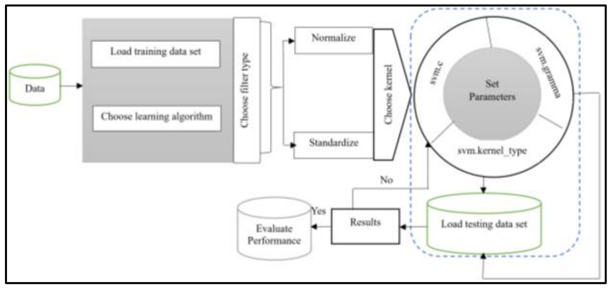

Fig. 3: Flow chat for support vector machine

Decision Trees (DT)

In the realm of regression applications, decision tree regression is a highly effective and understandable method for forecasting numerical values. It works by iteratively partitioning the dataset into subsets based on the values of the features, which then helps to construct a model that takes the shape of a tree. The feature and split point that minimizes the variance or mean squared error of the target variable in the subset are selected at each tree node. This process continues until a predetermined threshold is reached, such as a minimum number of samples per leaf or a maximum tree depth. To predict every data point, the decision tree is traversed from the root node to a leaf node and the numerical value linked to the leaf node determines the final prediction of the target variable. One of the most significant advantages of this type of decision tree is its exceptional interpretability, which makes it an excellent tool for elucidating the different elements that influence the expected numerical outcomes. Additionally, decision trees for regression are extremely useful in a wide range of applications, including financial forecasting and medical diagnostics, as they can adjust to nonlinear interactions and produce accurate numerical predictions. However, to ensure strong generalization performance and manage the complexity of the model, pruning techniques and hyperparameter tuning are frequently used to avoid overfitting. Overall, decision tree regression is a versatile and efficient method that clearly understands decision-making while delivering accurate numerical predictions.

The tree-based algorithm follows a systematic approach that begins with the root node, representing the modal features of the structure from the initial or earlier routine study. Typically, the initial layer of the algorithm comprises a few child nodes connected to this root node. These child nodes are supplemented with the original values and modifications to the modal properties of various settings in this first layer. As a result, the algorithm's starting point is the collection of child nodes. One of the critical parameters in the method settings that needs to be optimized is the number of child nodes (S), contingent on the bridge conditions. Once the first layer's

DOI: 10.3844/jcssp.2024.1357.1375

child nodes are roughly estimated, the algorithm creates a subtree structure to assess potential harm. The algorithm calculates the harm by determining the type of damage situation, which determines the substructure's depth (floor count). The algorithm also considers real damage and loudness levels to determine the most likely volume, chosen and measured using the subtree structure's leaf nodes. The indicator summarizes each possible damage's simultaneity to estimate the damage scenario. The diagram shows the algorithm's flow in Fig. 4, (Cihan, 2021). Overall, the algorithm's ability to assess potential harm in a systematic and detailed manner makes it a valuable tool in evaluating the safety and reliability of structures.

Random Forest (RF)

RF is a powerful technique that has recently gained immense popularity. It significantly improves over traditional decision tree methods, combining multiple trees to create robust classification and prediction. This technique helps reduce the model's overfitting and enhance its predictive power. Each decision tree in the forest makes predictions on its own and the ultimate forecast is derived by taking the median or averaging the predictions made by each tree in the forest. RF's ensemble technique improves model generalization and reduces overfitting risk. Additionally, adding randomization throughout the tree-building process makes the model more stable overall and resistant to outliers. Random forest is instrumental in handling nonlinear relationships between characteristics and the target variable, capturing intricate interactions and reducing the influence of noisy data. One of the random forest's most significant advantages is that it offers a feature importance score that indicates the significance of each feature in the prediction process. Understanding which features have the most significant impact on the regression's result can be done using this information. Random forest is a flexible and efficient approach for tasks involving regression. It has become popular in industries such as retail, healthcare and finance, where precise numerical forecasts are essential for making decisions. To better understand the algorithm's flow, the flow diagram in Fig. 5, (Kanchidurai *et al*., 2020) illustrates the random forest's step-by-step process.

Logistic Regression (LR)

LR is a statistical technique that predicts the likelihood of an event based on multiple independent variables and a categorical dependent variable. It is a powerful method that allows us to analyze the relationship between these variables and estimate the probability of an outcome. The LR is a Generalized Linear Model (GLM) that performs better than standard multiple regression. LG is a statistical method used to estimate the probability of an event by analyzing the relationship between independent factors and a dependent variable. It can be used for binary or multinomial variables and aims to identify the best model to explain the relationship between the variables. This is accomplished by analyzing the data and selecting the most appropriate model that fits the data. Binary logistic analysis is an iterative algorithm determining the relationship between independent and dependent variables (Ganesh *et al*., 2022; Valero-Carreras *et al*., 2023). The algorithm's flow is shown in the flow diagram Fig. 6.

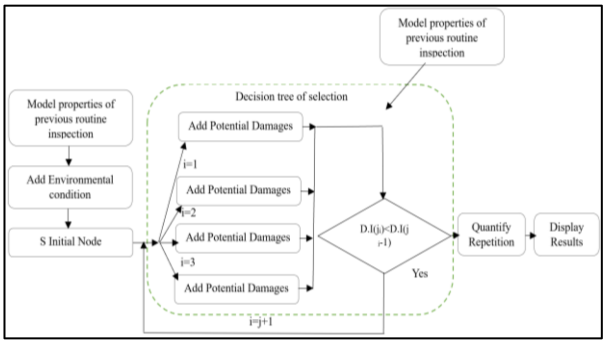

Fig. 4: Flow chart for decision trees

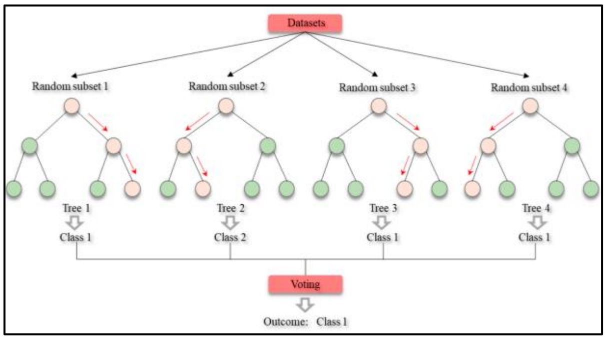

Fig. 5: Flow chat for random forest (Kanchidurai *et al*., 2020)

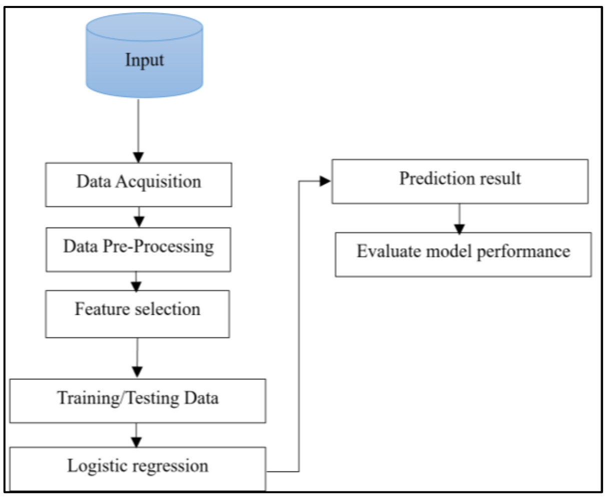

Fig. 6: Flow chart for logistic regression

Table 3: Optimal hyperparameter setting

Models	Optimal hyperparameter
Naïve Baye (NB)	Number of folds (10) sampling type (stratified sampling) Minimal ginn (0.01)
Support Vector Machines (SVM)	SVM. Kernel_type, Kernel_epachnenikov, kernel sigma1 = 1.0, kernel degree = 2.0, kernel cache = 200, C = 0.0, convergence epsion = 0.001, max iteration = 250
Decision Trees (DT)	Local random seed hidden (70/30)
Random Forest (RF)	Number of folds (10) Number of trees (100) Criterion (gain_ratio) Maximal depth (10)
Logistic Regression (LR)	Number of treads (10) reproducible regularization

Optimal Hyperparameter Setting

Once the pre-processing stage of the data was completed, a comparative analysis was performed to evaluate the performance of the baseline classifiers. This involved assessing the performance of each model and identifying the optimal hyperparameters for each one. To ensure the accuracy of the results, two separate subsets were randomly chosen from the source dataset, which included a training and testing set. The primary goal was to optimize the baseline model's hyperparameter settings for better performance. The results of this analysis are presented in Table 3, which outlines the critical findings and provides insights into the effectiveness of different hyperparameter settings. This analysis helps to identify the most promising models and hyperparameter settings for further refinement and improvement.

Several validation methods have been used to assess and contrast the effectiveness of the initial models. These techniques are well-known for their proven ability to carry out verification-related tasks. They have been employed to enhance the crucial parameters for each model, thereby improving their overall performance. By implementing these techniques, we can ensure that our models are robust and consistently deliver accurate results.

Performance Evaluation

To assess the efficiency of every proposed model, we have implemented a variety of evaluation metrics. This subsection will examine the most popular criteria for identifying fake news. These criteria include Accuracy (Acc), F-measure, Recall (R) and Precision (P), which are assessed according to the following equation (Khan *et al.*, 2023). By using these metrics, we can better understand the performance of each model and make informed decisions about which ones are most effective in identifying fake news.

Confusion Matrix

Several measures are used to gauge a model's effectiveness, including Sensitivity (Sen), Specificity (Spec), Accuracy (Acc), F-measure (Fm), Recall (R) and Precision (P). However, when the dataset's distribution of classes is unbalanced, evaluating a model's effectiveness using these matrices may not be appropriate (Ismael and Şengür, 2021; Roshani *et al.*, 2021). A model can achieve high accuracy in machine learning even if biased towards the dominant class. This is particularly true in unbalanced domains, where one class may be overrepresented in the dataset. A confusion matrix provides a detailed breakdown of a model's predictions. For instance, let's consider the scenario of a model that aims to identify fabricated news stories. In this case, the dataset may be heavily skewed towards fabricated news, making it difficult to evaluate the model's performance based solely on its overall accuracy. The model performs well because it can correctly classify most fabricated news stories. However, this does not accurately represent the model's performance because it may need to misclassify many real news stories. Using a confusion matrix, we can better understand the model's actual performance by examining the number of True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN). This will help us to identify any underlying issues with the model and make improvements accordingly. The following are some key terms used to evaluate the accuracy of a model's classification against a set of test data:

TP (True Positive) : The model predicts a positive outcome and it is actually positive
TN (True Negative) : The model predicts a negative outcome and it is actually negative
FP (False Positive) : The model predicts a positive outcome but it is actually negative
FN (False Negative): The model predicts a negative outcome but it is actually positive

Accuracy is a crucial metric when evaluating the effectiveness of a prediction classifier. It represents the percentage of predictions the classifier correctly made from the total predictions. We utilize Eq. (6) (Nawaz *et al.*, 2021) to calculate the accuracy score. The accuracy score is represented by the letter '*A*' in the formula. We can determine how well the model predicts the target variable by computing the accuracy score. A higher accuracy score indicates that the classifier is making accurate classifications, while a lower value implies that the classifier is making more errors. Therefore, accuracy is a critical measure of a model's performance in classification tasks:

$$Accurracy = \frac{True\ Position + True\ Negative}{Total\ Number\ of\ Prediction} \times 100 \quad (6)$$

Precision is a crucial metric to determine the accuracy of a prediction classifier. Precision refers to the number of correctly classified positive results out of all the positive ones, including those wrongly identified. We can calculate precision using the formula Eq. (7) defined in reference (Albahli and Nawaz, 2022). This equation considers the number of TP and *false positives* generated by the method and provides a value between 0 and 1, where higher values indicate better precision:

$$Precision = \frac{True\ Positive}{Positive + False\ Positive} \times 100 \quad (7)$$

Recall is a metric used to evaluate a model's ability to correctly identify all positive instances in a dataset. It represents the fraction of correctly identified positive cases out of all the actual positive instances in the dataset. Hence, a high recall value indicates a model's effectiveness in identifying positive instances. This measure is computed using a formula defined by Eq. (8) (Rustam *et al*., 2021), which considers the number of TP, FN and TN in the dataset. A high recall score is good at recognizing positive instances, but a low recall score is prone to missing some positive ones. Hence, it's essential to ensure that the recall score is high enough to achieve the desired level of accuracy in the model's predictions:

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \times 100 \quad (8)$$

The F-measure is a statistical method to evaluate a predictive model's accuracy in classification. It considers precision and recall and gives a complete picture of the performance, described in reference (Anoop *et al*., 2020). By calculating the F-measure, we can determine how accurate the model is in making correct predictions (precision) while ensuring that it is not missing any relevant predictions (recall). This metric is widely used in data science to evaluate the effectiveness of predictive models:

$$F\text{ - }measure = \frac{2 \times Recall \times Precision}{Recall + Precision} \times 100 \quad (9)$$

It is essential to understand that identifying bogus news requires a thorough and systematic approach involving several steps. The first step is selecting the appropriate data, followed by the pre-processing of this data to ensure its accuracy. Once the data is pre-processed, it can be used to train models and choose algorithms that are best suited for detecting false information. By following this comprehensive method, experts in spotting fake news can confidently make precise analytical conclusions. Analysts who rely on accurate and reliable results can also benefit from this process, as it allows them to make informed decisions based on the data at hand. Ultimately, the success of identifying bogus news depends on implementing a sound methodology that considers all these critical steps.

Results

We divided the dataset into two parts during our experiment, with a 30/70 split ratio. The first part, consisting of 30% of the data, was reserved for training purposes, while the second part, comprising 70% of the data, was allocated for testing and validation. We experimented with two datasets (Kaggle Data World and UCI dataset) to evaluate recommendation systems in different contexts. We also compared the results obtained with and without preprocessing the data. The outcomes of these experiments are presented in Tables 4-6, which provide a detailed analysis of the performance of the recommendation systems on each dataset.

Table 4: Shows the results of fake news detection using an SVM with and without preprocessing

	Performance			
Model	Accuracy (%)	Precision (%)	Recall (%)	F-measure (%)
NB	91.29	94.89	94.79	94.84
SVM	99.67	99.61	100.00	99.81
DT	97.35	97.06	99.57	98.45
RF	98.05	97.74	100.00	98.86
LR	94.49	95.36	98.26	96.78

Table 5: Shows the performance metrics for the data word dataset, comparing the outcomes obtained using SVM with and without preprocessing

	Performance			
Model	Accuracy (%)	Precision (%)	Recall (%)	F-measure (%)
NB	71.84	75.26	87.12	80.75
SVM	73.28	74.36	92.50	82.44
DT	75.52	73.87	98.91	84.57
RF	74.13	74.87	93.10	82.99
LR	74.13	74.87	93.10	82.99

Table 6: Shows the performance metrics for the UCI dataset, comparing the outcomes obtained using SVM with and without preprocessing

	Performance			
Model	Accuracy (%)	Precision (%)	Recall (%)	F-measure (%)
NB	81.10	86.20	98.70	87.91
SVM	82.17	84.62	93.81	88.96
DT	88.74	87.96	98.85	93.08
RF	89.22	88.35	98.99	93.36
LR	84.29	85.31	96.03	90.35

Results

This section presents the modeling outcomes and subsequent training with the preprocessed dataset. We have thoroughly compared five machine learning algorithms based on their performance. We have made significant alterations to the parameters to obtain a comprehensive understanding of the results. Furthermore, we provide perspectives on thoroughly analyzing the experimental setting employed in each study. By doing so, we aim to offer a clear and concise evaluation of the machine learning models' effectiveness, which can help make informed decisions for future work. The aim of this study is to enhance the ability to identify deceptive news using various machine learning algorithms. The researchers employed several machine learning models to achieve the desired objective. Various models' accuracy and effectiveness were assessed using Tables 4-6.

Table 4, specifically, was utilized to measure the accuracy of each model individually. After thoroughly analyzing the presented results in Table 4, it was concluded that the SVM model outperformed all the baseline models with an impressive accuracy score of 99.67%. RF classifier secured the second position with an accuracy score of 98.05%, while the Decision Tree (DT) model stood at the third position with a score of 97.35%. On the other hand, the LR classifier obtained an accuracy score of 94.49% and the Naive Bayes (NB) model showed an accuracy score of 91.29%.

Comparison with Related Work

Numerous researchers have dedicated their efforts to developing various classifiers and performing experiments aimed at identifying and flagging fake news using machine learning algorithms.

To evaluate the effectiveness of these methods, we conducted a thorough comparison of state-of-the-art methods for fake news classification, with the findings presented in Table 7. Anoop *et al*. (2020); Faustini and Covões (2020) have reported relatively good experimental results for their method that improves health fake news identification, achieving a 94.0% accuracy for DT. Faustini and Covões (Nagaraja *et al*., 2021) used NB, KNN, SVM and RF methods for the classification of fake news in multiple platforms and languages. They reported a 75.00% accuracy for detecting fake news, with K-Nearest Neighbors (KNN) achieving an FM score of 81.0%. Nagaraja *et al.* (2021); Truică *et al.* (2023) introduced a combination of Naïve Bayes and SVM algorithms to identify fake news, providing an excellent experimental result of 75.0% accuracy for SVM. Truică *et al*. (2023); Reis *et al*. (2019) used Convolutional Neural Network (CNN) and Bidirectional LSTM (BiLSTM) methods for real-time fake news mitigation in social platforms. According to their report, the results were impressive, with an accuracy rate of 96.38, precision of 96.39, recall of 96.38 and FM score of 96.38%. These numbers suggest that the task was performed efficiently and accurately. Reis *et al*. (2019); Sitaula *et al*. (2020) developed a supervised learning method for fake news classification and reported an accuracy of 85.0% and an FM score of 81.0% for RF. Sitaula *et al*. (2020); Jose *et al*. (2021) introduced a credibility-based fake news detection method using SVM, linear SVM, LR, RF, AdaBoost, NB and Gradient Boosted Decision Tree (GBDT) algorithms. They reported that linear SVM achieved an FM score of 82.0%, LR achieved an FM score of 82.0% and linear SVM reached an FM macro score of 77.0%. Lastly, Jose *et al*. (2021); Park and Chai (2023) introduced a detection model for fake news in online social media networks, including LR, DT, RF, Multinomial Naive Bayes (MNB) and SVM algorithms. They reported LR with an accuracy of 90.30, DT with 83.10, RF with 88.10%, MNB with 58.40 and SVM with 90.20%. In their research, Park and Chai (2023) developed a user-centric model for detecting fake news using various models, such as LR CART (Classification and Regression Trees), neural networks (NN), SVM and RF-the model aimed to classify news articles as either fake or real based on their content and characteristics. They reported LR achieving an accuracy of 91.80, CART of 96.70, NN with 91.80, SVM with 91.70 and RF with 95.10%, respectively.

Table 7: A comparative study of the latest and most advanced models for detecting fake news. It showcases an analysis of the state-of-the-art methods applied in identifying and classifying fabricated news stories from legitimate ones

Author	Year	Dataset	Model	Performance (%)
Anoop *et al*.	(2020)	HWB	NB, KNN, SVM, RF, DT, CNN, Adaboost, LSTM	Higher accuracy for DT, Acc = 94.0
Faustini and Covões	(2020)	Twitter Br	NB, KNN, SVM, RF	SVM: Acc = 79, KNN: FM = 81
Nagaraja *et al.*	(2021)	Merged four datasets	Naïve Bayes, SVM	SVM = Acc 75.0
Truică *et al.*	(2023)	Kaggle datasets	CNN, 3BiLSTM	Acc = 96.38, Precision = 96.39, Recall = 96.38, FM = 96.38
Reis *et al*.	(2019)	Buzz Feed	KNN, NB, RF, SVM, XGB	RF: Acc = 85.0, FM = 81.0
Sitaula *et al*.	(2020)	PolitiFact, Buzzfeed	SVM, linear SVM, LR, RF, AdaBoost, NB, GBDT	Linear SVM: macro = 82.0, LR: FM score = 82.0, Linear SVM: FM Macro = 77.0
Jose *et al*.	(2021)	Google News	LR, DT, RF, MNB, SVM	LR = 90.30, DT = 83.10, RF = 88.10, MNB = 58.40, SVM = 90.20
Park and Chai	(2023)	Twitter network	LR, CART, NN, SVM, RF	LR = 91.80, CART = 96.70, NN = 91.80, SVM = 91.70, RF = 95.10
Proposed methods	(2024)	Kaggle datasets	NB, SVM, DT, RF, LR	NB = 91.29, SVM = 99.67, DT = 97.35, RF = 98.05, LR = 94.49

DOI: 10.3844/jcssp.2024.1357.1375

Special Test (Kernels)

The evaluation of models in detecting fake news involves using different kernel functions of the SVM algorithm. In Table 8, the results of the model review are presented.

In the analysis conducted, it was found that the Root Mean Square Error (RMSE) is minimized when the Support Vector Machine (SVM) algorithm utilizes the radial kernel function. On the other hand, using the Epachnenikov kernel results in the highest RMSE value. This suggests that the radial kernel is more effective than the Epachnenikov kernel in minimizing prediction errors in the SVM algorithm. The following best-performing kernel functions are the SVM algorithm at the radial kernel and the SVM algorithm at the ANOVA kernel. These findings suggest that the radial kernel function of the SVM algorithm is the most suitable for detecting fake news, while the Epachnenikov kernel may not be the best choice in this context. Upon examining Tables 4 and 8, it can be observed that the SVM classifier was used as the baseline model, with an average accuracy score of 99.67%. However, it was noted that the SVM model still needed to be adjusted for parameters. When the SVM model was compared with the Epachnenikov kernel adjustment, it was found that the accuracy score increased to 99.70%, which is an increase of 0.03%. On the other hand, when comparing the SVM classifier with radial, the accuracy score decreased to 99.66%. Similarly, the SVM model with ANOVA decreased the accuracy score to 98.92%. Compared with polynomials, the SVM model saw a significant decrease in accuracy score to 91.63%. Additionally, the SVM model with multiquadric resulted in an even lower accuracy score of 84.43%. Finally, the SVM model with neural exhibited the lowest accuracy score of 70.48%.

Table 8: The models are evaluated at different kernel functions

	Performance			
Model + Kernels	Accuracy (%)	Precision (%)	Recall (%)	F-measure (%)
SVM + radial	99.66	99.60	100.00	99.80
SVM + polynomial	91.63	91.27	99.67	95.27
SVM + neural	70.48	86.56	76.89	80.01
SVM + Anova	98.92	99.39	99.33	99.36
SVM + Epachnenikov	99.70	99.61	100.00	99.81
SVM + multiquadric	84.43	84.43	100.00	91.56

Discussion

In today's digital age, accessing information online has become incredibly easy. The online world includes various sources, from reliable news outlets sharing verified information to unknown individuals spreading unverified viewpoints. Social media platforms have transformed how we communicate and consume information by providing instant access to news and various topics. These platforms allow effortlessly sharing valuable content, links, and personal opinions. However, the main challenge is distinguishing between accurate and inaccurate information, which poses significant risks to public perception and knowledge.

This study introduced an improved Support Vector Machine (SVM) model with hyperparameter tuning to detect fake news, specifically on Twitter. The proposed method involved two main phases: Check-worthiness identification and fact-checking. These phases included three critical tasks: Feature selection, fake news detection, and verifying the factuality of claims within tweets. By transforming complex optimization problems into more straightforward linear, the model used the approximation of the Gaussian kernel with Epanechnikov kernels. This approach enabled the selection of an optimal probability distribution from a predefined set of choices, and the minimax strategy was used to construct the most effective separating functions. The proposed method demonstrated high accuracy, precision, recall, and F-measure, achieving 99.67, 99.61, 100 and 99.81%, respectively. These findings not only suggest but also strongly affirm that the enhanced SVM model is efficient and highly effective in identifying fake news. The accuracy metrics indicate that the model can reliably distinguish between true and false information, making it a valuable and trustworthy tool in the fight against misinformation. This framework not only provides a robust solution to the pervasive problem of misinformation on social media but also holds the potential to significantly improve public discourse and information integrity in the future.

The following steps in our research could involve testing how well this model works on various social media platforms and datasets to ensure it can be applied effectively in different situations. We could also improve the model by carefully selecting which features to include and fine-tuning its parameters to make it even more effective.

Conclusion

The study employs five models to classify and detect false information. The predictive power of each model is analyzed to determine the one that performs best. The results reveal that the SVM is the most accurate model among ensemble methods and algorithms, with an astounding accuracy of 99.67%.

Random Forest (RF) is another model that demonstrates robust prediction capabilities, with an accuracy of 98.05%. This study emphasizes the importance of detecting fake news and highlights the need for expertise in various fields of study, including data science and social science. To effectively detect fake news, it is recommended to use a repository that detects it using different parameters. Additionally, collecting detailed network information and location data can help increase the accuracy of identifying fraudulent content. The goal is to fully understand fake news and devise strategies to stop such news from being created and spread. Students who use social media daily may need help understanding how misinformation spreads online. Therefore, this survey could be expanded to cover similar topics, such as identifying "clickbait," which involves using attention-grabbing language to persuade readers to click on a link to a fake or fake website. Additionally, recognizing the difference between headlines and news content can be used to predict fake news. Another study area is identifying spam in social networks spread by people, user groups, or social bots.

Acknowledgment

The authors would like to thank the anonymous reviewers and editors for their insightful comments and helpful suggestions, which improved the quality of our manuscript.

Funding Information

This research was financially supported by Mahasarakham University, Thailand.

Author's Contributions

Both the authors have equally contributed to this study.

Ethics

This article is an original work and includes previously unpublished material.

Conflicts of Interest

The authors declare no conflict of interest.

References

Adiba, F. I., Islam, T., Kaiser, M. S., Mahmud, M., & Rahman, M. A. (2020). Effect of corpora on classification of fake news using Naive Bayes classifier. *International Journal of Automation, Artificial Intelligence and Machine Learning*, *1*(1), 80–92. https://doi.org/10.61797/ijaaiml.v1i1.45.

Agarwal, A., & Dixit, A. (2020). Fake news detection: An ensemble learning approach. *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, 1178–1183. https://doi.org/10.1109/iciccs48265.2020.9121030.

Ahmed, S., Hinkelmann, K., & Corradini, F. (2022). Combining Machine Learning with Knowledge Engineering to detect Fake News in Social Networks-a survey. *Artificial Intelligence*, arXiv:2201.08032. https://doi.org/10.48550/arXiv.2201.08032

Albahli, S., & Nawaz, M. (2022). DCNet: DenseNet-77-based CornerNet model for the tomato plant leaf disease detection and classification. *Frontiers in Plant Science*, *13*, 957961. https://doi.org/10.3389/fpls.2022.957961

Ali, A., Li, J., Chen, H., & Bashir, A. K. (2022). Temporal pattern mining from user-generated content. *Digital Communications and Networks*, *8*(6), 1027–1039. https://doi.org/10.1016/j.dcan.2022.11.008

Anoop, K., Deepak, P., & Lajish, L. v. (2020). Emotion cognizance improves health fake news identification. *In Proceedings of 24th International Database Engineering & Applications Symposium (IDEAS 2020)*, 24. https://doi.org/10.48550/arXiv.1906.10365

Aphiwongsophon, S., & Chongstitvatana, P. (2018). Detecting fake news with machine learning method. *2018 15th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, 528–531. https://doi.org/10.1109/ecticon.2018.8620051

Arcuri, M. C., Gandolfi, G., & Russo, I. (2023). Does fake news impact stock returns? Evidence from US and EU stock markets. *Journal of Economics and Business*, *125*, 106130. https://doi.org/10.1016/j.jeconbus.2023.106130

Atodiresei, C.-S., Tănăselea, A., & Iftene, A. (2018). Identifying fake news and fake users on twitter. *Procedia Computer Science*, *126*, 451–461. https://doi.org/10.1016/j.procs.2018.07.279

Billones, P. J. R., Macasaet, D. D., & Arenas, S. U. (2022). Bilingual fake news detection algorithm using Naïve Bayes and support vector machine models. *2022 IET International Conference on Engineering Technologies and Applications (IET-ICETA)*, 1–2. https://doi.org/10.1109/ieticeta56553.2022.9971596.

Birzhandi, P., Kim, K. T., & Youn, H. Y. (2022). Reduction of training data for support vector machine: A survey. *Soft Computing*, *26*(8), 3729–3742. https://doi.org/10.1007/s00500-022-06787-5

Bondielli, A., & Marcelloni, F. (2019). A survey on fake news and rumor detection techniques. *Information Sciences*, *497*, 38–55. https://doi.org/10.1016/j.ins.2019.05.035

Capuano, N., Fenza, G., Loia, V., & Nota, F. D. (2023). Content-B-based fake news detection with machine and deep learning: A systematic review. *Neurocomputing*, *530*, 91–103. https://doi.org/10.1016/j.neucom.2023.02.005

Chan, S. W. K., & Chong, M. W. C. (2017). Sentiment analysis in financial texts. *Decision Support Systems*, *94*, 53–64. https://doi.org/10.1016/j.dss.2016.10.006.

Chen, P., Liu, H., Xin, R., Carval, T., Zhao, J., Xia, Y., & Zhao, Z. (2022). Effectively detecting operational anomalies in large-scale IoT data infrastructures by using a GAN-based predictive model. *The Computer Journal*, *65*(11), 2909–2925. https://doi.org/10.1093/comjnl/bxac085

Cihan, M. T. (2021). Comparison of artificial intelligence methods for predicting compressive strength of concrete. *Gradjevinar*, *73*(6), 617–632. https://doi.org/10.14256/JCE.3066.2020

Coronavirus. (2020). *5G and Microchip Conspiracies Around the World BBC News*.

Dwivedi, Y. K., Hughes, D. L., Coombs, C., Constantiou, I., Duan, Y., Edwards, J. S., Gupta, B., Lal, B., Misra, S., Prashant, P., Raman, R., Rana, N. P., Sharma, S. K., & Upadhyay, N. (2020). Impact of COVID-19 pandemic on information management research and practice: Transforming education, work and life. *International Journal of Information Management*, *55*, 102211. https://doi.org/10.1016/j.ijinfomgt.2020.102211

Efe, S., & Fadipe, O. (2023). Machine Learning Prediction of the Compressive Strength of Siliceous Fly Ash Modified Cementitious Concrete. *International Journal of Structural Engineering (IJSE*, *4*(1), 30–46. https://doi.org/10.17605/OSF.IO/CFHWJ

El-Azab, H.-A. I., Swief, R. A., El-Amary, N. H., & Temraz, H. K. (2023). Seasonal electric vehicle forecasting model based on machine learning and deep learning techniques. *Energy and AI*, *14*, 100285. https://doi.org/10.1016/j.egyai.2023.100285

Faustini, P. H. A., & Covões, T. F. (2020). Fake news detection in multiple platforms and languages. *Expert Systems with Applications*, *158*, 113503. https://doi.org/10.1016/j.eswa.2020.113503

Ganesh, A., B., Ganesh, A., Srinivas, C., Dhanraj, & Mensinkal, K. (2022). Logistic regression technique for prediction of cardiovascular disease. *Global Transitions Proceedings*, *3*(1), 127–130. https://doi.org/10.1016/j.gltp.2022.04.008

Garg, B., & Program, M. (2013). *Design and Development of Naïve bayes Classifier*.

Gaurav, A., Gupta, B. B., & Panigrahi, P. K. (2023). A comprehensive survey on machine learning approaches for malware detection in IoT-based enterprise information system. *Enterprise Information Systems*, *17*(3), 2023764. https://doi.org/10.1080/17517575.2021.2023764

Gebremariam, G. G., Panda, J., & Indu, S. (2023). Secure localization techniques in wireless sensor networks against routing attacks based on hybrid machine learning models. *Alexandria Engineering Journal*, *82*, 82–100. https://doi.org/10.1016/j.aej.2023.09.064

Gereme, F. B., & Zhu, W. (2019). Early detection of fake news "Before it flies high." *Proceedings of the 2nd International Conference on Big Data Technologies*, 142–148. https://doi.org/10.1145/3358528.3358567

Gupta, R., Patel, M. M., Shukla, A., & Tanwar, S. (2022). Deep learning-based malicious smart contract detection scheme for internet of things environment. *Computers & Electrical Engineering*, *97*, 107583. https://doi.org/10.1016/j.compeleceng.2021.107583

Hakak, S., Alazab, M., Khan, S., Gadekallu, T. R., Maddikunta, P. K. R., & Khan, W. Z. (2021). An ensemble machine learning approach through effective feature extraction to classify fake news. *Future Generation Computer Systems*, *117*, 47–58. https://doi.org/10.1016/j.future.2020.11.022

He, H., Pan, J., Lu, N., Chen, B., & Jiao, R. (2020). Short-term load probabilistic forecasting based on quantile regression convolutional neural network and Epanechnikov kernel density estimation. *Energy Reports*, *6*, 1550–1556. https://doi.org/10.1016/j.egyr.2020.10.053

Hiramath, C. K., & Deshpande, G. C. (2019). Fake news detection using deep learning techniques. *2019 1st International Conference on Advances in Information Technology (ICAIT)*, 411–415. https://doi.org/10.1109/icait47043.2019.8987258

Hogenboom, F., Frasincar, F., Kaymak, U., de Jong, F., & Caron, E. (2016). A Survey of event extraction methods from text for decision support systems. *Decision Support Systems*, *85*, 12–22. https://doi.org/10.1016/j.dss.2016.02.006

Ismael, A. M., & Șengür, A. (2021). Deep learning approaches for COVID-19 detection based on chest X-ray images. *Expert Systems with Applications*, *164*, 114054. https://doi.org/10.1016/j.eswa.2020.114054.

Jain, A., & Kasbe, A. (2018). Fake News Detection. *2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*, 1–5. https://doi.org/10.1109/sceecs.2018.8546944

Jain, A., Shakya, A., Khatter, H., & Gupta, A. K. (2019). A smart system for fake news detection using machine learning. *2019 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT)*, 1–4. https://doi.org/10.1109/icict46931.2019.8977659

Jing, J., Wu, H., Sun, J., Fang, X., & Zhang, H. (2023). Multimodal fake news detection via progressive fusion networks. *Information Processing & Management*, *60*(1), 103120. https://doi.org/10.1016/j.ipm.2022.103120

Jose, X., Kumar, S. D. M., & Chandran, P. (2021). Characterization, Classification and Detection of Fake News in Online Social Media Networks. *2021 IEEE Mysore Sub Section International Conference (MysuruCon)*, 759–765. https://doi.org/10.1109/mysurucon52639.2021.9641517

Kanchidurai, S., Krishnan, P. A., & Baskar, K. (2020). Compressive strength estimation of mesh embedded masonry prism using empirical and neural network models. *Journal of Soft Computing in Civil Engineering*, *4*(4), 24–35. https://doi.org/10.22115/SCCE.2020.228611.1213

Kar, A. K., Tripathi, S. N., Malik, N., Gupta, S., & Sivarajah, U. (2023). How Does Misinformation and Capricious Opinions Impact the Supply Chain - A Study on the Impacts During the Pandemic. *Annals of Operations Research*, *327*(2), 713–734. https://doi.org/10.1007/s10479-022-04997-6

Khaled, A. Y., Kabutey, A., Selvi, K. Ç., Mizera, Č., Hrabe, P., & Herák, D. (2020). Application of Computational Intelligence in Describing the Drying Kinetics of Persimmon Fruit (Diospyros kaki) During Vacuum and Hot Air-Drying Process. *Processes*, *8*(5), 544. https://doi.org/10.3390/pr8050544

Khan, S., Anwar, M., Qayyum, H., Ali, F., & Nawaz, M. (2023). Fake News Classification using Machine Learning: Count Vectorizer and Support Vector Machine. *Journal of Computing & Biomedical Informatics*, *4*(1), 54–63. https://doi.org/10.56979/401/2022/85

Kocherlakota, A., Saicharan, K., Gupta, S., Nidhi Singh, A., Teja Karri, S., Reddy, T., & Sujay Ashrith, K. (2022). Fake news detection using Naive Bayes, support vector machine, neural network and long short-term memory. *International Journal of All Research Education and Scientific Methods (IJARESM*, *10*(3), 746 – 752. http://doi.org/10.11591/ijai.v11.i2

Krishna, N. L. S. R., & Adimoolam, M. (2022). Fake news detection system using decision tree algorithm and compare textual property with Support Vector Machine algorithm. *2022 International Conference on Business Analytics for Technology and Security (ICBATS)*, 1–6. https://doi.org/10.1109/icbats54253.2022.9758999

Lee, Y. H., Hu, P. J. H., Zhu, H., & Chen, H. W. (2020). Discovering event episodes from sequences of online news articles: A time-adjoining frequent itemset-based clustering method. *Information & Management*, *57*(7), 103348. https://doi.org/10.1016/j.im.2020.103348

Leon, C. S., Bonilla, M., Brusco, L. I., Forcato, C., & Benítez, F. U. (2023). Fake news and false memory formation in the psychology debate. *IBRO Neuroscience Reports*, *15*, 24–30. https://doi.org/10.1016/j.ibneur.2023.06.002

Lin, Z., Wang, H., & Li, S. (2022). Pavement anomaly detection based on transformer and self-supervised learning. *Automation in Construction*, *143*, 104544. https://doi.org/10.1016/j.autcon.2022.104544

Lu, S., Ding, Y., Liu, M., Yin, Z., Yin, L., & Zheng, W. (2023). Multiscale Feature Extraction and Fusion of Image and Text in VQA. *International Journal of Computational Intelligence Systems*, *16*(1), 54. https://doi.org/10.1007/s44196-023-00233-6

Mishra, S., Shukla, P., & Agarwal, R. (2022). Analyzing Machine Learning Enabled Fake News Detection Techniques for Diversified Datasets. *Wireless Communications and Mobile Computing*, 1–18. https://doi.org/10.1155/2022/1575365

Monti, F., Frasca, F., Eynard, D., Mannion, D., & Bronstein, M. M. (2019). Fake news detection on social media using geometric deep learning. *Social and Information Networks*, arXiv:1902.06673. https://doi.org/10.48550/arXiv.1902.06673

Mumbai, F. C., Pune, H., & Kolkata. (2020). *No Military lockdown in Mumbai, Kolkata and Ahmedabad, Message is a Hoax*

Nadikattu, A. K. R. (2023). Machine learning vs deep learning models for detecting fake news: A comparative analysis on fake-newsnet dataset. *SSRN Electronic Journal*, SSRN 4382241, 1-16. https://doi.org/10.2139/ssrn.4382241

Naeem, M. (2021). Do social media platforms develop consumer panic buying during the fear of COVID-19 pandemic. *Journal of Retailing and Consumer Services*, *58*, 1 – 10 https://doi.org/10.1016/j.jretconser.2020.102226

Nagaraja, A., Soumya KN., Sinha, A., Prajwal N., Anubhav S., & Jain VR. (2021). Fake news detection using machine learning methods. *International Conference on Data Science, E-Learning and Information Systems 2021*, 185–192. https://doi.org/10.1145/3460620.3460753

Nawaz, M., Nazir, T., & Masood, M. (2021). Glaucoma detection using tetragonal local octa patterns and SVM from retinal images. *The International Arab Journal of Information Technology*, *18*(5), 686–693. https://doi.org/10.34028/iajit/18/5/8

Neeraj, S., Singh, L., Tripathi, S., & Malik, N. (2023). Detection of fake news using machine learning. *2023 13th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, 20–24. https://doi.org/10.1109/confluence56041.2023.10048819

Nozza, D., Manchanda, P., Fersini, E., Palmonari, M., & Messina, E. (2021). LearningToAdapt with word embeddings: Domain adaptation of Named Entity Recognition systems. *Information Processing & Management*, *58*(3), 102537. https://doi.org/10.1016/j.ipm.2021.102537

Odhiambo, O. E., Onyango O. G., & Waema K. M. (2021). Feature Selection for Classification using Principal Component Analysis and Information Gain. *Expert Systems with Applications*, *174*, 114765. https://doi.org/10.1016/j.eswa.2021.114765

Park, M., & Chai, S. (2023). Constructing a user-centered fake news detection model by using classification algorithms in machine learning techniques. *IEEE Access*, *11*, 71517–71527. https://doi.org/10.1109/access.2023.3294613

Patel, A., Tiwari, A. K., & Ahmad, S. S. (2021). Fake news detection using support vector machine. *Proceedings of the 3rd International Conference on Advanced Computing and Software Engineering*, 34–38. https://doi.org/10.5220/0010562000003161

Phan, H. T., Nguyen, N. T., & Hwang, D. (2023). Fake news detection: A survey of graph neural network methods. *Applied Soft Computing*, *139*, 110235. https://doi.org/10.1016/j.asoc.2023.110235

Qu, K., Xu, J., Hou, Q., Qu, K., & Sun, Y. (2023). Feature selection using information Gain and decision information in neighborhood decision system. *Applied Soft Computing*, *136*, 110100. https://doi.org/10.1016/j.asoc.2023.110100

Rajeswari, M. (2022). JETIRFN06035. *Journal of Emerging Technologies and Innovative Research*, *9*(6). www.jetir.org

Rasool, T., Butt, W. H., Shaukat, A., & Akram, M. U. (2019). Multi-Label Fake news detection using multi-layered supervised learning. *Proceedings of the 2019 11th International Conference on Computer and Automation Engineering*, 73–77. https://doi.org/10.1145/3313991.3314008

Reis, J. C. S., Correia, A., Murai, F., Veloso, A., & Benevenuto, F. (2019). Supervised learning for fake news detection. *IEEE Intelligent Systems*, *34*(2), 76–81. https://doi.org/10.1109/mis.2019.2899143

Renuka, M., & Anithaashri, T. P. (2022). Enhancing the detection of fake news in social media using support vector machine algorithms comparing over apriori algorithms. *2022 2nd International Conference on Technological Advancements in Computational Sciences (ICTACS)*, 51–54. https://doi.org/10.1109/ictacs56270.2022.9988701

Rohera, D., Shethna, H., Patel, K., Thakker, U., Tanwar, S., Gupta, R., Hong, W.-C., & Sharma, R. (2022). A taxonomy of fake news classification techniques: survey and implementation aspects. *IEEE Access*, *10*, 30367–30394. https://doi.org/10.1109/access.2022.3159651

Roshani, M., Phan, G. T. T., Jammal Muhammad Ali, P., Hossein Roshani, G., Hanus, R., Duong, T., Corniani, E., Nazemi, E., & Mostafa Kalmoun, E. (2021). Evaluation of flow pattern recognition and void fraction measurement in two phase flow independent of oil pipeline's scale layer thickness. *Alexandria Engineering Journal*, *60*(1), 1955–1966. https://doi.org/10.1016/j.aej.2020.11.043

Rustam, F., Khalid, M., Aslam, W., Rupapara, V., Mehmood, A., & Choi, G. S. (2021). A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis. *PLOS ONE*, *16*(2), e0245909. https://doi.org/10.1371/journal.pone.0245909

Sansonetti, G., Gasparetti, F., & Micarelli, A. (2021). Using social media for personalizing the cultural heritage experience. *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*, 189–193. https://doi.org/10.1145/3450614.3463387

Sansonetti, G., Gasparetti, F., D'aniello, G., & Micarelli, A. (2020). Unreliable users detection in social Media: deep learning techniques for automatic detection. *IEEE Access*, *8*, 213154–213167. https://doi.org/10.1109/access.2020.3040604

Sarraf, S., Kushwaha, A. K., Kar, A. K., Dwivedi, Y. K., & Giannakis, M. (2024). How did online misinformation impact stockouts in the e-commerce supply chain during COVID-19 – A mixed methods study. *International Journal of Production Economics*, *267*, 109064. https://doi.org/10.1016/j.ijpe.2023.109064

Shu, K., Mahudeswaran, D., & Liu, H. (2019). Fake news tracker: a tool for fake news collection, detection and visualization. *Computational and Mathematical Organization Theory*, *25*(1), 60–71. https://doi.org/10.1007/s10588-018-09280-3

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media. *ACM SIGKDD Explorations Newsletter*, *19*(1), 22–36. https://doi.org/10.1145/3137597.3137600

Singh, D., & Singh, B. (2020). Investigating the impact of data normalization on classification performance. *Applied Soft Computing*, *97*, 105524. https://doi.org/10.1016/j.asoc.2019.105524

Singh, D., & Singh, B. (2022). Feature wise normalization: An effective way of normalizing data. *Pattern Recognition*, *122*, 108307. https://doi.org/10.1016/j.patcog.2021.108307

Sitaula, N., Mohan, C. K., Grygiel, J., Zhou, X., & Zafarani, R. (2020). Credibility-based fake news detection. In *Disinformation, Misinformation and Fake News in social media* (pp. 163–182). Springer. https://doi.org/10.1007/978-3-030-42699-6_9

Smitha, N., & Bharath, R. (2020). Performance comparison of machine learning classifiers for fake news detection. *2020 2nd International Conference on Inventive Research in Computing Applications (ICIRCA)*, 696–700. https://doi.org/10.1109/icirca48905.2020.9183072

Sneha, N., & Gangil, T. (2019). Analysis of diabetes mellitus for early prediction using optimal features selection. *Journal of Big Data*, *6*(1), 1–19. https://doi.org/10.1186/s40537-019-0175-6

St, P. S., Pallotti, V., Gole, A., Kanherkar, P., Singh, S., Abhishek, P. R., & Wankhede, P. (2023). Comparative analysis of machine learning algorithms: random forest algorithm, Naive Bayes classifier and KNN-A survey. *International Journal for Research Publication and Seminar*, *13*(3), 194–197. https://www.researchgate.net/publication/369417666

Sultana, A., Islam, M., Hasan, M., & Ahmed, F. (2023). Fake news detection using machine learning techniques. *2023 IEEE/ACIS 21st International Conference on Software Engineering Research, Management and Applications (SERA)*, 98–103. https://doi.org/10.1109/sera57763.2023.10197712

Syed, L., Alsaeedi, A., Alhuri, L. A., & Aljohani, H. R. (2023). Hybrid weakly supervised learning with deep learning technique for detection of fake news from cyber propaganda. *Array*, *19*, 100309. https://doi.org/10.1016/j.array.2023.100309

Talwar, S., Dhir, A., Singh, D., Virk, G. S., & Salo, J. (2020). Sharing of fake news on social media: Application of the honeycomb framework and the third-person effect hypothesis. *Journal of Retailing and Consumer Services*, *57*, 102197. https://doi.org/10.1016/j.jretconser.2020.102197

Thakkar, P., Varma, K., Ukani, V., Mankad, S., & Tanwar, S. (2019). Combining user-based and item-based collaborative filtering using machine learning. In *Information and Communication Technology for Intelligent Systems* (pp. 173–180). Springer. https://doi.org/10.1007/978-981-13-1747-7_17

Truică, C.-O., Apostol, E.-S., Nicolescu, R.-C., & Karras, P. (2023). MCWDST: A minimum-cost weighted directed spanning tree algorithm for real-time fake news mitigation in social media. *IEEE Access*, *11*, 125861–125873. https://doi.org/10.1109/access.2023.3331220

Vadlamudi, P. S., Gunasekaran, M., & Nagalakshmi, T. J. (2023). An analysis of the effectiveness of the Naive Bayes algorithm and the support vector machine for detecting fake news on social media. *2023 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE)*, 726–731. https://doi.org/10.1109/iitcee57236.2023.10090978

Valero-Carreras, D., Alcaraz, J., & Landete, M. (2023). Comparing two SVM models through different metrics based on the confusion matrix. *Computers & Operations Research*, *152*, 106131. https://doi.org/10.1016/j.cor.2022.106131

Wang, S., Lei, Y., Yang, B., Li, X., Shu, Y., & Lu, N. (2023). A graph neural network-based data cleaning method to prevent intelligent fault diagnosis from data contamination. *Engineering Applications of Artificial Intelligence*, *126*, 107071. https://doi.org/10.1016/j.engappai.2023.107071

Wang, X., & Wang, C. (2020). Time series data cleaning: A survey. *IEEE Access*, *8*, 1866–1881. https://doi.org/10.1109/access.2019.2962152

Xiong, Z., Li, X., Zhang, X., Deng, M., Xu, F., Zhou, B., & Zeng, M. (2023). A comprehensive confirmation-based selfish node detection algorithm for socially aware networks. *Journal of Signal Processing Systems*, *95*(12), 1371–1389. https://doi.org/10.1007/s11265-023-01868-6

Yang, J., Gong, J., Tang, W., Shen, Y., Liu, C., & Gao, J. (2019). Delineation of urban growth boundaries using a patch-based cellular automata model under multiple spatial and socio-economic scenarios. *sustainability*, *11*(21), 6159. https://doi.org/10.3390/su11216159

Yogendra, P. P., Nagalakshmi, G., & Siva K. P. (2022). Fake news detection using machine learning. *Journal of Pharmaceutical Negative Results*, *13*(4), 1024–1030. https://doi.org/10.47750/pnr.2022.13.04.140

Zhang, J., Dong, B., & Yu, P. S. (2020). Fake detector: effective fake news detection with deep diffusive neural network. *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, 1826–1829. https://doi.org/10.1109/icde48307.2020.00180

Zhang, X., & Ghorbani, A. A. (2020). An overview of online fake news: characterization, detection and discussion. *Information Processing & Management*, *57*(2), 102025. https://doi.org/10.1016/j.ipm.2019.03.004