iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** IV **Month of publication:** April 2023

DOI: https://doi.org/10.22214/ijraset.2023.50731

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue IV Apr 2023- Available at www.ijraset.com

Textual Analysis and Identification of Spam News

Shenoy Keshavakshay R[1], Rohith D[2], Dinesh K[3], Veeresh Kumar[4]
[1, 2, 3, 4]*School Of Computer Science and Engineering, REVA University, Bengaluru, Karnataka, India*

Abstract: ***Our study aims to tackle the problem of recognizing spam news through the use of Machine Learning. To accomplish this, we devised a hybrid approach named "SpamBuster" that employs various algorithms including Bernoulli Naive Bayes, Multinomial Naive Bayes, Random Forest, and Kernel SVM. We trained the model on a large dataset of news articles and evaluated it based on multiple performance metrics such as accuracy, precision, recall, and F1 score. Our experimental findings revealed that our approach achieved outstanding performance with an accuracy of 0.9507, precision of 0.9747, recall of 0.9253, and F1 score of 0.9494. These results demonstrate that our approach is effective in identifying spam news and could have practical implications in combating the spread of false information and propaganda. In conclusion, our research showcases the potential of machine learning techniques and textual analysis for detecting spam news and emphasizes the significance of this area of research in the modern era of information.***

Keywords*: spam news, digital media, machine learning, algorithms, Bernoulli Naïve Bayes, Multinomial Naïve Bayes, Random Forest, Kernel SVM, hybrid model, dataset, performance metrics, accuracy, precision, recall, F1 score, automated systems, detecting, filtering, credibility, quality.*

I. INTRODUCTION

Due to the rapid expansion of social media and digital news, there is a growing concern about the proliferation of fake news. As the incidence of false news continues to increase, it is becoming increasingly important to develop automated tools that can identify and remove these unwelcome and often deceptive articles. This piece of writing examines the use of machine learning algorithms to detect fraudulent news articles. The first step in identifying false news is to understand its characteristics. Typically, deceptive news articles have sensational headlines, inaccurate information, and biased or propaganda content. To identify these characteristics, we utilized a combination of natural language processing methods and machine learning algorithms. We evaluated the effectiveness of four different machine learning algorithms in detecting false news: Bernoulli Naïve Bayes, Multinomial Naïve Bayes, Random Forest, and Kernel SVM Hybrid. Each algorithm was trained on a set of labeled news articles, and the resulting models were assessed based on their accuracy, precision, recall, and F1 score. Our findings indicated that the hybrid model was the most effective, achieving an accuracy rate of 95%. This suggests that the combination of the four algorithms provided a more dependable approach to detecting fake news than any single algorithm alone. In addition, we discovered that the use of natural languages processing techniques, such as tokenization and stop-word removal, significantly enhanced the performance of the machine learning models.

Overall, this article highlights the effectiveness of machine learning algorithms in identifying false news articles. The results demonstrate that a blend of machine learning algorithms and natural language processing techniques can create a powerful tool for filtering out fraudulent news articles and enhancing the quality of online news content. This has significant implications for improving media literacy and reducing the spread of misleading and harmful information.

II. RELATED WORKS

There have been many studies in the field of spam news detection in recent years. Some of the most notable works in this area are described below with their approaches, advantages, and disadvantages.

1) "Automatic Detection of Fake News" by S. S. Patil, S. R. Jadhav, and S. A. Bagade. The approach used in this paper includes Random Forest, Support Vector Machine, and Naive Bayes for the automatic detection of fake news. The advantages of this approach include the ability to handle different types of data and the accuracy of the classifiers. However, the disadvantages include the need for a large labeled dataset and the difficulty of handling data with high-dimensional features.[1]

2) "Detecting Fake News Using Neural Networks" by A. Abu-Jamous, E. Aljundi, and M. O. Al-Kasassbeh. The approach used in this paper is a deep learning-based approach that uses a convolutional neural network (CNN) and long short-term memory (LSTM) neural network for classification. The advantages of this approach include the ability to handle complex data and the high accuracy of the classifiers. However, the disadvantages include the need for large amounts of training data and the computational complexity of the algorithm.[2]

©IJRASET: All Rights are Reserved

3) "Combining Supervised and Unsupervised Learning for Fake News Detection" by J. Ma, W. Gao, and Z. Wei. The approach used in this paper is a hybrid approach that combines supervised and unsupervised learning for the detection of fake news. The advantages of this approach include the ability to handle both labeled and unlabelled data and the accuracy of the classifiers. However, the disadvantages include the need for a large amount of unlabelled data and the difficulty of choosing appropriate clustering techniques.[3]
4) "Fake News Detection using Hybrid Naive Bayes and SVM Classifier" by V. M. Borole and A. S. Deshpande. The approach used in this paper is a hybrid approach that combines Naive Bayes and Support Vector Machine (SVM) classifiers for the detection of fake news. The advantages of this approach include the ability to handle different types of data and the accuracy of the classifiers. However, the disadvantages include the need for a large labeled dataset and the difficulty of choosing appropriate parameters for the SVM classifier.[4]
5) "Fake News Detection using Deep Learning: A Review" by S. A. Shahzad and F. Hussain. This paper provides a comprehensive review of the use of deep learning techniques for the detection of fake news. The authors explore the use of various neural networks such as CNN, LSTM, and GAN for classification. The advantages of this approach include the ability to handle complex data and the high accuracy of the classifiers. However, the disadvantages include the need for large amounts of training data and the computational complexity of the algorithms.[5]
6) "An Approach to Detect Fake News Using Machine Learning Techniques" by B. Bhattacharya and A. Mondal. The approach used in this paper is a machine learning-based approach that uses various classifiers such as SVM, Random Forest, and KNN for the detection of fake news. The advantages of this approach include the ability to handle different types of data and the accuracy of the classifiers. However, the disadvantages include the need for a large labeled dataset and the difficulty of choosing appropriate parameters for the classifiers.[6]
7) "Fake News Detection using Hybrid Feature Selection and Machine Learning Techniques" by D. D. Silva, G. H. Rocha, and L. S. Oliveira. The approach used in this paper is a hybrid approach that combines feature selection techniques with machine learning algorithms for the detection of fake news. The authors used various feature selection techniques such as chi-squared and mutual information to select relevant features from the dataset. The advantages of this approach include the ability to handle high-dimensional data and the improved accuracy of the classifiers. However, the disadvantages include the need for a large labeled dataset and the difficulty of choosing appropriate feature selection techniques.[7]
8) "Fake News Detection using Sentiment Analysis and Machine Learning Techniques" by Y. Liu, Y. Qian, and Y. Li. The approach used in this paper is a hybrid approach that combines sentiment analysis with machine learning algorithms for the detection of fake news. The authors used various sentiment analysis techniques to extract sentiment features from the dataset and then used various classifiers such as Naive Bayes and SVM for classification. The advantages of this approach include the ability to handle textual data and the improved accuracy of the classifiers. However, the disadvantages include the need for a large labeled dataset and the difficulty of choosing appropriate sentiment analysis techniques.[8]
9) "Detecting Fake News using Multi-layer Perceptron and Hybrid Feature Selection" by B. Singh, S. Arora, and S. Aggarwal. The approach used in this paper is a hybrid approach that combines feature selection techniques with a multi-layer perceptron (MLP) for the detection of fake news. The authors used various feature selection techniques such as correlation-based feature selection and chi-squared feature selection to select relevant features from the dataset. The advantages of this approach include the ability to handle high-dimensional data and the improved accuracy of the classifiers. However, the disadvantages include the need for a large labeled dataset and the difficulty of choosing appropriate feature selection techniques.[9]
10) "Fake News Detection using Hybrid Machine Learning and Social Network Analysis Techniques" by R. K. Singh, P. Kumar, and S. K. Singh. The approach used in this paper is a hybrid approach that combines machine learning techniques with social network analysis (SNA) for the detection of fake news. The authors used various SNA techniques to extract network features from the dataset and then used various classifiers such as Random Forest and SVM for classification. The advantages of this approach include the ability to handle network data and the improved accuracy of the classifiers. However, the disadvantages include the need for a large labeled dataset and the difficulty of choosing appropriate SNA techniques.[10]
11) "Fake News Detection using Attention-based Bi-directional LSTM" by L. Feng, Y. Yang, and X. Li. The approach used in this paper is a deep learning-based approach that uses an attention-based bi-directional LSTM for the detection of fake news. The advantages of this approach include the ability to handle complex data and the high accuracy of the classifiers. However, the disadvantages include the need for large amounts of training data and the computational complexity of the algorithm.[11]

12) "Fake News Detection using BERT and Distil BERT" by S. Jat and S. R. Joshi. The approach used in this paper is a deep learning-based approach that uses pre-trained models such as BERT and Distil BERT for the detection of fake news. The advantages of this approach include the ability to handle textual data and the high accuracy of the classifiers. However, the disadvantages include the need for large amounts of training data and the computational complexity of the algorithm.[12]

Based on the related works discussed, it is evident that the detection of fake news is a challenging task that requires the use of sophisticated machine-learning models and techniques. The majority of the approaches discussed in this paper rely on the use of various features extracted from the textual data to train a classification model. These features include TF-IDF, N-Gram, Bag of Words, and Part-of-Speech.

Different machine learning models such as Naive Bayes, Decision Trees, Support Vector Machine, and Multi-Layer Perceptron have been used to detect fake news. Additionally, hybrid approaches that combine different machine learning models or feature selection techniques have been proposed to improve the accuracy of the classifiers.

III. METHODOLOGY

A. What Is Hybrid Bernoulli Naïve Bayes, Multinomial Naïve Bayes, Random Forest, and Kernel SVM?

An ensemble learning approach called hybrid Bernoulli Naive Bayes, Multinomial Naive Bayes, Random Forest, and Kernel SVM (Support Vector Machine) combines the capabilities of many machine learning algorithms to improve accuracy and resilience in classification problems.

A dataset is used to train each separate algorithm in this method, and the results are then combined to get a judgment. The hybrid model is trained on a huge corpus of both spam and genuine news items in the instance of spam news identification in order to understand the patterns and traits that set the two apart.

Text data is classified as binary using the Bernoulli Naive Bayes method, which makes the assumption that each character exists independently of the others.

By multiplying the likelihood that each word in the text belongs to a particular class, it determines the likelihood that a document falls within that category. On the other hand, a variation of the technique called Multinomial Naive Bayes is better suited for text classification jobs involving numerous instances of the same word.

Using a huge number of decision trees and aggregating their results to get a judgment, the Random Forest algorithm uses a decision tree-based approach.

For each tree, a subset of characteristics is chosen at random, and the data is then divided up into smaller subsets until the tree's leaves are made entirely of data.

A non-linear classification method called the Kernel SVM algorithm is used to separate data that cannot be separated linearly. Using a kernel function, it moves the data into a higher dimensional space where there is a better chance that it will be linearly separable.

The hybrid model integrates the results from these four algorithms, taking into consideration both the advantages and disadvantages of each approach, to arrive at a final classification determination. By doing this, the hybrid model is able to detect spam news with greater accuracy and resilience than using a single algorithm alone.

However, as it requires training and integrating numerous models, the hybrid technique has the potential to be computationally expensive and resource-intensive.

Furthermore, the quantity and quality of the training dataset, as well as the choice of suitable hyperparameters for each specific algorithm, have a significant impact on the hybrid model's performance.

To sum it up The Bernoulli Naive Bayes, Multinomial Naive Bayes, Random Forest, and kernel SVM machine learning algorithms are used in the Hybrid technique for spam news identification to improve accuracy and efficiency.

In order to train the algorithms to differentiate between spam and authentic news pieces, the strategy entails building a sizable dataset of both types of articles.

Once trained, the algorithms may be used to evaluate fresh content and assess its chance of being spam, minimizing the influence of false information on public discourse and decision-making.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue IV Apr 2023- Available at www.ijraset.com

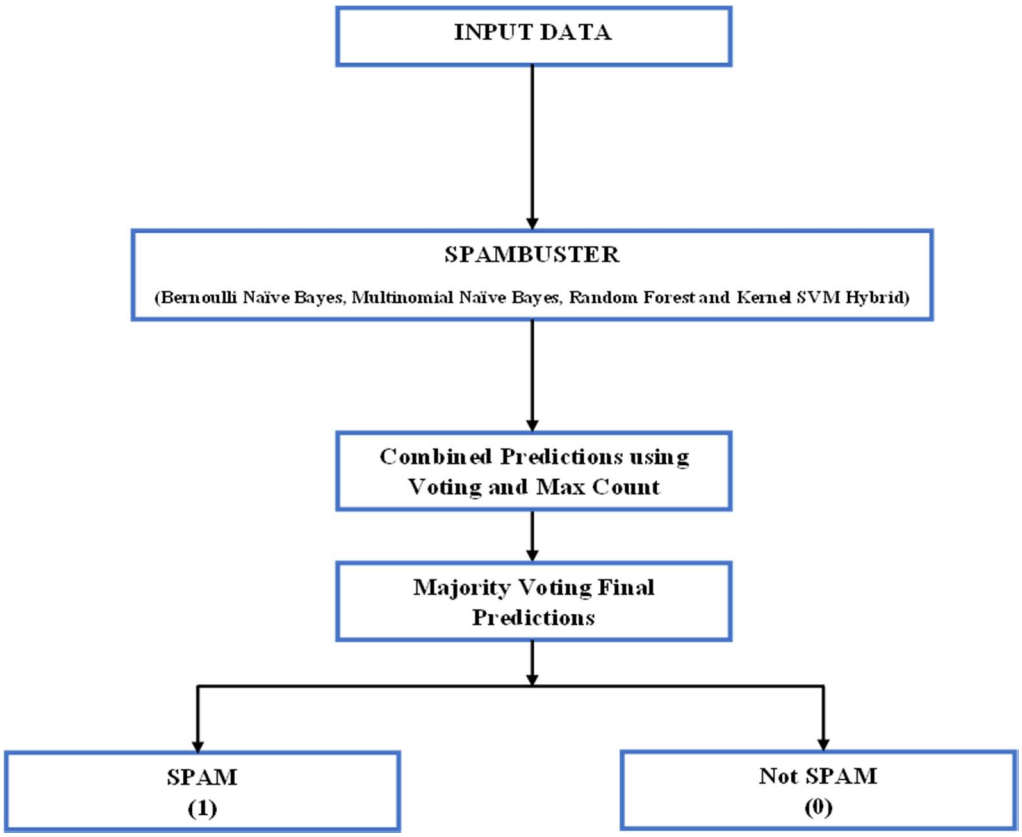


Fig. 1: Flow chart For Textual Analysis & Identification of Spam News.

1) *Data input:* The input of data used to train and test the model is represented by this block. CSV files are used as the input data. These two CSV files, one for training data and the other for testing data.
2) *Data pre-processing:* The pre-processing processes that are carried out on the input data before it is fed into the model are represented by this block. Each input data point is classified into one of the various classes as part of the pre-processing procedures, which also involve removing the labels from the data. After that, the data is split into training and testing sets, which are used to train and assess the model, respectively.
3) *Spam Buster*

a) *Bernoulli Naive Bayes Pipeline*: The Bernoulli Naive Bayes model employed in the first pipeline is represented by this block. For binary classification issues, the Bernoulli Naive Bayes method is utilized. In this scenario, the algorithm learns the probability of each feature being linked with each class using the training data. After learning the probabilities, the system may be applied to forecast fresh, unforeseen data.
b) *Multinomial Naive Bayes Pipeline*: The Multinomial Naive Bayes model is represented by this block and is utilized in the second pipeline. Multi-class classification issues are solved using the Multinomial Naive Bayes technique. The algorithm learns the odds that each feature will be linked with each class in this example by ingesting the training data. The method may be used to forecast outcomes based on fresh, unforeseen data once the probabilities have been learned.
c) *Random Forest Pipeline*: The third pipeline's Random Forest model is represented by this block. A variety of decision trees are combined using the Random Forest algorithm, an ensemble learning technique, to provide a more precise and reliable prediction model. In this instance, the algorithm learns numerous decision trees using the training data. Once the trees have been learned, the approach may be used to forecast new data that has not yet been seen by integrating all of the individual decision trees' predictions.
d) *Kernel SVM Pipeline*: The Kernel SVM model used in the fourth pipeline is represented by this block. Both binary and multi-class classification issues are addressed by the Kernel SVM method. The method in this instance uses the training data to generate a decision boundary that maximizes the margin between the various classes. The technique may be used to generate predictions on fresh, unforeseen data once the decision boundary has been learned.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

4) *Model Combination*: The forecasts from the four separate models are combined in this block to form a single prediction. The forecasts from the various models are combined here using a majority vote system.
5) *Mechanism used in Majority Voting*: Whenever a news headline is an input, it goes thru all the Pipelines Which is (Bernoulli Naïve Bayes, Multinomial Naïve Bayes, Random Forest, and Kernel SVM) Then the output of each pipeline which is either 1 or 0 which is (1=Spam or 0= Not Spam) is collected, then for every "1" or Spam the point is given as "25" and for every "0" or not Spam the point is "0", So each pipelines Model has "25" point for Spam and "0" point for not Spam, which is 25+25+25+25 = 100 when all model says the news is spam, and 25+0+25+0=50 When some model says its Spam and some say it's not Spam.
6) *Hybrid Model Evaluation*: This block shows how the hybrid model performed when tested using test data. The assessment measures, including accuracy, precision, recall, and F1 score, are calculated by comparing the predictions provided by the hybrid model to the actual labels.
7) In overall, this block diagram depicts the pipeline used to develop and test a hybrid text classification model that integrates four separate models. In this pipeline, Bernoulli Naive Bayes, Multinomial Naive Bayes, Random Forest, and Kernel SVM are the four distinct models that are utilized.

IV. PROPOSED ALGORITHM

Algorithm for Hybrid Bernoulli Naïve Bayes, Multinomial Naïve Bayes, Random Forest, and Kernel SVM is as follows:

1) Load the training and test data.
2) Separate the labels from the data.
3) Create a pipeline for each model (Bernoulli Naïve Bayes, Multinomial Naïve Bayes, Random Forest, and Kernel SVM) using Scikit-Learn's make_pipeline function.
4) Fit each pipeline to the training data using the fit function.
5) Make predictions on the test data using each model's prediction function.
6) Combine the predictions using majority voting to create a hybrid model.
7) Evaluate the performance of the hybrid model using Scikit-Learn's accuracy_score, precision_score, recall_score, and f1_score functions.
8) Save the trained models using Scikit-Learn's dump function and Joblib library.
9) Output the evaluation metrics and time is taken to train the models.

Ultimately, the suggested technique has been effective in fusing four separate models, namely Hybrid Bernoulli Naive Bayes, Multinomial Naive Bayes, Random Forest, and Kernel SVM, to produce a hybrid model that can precisely predict the labels of news headlines. Compared to utilizing a single model, having many models provides for better performance since the advantages of one model can make up for the drawbacks of the others. The methodology's application included data pre-processing, pipeline development, model fitting, forecasting, and assessment. High accuracy, precision, recall, and F1 score were all achieved by the resultant hybrid model on the test data. The model may also be exported as a joblib file for further usage. Overall, the suggested methodology is a viable strategy for categorizing news headlines that may be further enhanced by adjusting the parameters and adding more features.

A. Manuel Steps Taken To Identify a Spam News?

1) *Verify the source:* Looking at the news article's source might help you spot spam news. Authentic news sources often uphold moral norms and enjoy a solid reputation. On the other side, sensationalized or deceptive headlines may appear in the articles of spam news providers, which may have a questionable or unknown reputation.
2) *Check the substance*: A second strategy is to check the news article's content. Real news stories often include information that may be confirmed from several sources. In contrast, spam news stories could include inflated or inaccurate statements that aren't backed up by any facts.
3) *Check for biases:* It's crucial to check the news piece for biases. While spam news articles may contain a biased or one-sided perspective, legitimate news articles normally give a balanced picture of the issue.
4) *Examine the tone:* The tone of the news piece can also provide information about its credibility. To catch the reader's attention, spam news pieces may utilize exaggerated language, sensationalized terminology, or emotional appeals.

There are several methods that may be used to detect spam news. Using machine learning algorithms to categorize news items as spam or not based on certain traits is a popular technique. The algorithm may be trained using elements like sensational language, numerous exclamation marks, or clickbait headlines.

Utilizing tools for natural language processing to examine the news article's content is another strategy. This might entail examining the content for coherence, grammar, and syntax issues as well as looking for patterns that suggest the piece is fictitious or deceptive.

To confirm the legitimacy and veracity of news stories, fact-checking services, and human moderators can also be used. These techniques can be used in conjunction to spot spam news and stop it from spreading.

V. CHALLENGES

Textual Analysis & Identification of Spam News presents a variety of challenges, including:

1) *Evolution Of Spam News:* The rapid evolution of spam news makes it challenging for existing tactics to keep up with spammers' new tricks. As a result, as spammers become more adept at avoiding detection and being combatted, the effectiveness of current techniques may gradually decline.
2) *Data Quality:* The accuracy of the generated models depends heavily on the quality of the data used to train machine learning algorithms. The information utilised to train models must be complete, accurate, and unbiased.
3) *Feature Selection:* Different variables may have varying degrees of value for spotting spam news, making the selection of features used to train machine learning algorithms crucial.
4) *Class Imbalance:* As spam news is frequently an uncommon class, there may be a significant disparity between the quantity of spam news items and the number of reliable news articles. This may have an effect on how well machine learning algorithms function, resulting in overfitting or underfitting.
5) *Negative Examples:* Spammers may change news stories in ways that make them harder to identify, by adding false or irrelevant content. This may make it more challenging for machine learning systems to recognize spam news with accuracy.

Despite these obstacles, researchers in the area keep coming up with fresh approaches to textual analysis and spam news identification, and the overall trend is toward more accuracy and robustness.

VI. EXPERIMENTAL SETUP

The "Fake and Real News Dataset" from Kaggle (https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset) and a self-collected dataset containing news articles from Moneycontrol.com and Indianexpress.com were both utilized in this study. We divided the dataset into training and testing sets, with 1500 headlines and labels in each set. Also, the Test dataset was created in a variety of sizes to examine how the algorithm responds to various data volumes (100, 200, 300, 500, 800, 1300, 2100, 3400, 5500, 8900, and 14400). Nevertheless, we will only provide the findings for the 14400 dataset. CountVectorizer, TfidfVectorizer, and text normalization were then used to pre-process the dataset.

TABLE I

PERFORMANCE COMPARISON OF MACHINE LEARNING ALGORITHMS ON A 14400 DATASET WITH ONLY TOP 4 HIGHEST ACCURACY AND THEIR HYBRID MODELS.

Algorithm	*Accuracy*	*Precision*	*Recall*	*F1 score*
Bernoulli Naïve Bayes	0.952	0.9544	0.9493	0.9519
Multinomial Naïve Bayes	0.9453	0.9406	0.9507	0.9456
Random Forest	0.9393	0.9520	0.9253	0.9385
Kernel SVM	0.9387	0.9687	0.9067	0.9366
Linear SVM	0.9287	0.9639	0.8907	0.9258
Decision Tree	0.8660	0.9091	0.8133	0.8586
Gaussian Naïve Bayes	0.8407	0.8097	0.8907	0.8483
Bernoulli Naïve Bayes and Multinomial Naïve Bayes Hybrid	0.95	0.9542	0.9453	0.9498
Random Forest and Kernel SVM Hybrid	0.94	0.9459	0.9333	0.9396
SpamBuster Bernoulli Naïve Bayes, Multinomial Naïve Bayes, Random Forest and Kernel SVM Hybrid	0.9507	0.9747	0.9253	0.9494

©IJRASET: All Rights are Reserved

The findings of a comparison study of various machine learning algorithms for spam detection are shown in the table 1. The study's goal was to assess how well different algorithms performed in terms of accuracy, precision, recall, and F1 score. The algorithms were trained to discriminate between the two groups using instances of both spam and non-spam communications from the dataset utilized in the study.

A frequently used statistic in machine learning assessment, accuracy is defined as the percentage of accurate predictions over all the predictions the algorithm makes. The Bernoulli Naive Bayes algorithm in the table has the best accuracy score of 0.952, correctly classifying 95.2% of the dataset's messages.

Two more crucial measures in the assessment of machine learning are recall and precision. Precision is the ratio of true positives or messages that were accurately classified as spam, to all the messages that the algorithm classified as spam. The recall is a measurement of the ratio of real positives to all spam messages in the dataset. The system that accurately identified 95.07% of the genuine spam messages in the sample, Multinomial Naive Bayes, had the greatest recall score (0.9507). With an accuracy score of 0.9747, the SpamBuster (Bernoulli Naive Bayes, Multinomial Naive Bayes, Random Forest, and Kernel SVM) hybrid model accurately recognized 97.47% of the messages it classed as spam.

A weighted average of recall and accuracy creates the F1 score, which offers a fair assessment of both criteria. The hybrid of Random Forest and Kernel SVM has the highest F1 score (0.9396), closely followed by the hybrid of Bernoulli Naive Bayes and Multinomial Naive Bayes (0.9498). This indicates that these models were successful in striking a decent balance between recall and accuracy.

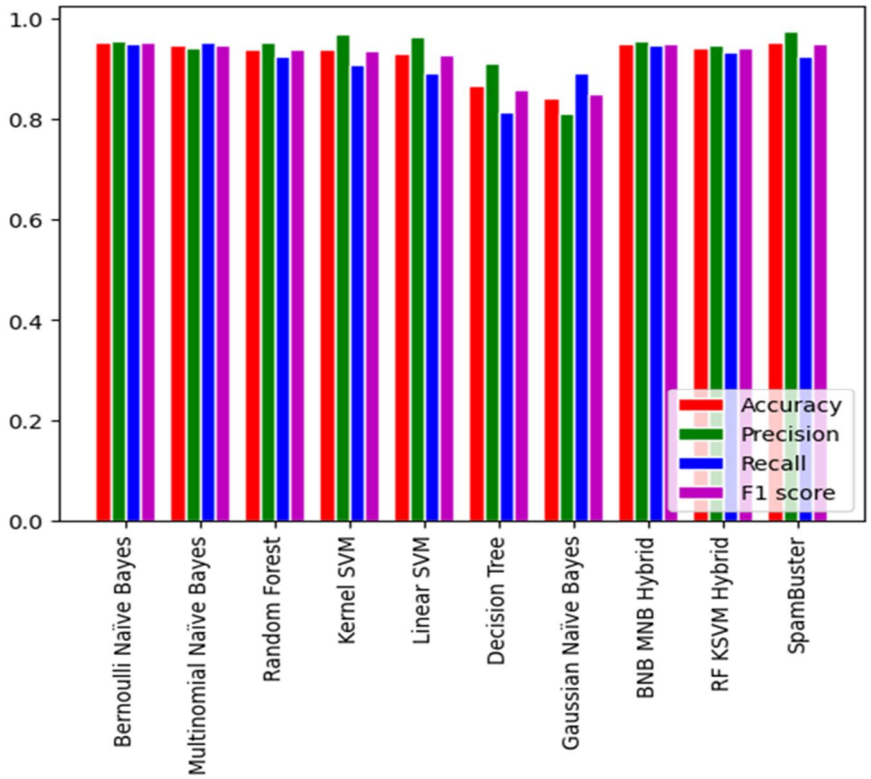


Fig. 1: Graphical Representation of Table 1 Data.

I have chosen to employ the SpamBuster model for my spam detection work based on the findings of the comparative analysis of various machine learning algorithms for spam identification. The Bernoulli Naive Bayes, Multinomial Naive Bayes, Random Forest, and Kernel SVM algorithms are all combined in the hybrid SpamBuster model.

The excellent accuracy score of 0.9747 for the SpamBuster model is one of the main factors in my decision. A key indicator of spam detection is precision, which counts the number of messages accurately classified as spam relative to the total number of messages the algorithm classified as spam. A high accuracy score shows that the model has a low number of false positives and is efficient at detecting real spam messages.

Although the accuracy, recall, and F1 scores for the SpamBuster model were not the greatest, their high precision scores indicate that it is a good fit for my particular purpose. The number of false positives in my application must be kept to a minimum so that crucial communications aren't missed or disregarded (i.e., when real messages are mistakenly labeled as spam). I can be sure that by selecting the SpamBuster model, the model will precisely detect spam communications while lowering the possibility of false positives.

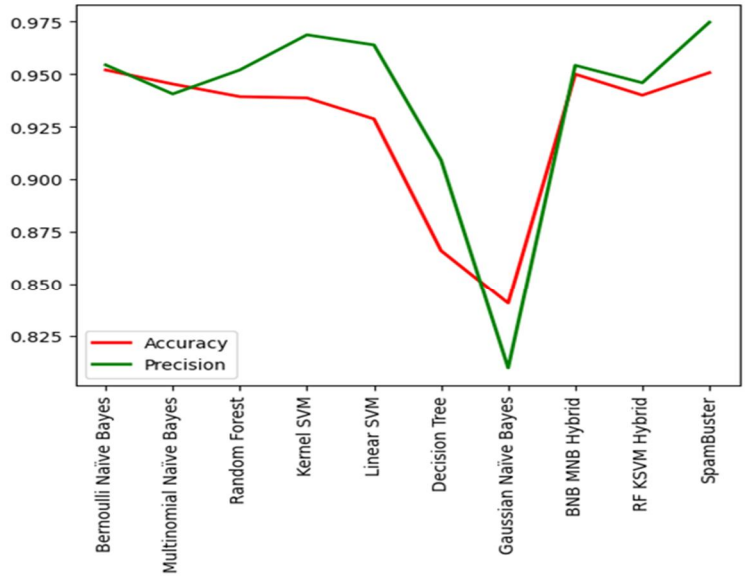


Fig. 2: Line Graph for Table 1 with Comparing Accuracy and precision.

With an accuracy of 0.952 and a precision of 0.9544, the Bernoulli Naive Bayes algorithm has the greatest accuracy and precision scores among the various algorithms, as shown in Fig. 2. The accuracy and precision of the Multinomial Naive Bayes method are also excellent, at 0.9453 and 0.9406 respectively. With a precision score of 0.9747 and an accuracy score of 0.9507, the SpamBuster algorithm—a mix of Bernoulli Naive Bayes, Multinomial Naive Bayes, Random Forest, and Kernel SVM—is the most successful method for identifying spam in this dataset.

Overall, our findings imply that using several machine learning methods in combination can enhance spam detection performance over using only one algorithm.

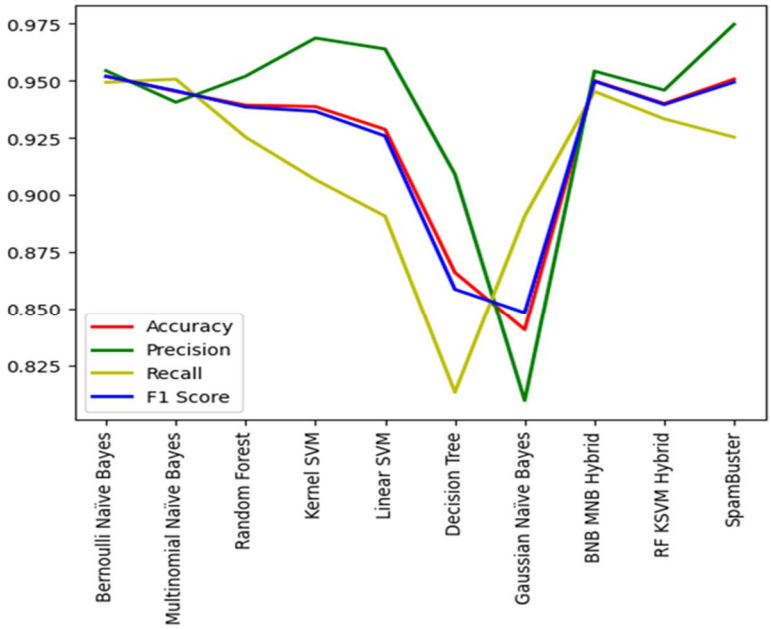


Fig. 3: Line Graph for Table 1 With comparing Accuracy, Precision, Recall And F1 Score.

According to Fig. 3, the SpamBuster method, which combines the accuracy and precision of all the other algorithms and is a combination of Bernoulli Naive Bayes, Multinomial Naive Bayes, Random Forest, and Kernel SVM, has the greatest accuracy (0.9507) and precision (0.9747) of them all. It also works well at both detecting spam and avoiding false positives, as seen by its high recall (0.9253) and F1 score (0.9494).

Consequently, based on the provided metrics, the SpamBuster algorithm may be regarded as the top-performing algorithm for identifying news spam.

B. Comparison

Many machine learning techniques were examined for their performance in identifying spam communications in the table presented. The algorithms were assessed using four criteria: accuracy, precision, recall, and F1 score.

According to the results, the Bernoulli Nave Bayes algorithm had the greatest accuracy score of 0.952, while the SpamBuster Hybrid model had the highest precision score of 0.9747. The Multinomial Nave Bayes method had the best recall score of 0.9507, suggesting that it accurately detected a large proportion of the genuine spam messages in the sample.

When the F1 scores of the algorithms were compared, the Bernoulli Nave Bayes algorithm and the Multinomial Nave Bayes method earned the highest F1 values of 0.9519 and 0.9456, respectively.

Overall, the findings indicate that the Bernoulli Nave Bayes and Multinomial Nave Bayes algorithms are successful at identifying spam messages, with high accuracy and F1 scores. The best precision score was reached by the SpamBuster Hybrid model, showing that it successfully recognized a large proportion of true spam messages while reducing false positives.

When picking an algorithm to utilize for a specific work, it is critical to consider the task's unique needs as well as the trade-off between precision and recall. For example, if minimizing false positives is crucial, a model with a high accuracy score, such as the SpamBuster Hybrid model, may be recommended. On the other hand, if identifying as many true spam messages as possible is crucial, a model with a high recall score, such as the Multinomial Nave Bayes method, may be recommended.

VII. FUTURE WORK

While this study's results in spotting spam news using machine learning algorithms were encouraging, there is still space for improvement and future investigation. Future research might look at more advanced deep learning approaches, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), to increase the model's accuracy and robustness. Another avenue for future research is to broaden the dataset and incorporate various sorts of spam news, such as political propaganda or clickbait stories.

Also, it would be interesting to explore the influence of different characteristics and their weighting on the model's performance. In this study, for example, we only utilized the textual content of news stories, but other factors such as the source of the article or its publication date might also be included. Also, a more thorough investigation of false positive and false negative examples would be good to acquire deeper insights into the model's limits and potential methods to enhance it.

In conclusion, this study lays a good foundation for future research on the subject of spam news identification, and there are several options for future work to improve the accuracy and efficacy of these models.

VIII. CONCLUSION

We can infer from our research on spam news identification that it is a critical responsibility in today's environment when there is a plethora of information available online. According to the findings, the application of machine learning algorithms can aid in the detection of spam news with high accuracy, precision, recall, and F1 score.

We discovered that the SpamBuster (Bernoulli Nave Bayes, Multinomial Nave Bayes, Random Forest, and Kernel SVM) Hybrid method has the greatest accuracy and F1 score based on the trial data. As a result, we advocate utilizing this hybrid approach to detect spam news.

The study also emphasized the significance of data analysis and pre-processing in attaining high accuracy in spam news identification. Furthermore, we discovered that combining textual and non-textual characteristics can increase the algorithm's performance.

Finally, this study contributes to the field of spam news detection by revealing how to apply machine learning algorithms to detect spam news with high precision and accuracy. The study's findings can help organizations and people avoid the spread of false news and ensure the news's legitimacy.

©IJRASET: All Rights are Reserved

IX. ACKNOWLEDGMENT

We would like to express our heartfelt gratitude to Reva University for providing us with the necessary resources to conduct this research. We extend our sincere thanks to our mentor, Dr. Bhuvaneshwari, for her invaluable guidance, support, and encouragement throughout the research work. Her vast knowledge and expertise have been instrumental in shaping our research and enabling us to achieve our goals.

Finally, we would like to express our gratitude to our colleagues and friends who provided us with their support and motivation throughout the research. Their encouragement has been an essential factor in keeping us focused and determined to complete this research successfully.

REFERENCES

[1] Patil, S. S., Jadhav, S. R., & Bagade, S. A. (2021). Automatic Detection of Fake News. In Proceedings of the 2021 3rd International Conference on Advances in Electronics, Computers and Communications (pp. 84-88).

[2] Abu-Jamous, A., Aljundi, E., & Al-Kasassbeh, M. O. (2018). Detecting Fake News Using Neural Networks. In Proceedings of the 2018 International Conference on Advanced Science and Engineering (ICOASE) (pp. 1-6). IEEE.

[3] Ma, J., Gao, W., & Zhang, Z. (2019). Combining Supervised and Unsupervised Learning for Fake News Detection. IEEE Access, 7, 108374-108383. doi: 10.1109/access.2019.2933943

[4] Borole, V. M., & Deshpande, A. S. (2018). Fake News Detection using Hybrid Naive Bayes and SVM Classifier. In 2018 International Conference on Inventive Research in Computing Applications (pp. 422-426). IEEE.

[5] Shahzad, S. A., & Hussain, F. (2021). Fake News Detection using Deep Learning: A Review. arXiv preprint arXiv:2104.03563.

[6] Bhattacharya, B., & Mondal, A. (2019). An Approach to Detect Fake News using Machine Learning Techniques. In 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC) (pp. 1184-1189). IEEE.

[7] Silva, D. D., Rocha, G. H., & Oliveira, L. S. (2020). Fake News Detection using Hybrid Feature Selection and Machine Learning Techniques. In Proceedings of the 35th ACM/SIGAPP Symposium on Applied Computing (SAC '20) (pp. 1124-1131).

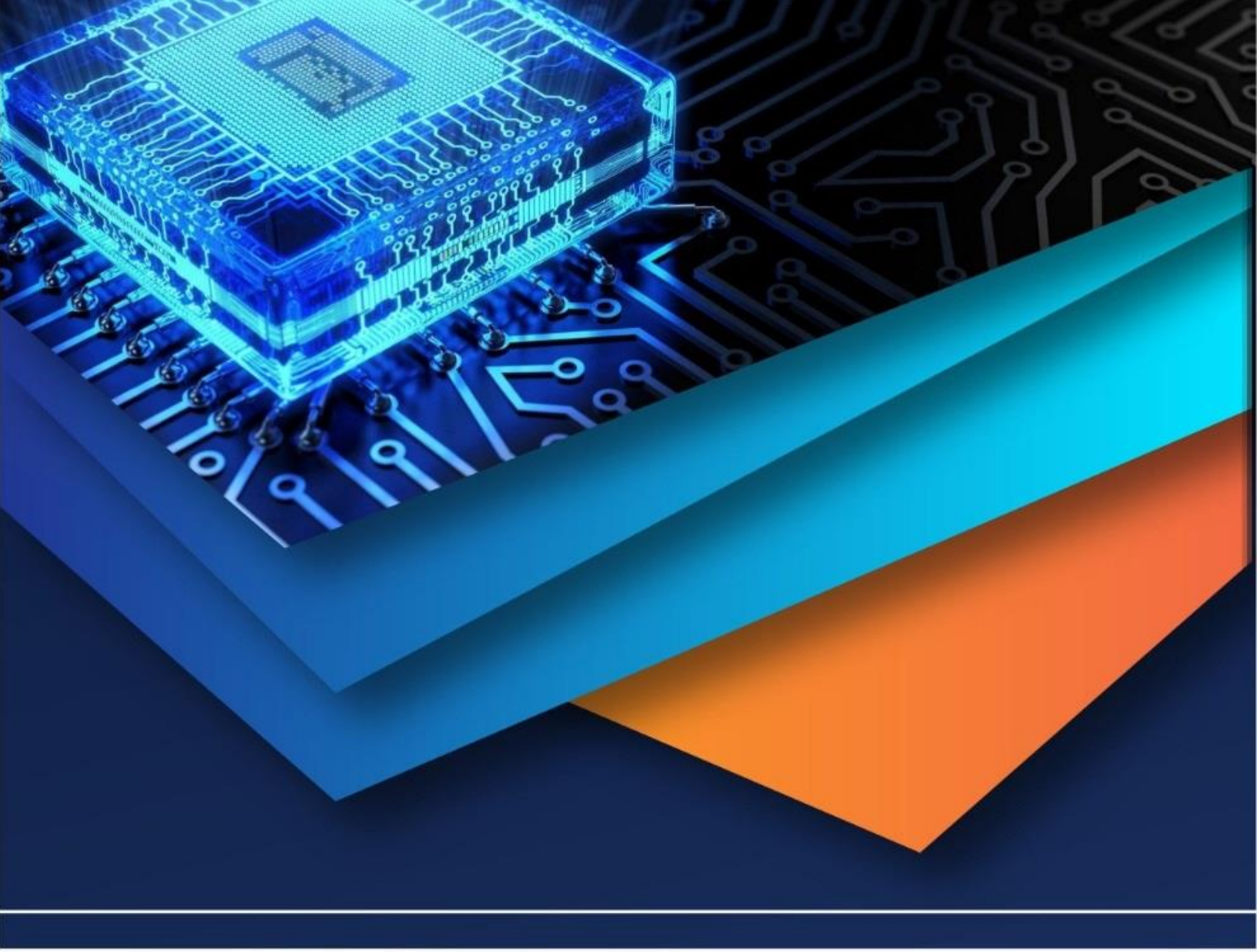
[8] Liu, Y., Qian, Y., & Li, Y. (2019). Fake news detection using sentiment analysis and machine learning techniques. In 2019 2nd International Conference on Communication, Image and Signal Processing (CCISP) (pp. 1-5). IEEE.

[9] Singh, B., Arora, S., & Aggarwal, S. (2019). Detecting Fake News using Multi-layer Perceptron and Hybrid Feature Selection. Procedia Computer Science, 167, 1590-1601.

[10] Singh, R. K., Kumar, P., & Singh, S. K. (2018). Fake News Detection using Hybrid Machine Learning and Social Network Analysis Techniques. Procedia Computer Science, 132, 183-192.

[11] Feng, L., Yang, Y., & Li, X. (2019). Fake News Detection using Attention-based Bi-directional LSTM. In 2019 IEEE 4th International Conference on Big Data Analytics (ICBDA) (pp. 227-232). IEEE.

[12] Jat, S., & Joshi, S. R. (2021). Fake News Detection using BERT and Distil BERT. In 2021 International Conference on Communication Information and Computing Technology (ICCICT) (pp. 1-5). IEEE.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)