FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 9 No. 1, January, 2025, pp 393 – 399
DOI: https://doi.org/10.33003/fjs-2025-0901-3226

ENHANCED SMS SPAM DETECTION USING BERNOULLI NAIVE BAYES WITH TF-IDF

***Abdullahi B. Ahmed and Khalid Haruna**

Department of Computer Science, Faculty of Computer and Information Technology, Bayero University Kano.

*Corresponding authors' email: burhanbobe@gmail.com

ABSTRACT

The use of mobile text messaging for communication is increasingly widespread, with Short Message Service (SMS) experiencing significant growth over the last decade. Consequently, the increase in SMS usage has led to a concerning rise in SMS spam, presenting substantial challenges for users and service providers. This study proposes a novel method for detecting SMS spam by combining Term Frequency-Inverse Document Frequency (TF-IDF) with Bernoulli Naïve Bayes (BNB) algorithm. The approach employs the use of TF-IDF for comprehensive feature extraction and the classification capabilities of the Bernoulli Naïve Bayes Algorithm. Through experimental validation employing TF-IDF for feature extraction and the BNB algorithm for classification, the results demonstrate high accuracy (98.36%), precision (99.19%), and a notable Matthews Correlation Coefficient (MCC) of 0.93, showcasing superior model performance compared to existing benchmarks. Likewise, the proposed model shows efficient processing time (0.22 seconds). By combining strengths of TF-IDF and BNB, the approach offers effective SMS spam detection, surpassing the performance of traditional and deep learning classifiers. This research contributes valuable insights towards enhancing SMS security, thereby increasing trust between users and service providers.

Keywords: SMS Spam Detection, TF-IDF, Bernoulli Naïve Bayes, Machine Learning, Text Classification, Feature Extraction

INTRODUCTION

Text messaging, or Short Message Service (SMS), has evolved from a simple tool for personal communication into a widely used medium for business and global interactions (Lubis et al., 2019). Initially designed for brief exchanges, SMS has expanded to serve various purposes, including customer engagement, promotional campaigns, and critical alerts, thanks to its high open rate and instant delivery (Oswald et al., 2022). Additionally, SMS remains highly accessible, functioning on nearly all mobile devices without requiring an internet connection, making it an essential communication tool in regions with limited internet access (Broadbent, 2020). The increasing integration of SMS with emerging technologies, such as social media and multimedia messaging, has further enhanced its functionality. A recent study by the Nigerian Communications Commission (NCC) reported that over 25.9 billion text messages were exchanged in Nigeria in 2022, reflecting a 28% increase from the previous year, emphasizing the growing reliance on SMS for both personal and professional communication (Li, 2019).

Despite its benefits, SMS is increasingly exploited by spammers, leading to security concerns and privacy risks (Abayomi-Alli et al., 2019). Spam messages, including financial fraud, phishing attempts, and job-related scams, disrupt communication and pose significant threats, such as identity theft and financial loss (Poster, 2022). Traditional rule-based filtering techniques have proven ineffective against evolving spam tactics, necessitating the adoption of machine learning-based approaches (Karim et al., 2019). Supervised learning models, such as Bernoulli Naïve Bayes (BNB) and Term Frequency-Inverse Document Frequency (TF-IDF), have demonstrated improved spam detection accuracy by analyzing message patterns and word frequencies (Rodrigues et al., 2022). However, challenges such as high computational costs and scalability issues persist. This research aims to enhance SMS spam detection by integrating BNB with TF-IDF to improve classification accuracy, optimize processing time, and refine Matthews Correlation Coefficient (MCC) performance, ultimately contributing to more secure and reliable mobile communication systems (Al Saidat et al., 2024).

The dynamic evolution of spam message formats has increases significant research interest. Researchers have explored various machine learning and deep learning techniques for spam classification to achieve accurate results. However, the effectiveness of classification heavily relies on the quality of embedding generated. Natural Language Processing (NLP) and Artificial Intelligence (AI) models play a crucial role in text classification tasks. These models analyze and categorize textual data by extracting meaningful features, which can enhance classification accuracy (Ajueyitsi & Ekuobase, 2024).

Nonetheless, there's been limited research into different embedding techniques or the integration of different techniques for embedding with machine learning for classification.

A study by Sjarif, (2020) found that spam messages remain a challenge despite filtering systems, with the Random Forest algorithm demonstrating superior performance in spam classification (accuracy of 97.50%)

Similarly, Gadde et al., (2021) integrated diverse embedding techniques such as count vectorizer, TF-IDF (term frequency and inverse document frequency) vectorizer, and hashing vectorizer with machine learning classification models like naive Bayes, logistic regression, K-nearest neighbor, random forest, support vector machine, and decision tree, alongside LSTM. Their study, utilizing the UCI SMS spam collection dataset, demonstrated that LSTM yielded the highest accuracy of 98.5%.

In a comparative study, Gupta et al., (2019) investigated SMS detection employing machine learning classifiers, artificial neural networks, and convolutional neural networks (CNN) using two datasets: SMS Spam Collection v.1 and Spam SMS Dataset 2011-12. However, their focus was not on embeddings. Furthermore, Roy et al., (2020) proposed a deep learning model employing CNN and LSTM models for classifying spam and ham text messages using the SMS Spam Collection dataset.

In a study made by Jain et al. (2019), the suthors introduced a semantic LSTM method for detecting and classifying spam SMS using the SMS Spam Collection dataset and Twitter dataset leveraging Google's Word2Vec for semantic layer creation.

Wei & Nguyen (2020) enhanced Jain et al.'s work by proposing Lightweight Gated Recurrent Unit (LGRU) for SMS spam detection. They replaced the LSTM layer with LGRU and used external knowledge (WordNet) to enrich semantics.

Moreover, Gupta et al. (2019) introduced a spam detection model based on ensemble learning, combining weak classifiers like Gaussian naive Bayes, Bernoulli naive Bayes, multinomial naive Bayes, and decision tree. The ensemble method, using a voting classifier, outperformed individual classifiers, achieving an accuracy of 98.29% on the SMS spam collection dataset.

Dada et al. (2019) conducted a comprehensive review of machine learning methods for spam email detection and classification. They explored various spam filtering techniques and recommended count-based approaches using machine learning for effective classification, while suggesting deep learning and deep adversarial learning as future techniques to combat spam emails.

Furthermore, Wang et al. (2019) developed a predictive scheme using Bayesian linear regression and random forest regression for numerical prediction on the spam dataset from the UCI Machine Learning Repository, achieving better accuracy with random forest regression. However, there remains a need to extract more relevant attributes (embeddings) to enhance spam detection models. In this work, we are proposing a classification model that will not only focus on the accuracy of the classification but will attempt to improve the classification of complex SMS sentences into either Spam or Ham.

MATERIALS AND METHODS

We have presented the research design, system architecture, the proposed method for SMS spam detection, data acquisition, preprocessing techniques, tools used for analysis, and performance evaluation metrics. The study leverages advanced feature extraction techniques and machine learning algorithms to enhance SMS spam detection accuracy.

Research Design

The research framework for the SMS spam detection project is illustrated in Figure 1. It encompasses existing approaches to SMS spam detection, the proposed methodology, and the overarching objective function guiding the research. The process begins with data collection, where a dataset comprising both spam and legitimate SMS messages is gathered. This is followed by data preprocessing, which involves text cleaning techniques such as removing punctuation, numbers, and special characters, converting text to lowercase, eliminating stop words, and applying stemming or lemmatization to standardize textual data. Next, feature extraction is performed using the Term Frequency-Inverse Document Frequency (TF-IDF) method to quantify the importance of words within the dataset, thereby improving the representation of textual features. The model training phase employs the Bernoulli Naïve Bayes (BNB) algorithm, where the dataset is split into training and testing subsets. The BNB model is then trained by calculating class priors and conditional probabilities to classify messages as spam or non-spam. Finally, model evaluation assesses the system's performance using standard classification metrics. By leveraging the strengths of both techniques, the proposed framework aims to optimize accuracy and efficiency in distinguishing spam from legitimate messages.

The SMS spam detection system architecture is designed to effectively process incoming messages and accurately classify them as either spam or legitimate. The integration of TF-IDF and the BNB algorithm forms the backbone of the system's classification mechanism. TF-IDF quantifies the significance of terms within SMS messages, considering their frequency of occurrence and rarity across the message corpus, while the BNB algorithm applies probabilistic principles to categorize messages based on the likelihood of their belonging to the spam category.

TF-IDF and Bernoulli Naïve Bayes (BNB) Algorithm

TF-IDF plays a crucial role in SMS spam detection by transforming textual data into numerical representations that highlight the significance of words in a message. It assigns greater importance to words that appear frequently within an individual message but are less common across the entire dataset. This weighting mechanism helps differentiate between commonly used terms and those that are more indicative of spam content. By emphasizing distinctive words, TF-IDF enhances feature extraction, enabling the classification model to make more informed decisions. The formulation for TF-IDF is given as:

Let t be a term (word) in a document d.

i. Term Frequency (TF):

$$\mathrm{TF}\ (\mathrm{t}, \mathrm{d}) = \frac{Number\ of\ times\ t\ appears\ in\ document\ d}{Total\ number\ of\ terms\ in\ document} \quad (1)$$

ii. Inverse Document Frequency (IDF):

$$IDF\ (t, D) = \log\left(\frac{Total\ number\ of\ documents\ in\ corpus\ N}{Number\ of\ documents\ containing\ term\ t+1}\right) \quad (2)$$

iii. TF-IDF:

$$TF - IDF\ (t, d, D) = TF\ (t, d)\ \text{x IDF}\ (\mathrm{t}, \mathrm{D}) \quad (3)$$

where: TF (t, d) represents the frequency of term t in document d, while IDF (t, D) represents the inverse document frequency of tern t across documents D.

The Bernoulli Naive Bayes (BNB) algorithm plays a crucial role in SMS spam detection by probabilistically classifying messages based on the presence or absence of specific terms. When combined with TF-IDF, which quantifies the importance of words within a message, BNB effectively distinguishes spam from legitimate messages. The algorithm applies conditional probability to assess how indicative certain words are of spam, assigning higher probabilities to messages containing spam-associated terms.

Mathematically, the posterior probability of a message belonging to a class C_k given a feature vector $x = (x_1, x_2, \dots, x_n)$ is expressed as:

$$P(C_k \mid x) \propto P(C_k) \prod_{i=1}^{n} P(x_i \mid C_k) \quad (4)$$

For Bernoulli Naive Bayes, the probability of a feature being present or absent is modeled as:

$$P(x_i = 1 \mid C_k) = \theta_{ik} \quad (5)$$

$$P(x_i = 0 \mid C_k) = 1 - \theta_{ik} \quad (6)$$

where θ_{ik} represents the likelihood of feature x_i appearing in class C_k . Using the log transformation to avoid numerical underflow, the classification decision is made by selecting the class with the highest log-posterior probability:

$$\log P(C_k \mid x) \propto \log P(C_k) + \sum_{i=1}^{n} [x_i \log \theta_{ik} + (1 - x_i) \log (1 - \theta_{ik})] \quad (7)$$

By deploying TF-IDF's ability to extract key features and BNB's probabilistic classification, this methodology enhances the accuracy and efficiency of SMS spam detection while minimizing false positives and negatives.

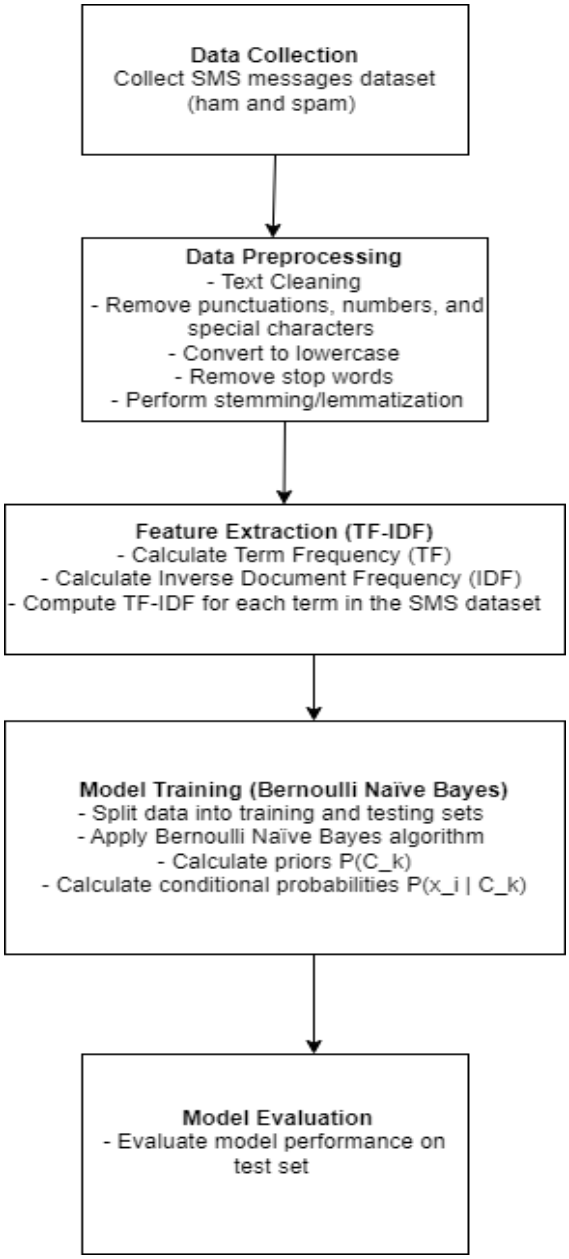

Figure 1: Illustrates the research design, showcasing the integration of TF-IDF and the BNB algorithm within the SMS spam detection framework

Data Acquisition

The SMS spam collection dataset used for this research is available online from the machine learning repository on the University of California Irvine (UCI) website (Almeida, 2011). The SMS spam collection is a public set of labeled SMS messages collected for mobile phone spam research. The collection contains 5,574 total messages of both spam and ham. It comprises four datasets from different sources. First is a collection of 425 SMS spam messages that were manually extracted from the Grumbletext Web site, a United Kingdom forum used for making public claims by cell phone users about SMS spam messages (UCI, 2012). The second is randomly chosen ham SMS messages of the NUS SMS Corpus (NSC) collected for research at the Department of Computer Science at the National University of Singapore (UCI, 2012). It is a subset of 3,375 of about 10,000 legitimate messages A list of 450 SMS ham messages collected from Caroline Tag's PhD Thesis is another source of data, and finally, the UCI collection incorporates the SMS Spam Corpus v.0.1 Big. It has 1,002 SMS ham messages and 322 spam messages (UCI, 2012). This dataset contains one message per line.

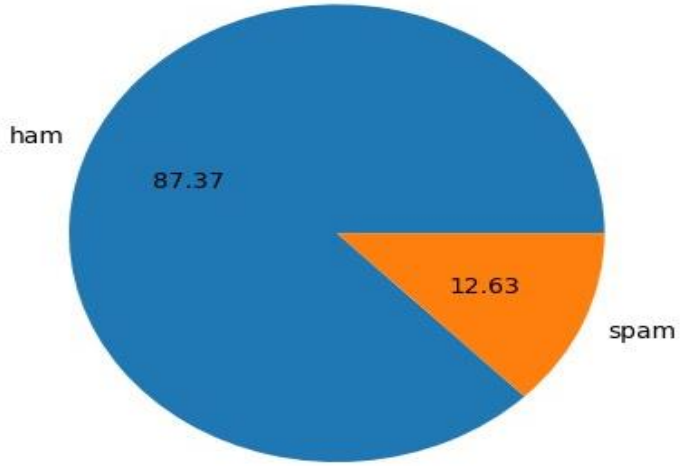

Figure 2: Pie chart of ham and spam in the dataset

Data Preprocessing

The first step in preprocessing is lowercasing, where all text is converted to lowercase to ensure uniformity and prevent duplicate representations of words with different capitalizations. Tokenization is then performed to break the text into individual words or tokens, making it easier to process. Punctuation removal follows, as punctuation marks hold little significance in SMS spam detection. Stop words, which are common words with minimal impact on classification, are also removed to improve efficiency. Finally, the text data is converted into numerical form using Term Frequency-Inverse Document Frequency (TF-IDF) vectorization, which helps in transforming text into a structured format suitable for machine learning models.

RESULTS AND DISCUSSION

In this section, presents the outcome of the work, using UCI datasets. It will also show comparison between the benchmark work and the proposed work results.

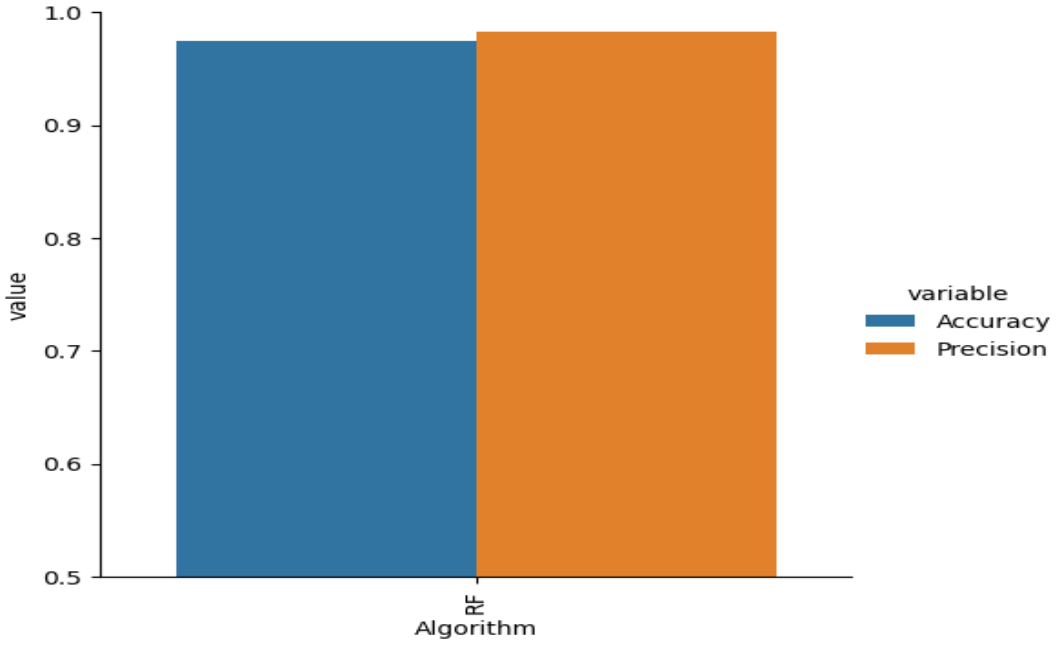

Figure 3: Benchmark Work (Sjarif, 2020)

Figure 3 above is an illustration of the benchmark paper which deployed the use of random forest and TFIDF which was able to attain an accuracy of 97% and precision of 98%.

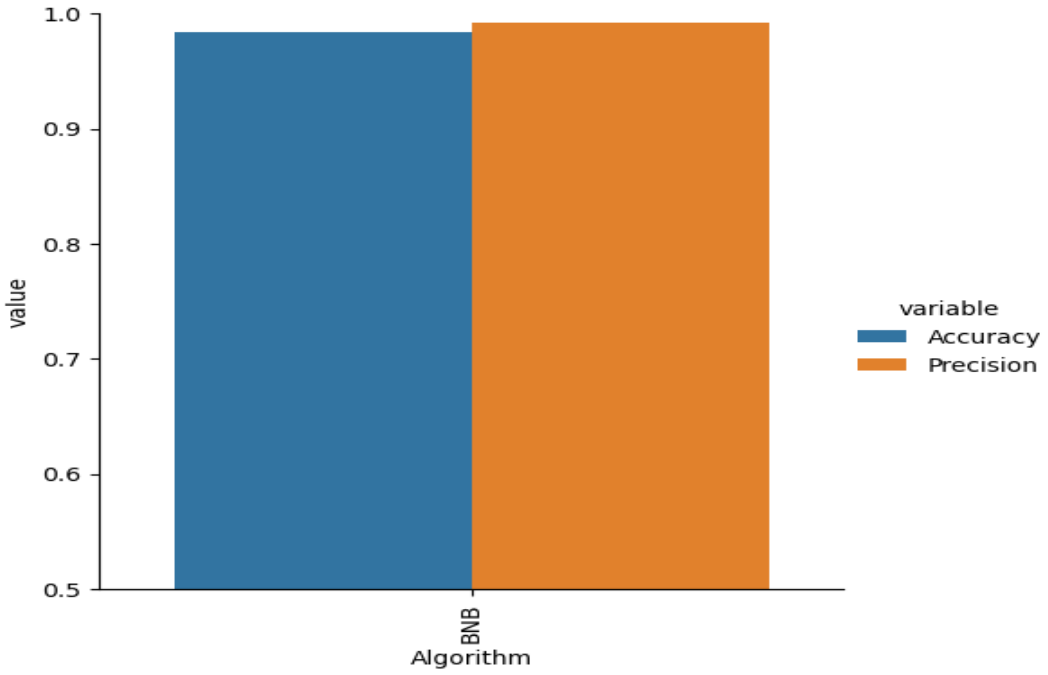

Figure 4: Proposed work

Figure 4 above is the proposed work which as seen in the figure above have an accuracy of 98% and likewise precision of 99%. The proposed work has a better result of 1 % on each of the evaluation metrics used.

Recent research works have demonstrated improvements in enhancing SMS spam detection accuracy while minimizing false positives. By analyzing the unique characteristics of spam messages and using statistical learning techniques, researchers have achieved notable success in developing robust and adaptive spam detection systems. This research not only improves the reliability of SMS communication but also contributes to the ongoing efforts to mitigate the impact of spam-related activities on mobile users and network infrastructure.

Table 1: Comparison with other Algorithms

Algorithm	Accuracy	Precision	Specificity	Sensitivity	MCC	Processing Time (seconds)
KN	0.905222	1.000000	1.000000	0.289855	0.511153	1.227311
MNB	0.970986	1.000000	1.000000	0.782609	0.870204	0.070284
BNB	0.983559	0.991870	0.998884	0.884058	0.927500	0.223687
RF	0.974855	0.982759	0.997768	0.826087	0.887757	13.963526
SVC	0.975822	0.974790	0.996652	0.840580	0.892185	9.810546
ETC	0.974855	0.974576	0.996652	0.833333	0.887718	42.143866
LR	0.958414	0.970297	0.996652	0.710145	0.809630	0.232018
XGB	0.970986	0.935484	0.991071	0.840580	0.870569	5.541015
AdaBoost	0.960348	0.929204	0.991071	0.760870	0.819612	26.491429
GBDT	0.947776	0.920000	0.991071	0.666667	0.756786	48.504141
BgC	0.957447	0.867188	0.981027	0.804348	0.810964	147.434996
DT	0.930368	0.817308	0.978795	0.615942	0.672450	1.620295

The provided table offers a clear overview of the performance metrics of various machine learning algorithms employed for this study SMS spam detection. Each algorithm is represented numerically, accompanied by its corresponding accuracy, precision, specificity, sensitivity, Matthews Correlation Coefficient (MCC), and processing time in seconds. Among the algorithms assessed, the Bernoulli Naïve Bayes (BNB) approach stands out prominently, achieving an accuracy of 98.36%. This indicates its ability to effectively classify SMS messages as either spam or legitimate with a high degree of accuracy. Moreover, the BNB algorithm exhibits impressive precision and specificity, measuring at 99.19% and 99.89%, respectively. These metrics underscore its capability to minimize false positives and accurately identify legitimate messages, contributing to the reliability of the spam detection system.

In terms of sensitivity, the BNB algorithm achieves a commendable score of 88.41%, indicating its effectiveness in detecting the majority of spam messages within the dataset. Furthermore, the high Matthews Correlation Coefficient (MCC) of 92.75% reinforces the robustness of the BNB approach by providing a balanced assessment of its classification performance, accounting for both true and false positives and negatives.

In contrast, other algorithms such as the K-Nearest Neighbors (KN) and Decision Trees (DT) exhibit comparatively lower performance across multiple metrics. For instance, the KN algorithm demonstrates a sensitivity of only 28.99%, indicating its limitations in accurately identifying spam

messages. Similarly, the DT algorithm achieves an accuracy of 93.04% and a precision of 81.73%, suggesting a higher rate of misclassification and false positives compared to the BNB approach.

Additionally, the processing time of each algorithm provides insight into its computational efficiency, crucial for real-time spam detection applications. While some algorithms such as Multinomial Naïve Bayes (MNB) and Logistic Regression (LR) exhibit minimal processing times, others such as Bagging Classifier (BgC) require significantly more time for computations.

The results obtained from the research using Term Frequency-Inverse Document Frequency (TF-IDF) and the Bernoulli Naïve Bayes (BNB) algorithm for SMS spam detection are highly promising and significant. The primary aim of this study was to evaluate the performance of this combined approach and compare it against established benchmarks to ascertain its effectiveness in addressing the challenges posed by spam messages in SMS communication. The evaluation metrics utilized for performance assessment include accuracy, precision, specificity, sensitivity, Matthews Correlation Coefficient (MCC), and processing time. These metrics provide comprehensive insights into the algorithm's capability to accurately classify messages as spam or legitimate and its computational efficiency.

Finally, the processing time of 0.224 seconds demonstrates that the algorithm is computationally efficient, capable of real-time spam detection without causing significant delays in message processing.

CONCLUSION

The study introduced a novel approach to SMS spam detection by combining TF-IDF for feature extraction with the Bernoulli Naïve Bayes (BNB) algorithm for classification. The results demonstrated the effectiveness of this approach, achieving an accuracy of 98.36%, precision of 99.19%, and an MCC of 0.93, surpassing other machine learning models such as Support Vector Machines (SVM), Random Forest (RF), and Decision Trees (DT). The BNB model also exhibited a low processing time of 0.22 seconds, making it efficient for real-time spam detection. Compared to benchmark models, which achieved slightly lower accuracy and precision, the proposed approach demonstrated superior performance in distinguishing between spam and legitimate messages. However, the study was conducted in a controlled simulation environment, which presents a limitation in real-world applicability. The need for updated datasets is also critical, as spammers continuously evolve their techniques. Future research should focus on implementing the model in practical settings, incorporating real-time spam messages, and expanding its use to other digital communication platforms such as social media.

REFERENCES

Abayomi-Alli, O., Misra, S., Abayomi-Alli, A., & Odusami, M. (2019). A review of soft techniques for SMS spam classification: Methods, approaches and applications. *Engineering Applications of Artificial Intelligence*, *86*, 197-212.

Abid, M. A., Ullah, S., Siddique, M. A., Mushtaq, M. F., Aljedaani, W., & Rustam, F. (2022). Spam SMS filtering based on text features and supervised machine learning techniques. *Multimedia Tools and Applications*, *81*(28), 39853-39871.

Ajueyitsi, O., & Ekuobase, G. O. (2024). A MULTIFACETED SENTIMENT ANALYSIS APPROACH TO THE ESTIMATION OF THE STRENGTH OF ONLINE SUPPORT FOR POLITICAL CANDIDATES IN NIGERIA'S ELECTIONS: Online Support Strength of Political Candidates in Nigeria's Elections. *FUDMA JOURNAL OF SCIENCES*, *8*(6), 184 - 192. https://doi.org/10.33003/fjs-2024-0806-2896

Al Saidat, M. R., Yerima, S. Y., & Shaalan, K. (2024). Advancements of SMS Spam Detection: A Comprehensive Survey of NLP and ML Techniques. *Procedia Computer Science*, *244*, 248-259.

Almeida, T. (2012). *SMS Spam Collection*. UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/SMS%2BSpam%2BCollection

Broadbent, S. (2020). Approaches to personal communication. In *Digital anthropology* (pp. 127-145). Routledge.

Chakraborty, A., Chattaraj, S., Karmakar, S., & Mishrra, S. (2021). A robust approach for effective spam detection using supervised learning techniques. *Machine Learning Techniques and Analytics for Cloud Security*, 171-191.

Dada, E. G., Bassi, J. S., Chiroma, H., Adetunmbi, A. O., & Ajibuwa, O. E. (2019). Machine learning for email spam filtering: Review, approaches and open research problems. *Heliyon*, *5*(6), e01802.

Gadde, S., Lakshmanarao, A., & Satyanarayana, S. (2021). SMS spam detection using machine learning and deep learning techniques. In *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)* (pp. 1-6). IEEE.

Gupta, V., Mehta, A., Goel, A., Dixit, U., & Pandey, A. C. (2019). Spam detection using ensemble learning. In *Harmony Search and Nature Inspired Optimization Algorithms: Theory and Applications, ICHSA 2018* (pp. 173-184). Springer.

Imran, M., Castillo, C., Diaz, F., & Vieweg, S. (2015). Processing social media messages in mass emergency: A survey. *ACM Computing Surveys (Csur)*, *47*(4), 1-38.

Jain, G., Sharma, M., & Agarwal, B. (2019). Spam detection in social media using convolutional and long short term memory neural network. *Annals of Mathematics and Artificial Intelligence*, *85*(1), 21-44.

Karim, A., Azam, S., Shanmugam, B., Kannoorpatti, K., & Alazab, M. (2019). A comprehensive survey for intelligent spam email detection. *IEEE Access*, *7*, 168261-168295.

Li, C.-Y. (2019). How social commerce constructs influence customers' social shopping intention? An empirical study of a social commerce website. *Technological Forecasting and Social Change*, *144*, 282-294.

Liu, X., Lu, H., & Nayak, A. (2021). A spam transformer model for SMS spam detection. *IEEE Access*, *9*, 80253-80263. https://doi.org/10.1109/access.2021.3081479

Lubis, A. R., Lubis, M., & Azhar, C. D. (2019). The effect of social media to the sustainability of short message service (SMS) and phone call. *Procedia Computer Science*, *161*, 687-695.

Oswald, C., Simon, S. E., & Bhattacharya, A. (2022). Spotspam: Intention analysis–driven sms spam detection using bert embeddings. *ACM Transactions on the Web (TWEB)*, *16*(3), 1-27.

Poster, W. R. (2022). Introduction to special issue on scams, fakes, and frauds. *new media & society*, *24*(7), 1535-1547.

Qabasiyu, M. G., Zayyad, M. A., & Abdullahi, S. (2023). An Ensembled Based Machine Learning Technique of Sentiment Analysis. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, *15*(1), 23-28.

Rodrigues, A. P., Fernandes, R., Shetty, A., K, A., Lakshmanna, K., & Shafi, R. M. (2022). [Retracted] Real-Time Twitter Spam Detection and Sentiment Analysis using Machine Learning and Deep Learning Techniques. *Computational Intelligence and Neuroscience*, *2022*(1), 5211949.

Roy, P. K., Singh, J. P., & Banerjee, S. (2020). Deep learning to filter SMS spam. *Future Generation Computer Systems*, *102*, 524-533.

Shafi'I, M. A., Abd Latiff, M. S., Chiroma, H., Osho, O., Abdul-Salaam, G., Abubakar, A. I., & Herawan, T. (2017). A review on mobile SMS spam filtering techniques. *IEEE Access*, *5*, 15650-15666.

Sjarif, N. N. A., Azmi, N. F. M., Chuprat, S., Sarkan, H. M., Yahya, Y., & Sam, S. M. (2019). SMS spam message detection using term frequency-inverse document frequency and random forest algorithm. *Procedia Computer Science*, *161*, 509-515.

Silva, R. M., Alberto, T. C., Almeida, T. A., & Yamakami, A. (2017). Towards filtering undesired short text messages using an online learning approach with semantic indexing. *Expert Systems with Applications*, *83*, 314-325.

Wang, H., Dai, B., & Yang, D. (2019). A comparative study of two different spam detection methods. In *Dependability in Sensor, Cloud, and Big Data Systems and Applications: 5th International Conference, DependSys 2019, Guangzhou, China, November 12–15, 2019, Proceedings* (pp. 450-459). Springer.

Wei, F., & Nguyen, T. (2020). A lightweight deep neural model for SMS spam detection. In *2020 International Symposium on Networks, Computers and Communications (ISNCC)* (pp. 1-6). IEEE