ISBN: 978-981-18-7950-0
2023 the 13th International Workshop on Computer Science and Engineering (WCSE 2023)
doi: 10.18178/wcse.2023.06.021

Comparative Analysis of Various SMS Spam Detection Methods using Machine Learning

Kartik Ahluwalia [1], Gururaj H L[1+], Rashmi R[1] and Hong Lin[2]

[1]Manipal Institute of Technology Bengaluru, Manipal Academy of Higher Education, Manipal, India
[2]University of Houston, Downtown, USA

Abstract. The term SMS (Short Message Service) refers to a popular text messaging service that is commonly used in telephone, internet, and mobile device systems. This service relies on standardized communication protocols that enable short text messages to be exchanged between mobile devices. The increase in SMS spam messages can be attributed to the higher limit of free SMS allowed by Internet Service Providers (ISPs) SMS spam detection relies heavily on the presence of known words, phrases, abbreviations, and idioms commonly used in spam messages. Studies have developed various datasets to train and test SMS spam detection models and have used different classification techniques to improve the accuracy and efficiency of these models. In the present study, various classification techniques for SMS spam detection have been explored such as Naive Bayes, Support Vector Machines (SVM), Decision Trees, Random Forest, and Neural Networks. These techniques use different approaches to identify patterns and features in the messages that distinguish spam from legitimate messages. Among the various algorithms Na ïve Bayes Classifier achieved a highest accuracy of 98.44% and Matthew Correlation Coefficients value of 0.93 for the dataset.

Keywords: SMS, Spam Detection, Machine Learning, Legitimate Messages.

1. Introduction

SMS relies on standardized communication protocols that enable short text messages to be exchanged between mobile devices. One of the most significant benefits is the ability to send messages to thousands of people at once, with immediate delivery. SMS can be customized to target an individual or a group of people, making it a versatile and effective communication tool. The high open rate of SMS messages, which is estimated to be around 99%, is another benefit that makes it a reliable marketing strategy. SMS can be easily automated using an SMS API and integrated into other applications to enable more complex processes. Many retailers consider SMS marketing to be a highly effective strategy, with around 65% of retailers reporting positive results from SMS campaigns. Another advantage of SMS is that customers can receive and view messages instantly, as notifications appear on their mobile device, eliminating the need for them to check their inbox. Overall, SMS offers businesses an effective and efficient means of communication that is easily customizable and accessible to a large audience.

According to a survey conducted by online community platform Local Circles 96% of Indians receive unwanted text messages every day, with almost half of the respondents receiving between four to seven such messages daily [6]. Most of the SMS spam originates from real estate companies, followed by banking and insurance offers.

SMS has become a popular target for spammers and hackers, making it more vulnerable to compromises. Clicking on a malicious link or message can automatically compromise a user's mobile device. Hence, limiting the content a user receives is crucial. One solution to this problem is the implementation of a system that can inform the user whether a message is spam or not. The term "HAM" is used for non-spam messages. To address these challenges, systems have been trained using machine learning algorithms to identify whether a message is spam or not based on its content [8]. By comparing such systems, we can determine the most effective approach to address the issue of spam messages in SMS.

[+] Corresponding author. Tel.: + 9686418942; fax: +9686418942
E-mail address: gururaj.hl@manipal.edu.

SMS fraud is one common form of telecom fraud, which can have devastating consequences. Fraudsters can send fake text messages that appear to be legitimate and trick people into providing their personal and financial information. The consequences of SMS fraud can be severe, including financial loss, damage to business reputation, and exposure of sensitive information to hackers. There are several types of SMS fraud that fraudsters use to deceive mobile users. These include:

- **Spam messages**: These types of fraudulent messages may ask the recipient to respond with personal information. In some cases, spam messages may contain a web link that redirects the recipient to a fraudulent website where they are asked to enter their personal information in exchange for a service.
- Grey routes can be attractive to messaging service providers and resellers because they can offer lower costs and greater flexibility compared to traditional messaging routes.
- **SIM farms** are a type of SMS fraud that can be used by businesses to send bulk SMS messages to customers, but they do so use unsecured delivery methods that can compromise personal information and leave it vulnerable to exploitation by fraudsters. One of the key risks associated with SIM farms is that they can be used to send messages that appear to come from a trusted source, such as a bank or other financial institution, and may include links or other requests for personal information.
- SMS phishing, also known as "**smishing**," is a form of social engineering that involves sending text messages to trick recipients into revealing sensitive information or performing an action that benefits the attacker.
- SIM swap fraud, also known as "**SIM splitting**," is a type of fraud where a criminal gains access to a victim's mobile phone number by impersonating the victim and convincing the mobile phone provider to transfer the number to a new SIM card.
- In a roaming fraud attack, criminals may exploit vulnerabilities in the billing systems of mobile network operators to make unauthorized charges on a victim's account or to gain access to the victim's personal information.
- In an SMS originator spoofing attack, the attacker may use software or online services to alter the information that appears in the messages "From" field, which is typically used to display the phone number or name of the sender. By changing the "From" field, the attacker can make it appear as though the message was sent from a different phone number or identity, such as a trusted friend or family member.

The paper is structured as follows: Section-2 describes the literature work, Section-3 explains the methodology, Section-4 depicts the result analysis and conclusion is drawn in Section-5.

2. Related Work

The progress in technology and communication tools has facilitated people's ability to communicate with one another, regardless of where they are located geographically. However, the misuse of these technologies by sending spam messages has become a significant concern [19-22].

SMS spam messages are often difficult to identify and filter out due to their limited format and the fact that they are typically sent from unknown or non-standard phone numbers [10]. Here are some of the reasons why identifying spam messages in the context of SMS is particularly challenging:

- **Limited Character Count:** SMS messages have a limited character count, typically 160 characters per message. This means that spammers have to be concise in their messages and may use shorthand or abbreviations that can be difficult to interpret.
- **No Standard Formatting:** There is no standard format for SMS messages, which makes it difficult to identify spam based on formatting alone.
- **Non-Standard Phone Numbers:** Spammers often use non-standard phone numbers to send SMS messages, making it difficult to filter out messages from known spammers or block certain numbers.
- **Lack of Context:** SMS messages are often standalone messages without any context or background information.
- **User Behavior:** Users may not report SMS spam as frequently, which makes it difficult for carriers and spam detection systems to identify and filter out spam messages.

Overall, the limited format of SMS messages and the lack of standard formatting make it challenging to identify and filter out spam messages. To combat this, carriers and spam detection systems use a combination of Machine Learning (ML) algorithms, user feedback, and other techniques to identify and block spam messages.

A combination of content-based and rule-based filtering techniques can be used to effectively detect and filter spam in SMS messages [17]. By continually refining the filtering rules and language model, the system can improve its accuracy and provide a better user experience [16].

Delany et. al.,[1] performed spam filtering by using content-based technologies. They analyzed the spam messages and identified nine to ten clearly defined clusters. Ahmed et.al., [3] performed text mining on large datasets using various methods like decision tree, neural network, SVM and compared each of the classifiers based on computational efficiency and accuracy. Emphasized and proposed detection of messages based on supervised, unsupervised, reinforcement learning. Various Deep Neural Network (DNN) techniques in identifying the SPAM and HAM discussed by Nivaashini et.al., [2]. Gupta et.al., [5] compared different ML techniques based on their accuracies, precision, recall and CAP Curve.

3. Methodology

In the present work various machine learning classifiers are analyzed for SMS spam detection and details of the various machine learning classifiers are explained subsections.

3.1. Classification

Classification is a fundamental problem in ML and statistics, and it involves assigning new observations to predefined categories or subpopulations based on some set of features or attributes. To build a classification model, one typically uses a training set of data that includes labeled examples of the different categories or subpopulations. The model is trained to analyze the patterns and relate the characteristics with labels. Once the model is trained, it can be used to predict the class labels for unseen records based on their characteristics. There are many different techniques that can be used for classification, such as Logistic Regression, Decision Trees, SVM and Neural Networks.

Two types of Classification are:

- **Binary Classification:** The technique used to divide the provided data into two different classifications is called binary classification, such as spam or not spam. In this case, we would use a binary classification algorithm to learn from the data and predict whether the message is spam or not based on the individual's health status. Examples of binary classification algorithms include logistic regression, decision trees, and SVM.
- **Multiclass Classification: T**he problem of identifying the species when there are more than two classes is known as a multi-class classification problem.

The present study aims at classifying the objects into one of two categories they are Ham/spam.

Creating a categorization model involves several steps, some of which are:

- **Dataset:** In the present work 5572 dataset of various SMS are used. The training dataset has been taken from the Kaggle repository (SMS spam collection dataset [4]).
- **Preprocessing:** It involves cleaning, normalizing, and transforming the data into a format that is ready to be processed by the model. This includes removing stop words, stemming, and converting text into numerical vectors [18].
- **Classifiers:** Classification is the process of grouping data into predefined categories or classes based on specific features or characteristics. Regardless of the type of data, classification involves dividing it into subcategories or classes based on specific criteria. The classification algorithm used in the present study is given below.

1. **Naive Bayes** is a popular classification algorithm based on Bayes' theorem, which allows us to calculate the probability of a category. Naive Bayes assumes that the features (or words) in the input text are independent of each other, which simplifies the calculations and makes the algorithm fast and scalable.
2. **Decision Tree:** works by dividing the data into smaller subsets based on different data attributes. The decision tree is a supervised learning method, which means that it requires a labeled dataset as input [9].
3. **Support Vector Machines:** are a type of supervised machine learning algorithm which work by

finding a hyperplane or a line that separates the different classes of data with a maximum distance between the hyperplane and the nearest data points from each class.

4. **K-Nearest Neighbor (KNN):** is a supervised ML algorithm used for classification and regression tasks. In KNN, the prediction of a new data point is based on the class or value of its nearest neighbors in the training data.
5. **Logistic Regression:** is a logistic function that takes the input features of an instance and returns a value between 0 and 1, which represents the probability of the instance belonging to class 1.
6. **Random Forest (RF):** In this model, multiple decision trees are trained on different subsets of the training data and with different subsets of the features.
7. **Ada-boost or Adaptive Boosting [11]:** is an ensemble learning method that combines multiple feeble classifiers to create a powerful classifier. It works by repetitive training the classifiers on weighted versions of the training dataset.
8. **Bagging (short for Bootstrap Aggregating) [12][13]:** is a method that uses randomization to reduce the variance of a ML algorithm. The randomization introduced by bagging reduces overfitting and improves accuracy and robustness.
9. **Extremely Randomized Trees Classifier [15]: (**Extra Trees Classifier) In the Extra Trees Classifier, the decision trees are built using a random subset of the features, rather than considering all the features at each split point.

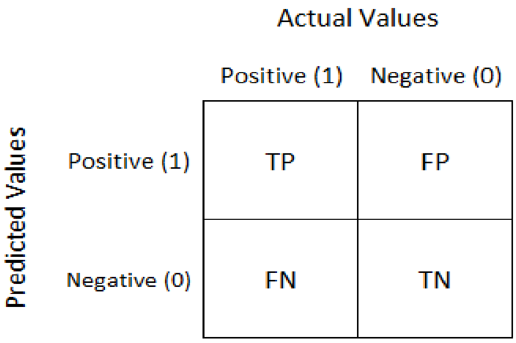
3.2. Performance Metrics for Classification Problems

Accuracy: This metric is often used to evaluate the performance of a classification model, where the model is trained with predicting the correct class label for each observation in a dataset.

Accuracy = correct predictions/ total predictions

The accuracy score ranges from 0 to 1, with a value of 1 indicating a perfect model that makes no errors in its predictions.

Confusion Matrix:



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Fig 1: Confusion Matrix

The Fig 1. summarizes the number of correct and incorrect predictions made by the model, by comparing the predicted class labels to the actual class labels in the test dataset.

The confusion matrix helps in visualizing the performance of a classification model and calculating various performance metrics like accuracy, precision, recall, F1-score, etc. from the predicted and actual values.

Recall, Precision, F-Score

- Recall = TP / (TP+FN)
- Precision = TP / (TP+FP)
- F1-score= 2 * (Recall * Precision) / (Recall + Precision)

Cohen Kappa Score:

Kappa = (Observed Accuracy – Expected Accuracy)/ (1-Expected Accuracy)

Table 1: Kapa Score and Interpretation

K	**Interpretation**
<0	Poor agreement
0.1 – 2.0	Slight agreement
0.21 – 0.40	Fair agreement
0.41 – 0.60	Moderate agreement
0.61 – 0.80	Substantial agreement
0.81 – 1.0	Almost perfect agreement

Matthews correlation coefficient (MCC): [14]

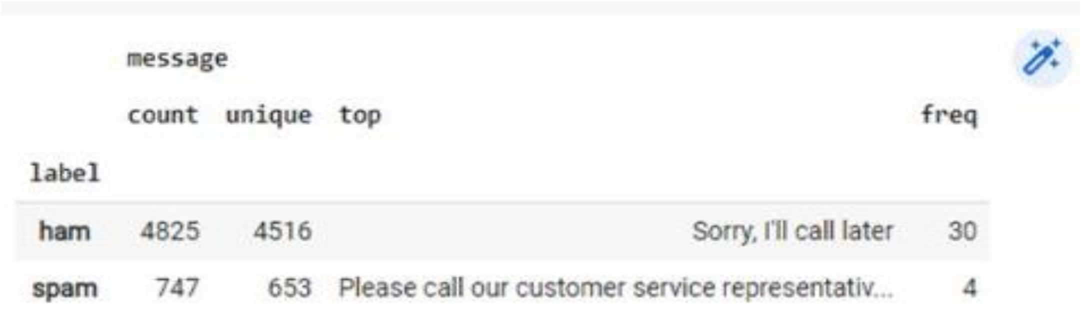
MCC = (TP*TN – FP*FN) /√(TP+FP)(TP+FN)(TN+FP)(TN+FN)

The value for MCC ranges from -1 to 1 where:

- -1 indicates total disagreement between predicted classes and actual classes.
- 0 is synonymous with completely random guessing.
- 1 indicates total agreement between predicted classes and actual classes.

4. Results and Discussion

This section discusses the outcomes of different machine learning classifiers and compares them using performance indicators. The amount of spam and junk mail in the dataset under consideration is depicted in Fig. 2. Fig 3 depicts the distribution of Spam and Ham messages based on message length.



```
       message
       count  unique  top                                              freq
label
ham    4825   4516                                  Sorry, I'll call later   30
spam    747    653    Please call our customer service representativ...    4
```

Fig 2: Number of Spam and Ham messages in dataset

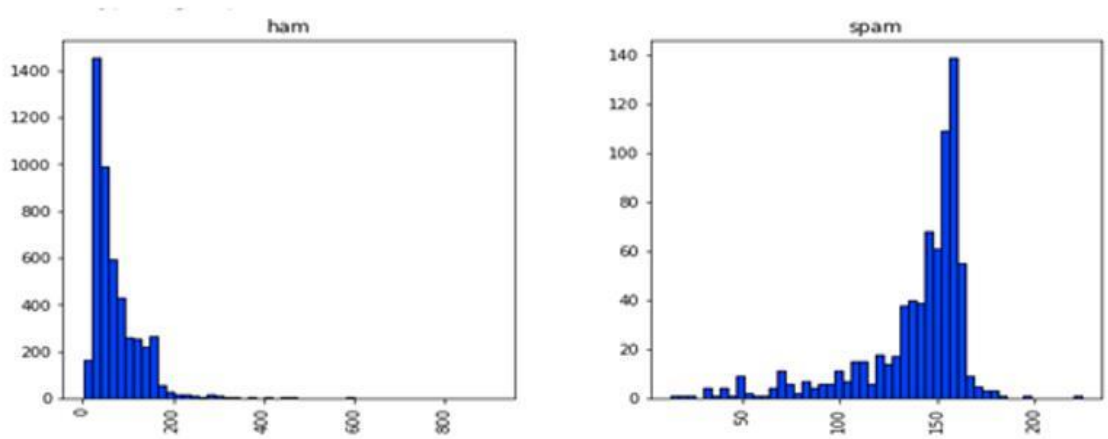
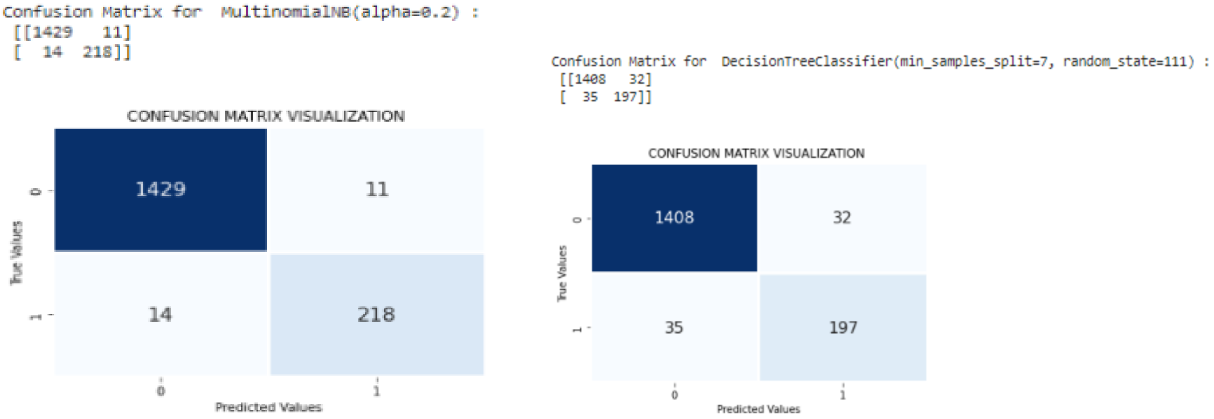


Fig 3: Number of Spam and Ham messages based on message length.

4.2 Confusion Matrix Visualization for different Classifiers

Fig. 4, 5, 6, 7 and 8 display confusion matrix visualization for several classifiers. It has been observed that decision tree classifiers create more false positives than other techniques. In a similar vein, the KNN algorithm generates more false positives than other techniques.

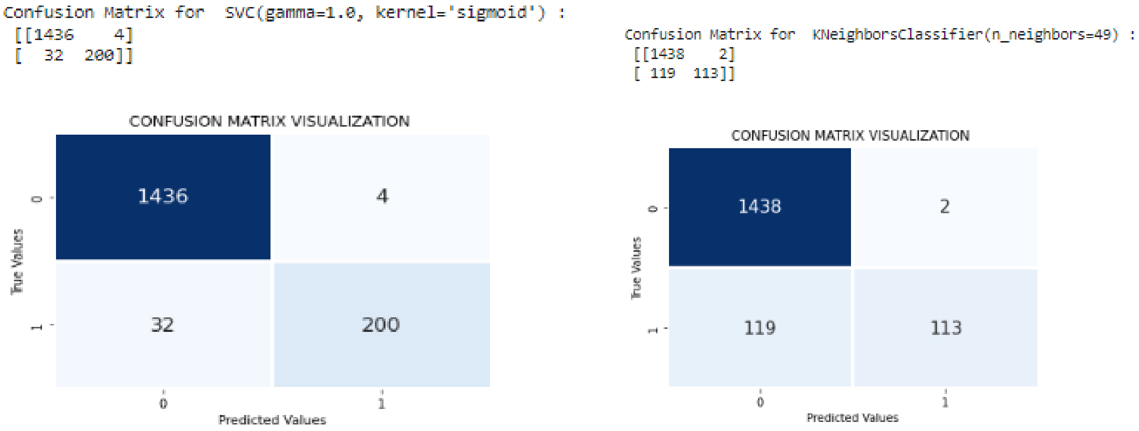


```
Confusion Matrix for  MultinomialNB(alpha=0.2) :
 [[1429   11]
 [  14  218]]
```

```
Confusion Matrix for  DecisionTreeClassifier(min_samples_split=7, random_state=111) :
 [[1408   32]
 [  35  197]]
```

(a) (b)

Fig 4: Confusion matrix for (a) Na ïve Bayes and (b) Decision Tree classifiers.

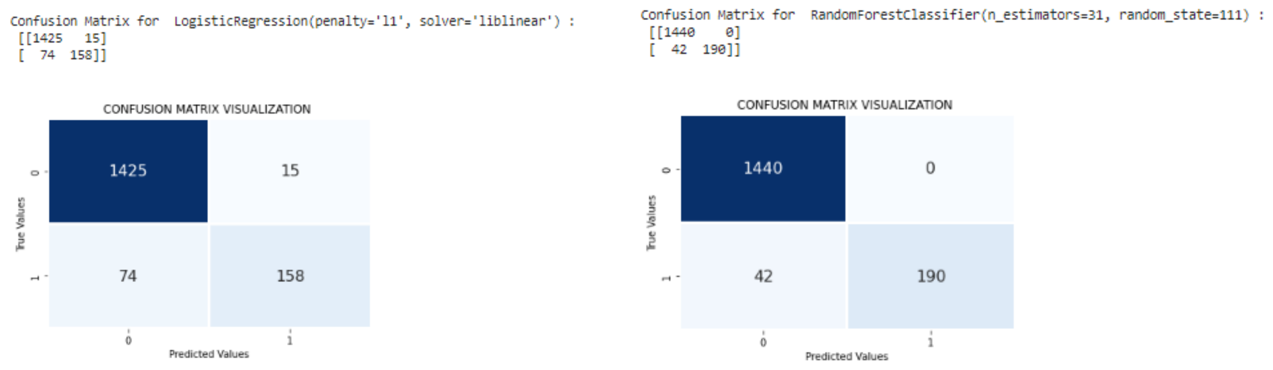


```
Confusion Matrix for  SVC(gamma=1.0, kernel='sigmoid') :
 [[1436    4]
 [  32  200]]
```

```
Confusion Matrix for  KNeighborsClassifier(n_neighbors=49) :
 [[1438    2]
 [ 119  113]]
```

(a) (b)

Fig 5: Confusion matrix for (a) SVM and (b) KNN classifier.

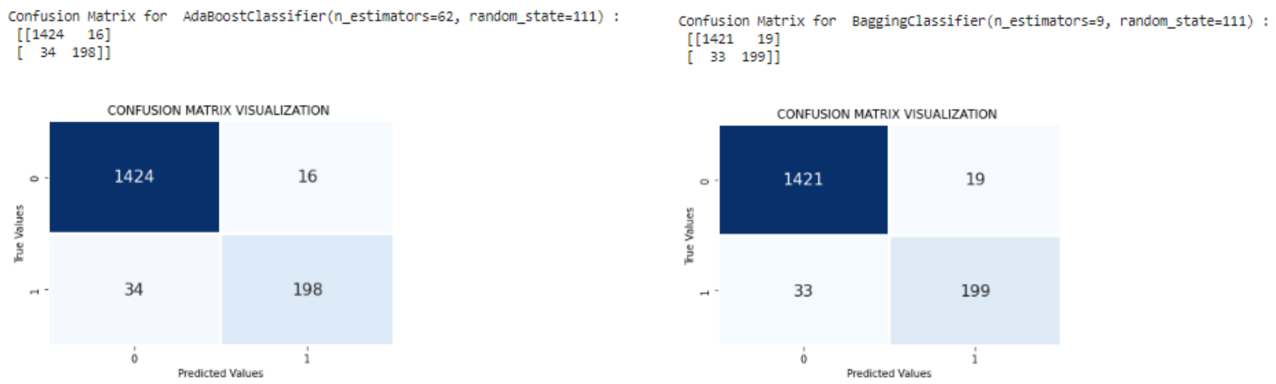


```
Confusion Matrix for  LogisticRegression(penalty='l1', solver='liblinear') :
 [[1425   15]
 [  74  158]]
```

```
Confusion Matrix for  RandomForestClassifier(n_estimators=31, random_state=111) :
 [[1440    0]
 [  42  190]]
```

(a) (b)

Fig 6: Confusion matrix for (a) Logistic Regression and (b) RF classifier.

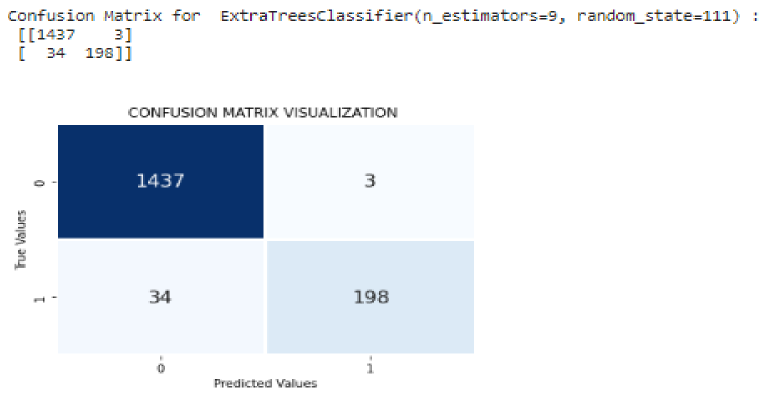


```
Confusion Matrix for  AdaBoostClassifier(n_estimators=62, random_state=111) :
 [[1424   16]
 [  34  198]]
```

```
Confusion Matrix for  BaggingClassifier(n_estimators=9, random_state=111) :
 [[1421   19]
 [  33  199]]
```

(a) (b)

Fig 7: Confusion matrix for (a) Ada-boost and (b) Bagging classifier.



```
Confusion Matrix for  ExtraTreesClassifier(n_estimators=9, random_state=111) :
 [[1437    3]
 [  34  198]]
```

Fig 8: Confusion matrix for Extra Tree Classifier.

Table 2: Comparison Table of different classifiers based on Accuracy, Precision, Recall, F1 Score, Kappa Coefficient and Matthew Correlation Coefficients.

S.No	**Classifier**	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**Kappa Coefficient**	**Matthew Correlation Coefficients**
1	**Naive Bayes (NB)**	**0.98**	**0.99**	**0.99**	**0.99**	**0.93**	**0.93**
2	**Decision Tree (DT)**	0.95	0.97	0.97	0.97	0.82	0.82
3	**Support Vector Machine (SVM)**	0.97	0.97	0.99	0.98	0.90	0.90
4	**K-Nearest Neighbor (KN)**	0.92	0.92	0.99	0.95	0.59	0.64
5	**Logistic Regression (LR)**	0.94	0.94	0.98	0.96	0.73	0.74

6	**Random Forest (RF)**	0.97	0.96	1.00	0.98	0.86	0.87
7	**AdaBoost**	0.96	0.96	0.99	0.98	0.85	0.85
8	**Bagging Classifier (BgC)**	0.96	0.97	0.98	0.98	0.85	0.85
9	**Extra Trees Classifier (ETC)**	0.97	0.97	0.99	0.98	0.89	0.90

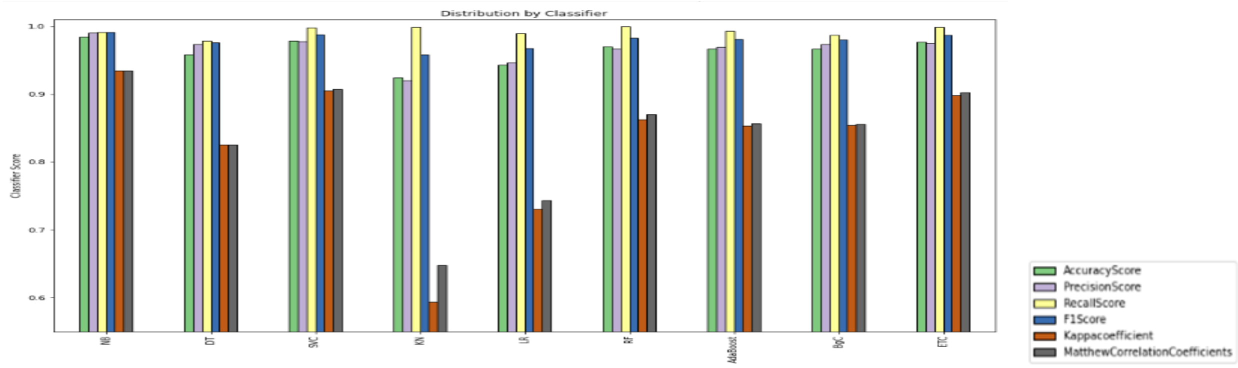


Fig 9: Comparison of different algorithms and performance metrics.

From the studies, it can be observed that the Naive Bayes algorithm performs better than all other algorithms, with a F1-score of 0.99 and a Kappa coefficient of 0.93. Additionally, it can be noted that the SVM and ETC algorithms performed favorably on the dataset. With an F1-score of 0.89, the RF, AdaBoost, and Bagging classifier algorithms perform similarly to one another. This finding implies that several ensemble approaches can detect spam with comparable success. Furthermore, it can be noted that the RF algorithm generated a recall value of 1, suggesting that it has no false negatives. Additionally, Table 2 lists the various performance metrics comparisons of these methods.

5. Conclusion

Spam detection is crucial for protecting message communication. The present study focussed on evaluating ML techniques for spam SMS detection where different classifiers namely, Naive Bayes, Decision Tree, SVM, K-Nearest Neighbour, Logistic Regression, Random Forest, AdaBoost, Bagging Classifier and Extra Trees Classifier are analysed with six different evaluation methods such as Accuracy, Precision, Recall, F1Score, Kappa Coefficient and Matthew Correlation Coefficients. From the dataset 80% of SMS are used for training the classifier and remaining is used for testing. The results obtained from our evaluation of the classifiers shows that K-Nearest Neighbour achieved the lowest accuracy of 92.4%, with precision of 0.92 and Matthew Correlation Coefficients of 0.64.

Na ïve Bayes Classifier achieves the highest accuracy of 98.44% and Matthew Correlation Coefficients value of 0.93. The findings imply that the Naive Bayes is more efficient for precise and prompt identification of spam and will secure messaging systems. We can continuously adapt the techniques to evade detection, using a diverse set of machine learning models can help identify previously unseen types of spam messages. In future we can explore Deep Learning techniques to accurately detect spam SMS.

6. Acknowledgements

I am writing to express my sincerest gratitude for the invaluable contributions to the completion of my research paper to Dr. Gururaj H L and Dr. Rashmi R. Your guidance, encouragement, and expertise were instrumental in shaping this work into its final form.

Dr. Gururaj H L, your extensive knowledge and insightful feedback were truly invaluable throughout the entire research process. Your guidance helped me to develop a deeper understanding of the subject matter,

and your constructive criticism and feedback were crucial in refining my arguments and ideas. I am deeply grateful for the time and effort you invested in my project.

Dr. Rashmi R, your meticulous attention to detail and unwavering support were integral to the success of this research paper. Your technical assistance and help with data analysis were invaluable, and your enthusiasm and dedication kept me motivated throughout the project.

I am honoured to have had the opportunity to work with such talented and dedicated professionals, and I am grateful for the knowledge and skills that I have gained under your mentorship. Your contributions have been instrumental in shaping my academic and professional development. Once again, I express my sincere thanks and appreciation for your unwavering support and guidance.

7. References

[1] Delany, Sarah Jane, Mark Buckley, and Derek Greene. "SMS spam filtering: Methods and data." Expert Systems with Applications 39.10 (2012): 9899-9908.

[2] Nivaashini, M., et al. "SMS spam detection using deep neural network." International Journal of Pure and Applied Mathematics 119.18 (2018): 2425-2436.

[3] Ahmed, Naeem, et al. "Machine learning techniques for spam detection in email and IoT platforms: analysis and research challenges." Security and Communication Networks 2022 (2022): 1-19.

[4] SMS spam collection dataset URL:https://www.kaggle.com/datasets/uciml/sms-spam-collection-dataset

[5] Gupta, Mehul, et al. "A comparative study of spam SMS detection using machine learning classifiers." 2018 Eleventh International Conference on Contemporary Computing (IC3). IEEE, 2018.

[6] https://scroll.in/article/916575/spam-alert-96-indians-with-mobile-phones-get-unwanted-text-messages-every-day

[7] Pereira, Francisco, Tom Mitchell, and Matthew Botvinick. "Machine learning classifiers and fMRI: a tutorial overview." Neuroimage 45.1 (2009): S199-S209.

[8] Gómez Hidalgo, José María, et al. "Content based SMS spam filtering." Proceedings of the 2006 ACM symposium on Document engineering. 2006.

[9] Navaney, Pavas, Gaurav Dubey, and Ajay Rana. "SMS spam filtering using supervised machine learning algorithms." 2018 8th international conference on cloud computing, data science & engineering (confluence). IEEE, 2018.

[10] Shafi’I, Muhammad Abdulhamid, et al. "A review on mobile SMS spam filtering techniques." IEEE Access 5 (2017): 15650-15666.

[11] Li, Xuchun, Lei Wang, and Eric Sung. "AdaBoost with SVM-based component classifiers." Engineering Applications of Artificial Intelligence 21.5 (2008): 785-795.

[12] Skurichina, Marina, and Robert PW Duin. "Bagging for linear classifiers." Pattern Recognition 31.7 (1998): 909-930.

[13] Zareapoor, Masoumeh, and Pourya Shamsolmoali. "Application of credit card fraud detection: Based on bagging ensemble classifier." Procedia computer science 48.2015 (2015): 679-685.

[14] Chicco, Davide, and Giuseppe Jurman. "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation." BMC genomics 21 (2020): 1-13.

[15] Sharaff, Aakanksha, and Harshil Gupta. "Extra-tree classifier with metaheuristics approach for email classification." Advances in Computer Communication and Computational Sciences: Proceedings of IC4S 2018. Springer Singapore, 2019.

[16] Kontsewaya, Yuliya, Evgeniy Antonov, and Alexey Artamonov. "Evaluating the effectiveness of machine learning methods for spam detection." Procedia Computer Science 190 (2021): 479-486.

[17] Navaney, Pavas, Gaurav Dubey, and Ajay Rana. "SMS spam filtering using supervised machine learning algorithms." 2018 8th international conference on cloud computing, data science & engineering (confluence). IEEE, 2018.

[18] Almeida, Tiago A., et al. "Text normalization and semantic indexing to enhance instant messaging and SMS spam

filtering." Knowledge-Based Systems 108 (2016): 25-32.

[19] Mendez, Jose R., Tomas R. Cotos-Yanez, and David Ruano-Ordas. "A new semantic-based feature selection method for spam filtering." Applied Soft Computing 76 (2019): 89-104.

[20] Dada, Emmanuel Gbenga, et al. "Machine learning for email spam filtering: review, approaches and open research problems." Heliyon 5.6 (2019): e01802.

[21] Méndez, José Ramon, et al. "A comparative performance study of feature selection methods for the anti-spam filtering domain." Advances in Data Mining. Applications in Medicine, Web Mining, Marketing, Image and Signal Mining: 6th Industrial Conference on Data Mining, ICDM 2006, Leipzig, Germany, July 14-15, 2006. Proceedings 6. Springer Berlin Heidelberg, 2006.

[22] Mathew, Kuruvilla, and Biju Issac. "Intelligent spam classification for mobile text message." Proceedings of 2011 International Conference on Computer Science and Network Technology. Vol. 1. IEEE, 2011.