iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** VI **Month of publication:** June 2023

DOI: https://doi.org/10.22214/ijraset.2023.54341

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 VI Jun 2023- Available at www.ijraset.com

Mobile Message Classification Using Natural Language Processing and Machine Learning Algorithms

G. Vikas[1], M.V. S Koushik[2], M. Nithya[3], CH. Sudha[4]
[1, 2, 3, 4]*Mahatma Gandhi Institute of Technology*

Abstract: ***SPAM: Stupid Pointless Annoying Malware is any unwanted, unsolicited digital communication sent out in bulk Although email is the most common method of spreading spam, it can also be communicated via social media, text messages, and phone calls. Sadly, whether we like it or not, spam messages must irritate everyone with a mobile device. Here this project classifies spam messages. Understanding different spam text classification techniques like extraction, text preprocessing, and NLTK stop words is vital. This project mainly focuses on the spam classification approach using machine learning algorithms such as Random Forest, KNN, Naïve Bayes, Support Vector Machine, decision tree, and NLP algorithms Count Vectorization and TF-IDF***

Keywords: Support Vector Machine (SVM), Random Forest (RF), Naïve Bayes (NB), Decision Tree (DT), TF-IDF, Count Vectorization (cv).

I. INTRODUCTION

The Short Messaging Service (SMS) is primarily used for informal communication, such as advertising new goods and services, but it is also occasionally used for formal communication, such as confirming an order placed on an online store or receiving information about a bank transaction. Technology developments have substantially lowered the cost of sending a text message. This has turned out to be a blessing for some people and a curse for countless others. People are abusing the SMS feature to advertise goods, services, deals, and other things. Twenty to thirty percent of all SMS received are spam, therefore it's easy to see how bothersome this has gotten by the fact that people have started disregarding the messages they get (Kim et al., 2015).

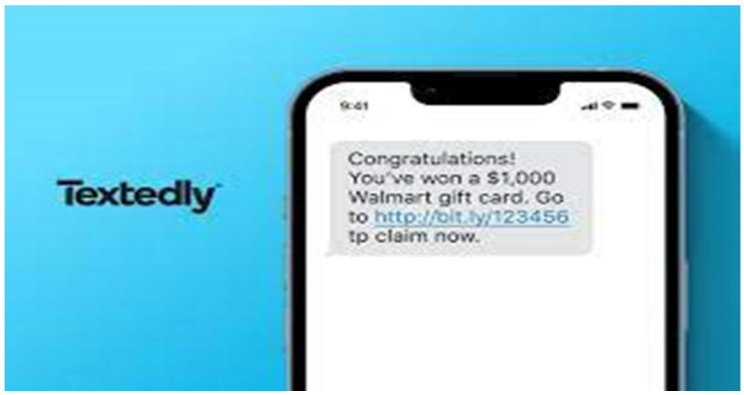


Figure 1 Sample Spam text

This study aims to implement machine learning methods to distinguish between spam and legitimate messages. Machine learning and natural language processing techniques were combined to make the procedure more fluid and effective. Consumers who receive spam messages run a variety of hazards, including unwelcome advertising, the revelation of personal information, being a victim of financial fraud or other schemes, falling for the traps of malware and phishing websites, unintended exposure to unpleasant content, etc. The network operator incurs higher operational costs as a result of spam messages.

II. LITERATURE SURVEY

1) Contributions To The Study of SMS Spam Filtering: New Collections And Results (2018)
Authors: Tiago A. Almeida & Jose Maria Gomez Hidalgo.

©IJRASET: All Rights are Reserved

Employing machine learning and feature selection to classify SMS spam. In COS (pp 5):2021 IEEE International Conference on Intelligent Techniques in Control, Optimisation, and Signal Processing.
This article offered a feature selection methodology integrated with machine learning methods like Naive Bayes, decision trees, and random forests to classify SMS spam. The feature selection approach aimed to find the most pertinent features to enhance classification performance.

2) Ham or Spam? A comparative Study for Content-Based Classification (2017)
Authors: Mariette Awadh, Salwa Adriana Saab
Comparative study on the use of machine learning for SMS spam detection. (pp. 1-6) in 2019 IEEE 7th International Conference on Control, Engineering & Information Technology (CEIT). This study compared machine learning methods, such as SVM, KNN, and gradient boosting, to detect SMS spam. The study assessed these algorithms' performance with various feature sets and determined how effective they were.

III. PROPOSED SYSTEM

A. Data Pre-processing
Pre-processing operations like tokenization, stop word removal, stemming/lemmatization and other essential functions are carried out to clean up the dataset by removing irrelevant or duplicate messages. Create training, validation, and testing sets from the dataset.

B. Feature Extraction:
1) Apply NLP techniques to extract relevant features from text messages.
2) Utilise strategies like TF-IDF (Term Frequency-Inverse Document Frequency), bag-of-words, and word cloud to properly represent the textual content.

C. Model Selection and Training
Try out several ML techniques like Naive Bayes, decision trees, random forests, support vector machines, and KNN Train. Then, use the training dataset to fine-tune the chosen models.

D. Evaluation and Model Validation:
1) Apply the validation dataset to the trained models' evaluation.
2) Assess and contrast the effectiveness of various models using parameters including accuracy, precision, recall, and F1-score.
3) Choose the model that performs the best for further analysis.

E. System Integration & Deployment:
1) Create a user-friendly interface so that users may enter text messages for classification.
2) To handle user requests in real time, include the learned model into the system. Deploy the system so people can access it via a web interface or an API.

IV. ALGORITHM

A. K Nearest Neighbours
The KNN algorithm considers the neighbors' class labels (such as spam or ham) and chooses the new message's class using a majority vote. The new communication is categorized as spam if more of its neighbors are marked as spam. If additional neighbors carry the ham classification, the new message is categorized as ham.

B. ID3
The algorithm creates a decision tree using ID3 and learns patterns and rules from the examples that have been labeled in order to categorize new communications as spam or ham depending on their characteristics. The algorithm's goal is to produce a tree that successfully distinguishes between the two classes and offers precise forecasts for unread messages.
It's critical to remember that ID3 has significant drawbacks, including its sensitivity to training data and propensity for overfitting.

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

C. Random Forest

The Random Forest algorithm runs a fresh message through each decision tree in the forest and gathers predictions from each tree in order to classify the message.

The decision tree predictions are combined to get the final categorization. It is possible to choose the class with the highest likelihood by utilizing probabilities rather than a majority vote (the class that is predicted by the majority of trees).

D. Naive Bayes

For a new message, Naive Bayes calculates the most probable class. The computations are made more accessible by the assumption of attribute independence, albeit this assumption may not always hold true in practice. It functions well even with a few training samples and can handle huge data sets. Nevertheless, it can have a hard time dealing with intricate interactions between attributes and could be pretty sensitive to the calibre and representativeness of the training data.

E. Support Vector Machine (SVM)

SVM is renowned for its prowess with small to medium-sized datasets, high-dimensional data handling, and good generalization to new cases. Large datasets may make it computationally demanding, and choosing the right kernel functions and hyperparameters must be done with care.

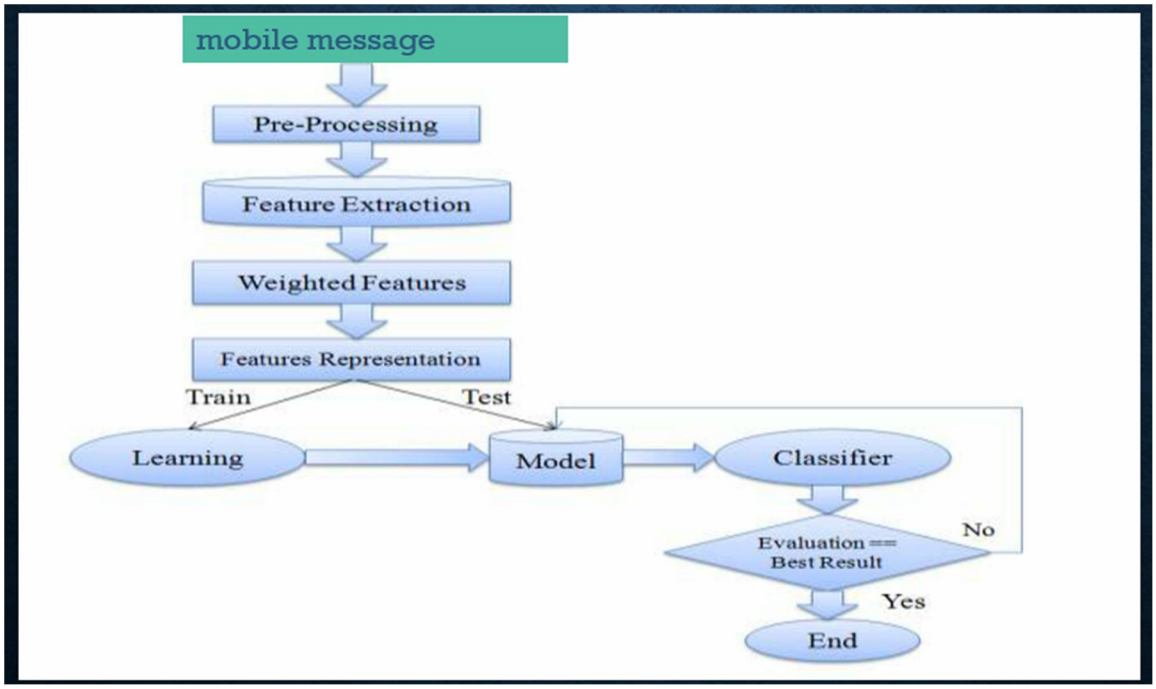


Figure 2 System Architecture

V. RESULTS

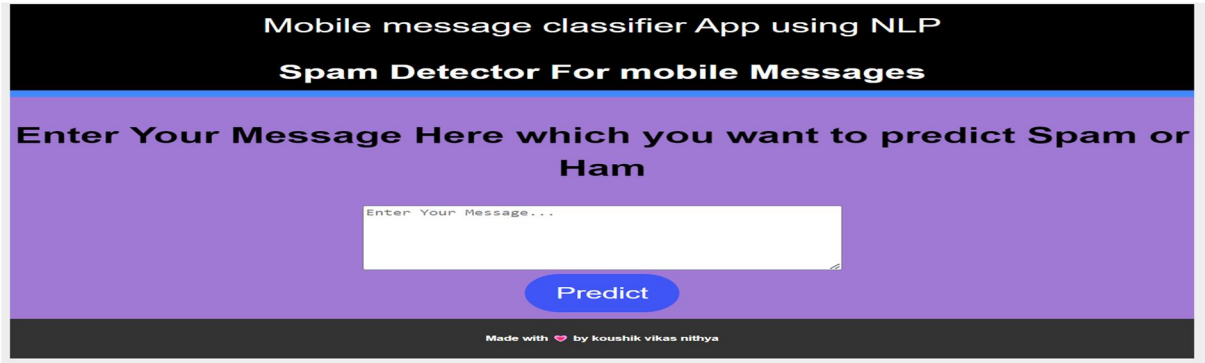


Figure 3 User interface of the classifier

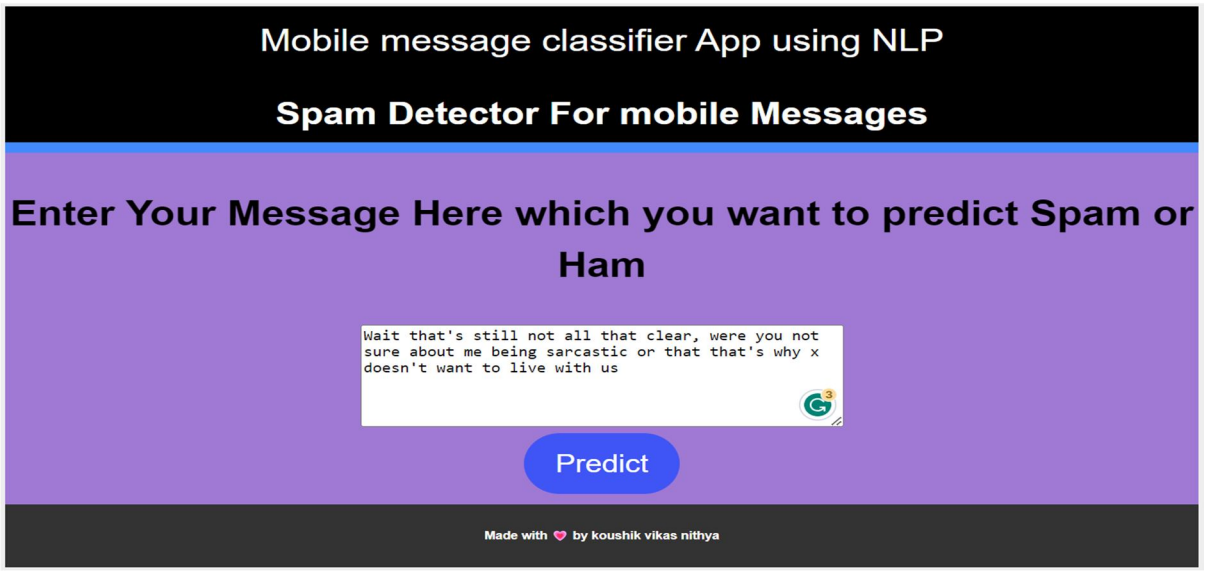


Figure 4 ham text-example

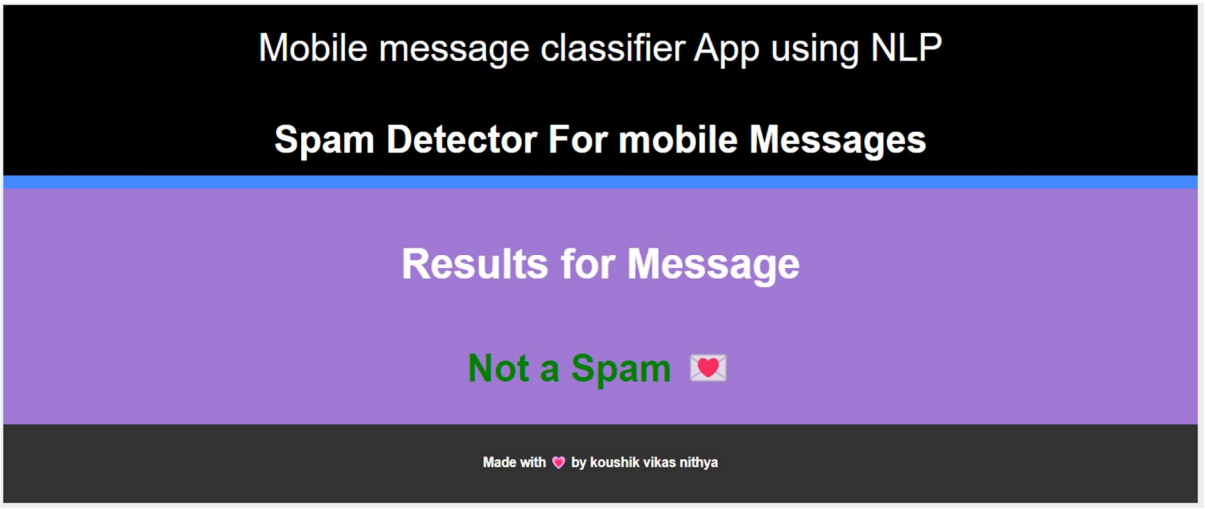


Figure 5 not a spam / ham

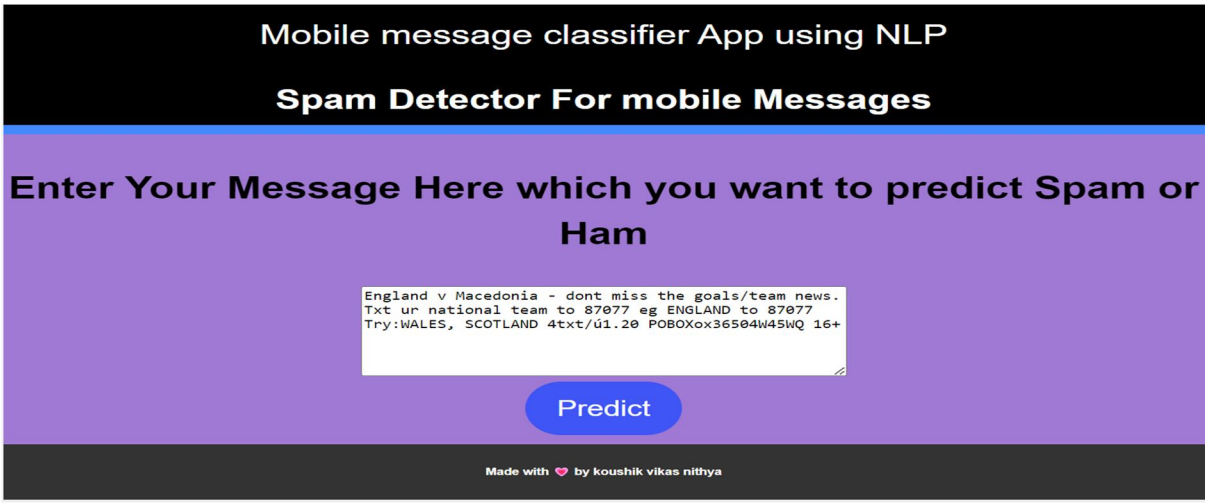


Figure 6 spam text example

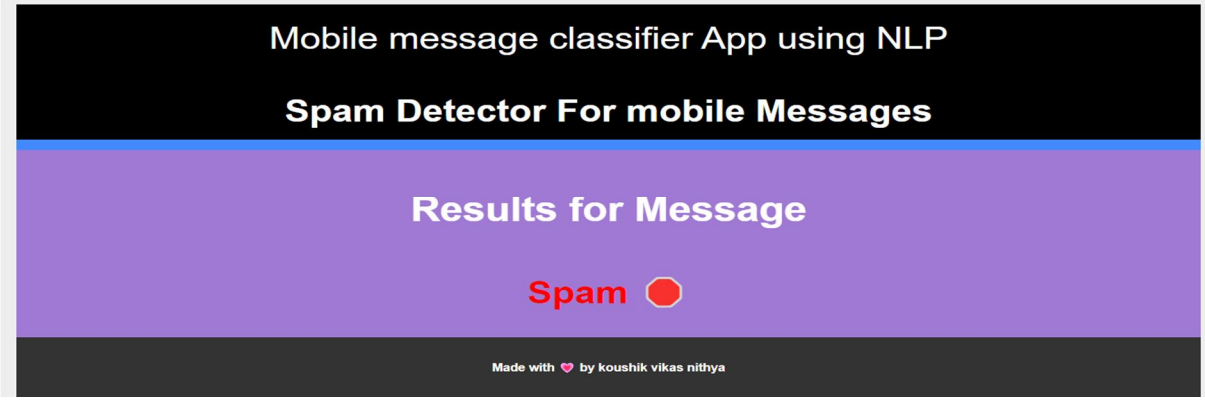


Figure 7 spam example

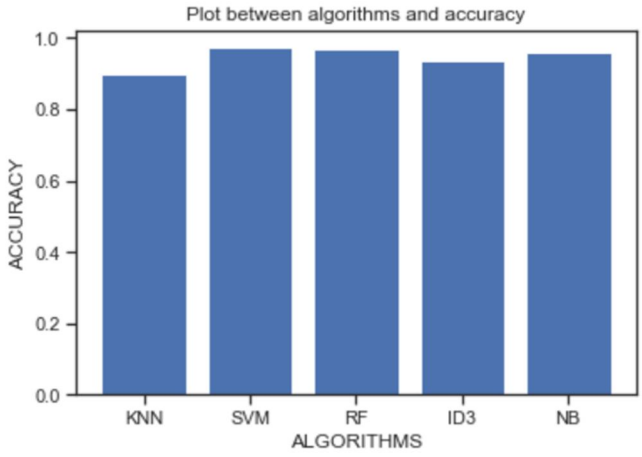


Figure 8 Plot between algorithms and accuracy

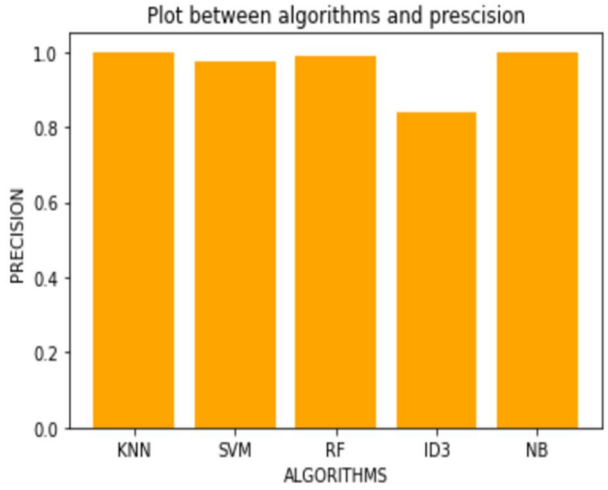


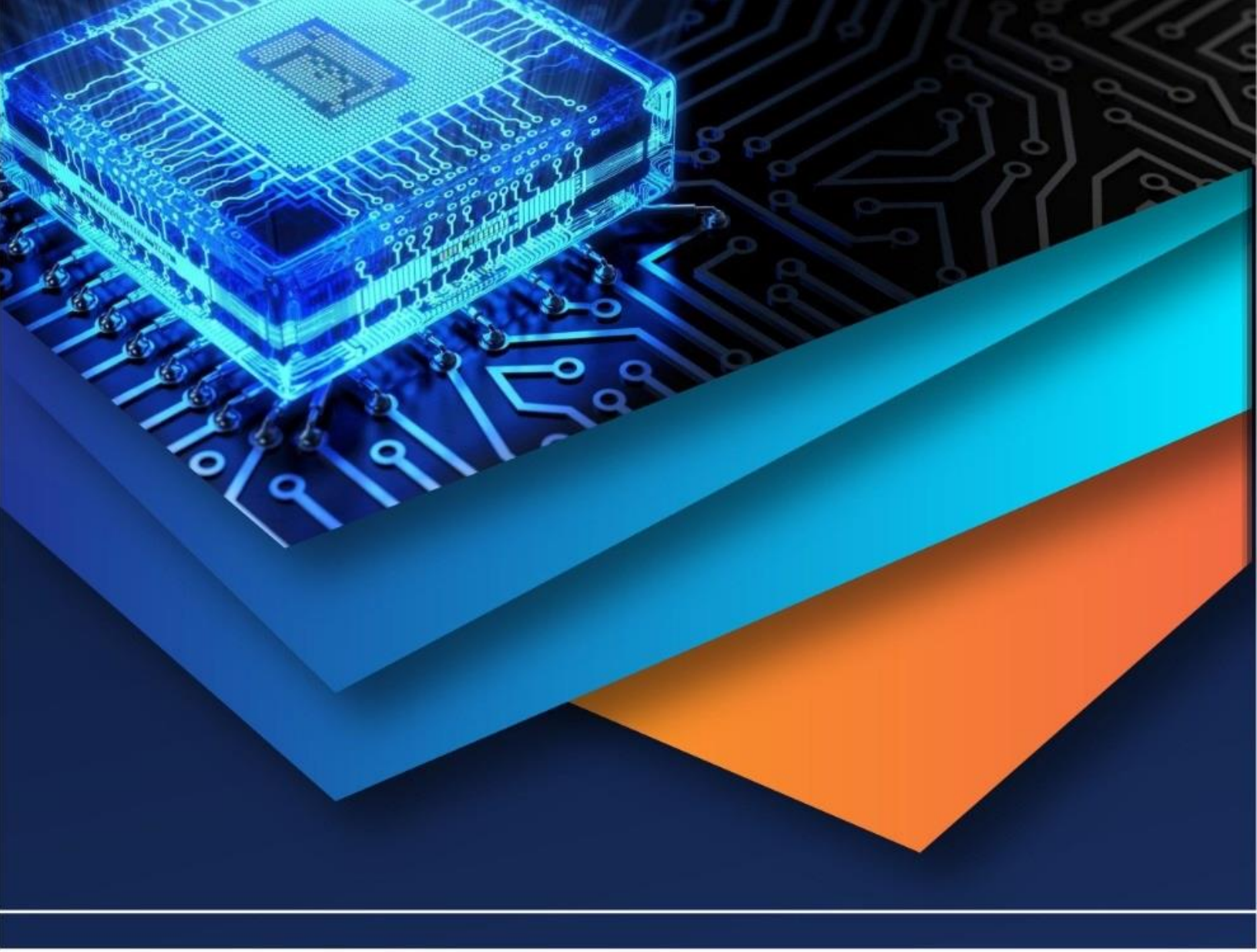
Figure 9 plots between algorithms and precision

VI. CONCLUSION

In summary, spam/ham classifiers are essential for separating spam from valid messages. For this task, a variety of machine learning algorithms can be used, including SVM, Naive Bayes, KNN, and Random Forest. These algorithms employ a variety of strategies, such as feature extraction, text analysis, and classification based on decision trees. Effective spam/ham categorization systems may be created by utilising the strengths of these algorithms, which will improve email and message filtering, lessen the impact of spam, and improve user experience and security in communication platforms.

REFERENCES

[1] Jain, N., Khanna, A., & Singh, N. (2020). SMS spam classification using machine learning algorithms and feature selection techniques. In Proceedings of the 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 1-5.
[2] Fernandes, G. F. A., Almeida, T. A., & Gonçalves, M. A. (2018). SMS Spam Collection: A public set of SMS spam messages. ACM Transactions on Asian and Low-Resource Language Information Processing, 17(1), Article 3.,
[3] Das, S., & Choudhury, G. (2013). SMS spam filtering using machine learning techniques. International Journal of Computer Applications, 65(18), 12-17.
[4] SMS spam classification using machine learning techniques and feature selection. International Journal of Computer Applications, 182(30), 13-18.
[5] Siddiqui, A., & Naik, R. (2019). SMS spam classification using machine learning techniques and feature selection. International Journal of Computer Applications, 182(30), 13-18.



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)