ADVANCES IN SPAM DETECTION TECHNIQUES AND APPROACHES

Abstract

Short Message Service (SMS) is an application that allows users to send short messages via mobile phones. Over the past few years, SMS has grown into a highly profitable industry as mobile phones have become more prevalent. Unwanted, unexpected, and potentially malicious messages are spam. An area of artificial intelligence is referred to as machine learning. Using algorithms and data, this area aims to steadily increase the accuracy of its models by mimicking how people learn. Classifiers based on machine learning are usually used to predict spam SMS. Based on the wording in the message, these algorithms can categorize SMS messages as spam or ham. The objective is to demonstrate a machine learning model that is effective at classifying messages as ham and spam.

Key words: Spam, Ham, SMS message

Author

Seethal Prince E
Research Scholar
Nehru Arts and Science College

Mahesh K.M
SCMS College

I. INTRODUCTION

A branch of artificial intelligence, known as machine learning, utilizes algorithms to analyze and gain insights from data. It allows systems to improve performance by learning from previous experiences. NLP is a specific area of machine learning that deals with analyzing and understanding human language. Language translation, text generation, and sentiment analysis are the application of it. Both machine learning and NLP are widely used in various industries, including healthcare, finance, and e-commerce.

Machine learning can be utilized to study SMS or short message services, which is a widely-used and popular form of communication. SMS or short message service comprises of small text messages. SMS has been a popular method for various purposes like advertising products, offers, for banking updates, it is commonly used for OTP system. It is also commonly used for SMS marketing. However, some users may find SMS marketing to be disruptive, and these types of messages are referred to as spam SMS. Spam SMS is unsolicited messages that users do not want to receive, and it is often sent with a wide collection of similar messages. The primary reason for receiving SMS spam include promotion of various products, inappropriate adultcontent, policy concerns and offers by the data provider. As a result of this, spam overflowing has become a serious issue globally. NLP, or Natural Language Processing, is used for content-based analysisof SMS messages.

1. OBJECTIVE

On average, text messages have an open rate of around 90%, and nearly all of them are read within 3 minutes of being received. However, these figures may fluctuate depending on the target audience, message content, and other elements. The project sets its goal to demonstrate a machine- learning model that is effective in classifying messages as spam or ham. The terms "spam" and "ham" are used to differentiate between unwanted and legitimate messages respectively. This project aims to use various machine learning algorithms for the classificationof SMS spam, comparing their performance to gain insight and further understand the problem. Theultimate goal is to design a model based on these algorithms that can accurately filter out SMS spam messages. This project uses a dataset of 5574 text messages taken from the UCI Machine Learning repository. The dataset is preprocessed and features are extracted, then different machinelearning techniques like Naive Bayes, SVM and other methods are implemented on the dataset, their performance is compared to evaluate their efficiency in identifying spam SMS.

2. SCOPE

Scope of project was to design and implement an SMS classifier using machine learning methods. The model is trained on a dataset of SMS messages and will be able to accurately identify and classify incoming SMS messages as spam or ham. The classifier will be developed using machinelearning and NLP techniques. The project is limited to classifying only SMS messages in English

II. LITERATURE SURVEY

SMS classification into spam and ham (legitimate messages) is important because it

helps to protect individuals and organizations from unwanted or potentially harmful communications. Spam SMS can contain phishing scams, malware, or attempts to sell fraudulent products or services, which impacts on the device or personal information of recipient. Additionally, spam

SMS can be a nuisance and can take up valuable storage space on a device. Classifying SMS messages into spam and ham allows for the efficient filtering and blocking of unwanted messages, which can improve the overall user experience and protect individuals and organizations from potential harm. Additionally, SMS classification can also be used to improve the effectiveness of SMS-based marketing campaigns by ensuring that messages are only sent to individuals who have opted-in to receive them.

SMS spam classification involves using algorithms to determine whether SMS messages as spam or not spam. Decision trees, random forests, and support vector machines are classifiers that can be used to train the model on a dataset of SMS messages and their labels, andthen make predictions on new incoming messages. The accuracy score of the model can be enhanced by using techniques such as feature selection and hyperparameter tuning. In the paper “Mobile SMSspam detection using machine learning techniques” authors Nagre et al., has observed the SpamFilter based on Naïve Bayes outperforms the Spam Filter based on Support Vector Machine[1]. The study has gone through the empirical analysis of both the Spam filters (Support Vector

Machine and Naïve Bayes) for messages. Extensive tests have been performed with varyingnumbers of dataset size. The success rates reach maximum using all messages and all words in a scorpus.

In paper “Spam detection approach for secure mobile message communication using machine learning algorithms” authors GuangJun et al., has used Logistic regression, Decision tree classifier, K-nn classifier to detect spam messages [2]. This paper proposes a method for detecting messages that are spam in a secure mobile system. The approach involves using machinelearning classifiers like Logistic regression, Decision tree classifier, K-nn classifierclassify messages as spam or non- spam. Model performance is evaluated on a dataset of SMS messages and compared with traditional methods. The study finds that machine learning algorithms can effectively detect spam messages and provide an accuracy of high magnitude in the classification. The approach also incorporates security measures to enhance the privacy of themessages being transmitted. Although each classifier performed well in classifying messages the into spam and ham, it has been observed that Logistic regression detectedspam messages with more accuracy in less processing time.

In paper “Machine Learning SMS Spam Detection Model” authors Kipkebut1et al., has conducted study of spam detection using Naïve bayes classifier [3]. The classifier accurately identifiedthe message as either spam or ham by examining the words that were contained in it. The study uses various classifiers like as Naive Bayes, Random Forest, and Support Vector Machines to train the model on a dataset of SMS messages and their labels. Accuracy, precision, F1-score were used to evaluate performance. The results show that the machine learning models outperform traditional methods in detecting spam messages and can be used to effectively filter out unwanted messages. The study also highlights the importance of feature selection and pre-processing in improving the performance of the model.

According to the paper "Mobile SMS spam filter techniques using machine learning techniques" authors Sravya et al., has done classification of SMS using various classifiers such as Linear regression, Support vector machine, K nn classifier, Decision tree classifier, Random Forest, Logistic regression and Naïve bayes [4]. SVM with linear kernel gave the best result among all algorithms. The study investigates the use of machine learning approach for filtering spam messages in mobile SMS communication. The study explores several algorithms including Linear regression, Support vector machine, K nn classifier, Decision tree classifier, Random Forest, Logistic regression and Naïve bayes and compares their performance on a dataset of SMS messages. The results show that machine learning techniques can effectively identify spam messages with a high degree of accuracy. The study also highlights the importance of feature selection and pre-processing in improving the performance of the models. The authors conclude that machine learning-based approaches are promising for filtering spam messages in mobile SMS communication.

III. PROPOSED SYSTEM

To check whether an SMS is ham or spam, a SMS classification system must go through several phases, including data pre-processing, feature extraction, training and classification. Through pre- processing text messages, we clean the text. Then the features of the pre-processed messagesare extracted. Machine learning classifiers such as Naïve bayes, Decision tree classifier, Random Forest and Support Vector Machine is used to classify these messages. Performance of these classifiers are compared and an ensemble method called stacking is used to create an effective model for classifying SMS into spam and ham.

1. **Spam Ham Sms Classification:** Dataset of the SMS messages was collected from UCI machine learning repository. The dataset was converted into csv file containing label followed by message. The label was either spam or ham. Then several sort of natural language processing methods are used to analyze the text in order to extract features which can be used for later analysis. The following figure shows a step-by-step approach towards analyzing and detecting ham spam SMS.

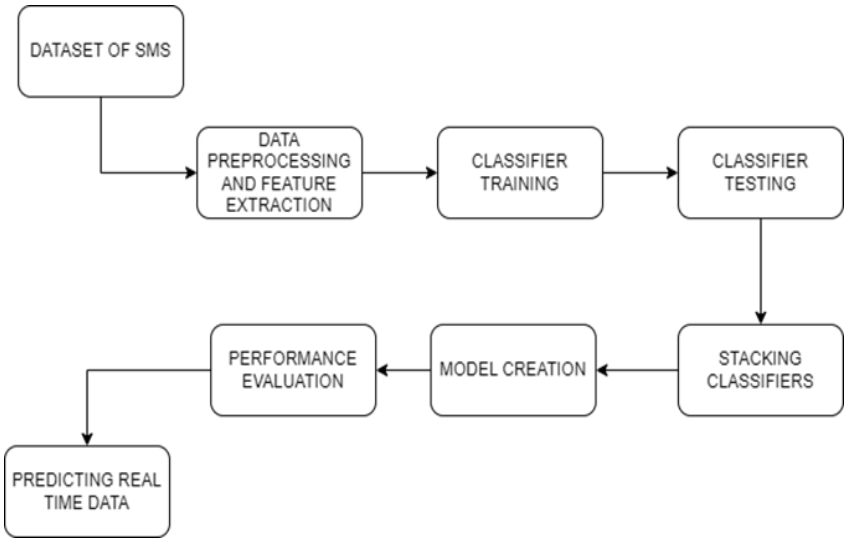
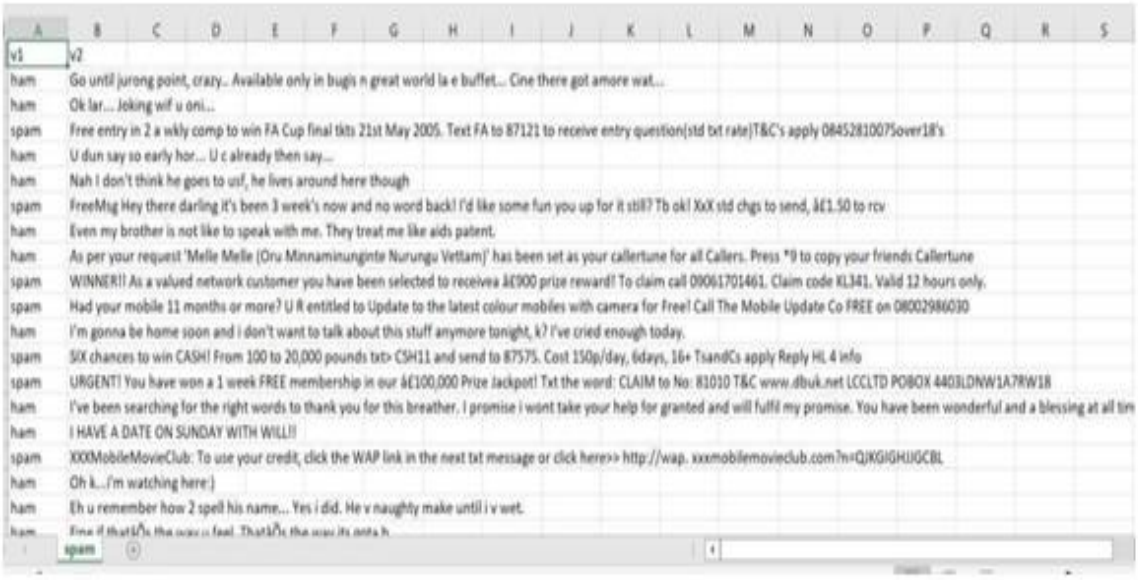


Figure 1: Proposed system architecture

This involves different steps as in the figure data preprocessing, feature extraction, classifier training and classifier testing. After comparison between different classifiers stacking is used as ensemble method. Then user interface was created to check real time data.



v1	v2
ham	Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got amore wat...
ham	Ok lar... Joking wif u oni...
spam	Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt rate)T&C's apply 08452810075over18's
ham	U dun say so early hor... U c already then say...
ham	Nah I don't think he goes to usf, he lives around here though
spam	FreeMsg Hey there darling it's been 3 week's now and no word back! I'd like some fun you up for it still? Tb ok! XxX std chgs to send, â£1.50 to rcv
ham	Even my brother is not like to speak with me. They treat me like aids patent.
ham	As per your request 'Melle Melle (Oru Minnaminunginte Nurungu Vettam)' has been set as your callertune for all Callers. Press *9 to copy your friends Callertune
spam	WINNER!! As a valued network customer you have been selected to receivea â£900 prize reward! To claim call 09061701461. Claim code KL341. Valid 12 hours only.
spam	Had your mobile 11 months or more? U R entitled to Update to the latest colour mobiles with camera for Free! Call The Mobile Update Co FREE on 08002986030
ham	I'm gonna be home soon and i don't want to talk about this stuff anymore tonight, k? I've cried enough today.
spam	SIX chances to win CASH! From 100 to 20,000 pounds txt> CSH11 and send to 87575. Cost 150p/day, 6days, 16+ TsandCs apply Reply HL 4 info
spam	URGENT! You have won a 1 week FREE membership in our â£100,000 Prize Jackpot! Txt the word: CLAIM to No: 81010 T&C www.dbuk.net LCCLTD POBOX 4403LDNW1A7RW18
ham	I've been searching for the right words to thank you for this breather. I promise i wont take your help for granted and will fulfil my promise. You have been wonderful and a blessing at all tim
ham	I HAVE A DATE ON SUNDAY WITH WILL!!
spam	XXXMobileMovieClub: To use your credit, click the WAP link in the next txt message or click here>> http://wap. xxxmobilemovieclub.com?n=QJKGIGHJJGCBL
ham	Oh k...i'm watching here:)
ham	Eh u remember how 2 spell his name... Yes i did. He v naughty make until i v wet.
ham	Fine if thatâ€™s the way u feel. Thatâ€™s the way its gota b

spam

Figure 2 : Dataset

2. **Data Preprocessing:** Zero and one is used to label ham and spam respectively. The data is cleaned and preprocessed by removing stop words, special characters, lemmatizing words and converting data to lower case. Nltk library was useful in removing stop words, tokenization and in lemmatization. Lemmatizationis a text normalization technique in Natural Language Processing (NLP) that groups different forms of a word (such as verb conjugations, plural forms, and different tenses) into one base form called a "lemma". The purpose of lemmatization is to change dimensionality of data by convertingall forms of a word into a single word, making it simpler to perform text analysis and processing.Tokenization is the process of converting a sequence of characters into smaller units, known as tokens. In NLP, tokenization is the first step in text pre-processing and refers to splittinga sentence or paragraph into words or tokens. The goal of tokenization is to obtain a sequence of tokens that can be used for further processing, such as building a vocabulary or performing NLP tasks like text classification, sentiment analysis, and named entity recognition. The resulting tokens can be either words or sub words, depending on the choice of tokenization method and the specific use case.

3. **Feature Extraction:** A phase where the relevant features of the text are extracted and converted into numerical representations that can be used for classification is feature extraction. TF-IDF vectorization is used for featureextraction process [5]. TF-IDF or “Term Frequency-Inverse Document Frequency”is a common technique used in NLP and information retrieval for classification and featureextraction of text. It is a numerical representation of the importance of a word in a corpus (a collection of documents). In a document the TF-IDF of a word is the product of its term frequency and inverse

document frequency [6]. The TF-IDF represents the final importance of words in a document, taking into account both its raw frequency and its general importance in the corpus.

The raw importance of one word in the document is the frequency. The inverse document frequency of a word is the logarithm of the total number of documents in the corpus divided by thenumber of documents that contain the word. The inverse document frequency measures the generalimportance of a word across the entire corpus. TF-IDF can be used to represent the content of a document in a numerical vector format, which can is the input to machine learning classifiers for text classification or clustering.

4. **Training & Testing:** For improved performance training different machine learning classifiers and comparing their results is a better choice. The machine learning classifiers used in this project are Multinominal naïve bayes, Decision tree classifier, Random Forest classifier and Support vector machine. Dataset is divided into two, training and testing dataset in a ratio of 80:20.

Multinomial naïve bayes is a type of Bayesian algorithm. Naive Bayes is a probabilistic algorithmthat uses Bayes' theorem to predict the class of an input by considering the prior probabilities of each class and the likelihood of the features given each class [7]. The "naive" assumption in Naive Bayes is the independence assumption between features, which means that it assumes that the presence of one feature is not related to the presence of any other feature [8]. Multinomial Naive Bayes is a type of Naive Bayes algorithm that is specifically designed for text classification problems. In Multinomial Naive Bayes, each feature is treated as a count for how many times a particular word appears in the text. For example, if the input is a sentence, the features could bethe frequency of each word in the sentence. These feature counts are then used to calculate the likelihood of each class given the input text.

Decision Tree classifier is a supervised learning method that is used for binary classification tasks such as spam SMS detection. The decision tree works by creating a tree model for decisions and for consequences. Every internal node in the tree represents a decision based on feature of input data, and every leaf node represents a prediction for the class of the input [9]. In the case of spam SMS detection, the decision tree would consider various features of the text messages, such as the presence of certain words, the length of the message, or the numberof links contained in the message. Based on these features, the tree would make decisions and determine whether the message is likely to be spam or not.

An ensemble method known as random forest used for classification problems, including spam SMS classification. It combines multiple decision trees to make a prediction by taking the majorityvote of all trees, making it more robust and less likely to overfit the data than a single decision tree[10]. Random Forest has several advantages over other classification algorithms, including improved accuracy, better handling of noisy and missing data, and the ability to estimate feature importance. However, the model can be computationally expensive and slow to train on large datasets. Also, it requires a large number of trees to make a robust prediction, which can increase the memory usage.

Support Vector Machine (SVM) is a popular machine learning technique that is used for text classification tasks, including spam SMS classification [11]. SVM tries to find best hyperplane which separates the classes in the feature space. In the case of spam SMS classification, the hyperplaneseparates the spam SMS from the ham SMS. By mapping the input data into a high- dimensional feature space, where a maximum margin hyperplane is created to separate the classes SVM works.The data points most close to hyperplane are support vectors, and they define the hyperplane [12]. SVM is well-suited for text classification problems as it can handle high- dimensional data, such as text data, efficiently. SVM is also highly flexible and can work well with a variety of kernel functions, which can be used to transform the data into a higher-dimensional space were the hyper plane can be created.

After comparing the accuracy and results of each classifier stacking is introduced as an ensemble method [13]. Stacking is an ensemble learning method that combines the output of multiple base models to improve the overall accuracy of our model. The idea behind stacking is to train a setof base models on the input data and then use these base models' predictions as inputs to a final meta- model, which makes the final prediction. This can lead to improved performance comparedtousinga single base model, as the meta-model can learn how to effectively weight the predictions ofthe base models to optimize performance.

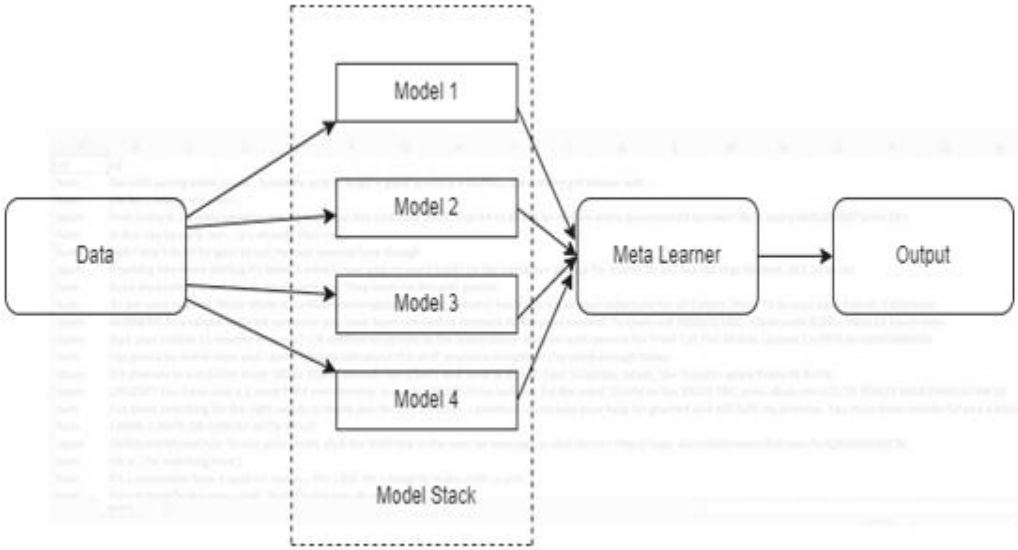


Figure 3: Stacking ensemble method

Stacking is a machine learning ensemble method that combines multiple base models to form a more accurate and robust prediction. The idea behind stacking is to train several base models on the input data and use their predictions as input features for a final meta-model, which makes the final prediction. In this project the meta classifier used is SVM. Stacked model is used to implement user interface for predicting real time data.

IV. RESULTS AND DISCUSSION

The project of spam ham SMS classification was able to construct a model that can classify SMS messages into spam and ham based on their text content. The model was build using ensemble method called stacking, in which four machine learning classifiers were stacked. After defining the model, it was trained with a large dataset of SMS messages which were ham and spam. Thus, when a message is given as input the model predicted it to be spam or ham based on its content. For comparing the performances of four classifiers and the stacked model performance metrics such as accuracy score, confusion metrics were used.

	v1	v2	Unnamed: 2	Unnamed: 3	Unnamed: 4
0	ham	Go until jurong point, crazy.. Available only ...	NaN	NaN	NaN
1	ham	Ok lar... Joking wif u oni...	NaN	NaN	NaN
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...	NaN	NaN	NaN
3	ham	U dun say so early hor... U c already then say...	NaN	NaN	NaN
4	ham	Nah I don't think he goes to usf, he lives aro...	NaN	NaN	NaN

Figure 4: Loaded Dataset

Dataset was then visualized with the help of seaborn library to check the count of spam and ham SMS. Seaborn is a Python library used for data visualization and statistical plotting

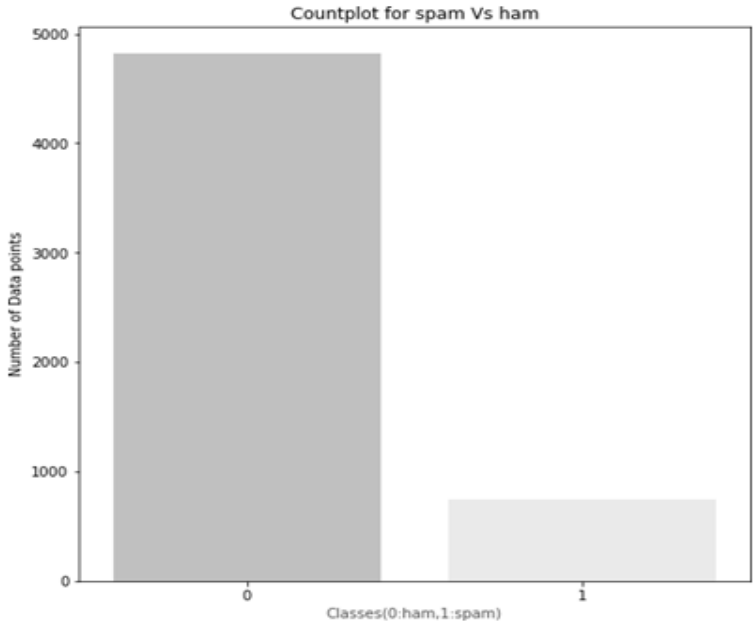


Figure 5: Statistical distribution of ham and spam SMS in dataset

Table 1 : Accuracy and Precision of Different Classifiers

Classifier	Accuracy	Precision
Naïve Bayes	0.976682	0.941606
Decision tree	0.959641	0.859155
Random Forest	0.983857	0.951049
SVM	0.984753	0.977941

Confusion matrix helps to visualize overall performance of each classifier.

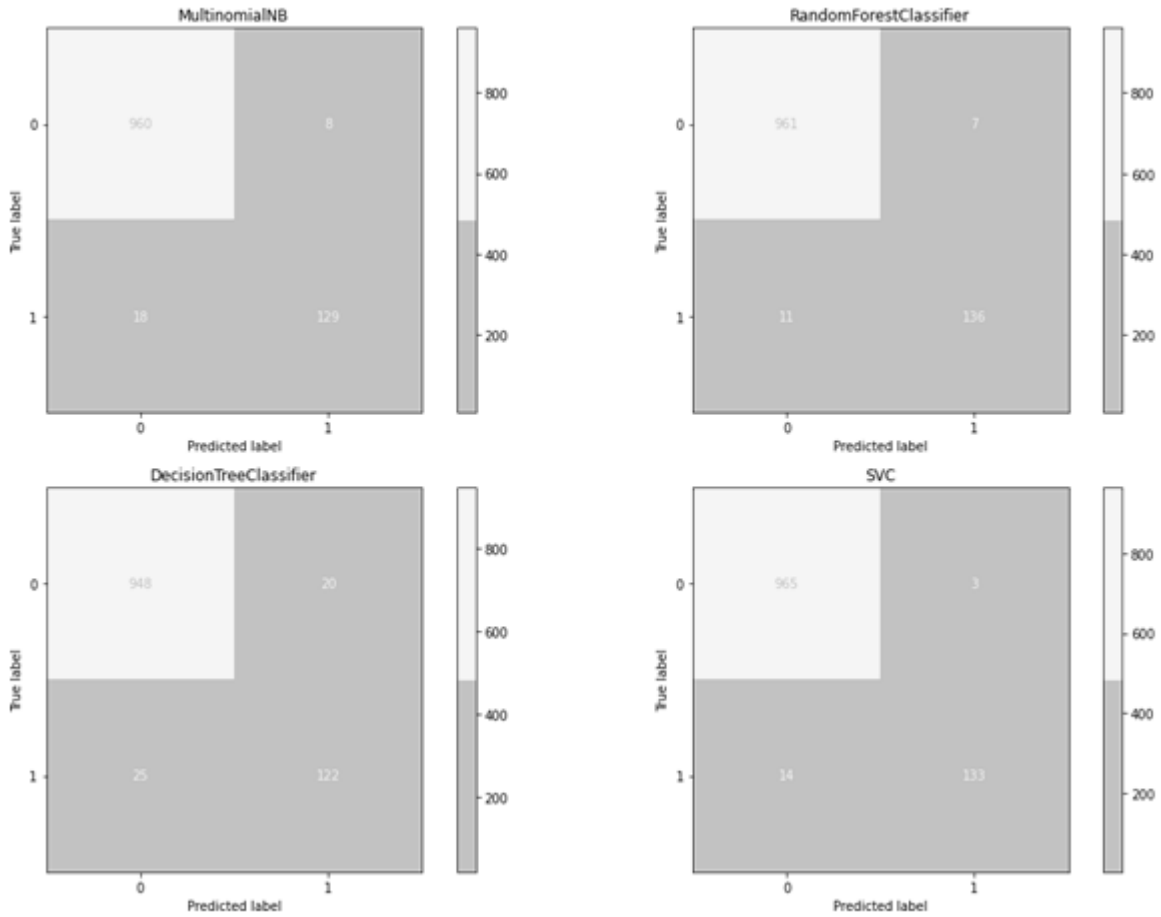


Figure 6: Confusion matrix of classifiers

The accuracy score is a metric used in machine learning to check the efficiency of a classifier ormodel. It represents the proportion of correct predictions made by the model compared to the totalnumber of predictions. Precision represents the capacity of the model to

identify correctly positive instances and avoid false alarms. The confusion matrix is a matrix used to evaluate the performance of a classifier. It contains four elements: true positives (TP), false positives (FP), false negatives (FN),and true negatives (TN). The confusion matrix is used to calculate various evaluation metrics such as accuracy, precision, recall, and F1 score.

Table 2 : Accuracy and Precision of Stacked model

Accuracy	Precision
0.9865	0.9647

The classification report is a summary of the efficiencyof a classifier. It providesevaluation metrics such as precision, recall, f1-score, and support for each class in the classification problem. The report helps to understand how well the classifier is performing in terms of identifyingthe correct class for each instance. The f1-score, which is the harmonic mean of precision and recall,implies overall measure of the classifier's performance. The support is the number of instances of each classin the test dataset. A classification report is an important tool for evaluating the performance of a classifier.

Table 3 : Classification Report

	Precision	Recall	F1 score	Support
0	0.99	0.99	0.99	968
1	0.96	0.93	0.95	147
Average			0.99	1115
Macro avg	0.98	0.96	0.97	1115
Weighted avg	0.99	0.99	0.99	1115

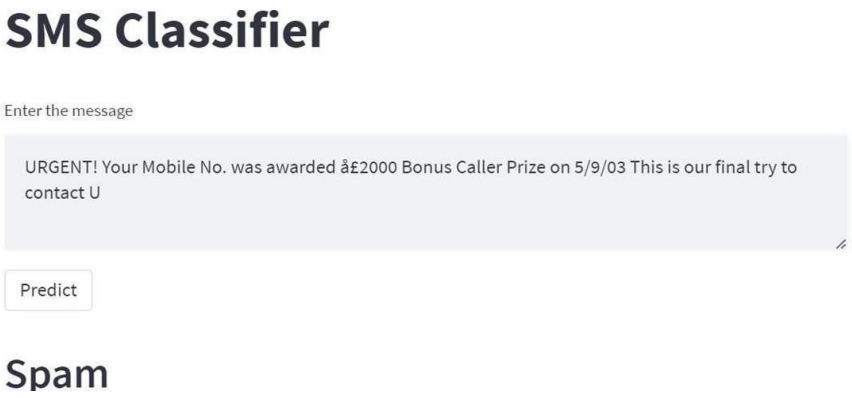


Figure 7: Classification of SMS in user interface

System is used to classify SMS into spam and ham. Based on its classification the accuracy obtained is 98.65%. Certain factors that affect accuracy of the system may include

dataset and meta classifier used. Large and balanced data set can improve the performance as the system used an imbalanced dataset for training. Also, the choice of meta classifier used in stacking can result in a difference in accuracy.

V. CONCLUSION & FUTURE SCOPE

SMS classification is an important challenge in field of NLP and machinelearning. To develop a system that can accurately classify SMS messagesas spam or ham was project aim. To check whether a SMS is ham or spam we have to go through many steps including cleaning and preprocessing the data, feature extraction using vectorization then it is given to the model to predict whether the SMS is ham or spam.

To reach this, various machine learning algorithms were applied and evaluated, including Naive Bayes, Decision Tree, Random Forest, Support Vector Machine, and Stacking Ensemble Method.

The results showed that the Stacking Ensemble Method outperformed the other algorithms in terms of accuracy, precision, and recall. This suggests that the Stacking Ensemble Method is a suitable choice for SMS classification, especially when combined with appropriate feature engineering and data pre-processing techniques. Stacking is a powerful ensemble method in machine learning that can lead to improved performance compared to individual models. The reason stacking is often better than other methods is that it allows for the combination of the strengths of multiple models. Stacking uses the predictions of multiple base models to train a higher-level meta-model, which is then used to make the final prediction. This allows for the strengths of each base model to be combined and the weaknesses to be compensated for, leading to a more robust and accurate final model.

In addition, stacking can also be used to address the issue of model diversity. The key to a successful ensemble is having a diverse set of models that capture different aspects of the data.

Stacking provides a way to incorporate different models into the ensemble, which increases the diversity of the models and leads to improved performance. This is because the stacking algorithm trains a meta-model to make the final prediction based on predictions of the base models, so it effectively combines the strengths of all the models into a single, more accurate prediction.

However, it should be noted that the results of the project may not be generalizable to other datasets or real-world applications, as the performance of machine learning algorithms depends on many factors like the size and quality of the training data. Further studies may be necessary tovalidatethe results and to enhance performance of SMS classification system. In conclusion, SMS classification is one of the challenging problem that requires a combination of domain knowledge, NLP techniques, and machine learning algorithms. Also, we can use non-content-based SMS classification techniques [14].

The results of this project provide some insights into the capabilities and limitations of different machine learning algorithms for SMS classification and highlight the importance of feature engineering and data pre-processing in achieving high accuracy.

