Mobile Sms Spam Filter Techniques Using Machine Learning Techniques

Gomatham Sai Sravya, G Pradeepini, Vaddeswaram, Guntur

Abstract: SMS spam is a contemporary issue fundamentally because of the accessibility of very modest mass SMS bundles and the way that SMS induces higher reaction rates as it's far a depended on and personal service. In this paper, we will be differentiating the messages into two categories: Ham and Spam. Ham is described as the dataset that includes the textual content of SMS messages at the side of the label indicating whether the records is legitimate message or now not. Spam is defined as the dataset that includes the textual content of SMS messages along with the label indicating the junk messages. In SMS Spam messages, the advertisers utilize the SMS text messages to target the customers with unwanted advertising. But it is troublesome, because the users pay a fee per SMS received. To overcome this, we perform a comparison between the machine learning algorithms to predict the messages and calculate the accuracy criterion by using the SMS spam dataset.

Index Terms: Mobile SMS Spam, Ham, Spam, Machine Learning, SMS Spam dataset, Algorithms, Messages, Accuracy.

1 INTRODUCTION

Globally, SMS is maximum famous and lower priced telecommunication service. Spam is unwanted messages sent electronically. However, cell customers have emerged as more and more worried about the safety for differentiating the junk and regular messages. SMS Spamming is a major nuisance to mobile users as they get lots of junk messages rather than the legitimate messages. The messages are sorted as electronic, spontaneous, business comprises a developing danger for the most part because of following components especially accessibility of low-expense mass SMS, unwavering quality and generally execution, okay of getting reaction from obscure clients.Evolution of mobile conversation technology and the growth of cell telephones, SMS has grown to be the maximum major transmission modes consistent with its basic operation and less price. SMS spams are ensuing in higher reaction charges than Email spams as SMS is a trusted carrier with customers at ease using exclusive statistics. As many people are using SMS, it became very easy to everyone to use the latest technology for easy communication. It is becoming cost powerful for SMS to goal the provision of the resources which increases the limitless SMS programs for higher use. Although, SMS messages are not free, there is a high intensity for SMS spams, because the filtering techniques are still improving day-by-day to reduce the spam messages and get the solicited messages.Spam SMS is an emerging problem in the society as there are number of unwanted messages which are not useful for the users and occupies much storage space which increases junk messages rather than the legitimate messages. Anti-unsolicited mail measures are also along with anti-spoofing and faking measures that may correctly become aware of SMS messages which have been controlled to manufacture the first subtleties.

The spam in SMS reasons exasperation at the end-user, useful capital intake of the cell gadget and in a few groups even the recipient is charged for purchasing the SMS. The spam in SMS causes inconvenience at the individual, helpful asset utilization of the cell gadget and in certain networks even the recipient is charged for buying the SMS. Subsequently it is of especially centrality that these garbage mail messages be evacuated when they are acquired at the versatile station, if no longer sooner than that. The trouble of cell smartphone unsolicited mail has not acquired that ton's attention through the research network because the more acquainted electronic mail junk mail.Spam filters offer a few levels of performance as they lie to receivers with the aid of the fact's samples. SMS is used as an alternative for voice calls where communication is either not possible or not desired between end phone users. The drawback of Spam SMS is it's far unwanted and meaningless messages which can't be beneficial for the customers, content material traffic. It occupies storage space in the phone and computational power. The procedure provided directly here depends on machine examining, and the effortlessness of the method is that it calls for just four capacities extricated from the SMS messages. It has been demonstrated exactly that these 4 capacities are sufficient to sift through the spontaneous mail SMS from the non-garbage mail or 'ham' messages. It might do the sort progressively.Many methods are applied for filtering SMS spams. It is divided into content-based approaches and non-content-based approaches. Social network is the best example for non-content-based approaches. It is used by telecom users rather than the mobile users. We use classification techniques for predicting the text messages whether it is Ham or Spam. The text classification techniques which we are using in this paper are content-based. It is based at the truth that spam SMS is essentially an issue of content kind. It must be remembered that straight-forwardness of the plan is extremely crucial because of the obliged sources accessible on a cell gadget. It is reasonable to have a machine with significantly less computational burden and less memory and battery necessities despite the fact that there is a touch penance at the precision. Regardless of whether an endorser gets a spam SMS now not separated through the device from time to time on the cost of state 5-

- *Gomatham Sai Sravya is currently pursuing masters degree program in computer science and engineering in Koneru Lakshmaiah Education Foundation, India, PH-9030950258. E-mail: sravyasaaradhi@gmail.com*

10 % that would not be as bounty a difficulty as a mind-boggling SPAM get out that is ingesting resources at an extreme cost.In this paper, we will be providing the structured view of the present machine learning algorithms for spam filter. We use similar datasets for spam filter to predict the ham and spam messages by using different machine learning algorithms for better accuracy rate. The basic concept of content-based approach is that it calculates the ham and spam messages and how does it affect in means of algorithms.
Taking a SMS dataset and using different machine learning algorithms, we will be getting the output with different accuracy scores so that we can easily find the better algorithm to predict whether the message is ham or spam.

1.1 Purpose of Spam

- Multi-level Marketing
- Advertisements
- Stock markets
- Political emails
- Chain letters

1.2 Characteristics of SMS Spam Filtering System

Our research goal is to broaden an SMS spam filtering machine on a smartphone with the following crucial characteristics.

1. **Independent:** It does now not want a assisting laptop. Thus, training and updating the filtering machine will all be achieved on the mobile phone. This will reduce communication costs between the cell smartphone and computer. It moreover reduces hardware upkeep and infrastructure expenses.

2. **Private:** The filtering machine ought so as to make sure the man or woman's privacy. The filtering gadget ought to no longer keep the consumer's SMS to anywhere. Storing SMS at a third birthday celebration can raise private worries, especially with SMS ham, which can also consist of private statistics.

3. Secure: The spammer has to now not be capable of access the filtering device because the spammer can create SMS spam that could fool the filtering system.

4. **Personal:** Each consumer has a special perception of SMS junk mail eight. Some users may additionally say that a particular SMS is unsolicited mail, others won't. As such, customers need to have the chance to create their own filtering device, deciding on the schooling data set via themselves.

5. **Simple**: The customers cannot wait until they've a massive amount of SMS information for a schooling records set earlier than they begin filtering SMS unsolicited mail. Thus, the filtering device must be capable of begin filtering SMS direct mail using a small training information set. Simple also technique that clients do no longer need to configure a connection between their mobile telephone and a laptop.

6. **Updatable:** The filtering machine need for you to adapt to new SMS characteristics with the aid of constantly updating the filtering device whilst receiving new incoming SMS.

1.3 Flowchart

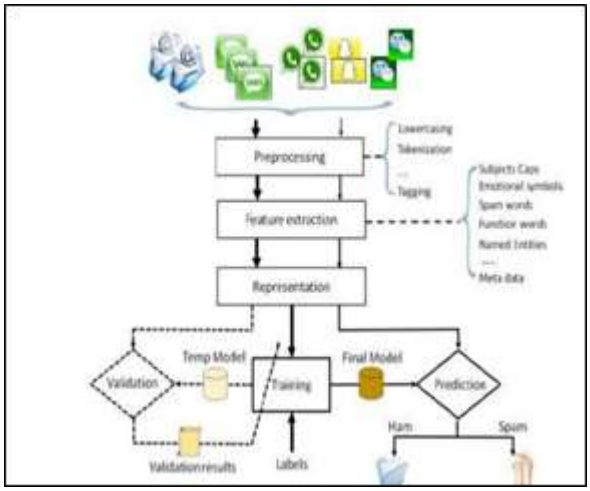

***Fig. 1**. Flowchart Of SMS Spam Filtering*

2 LITERATURE SURVEY

TABLE 1
RELATED WORK

S.No	AUTHOR	PROBLEM	SOLUTION
1	Shafi'i Muhammad Abdulhamid, Muhammad Shafie Abd Latiff, Haruna Chiroma, Oluwafemi Osho, Gaddafi Abdul-Salaam, Adamu I. Abubakar, and Tutut Herawan	Detection and filtering of SMS Spams	By using different machine learning algorithms like SVM and Bayesian classifiers.
2	Lutfun Nahar Lota B M Mainul Hossain	Increasing the Accuracy and decreasing the complexity	By having the SVM algorithm, it gives better accuracy but suffers from implementation complexity.
3	Neelam Choudhary, Ankit Kumar Jain	For better accuracy	Used five machine learning algorithms namely Logistic Regression, Naive Bayes,J48, Decision Trees and Random Forest.

4	Houshmand Shirani-Mehr	Reduce the spam messages and for better accuracy	By using UCI Repository dataset with different machine learning algorithms
5	Naresh Kumar Nagwani, Aakanksha Sharaff	Detecting the spam and non-spam messages	By using clustering techniques, we can detect the SMS spams and find the better accuracy
6	Kuldeep Yadav, Swetank K. Saha, Ponnurangam Kumaraguru, Rohit Kumra	To detect the SMS spam messages and calls.	By using SVM, we can detect the SMS and give the better accuracy
7	Sakshi Agarwal, Sanmeet Kaur, Sunita Garhwal	To give the better accuracy	SVM and Multinomial Naive Bayes are used by having the dataset and calculated the accuracy for better score
8	Tej Bahadur Shahi, Abhimanu Yadav	By using machine learning algorithm like SVM it is found that the accuracy is 87.15%	Used the naïve bayes algorithm which gives the accuracy of 92.74%
9	Dr. Ghulam Mujtaba, Majid Yasin	For better accuracy and performance	Used Naive Bayes algorithm for better performance.

3 METHODOLOGY

In this paper, the purpose is to explore the results of applying machine learning techniques to detect message spam detection. We are going to make a version to classify a message as unsolicited message or ham. In that model, we will train and test data using different machine learning algorithms and find out which algorithm works best in the dataset. In this, we will be using classification algorithms like Logistic Regression, Kneighbors Classifier, Random Forest Classifier, Decision Tree Classifier and Support Vector Machines.

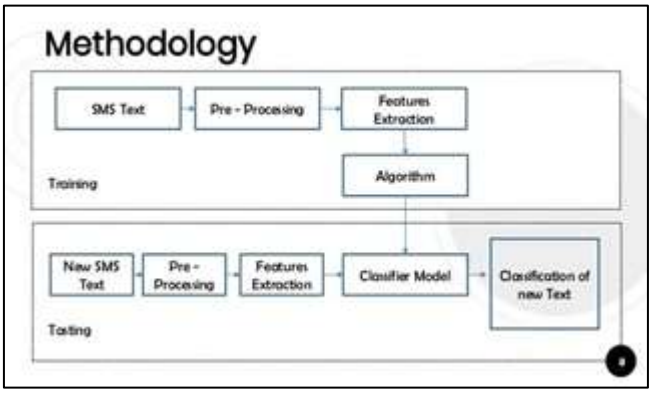

***Fig. 2**. Methodology Description For Training And Testing Cases*

Sr.#	SMS Status	SMS text
1	Ham	Fine if thats the way you feel.
2	Spam	Jazz Karaoke! Mobilink Jazz presents a unique technology with which you can sing along with your favorite songs.
3	Ham	I'm going to try for 2 months ha ha only joking
4	Ham	Just forced myself to eat a slice. I'm really not hungry tho. This sucks
5	Ham	Lol you are always so convincing.
6	Ham	K tell me anything about you.
7	Spam	*Jazz New SIM Offer* New family members of Jazz can enjoy 100 FREE Minutes and 100 free sms on daily usage of just Rs. 10+tax.
8	...	.....
9		And so on

***Fig. 3**. Spam SMS Dataset*

Using different approaches to establish relation between the text and category SPAM or HAM like, based on the size of message, word count, and special keywords. We compare the accuracy of each technique and plot the accuracy graphs in single bar plot.

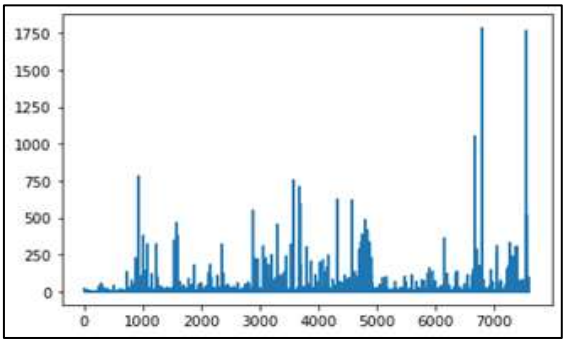

***Fig. 4**. Word Frequencies*

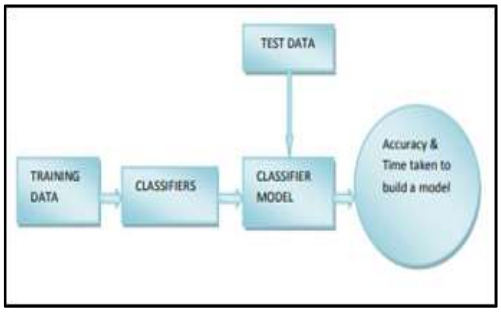

***Fig. 5**. Spam Classification*

3.1 Spam Filtering Process:

SMS Spam is manually classified into ham and spam messages which are given as the input or the training data for the spam filter algorithms. These algorithms consist of following stages:

1. **Data Pre-processing:** Data pre-processing is used to transform the raw data to predictable format. In this stage, the immaterial data such as stop words are eliminated.

2. **Tokenization:** Segmenting the message in keeping with phrases, characters or symbols known as tokens. Word tokenization is one type of approach in tokenization techniques. In order to train the model, we need to convert the text into appropriate numerical values using vectorization.

3. **Representation:** It converts the attribute value pairs in the training set.

4. **Selection:** Rather than choosing all the attribute value pairs, we can select the important attribute value pairs for classification.

5. **Training:** It trains the algorithms for the important attribute values.

6. **Testing:** It tests the new data with the training model.

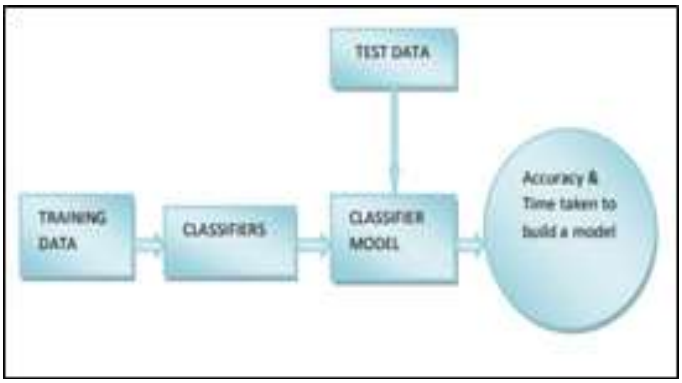

Fig. 5. Spam Classification

3.2 Algorithms

1. Logistic Regression:

It is a category set of rules used to predict the chance of established variables. It is a simple approach to categorise the information- be it ham or no longer ham, unsolicited mail or not junk mail. It classifies the ham and unsolicited mail messages and gives the confusion matrix for the dataset. Here, we will split educate and check the dataset and gives the accuracy score. We have our education facts in columns (v1 and v2). We will use depend vectorizer to convert the gathering of text documents to a matrix of token of specific word counts.

$$f(x) = \frac{1}{1+e^{-\beta x}}$$

2. KNeighbors Classifier:

It is a pattern classifier and it is non-parametric method used for classification and regression. Here, we classify the train and test data and perform testing for better accuracy. We will fit every classifier by way of enforcing the KNeighbors classifier and discover the accuracy criterion. Cluster centres are again and again adjusted to the mean in their currently acquired facts factors. The classification algorithm attempts to locate the K-Nearest Neighbor of a take a look at data point and uses a majority vote to decide its elegance label. The performance of KNN classifier is often determined by means of (i) the proper desire of K, and (ii) the space metric implemented. We use "minkowski" as a metric. We check the data frame for the classifier and we calculate confusion matrix for test dataset and finds the accuracy for the dataset. We have our training data in two columns (v1 and v2). This could be summarized in:

$$y(d, C_j) = \sum_{d_i \text{ is in } KNN} Sim(d, d_i) \times y(d_i, C_j) - b_j ;$$

$$y(d, c) = \begin{cases} 0, \text{if } d \text{ is in class C} \\ 1, \text{if } d \text{ is not in class C} \end{cases}$$

3. Decision Tree Classifier:

This is a divide and conquer algorithm. Here, we classify the data and divide the train and test dataset for better accuracy. We include data frame, calculate the confusion matrix and discover the higher accuracy criterion. It divides the records and compares with each and each algorithm. We have our training data in two columns (v1 and v2).

4. Random Forest Classifier:

It is an ensemble learning method for classification and uses averaging to improve the predictive accuracy criterion. In this, it calculates the confusion matrix and finds the accuracy for the given dataset. We use RandomForestClassifier method to run the python code and takes the input from the given dataset. We have our training data in two columns (v1 and v2).

5. Support Vector Machines:

This is a tool for data classification. In this, SVM is used to determine whether the SMS is spam or ham. An SVM method is based totally on structural threat minimization. It avoids the use of many schooling documents, using simplest those close to the class border, to construct an abnormal border isolating fine and negative examples. By employing a suitable kernel functions, it could learn polynomial classifiers, radial foundation capabilities, and three-layered sigmoid neural nets, accordingly acquiring widespread gaining knowledge of potential. In this study, we have used SVM with linear and RBF kernels to get the better accuracy and confusion matrix. We have our training data in two columns (v1 and v2).

a) SVM with Linear:
In linear kernel, we will divide the train and test dataset and calculate the accuracy for the given dataset. We represent the dataset with plot diagram.

b) SVM with RBF:
In RBF kernel, we will divide the train and test dataset and calculate the accuracy for the given dataset. It will differentiate the dataset with plot diagram representation.

6. Naïve Bayes Classifier:
The Naive Bayes set of rules creates a probabilistic version for type of SMS messages. Even although all capabilities make contributions towards the general probability of class, Naive Bayes set of rules assumes that the capabilities are statistically unbiased of every different. Although this assumption may not hold genuine for all cases, Naive Bayes set of rules has shown promising consequences in contrast with other famous type algorithms. A benefit of Naive Bayes is that its handiest calls for a small amount of education statistics to estimate the parameters necessary for type. Because unbiased variables are assumed, simplest the variances of the variables for every magnificence want to be determined and no longer the whole covariance matrix. The basic decision rule can be defined as follows:

$$f(\vec{x}) = \underset{y=\{c_{spam}, c_{leg}\}}{argmax} \left(\hat{P}(y) \prod_{j:x^j=1} \hat{P}(x^j = 1 | y) \right)$$

7. Linear Regression:
Linear regression is a linear perspective to model the connection among a variable response and one or extra descriptive variables. The case of one descriptive variable is called Simple Linear Regression. If there are more than one descriptive variable, then it is called as Multiple Linear Regression. This term is well-defined from multivariate linear regression, wherein more than one associated dependent variables are forecasted, as conflicting to a single constant variable. In this, we will predict the messages using this machine learning algorithm.

4 RESULTS
The cause of this Project is to discover the effects of making use of gadget gaining knowledge of techniques to discover Message spam detection. SMS unsolicited mail (every now and then known as cell smartphone junk mail) is any junk message brought to a cellular phone as textual content messaging via the Short Message Service (SMS). The dataset for this mission originates from the UCI Machine Learning Repository.Using special techniques to establish relation between the textual content and the category SPAM or HAM like, primarily based on length of message, word depend, unique keywords. Then construct class fashions the use of one-of-a-kind strategies to differentiate spam SMS. Compare accuracy of each method and plot the accuracy graphs in a single bar plot. Also generate a phrase-cloud for junk mail SMS.
For given dataset, we will compare with different algorithms and checks which algorithm gives the better accuracy whether the message contains HAM or SPAM.
Total Dataset: 5572
HAM messages: 4825
SPAM messages: 747
After training and testing the dataset, the messages will be split. The result will be as follows:
HAM: 3876
SPAM: 581

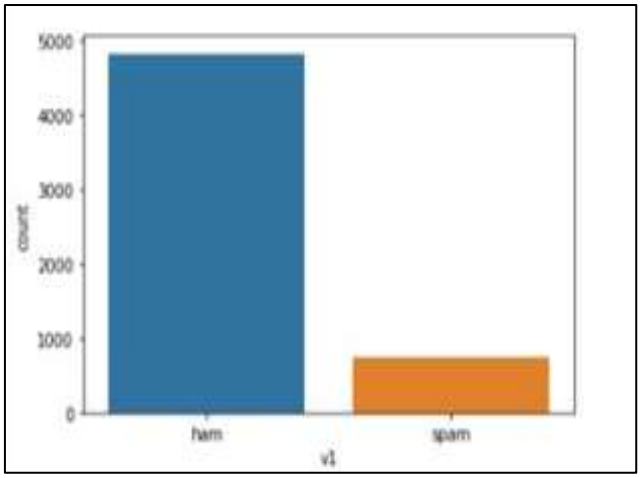

***Fig. 6**. Ham And Spam Count Bar Graph Output*

TABLE 2
ACCURACIES

S.NO	ALGORITHM	ACCURACY
1	Linear Regression	74.9%
2	Support Vector Machines (SVM)	85.3%
3	KNN Classifier	92.4%
4	Decision Tree Classifier	95.8%
5	Random Forest	96.2%
6	Logistic Regression	97.8%
7	Naïve Bayes Classifier	98.9%

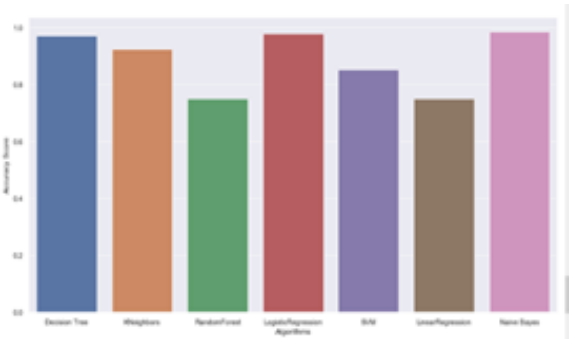

***Fig. 7**. Machine Learning Pyplot Accuracy Scores Comparative Graph*

By comparing different classification algorithms, it clearly shows that SVM with linear kernel gives the better accuracy with 98.1%.

5 CONCLUSION

The undertaking of automated filtering of SMS messages is still a trouble. There are two predominant challenges hindering the improvement of algorithms: lack of actual datasets, and occasional functions that extracts the messages. The main aim of this paper is to compare different machine learning algorithms i.e., SVM with RBF Kernel, KNeighbors Classifier, Decision Tree Classifier, Random Forest, Logistic Regression, SVM with Linear Kernel with better accuracy score. By using kernels in this, we can find the better output through SVM with linear kernel. The text messages are differentiated with Ham or Spam. It predicts whether the message in the dataset is Ham or Spam and predicts the performance through accuracy criterion. Using these algorithms, we can see whether the data in the dataset is predictable or not.

6 REFERENCES

[1]. S. M. Abdulhamid et al., “A Review on Mobile SMS Spam Filtering Techniques,” in IEEE Access, vol.5.

[2]. JOUR Lota, Lutfun, Hossain, “A Systematic Literature Review on SMS Spam Detection Techniques”, International Journal of Information Technology and Computer Science.

[3]. CHAP, Choudhary, Neelam, Jain, Ankit, “Towards Filtering of SMS Spam Messages Using Machine Learning Based Techniques”.

[4]. “SMS Spam Detection using Machine Learning Approach”, Houshmand Shirani-mehr,2013.

[5]. Nagwani, N. K., & Sharaff, A, “SMS spam filtering and thread identification using bi-level text classification and clustering techniques”, Journal of Information Science,2017.

[6]. Yadav, Kuldeep, Saha, Swetank, Kumaraguru, Ponnurangam, Kumra, Rohit, “Take control of your SMSes: Designing an usable spam SMS filtering system”, IEEE 13th International Conference on Mobile Data Management, MDM 2012

[7]. S. Agarwal, S. Kaur and S. Garhwal, “SMS spam detection for Indian messages,” 1st International Conference on Next Generation Computing Technologies (NGCT), Dehradun, 2015.

[8]. Tej Bahadur Shahi, Abhimanu Yadav, “Mobile SMS Spam Filtering for Nepali Text Using Naïve Bayesian and Support Vector Machine”, International Journal of Intelligence Science, Vol.4 No.1, 2014.

[9]. Ghulam Mujtaba and Majid Yasin, “SMS Spam Detection Using Simple Message Content Features”, 2014.Applications, vol. 39, pp. 9899-9908, 2014.