IJETIR
ISSN (Online) 2583-0554
Impact Factor: 5.731
International Journal of Emerging Technologies and Innovative Research (IJETIR)
Volume 5, Issue 12, December 2025

Machine Learning Approach for Natural Language Processing

Jaanaki S. M
Department of Electronics and Communication
East Point College of Engineering and Technology, Bangalore
jaanumohan88@gmil.com

Abstract: *SMS is a communication protocol designed for exchange of short, written messages used for various business purposes, despite of various other method of communication. Automated management of SMS is important in the today's context as the volume of SMS grows day by day. Out of total SMS, nearly a fair percentage is identified as spam. The spam actually consumes a lot of time of the users and the memory of the device. The spammers use SMS to share messages which can also contain links to malware, phishing, or scam websites that can infect the device or steal the data. So, it necessitates to have a proper mechanism to get devoid of spam messages. The motivation of this paper mainly focusing on the spam classification approach using machine learning algorithms. SVC approach was applied to analyze and perform data preprocessing. In this method a statistical measure is used to evaluate the importance of the word in a document related to collection of documents*

Keywords*:* Spam classification, Spam detection, Spam Filter, SMS, Supervised Learning, Machine Learning Algorithms, Spam Detection Using Machine Learning Algorithms

I. INTRODUCTION

SMS is a communication protocol designed for exchange of short, written messages upto 160 characters to mobile devices including cellular and smartphones. SMS messages are transmitted over the cellular networks. Almost 2/3rd of the world's population sends and receives SMS data predicts and 91% of business owners and marketing managers say they see higher conversion rates with integrated marketing campaigns that include SMS. SMS is very popular in the commercial industries for it is an instant method to send data. Out of total SMS, nearly a fair percentage is identified as spam messages. Spam messages are the unwanted messages that contains various information including mobile numbers, malicious links, ID numbers, phishing, or scam websites that can infect the device or steal the data etc. An effective spam detection should consider the following factors into account are

1. Privacy of user
2. Personalization of user
3. Accuracy
4. Self-learning capability
5. Real time filtering support

Machine learning is the area of computational science which focuses on analyzing and interpreting patterns and structures in data to enable learning, reasoning, and decision making outside of human interaction. Machine learning allows the user to feed a computer algorithm an immense amount of data and have the computer analyze and make data-driven recommendations and decisions based on only the input data. There are many spam detection techniques used in machine learning. Few methods are

- Logistic Regression: It is for the binary classification problems. This method chooses parameters that maximises the likelihood of observing the sample values.
- Random Forest method: It works on the basis that each branch of a tree makes a unique prediction that are different from the other tress. This difference gives a better generalization for the given problem.

- Naive Bayes method: It is a simplest classification method. In this method every single word is considered not only mutually independent but also single.
- Support vector machine (SVM): SVM gives a more optimal solution and best classifies the learning data.

The unsupervised machine learning algorithms are more preferably used as it can understand and resolve complex environmental issues. Some of their applications are grouping of specific datasets, reducing the dimensionality, virus detection, customer preference detection etc.

II. LITERATURE SURVEY

The spam messages are those which normally increases the message stream. Presently the user themselves segregate the ham and the spam messages by themselves manually. This becomes really tedious if a particular message has to be found. The emergence of machine learning algorithm made this segregation a simple one. In this regard the paper by P. Sethi, V. Bhandari and B. Kohli, on "SMS spam detection and comparison of various machine learning algorithms," [2] is comparing the various machine learning algorithms and its performances. Naive Bayes model performance is comparatively better is the conclusion of this work.

Paper by M. Healy, A. Zamolotskikh and S. J. Delany, on "An Assessment of Case Base Reasoning for Short Text Message Classification," [3] compares three methods, namely naive bayes, K-nearest neighbour and support vector machine and made a conclusion that the Naïve bayes and SVM is superior with its results.

Many researchers came out with various comparisons of ML algorithms.

Paper by M. E. Boujnouni, on "SMS Spam Filtering Using N - gram method, Information Gain Metric and an Improved Version of SVDD Classifier," Journal of engineering science and technology review [4], have proposed an SMS spam filter based on three components: N-gram to extract features from text messages, enhanced version of SVM and information gain ration to specify the most relevant features. This experiment result has shown that choosing an optimal value for suggested filter parameters can make it reach an accuracy of 95.13% in the training set & 89.32% in the testing set.

Q. Xu et al. [5] have shown concern that most ML filters depend on content-based algorithms which violate user's privacy in certain ways. They proposed a non-content filtering scheme that process on the service side by using graph data mining to recognize suspicious behaviour from ordinary senders.

In E. Ezpeleta et al. [6] paper, they have suggested adding the personality influence measure in the spam filtration process, which proved to improve accuracy up to 98.94%, and the false-positive rates.

In U. Erra, S. Senatore, F. Minnella and G. Caggianese,

[7] they have propose an approximate version of the TF–IDF measure suitable to work on continuous data stream (such as the exchange of messages, tweets and sensor-based log files), also a parallel implementation of the approximate tf–idf calculation using Graphical Processing Units (GPUs).

In Yishan Gong and Qiang Chen, [8], a spam filtering in email, based on the Naive Bayesian algorithm and designed a spam filtering model based on the naive Bayesian algorithm. The paper carried out an experiment on the Ling-spam Corpus with VC 6.0 platform, and got a higher accuracy and recall rate.

In I. Joe and H. Shim, [9] the paper describes a powerful and adaptive spam filtering system for SMS (Short Messaging Service) that uses SVM (Support Vector Machine) and a thesaurus. The system isolates words from sample data using a pre-processing device and integrates meanings of isolated words using a thesaurus, generates features of integrated words through chi-square statistics, and studies these features.

In MA Shafi'I, MS Abd Latiff. [10] paper has shown review of the currently available methods, challenges, and future research directions on spam detection techniques, filtering, and mitigation of mobile SMS spams. The most popular techniques for SMS spam detection, filtering, and mitigation are compared.

III. PROBLEM STATEMENT

As a follow up of the previous literatures discussed, the proposed work focusses on choosing a particular data set which contains the ham and the spam labelled SMS messages. The ham is replaced as 0 and spam as 1. Here spam is considered

IJETIR
ISSN (Online) 2583-0554
International Journal of Emerging Technologies and Innovative Research (IJETIR)
Impact Factor: 5.731
Volume 5, Issue 12, December 2025

as positive class for which tf and tf-id vectorization using count vectorizer and tfidf vectorizer using scikit-learn is performed.

Every supervised model needs a proper dataset to be modelled. In this the dataset is chosen from impartus and they are split as testing and training set. In this 5- fold cross validation with SVC classifier is used.

The dataset used in training and testing the current hybrid system was collected from impartus. It contains instances of SMS messages in English. Each data only includes two features, which is the label (spam or ham) and the content of the text message. Secondly, the data was converted from text form to CSV (Comma Separated Value) form for easier feature extraction. Then preprocessing is done by cleaning the data by removing punctuation marks and removing stop words as they appear very often and cannot help in concluding the final classification. Furthermore, the general data is analysed by implementing various feature extractions techniques. Term frequency inverse document is a feature extraction technique in which is used to evaluate importance of a word to a document [7]. Then the data is separated as training and test sets. Then a classifier is applied based on its accuracy and precision.

In the program flow the required modules are imported first. Then a data frame is generated from the spam database. It is further labelled in such a way that spam is 1 and ham is 0. The method by which an algorithm could be tested is by these following performance metrics Recall, Precision and Accuracy

These metrics are based on the confusion matrix parameters. Precision =True positive/ (True positive + False positive)

Accuracy = (True positive + True negative)/ Total

Recall = True positive/ (True positive + False negative) The steps in designing an algorithm are as follows

- Preparing the dataset
- Preprocessing of data
- Building the classifier
- Building a classified model
- Train model with the labelled training dataset
- Test given model with the trained dataset
- Analyse the final results using machine learning tools.

IV. METHODOLOGY

The process gets initiated by assigning token for each word so that it can be treated separately. All are converted to single case, either upper or lower case for easy comparison. The data has to undergo preprocessing, filling the null data, removing punctuations. Then it is to be noted that which words are common among ham and spam messages. The spam word cloud is given below

Fig. 1. Spam word cloud

IJETIR
ISSN (Online) 2583-0554
Impact Factor: 5.731
International Journal of Emerging Technologies and Innovative Research (IJETIR)
Volume 5, Issue 12, December 2025

Fig. 2. Ham word cloud

Next to proceed is to convert spam as 1 and ham as 0.Then the tf is calculated using count vectorization. tf means the number of times a word appears in a given sentence. tf-idf means Term Frequency - Inverse Document Frequency. This is a statistic that is based on the frequency of a word in the corpus but it also provides a numerical representation of how important a word is for statistical analysis. tf-idf is better than Count Vectorizers because it not only focuses on the frequency of words present in the corpus but also provides the importance of the words. The dataset is first read.Then spam is taken as 1 and Ham as 0.

	label	msg
0	ham	Go until jurong point, crazy.. Available only ...
1	ham	Ok lar... Joking wif u oni...
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...
3	ham	U dun say so early hor... U c already then say...
4	ham	Nah I don't think he goes to usf, he lives aro...
5	spam	FreeMsg Hey there darling it's been 3 week's n...
6	ham	Even my brother is not like to speak with me. ...
7	ham	As per your request 'Melle Melle (Oru Minnamin...
8	spam	WINNER!! As a valued network customer you have...
9	spam	Had your mobile 11 months or more? U R entitle...

Fig. 3. Dataset

	label	msg
0	0	Go until jurong point, crazy.. Available only ...
1	0	Ok lar... Joking wif u oni...
2	1	Free entry in 2 a wkly comp to win FA Cup fina...
3	0	U dun say so early hor... U c already then say...
4	0	Nah I don't think he goes to usf, he lives aro...
5	1	FreeMsg Hey there darling it's been 3 week's n...
6	0	Even my brother is not like to speak with me. ...
7	0	As per your request 'Melle Melle (Oru Minnamin...
8	1	WINNER!! As a valued network customer you have...
9	1	Had your mobile 11 months or more? U R entitle...

Fig. 4. Dataset with spam is taken as 1 and ham as 0

In these type of word problems a techniques called word embedding technique is used. As a part of it the tf, tfidf and count vectorizer methods are used. This converts word into numeric format. The count vectoriser converts words into matrix of words. This is called Document Text Matrix (DTM).This further converts into sparse matrix.

tf-idf stands for term frequency — inverse document frequency, where TF is just the frequency of the term in document term matrix and

IJETIR

ISSN (Online) 2583-0554

Impact Factor: 5.731

International Journal of Emerging Technologies and Innovative Research (IJETIR)

Volume 5, Issue 12, December 2025

IDF = log [(1 + D) / (1 + df(d, t))] + 1,

where D is the number of documents and df (d,t) is number of documents a term t has appeared in the DTM.

Then a function is created to calculate the tf and tf-idf values based on SVM parameters.The kernel values are linear, RBF (Radial Basis Kernel function) and sigmoid. The tf and tfidf vectorizer uses 3 classifiers Linear, RBF, Sigmoid. Each classifiers uses 3 values each C=0.1, C=1 and C=10. Further the Macroaverage precision gives the value for true positive case of the confusion matrix.The Macroaverage recall gives the value for true negative case and the Macroaverage F1 is the weighted average of both precision and recall.Similarly these classifiers are used with tf-idf.

	Vectorizer	Classifier	Macroaverage Precision	Macroaverage Recall	Macroaverage F1
0	tf,	C=0.1,linear	98.76981611216722	93.8659317731335	96.12788008074682
1	tf,	C=1,linear	98.76161085724793	94.25870106988444	96.35234394985591
2	tf,	C=10,linear	98.77160768726868	94.3253677365511	96.39502375635375
3	tf,	C=0.1,rbf	43.29684971540362	50.0	46.40762015762016
4	tf,	C=1,rbf	43.29684971540362	50.0	46.40762015762016
5	tf,	C=10,rbf	96.58787123428837	77.23169547124759	83.52063112014486
6	tf,	C=0.1,sigmoid	43.29684971540362	50.0	46.40762015762016
7	tf,	C=1,sigmoid	43.29684971540362	50.0	46.40762015762016
8	tf,	C=10,sigmoid	93.63387328881406	52.88008948545862	52.021104884266514

Fig. 5. Output for Tf vectoriser for 3 types of classifiers.

	Vectorizer	Classifier	Macroaverage Precision	Macroaverage Recall	Macroaverage F1
0	tf-idf,	C=0.1,linear	96.4781088618292	76.42766862560998	82.78171043053581
1	tf-idf,	C=1,linear	98.5393885800984	93.70106292960557	95.93173499183429
2	tf-idf,	C=10,linear	98.39837995778801	94.20599274379572	96.16453015612143
3	tf-idf,	C=0.1,rbf	43.29684971540362	50.0	46.40762015762016
4	tf-idf,	C=1,rbf	43.29684971540362	50.0	46.40762015762016
5	tf-idf,	C=10,rbf	43.29684971540362	50.0	46.40762015762016
6	tf-idf,	C=0.1,sigmoid	43.29684971540362	50.0	46.40762015762016
7	tf-idf,	C=1,sigmoid	43.29684971540362	50.0	46.40762015762016
8	tf-idf,	C=10,sigmoid	43.29684971540362	50.0	46.40762015762016

Fig. 6. Output for TFIDF vectoriser for 3 types of classifiers.

V. CONCLUSION

A method was proposed to effectively segregate Ham and Spam messages was initiated.from the data obtained the following observations could be made.

From the output using TF vectoriser, it is observed that for the classifier value C=10, Linear,the macroaverage precision, macroaverage recall and the weighted macroaverage F1 is fairly high and delivers maximum efficient percentage of output.

Similarly, From the output using tfidf vectoriser, it is observed that for the classifier value C=10, Linear, the macroaverage precision ,macroaverage recall and the weighted macroaverage F1 is fairly high and delivers maximum efficient percentage of output.

This defined model suits for other set of datasets as well. Further the work can be extended by using various deep learning and machine learning algorithms.

REFERENCES

[1] "statista," 2017. [Online]. Available: https://www.statista.com/statistics/483255/numberof- mobile-messaging-users-worldwide/. [Accessed February 2019].

[2] P. Sethi, V. Bhandari and B. Kohli, "SMS spam detection and comparison of various machine learning algorithms," in 2017 International Conference on Computing and Communication Technologies for Smart Nation (IC3TSN), 2017.

[3] M. Healy, A. Zamolotskikh and S. J. Delany, on "An Assessment of Case Base Reasoning for Short Text Message Classification," in Proceedings of the 15th. Irish Conference on Artificial Intelligence and Cognitive Sciences (AICS'04), Castlebar, 2004

[4] M. E. Boujnouni, on "SMS Spam Filtering Using N - gram method, Information Gain Metric and an Improved Version of SVDD Classifier," JOURNAL OF ENGINEERING SCIENCE AND TECHNOLOGY REVIEW, Vols. 10. 131-137. 10.25103/jestr.101.18. ,
2017.
[5] Q. Xu, E. W. Xiang, Q. Yang, J. Du and J. Zhong, "SMS Spam Detection Using non-content Features," IEEE Intelligent Systems, vol. 27, no. 6, 2012.
[6] E. Ezpeleta, U. Zurutuza and J. M. Gómez, "Short Messages Spam Filtering Using Personality Recognition," in the 4th Spanish Conference, 2016.
[7] U. Erra, S. Senatore, F. Minnella and G. Caggianese, "Approximate TF–IDF based on topic extraction from massive message stream using the GPU," Information Sciences, p. 292:143–161, 2015.
[8] Yishan Gong and Qiang Chen, "Research of spam filtering based on Bayesian algorithm," 2010 International Conference on Computer Application
[9] I. Joe and H. Shim, ‘‘An SMS spam filtering system using support vector machine,’’ in Future Generation Information Technology. Berlin, Germany: Springer, 2010, pp. 577–584.
[10] MA Shafi'I, MS Abd Latiff, H Chiroma, O Osho, G Abdul-Salaam, AI Abubakar, T Herawan, “A review on mobile SMS spam filtering techniques” IEEE Access, 2017, ieeexplore.ieee.org
[11] DENG, Wei-Wei., PENG, Hong. Research on a Naive Bayesian Based Short Message Filtering System. Machine Learning and Cybernetics, 2006 International Conference on, p. 1233-1237, Aug. 2006