ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 8, Issue 7, July 2021

Content-based SMS Spam Messages classification using Natural Language Processing and Machine Learning

[1] S. Sumahasan, [2]Uday Kumar Addanki, [3]Anjani Kintali, [4] Srilekha Bontu
[1] [2] Assistant Professor, [3][4] Students of B.Tech-CSE G.V.P.C.E.W, Visakhapatnam, India.

Abstract: Spam is any unwanted digital communication that is sent in bulk from compromised machines. In this work, the suggested strategy is a model that uses the Bag of Words technique to calculate the frequency of words and Supervised Machine Learning techniques such as Naïve Bayes and Support Vector Machine to categorise the message. The suggested system shows the message's categorization as well as the most prevalent spam terms discovered in the message. We compare the performance of the Naïve Bayes and Support Vector Machine algorithms in this study. The feature of adding additional spam messages to the collection improves accuracy as well.

Keywords: Support Vector Machine, machine learning, naïve bays, spam messages.

1. INTRODUCTION

Spam messages are sent to a large number of people at once [1]. SMS spam represents extra difficulties as SMS texts are restricted to 160 characters which are very little content to perform text classification and distinguish whether a message is a spam or ham [2]. According to Wikipedia, "the use of electronic messaging networks to send unwanted mass communications, particularly a mass advertisement, malicious links, etc." is called spam [3]. People tend to believe the content of the message if the message is not classified and would greatly affect their vital information. The proposed approach's frequency was determined using a bag of words, a Natural Processing technique, and categorizes messages using machine-learning-based algorithms live Support Vector Machine and Naïve Bayes.

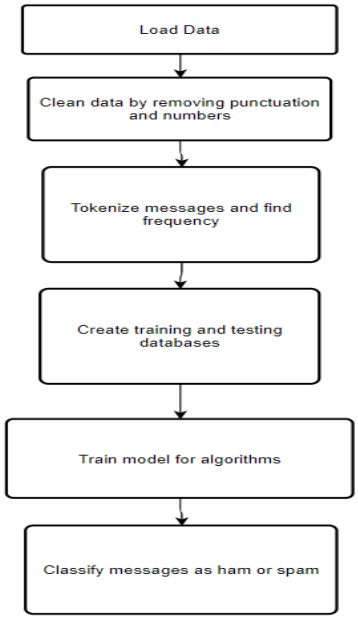

Fig 1: *Model Diagram*

2. LITERATURE SURVEY

The paper "SMS Spam Filtering Using Supervised Machine Learning Algorithms," released in 2018, outlines the methods for classifying spam messages using supervised machine learning algorithms such as max entropy and SVM algorithms. The paper describes the steps of classification which included pre-processing, feature extraction, training, and classification. and

ISSN (Online) 2394-6849

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 8, Issue 7, July 2021

performance has been evaluated. This paper has only been confined to one classification algorithm [1].

The paper "Email Spam Classification by Support Vector Machine" that was released in 2018 summarizes how to categorize junk mail using the Support Vector Machine algorithm. The paper describes the steps of classification which included pre-processing, feature extraction, training, and classification. Various types of kernels and performance have been evaluated. This paper has only been confined to one classification algorithm [2].

Published in 2020, "Email Spam Detection Using Mail Learning Algorithms" evaluates various machine learning algorithms for paper spam emails, such as Naïve Bayes, Support Vector Machine, Decision Tree, Neighborhood Neighbor, and Random Forest Classification. According to this paper, the Naïve Bayes algorithm performed well. The downside to this paper is that the application has not been tested on different data sets [3].

Identifies and classifies SMS spam using machine learning algorithms such as "A Fog-augmented Machine Learning-Based SMS Spam Detection and Classification System" paper, published in 2020. Algorithms are tested for different datasets and evaluate performance for each dataset. The paper summarizes that the performance has been best by the different algorithms for different datasets [4].

The paper "Content-Based Spam Detection in Email using Bayesian Classifier" published in 2015 summarized how to use the Bayesian classifier algorithm to classify spam. The paper describes the steps of classification which included pre-processing, feature extraction, training, and classification. and performance has been evaluated and also describes how the emails are classified based on the content of the email. This paper has only been confined to one classification algorithm [5].

The paper "A Machine Learning based Spam Detection Mechanism" was published in 2020. The paper describes email spam detection using the Naïve Bayes algorithm including pre-processing, URL checking, keyword checking, and tokenizing. This paper is confined to only one classification algorithm[8].

The paper "An Overview of Bag of Words; Importance, Implementation, Applications, and Challenges" which was published in 2019 summarizes the Bag of Words (BoW) or Bag of Features (BoF), its importance, implementation in classification, and the challenges. The advantages of this paper are the clear explanation of the BoW technique and its limitations as well [10].

3. PROPOSED SYSTEM

The proposed system is designed to identify the type of message by the content of the message. For tokenizing the message, the NLP technique called Bag of Words is employed, and to categorize Supervised Machine Learning algorithms are used.

After the category is determined, if the message is new, then the message is appended to the dataset to increase the accuracy of the system. The system also displays the common spam words that occur in the message for the user to get aware of the difference between spam and ham messages.

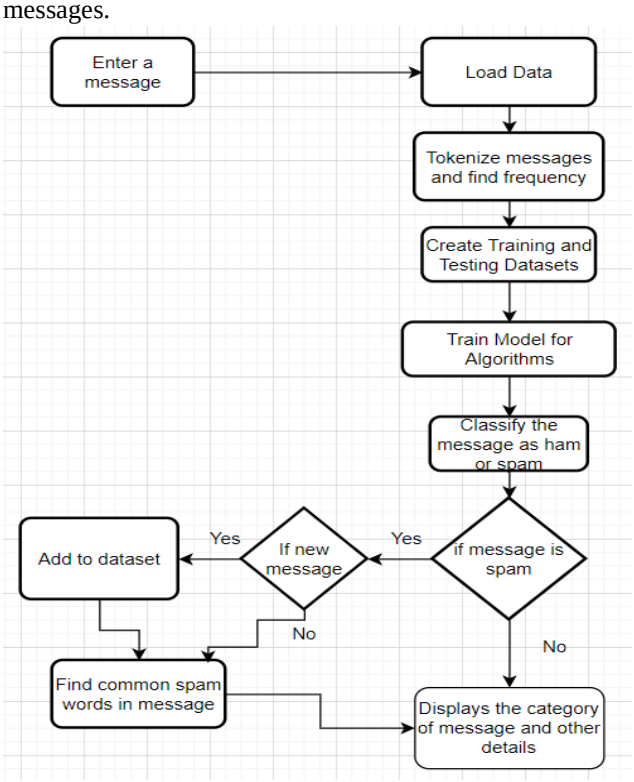

Fig 2: *Architecture*

A. Naïve Bayes Algorithm

Naïve Bayes algorithm is a supervised machine learning algorithm. This algorithm is based on the Bayes rule. The Bayes rule states that

$P(Y|X) = [P(X|Y) * P(Y)]/P(X)$

where P(Y|X) is a posterior probability, P(X|Y) is a likelihood, P(Y) is prior probability and P(X) is the probability of data that is independent of Y and can be ignored.

The algorithm is based on conditional probabilities and finds out the probability for the word to be spam as well as ham. The highest probability is considered and it belongs to the respective class. The probabilities for

ISSN (Online) 2394-6849

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 8, Issue 7, July 2021

words are calculated in the following manner:
*P(spam/word) = [P(word/spam) * P(spam)]/P(word)*
*P(word) = P(word/spam)*P(spam) + P(word/ham) * P(ham)*
Often in spam but not in ham, then that email is spam [3].

B. Support Vector Machine (SVM)

Support vector machine is one of the supervised machine learning algorithms. It is used for classification tasks. It classifies data into categories using hyperplanes. The hyperplane handles data with many features. In this algorithm, the margin, which is the interval between hyperplanes, is used to determine whether the points are classified correctly.

In the spam classification, the two classes that SVM should distinguish are spam and ham. Characteristics Different message words converted into frequency matrix.

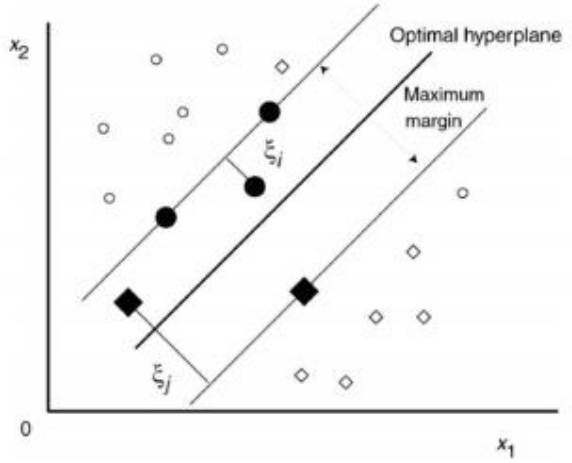

Fig 3: *SVM classifier*

C. The Dataset

In machine learning, datasets are important and should be carefully selected for appropriate training and testing of machine learning models. Unfortunately, important public SMS datasets are not available, which makes filtering spam SMS very difficult. SMS services lack private datasets and cannot share their customers' data for research purposes as they are privately operated. The dataset used in this study is a set of approximately 5574 SMS texts from Cogley, provided by the UCI Machine Learning Repository.

The dataset has been divided into two datasets, a training dataset of 4460 records and a testing dataset of 1114 records [1].

.

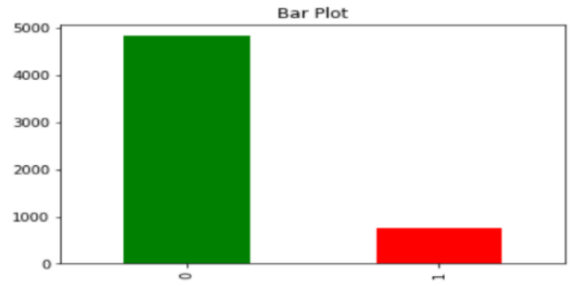

Fig 4: *dataset visualization*

In fig 4, the x-axis represents the type of message i.e, spam or ham, and the y-axis represents the count of messages.

D. Pre-processing and Feature Extraction

Data Pre-processing is done on the data frame by labeling ham as 0 and spam as 1. The problems with text are that it will have redundant words, punctuations, and are the string of words and machine learning algorithms prefer structured data and by using the Bag-of-Words technique, we will convert variable-length texts into a fixed-length vector. Machine learning algorithms work with numeric data rather than text data. Using bag-of-words technology, we convert a text into its equivalent vector numbers. Say we have 4 documents as follows:

['Hello, how are you!', 'Win money, win from home.', 'Call me now.', 'Hello, do you want to call tomorrow?']

Our goal is to convert this set of text to a matrix constituting the count of every unique word: Here, the rows are the indexes to sentences and each word is columns are the unique words in the sentences sorted alphabetically. with the corresponding value.

	are	call	from	hello	home	how	me	money	now	tomorrow	win	you
0	1	0	0	1	0	1	0	0	0	0	0	1
1	0	0	1	0	1	0	0	1	0	0	2	0
2	0	1	0	0	0	0	1	0	1	0	0	0
3	0	1	0	1	0	0	0	0	0	1	0	1

Fig 5: *Frequency Matrix*

The Bag of Words creates the vocabulary of all the specific words in all the documents in the training dataset [3].

E. Classification Algorithms

The classification algorithms classify the category of the message as spam or ham (not spam).

In this article, several naive Bayes algorithms are used to

ISSN (Online) 2394-6849

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 8, Issue 7, July 2021

classify messages, because several naive Bayes classifiers are suitable for classification with discrete characteristics. Multinomial nav bases consider a feature vector, where the given word appears many times or most often i.e. frequency.

The other famous Machine Learning classifier used to classify the category of the message is the Support Vector Machine. Messages are categorized using a set of data in the feature space on the polynomials of the original variables used in the kernel.

F. Updation of the dataset

The number of ham messages is less than spam messages in the considered dataset. The imbalance data can be handled using techniques like Upsampling or Downsampling. But this paper describes a different approach. The approach is if the user's message is a spam message and a new message, the message along with its category is appended to the dataset. This approach can not only handle the imbalance but also enhances the performance. The updating is as follows:

Before Appending:

4452	spam	Urgent UR awarded a complimentary trip to EuroDisinc Trav, Aco&Entry41 Or Â£1000.
4453	ham	Awesome, be there in a minute
4454	ham	And that is the problem. You walk around in "julianaland" oblivious to what is going
4455	ham	I've told you everything will stop. Just dont let her get dehydrated.
4456	ham	Or I guess <#> min
4457	ham	I'm home. Ard wat time will u reach?
4458	ham	Storming msg: Wen u lift d phne, u say "HELLO" Do u knw wt is d real meaning of HELI
4459	ham	If you want to mapquest it or something look up "usf dogwood drive", that's the tiny
4460	ham	Aight should I just plan to come up later tonight?
4461	ham	Die... I accidentally deleted e msg i suppose 2 put in e sim archive. Haiz... I so sad...

Fig 6: *Dataset before appending new message*

After appending:

4458	ham	Storming msg: Wen u lift d phne, u say "HELLO" Do u knw wt is d real meaning of HELLO??
4459	ham	If you want to mapquest it or something look up "usf dogwood drive", that's the tiny stre
4460	ham	Aight should I just plan to come up later tonight?
4461	ham	Die... I accidentally deleted e msg i suppose 2 put in e sim archive. Haiz... I so sad...
4462	spam	URGENT Your grandson was arrested last night in mexico.Need bail money immediately V

Fig 7: *Dataset after appending new message*

G. Analysis between Classification Algorithms

The algorithms are evaluated based on two aspects in this paper. The first aspect is performance. It is depicted in tabular form as follows:

Table-1 PERFORMANCE OF ALGORITHMS

Algorithm	Performance Measures			
	Accuracy	Precision	Recall	F1
Multinomial Navie Bayes	0.9847	0.9383	0.9448	0.9415
Support Vector Machine	0.9513	0.9595	0.6551	0.7786

The second aspect is the time taken for execution. The time taken by the Naïve Bayes algorithm is to classify messages is around 0.106 seconds whereas the time taken by the support vector machine algorithm to classify the message is 2.756 seconds approximately.

Therefore, the paper concludes that the Naïve Bayes algorithm is a better classifier as it performs well and takes very little time to determine the category of the message.

H. Common Spam Words

The proposed system identifies the common spam words that occur in a new message by considering the training dataset and determining the list of common spam words. Spam messages are treated, converted to lowercase, punctuation and finally stop words. Stop words are English words that do not give much meaning to a sentence. They can be safely ignored. The word in the user message can then be compared and identified with a list of common spam words [3].

The common spam words can be visualized as a word cloud. Word cloud is a method of externally describing the repetition of words in information. The cloud contains words that are randomly scattered around the toy. Words that appear more frequently in content appear in larger text styles, while less common words appear in lower text styles. This type of number has become popular in recent years as it gives a way to examine trending behaviors on social networking sites [1].

Fig 8: *Word Cloud*

ISSN (Online) 2394-6849

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 8, Issue 7, July 2021

4. Experimental Results

We used HTML and CSS to implement the proposed system and created a web application that categorizes user-entered messages and displays common spam words to the user. If the entered message is new and if it is spam, the system displays a message to a user called "New Message" and appends the message to the dataset as shown in the following figures:

Algorithm	Performance Measures			
	Accuracy	Precision	Recall	F1
Multinomial Naïve Bayes	0.9847	0.9383	0.9448	0.9415
Support Vector Machine	0.9513	0.9595	0.6551	0.7786

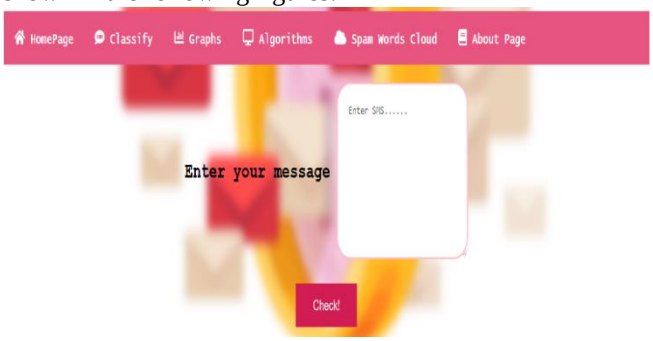

Fig 9: *Entering a message*

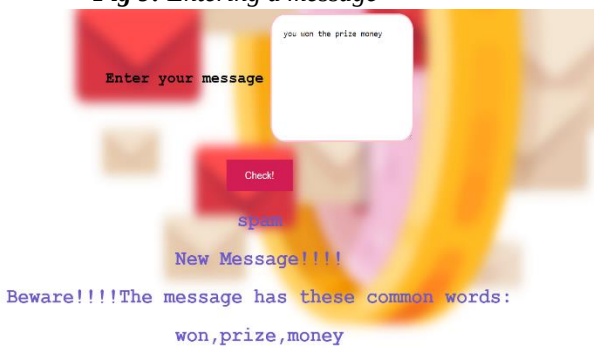

Fig 10: *displaying the category of message*

The common spam words in the message are displayed to the user in the following way, which makes the user aware of the difference between ham and spam messages:

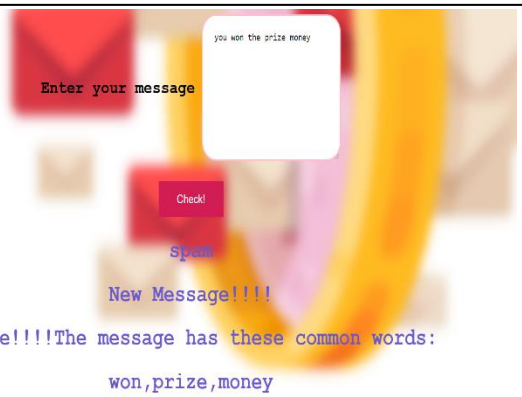

Fig 11: *displaying common spam words*

The performance of both the algorithms is as follows:

TABLE I
PERFORMANCE OF ALGORITHMS

As the new spam messages append to the training dataset, the performance is as follows:

TABLE II
PERFORMANCE OF ALGORITHM AS MESSAGES APPEND

	Performance Measures			
	Accuracy	Precision	Recall	F1
Before appending	0.9941	0.9864	0.9700	0.9782
After appending	0.9944	0.9885	0.9709	0.9796

This functionality of appending new spam messages enhanced the performance of the training dataset

5. CONCLUSION AND FUTURE SCOPE

The spam categorization technique is used in this model to assist manage incoming communications and avoid them from being inundated with non-essential emails/messages. Spam filters can also add extra layer protection. Given a message the proposed method predicts whether the message is spam or not, purely based on the content of the message. The proposed method is a web application that determines the category of messages, common spam words. The proposed system also enhances the model by appending new spam messages along with categories.

In this work, we compared Naïve Bayes and Support Vector Machine algorithms for SMS spam classification

problem. The two models have been proposed, trained, and tested using popular and often used standard databases.

The Empirical results of the simulation showed that the proposed scheme base on Naïve Bayes outperformed the Support Vector Machine in terms of precision and operating speed.

REFERENCES

[1] Pavas Navaney, Gaurav Dubey, Ajay Rana. "SMS Spam Filtering Using Supervised Machine Learning Algorithms", 2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2018.

[2] M. Singh, R. Pamula and S. k. Shekhar, "Email Spam Classification by Support Vector Machine," 2018 International Conference on Computing, Power and Communication Technologies (GUCON), 2018, pp. 878-882.

[3] N. Kumar, S. Sonowal and Nishant, "Email Spam Detection Using Machine Learning Algorithms," 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), 2020, pp. 108-113.

[4] S. Bosaeed, I. Katib, and R. Mehmood, "A Fog-Augmented Machine Learning based SMS Spam Detection and Classification System," 2020 Fifth International Conference on Fog and Mobile Edge Computing (FMEC), 2020, pp. 325-330.

[5] S. B. Rathod and T. M. Pattewar, "Content-based spam detection in email using Bayesian classifier," 2015 International Conference on Communications and Signal Processing (ICCSP), 2015, pp. 1257-1261.

[6] W. Etaiwi and A. Awajan, "The Effects of Features Selection Methods on Spam Review Detection Performance," 2017 International Conference on New Trends in Computing Sciences (ICTCS), 2017, pp. 116-120.

[7] P. Sethi, V. Bhandari and B. Kohli, "SMS spam detection and comparison of various machine learning algorithms," 2017 International Conference on Computing and Communication Technologies for Smart Nation (IC3TSN), 2017, pp. 28-31

[8] A. Junnarkar, S. Adhikari, J. Fagania, P. Chimurkar, and D. Karia, "E-Mail Spam Classification via Machine Learning and Natural Language Processing," 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), 2021, pp. 693-699.

[9] K. Jayamalini, M. Ponnavaikko, J. Kothandan. "A COMPARATIVE ANALYSIS OF VARIOUS MACHINE LEARNING BASED SOCIAL MEDIA SENTIMENT ANALYSIS AND OPINION MINING

[10] APPROACHES", Advances in Mathematics: Scientific Journal, 2020.

[11] W. A. Qader, M. M. Ameen and B. I. Ahmed, "An Overview of Bag of Words; Importance, Implementation, Applications, and Challenges," 2019 International Engineering Conference (IEC), 2019, pp. 200-204.