IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A Comprehensive Framework For Spam Sms Classifier Model Using Natural Language Processing And Machine Learning

[1]Mr. Adurti V S Praveen Kumar[1], [2]Dr. G. Sharmila Sujatha[2]

[1]PG Student, Dept. of Computer Science & Systems Engineering, Andhra University College of Engineering, Andhra University

[2]Assistant Professor, Dept. of Computer Science & Systems Engineering, Andhra University College of Engineering, Andhra University

Abstract: Rapid development of technologies and the widespread use of mobile phones have resulted in various risks, such as spam and phishing attacks. SMS (short message service) is a text message service available in smartphones as well as keypad phones. So, the traffic of SMS increased drastically. Many fraudulent URLs and short links are sent through SMS, and we don't know if they are safe or not. So, spam classification has special attention and it is very important to classify HAM and SPAM for the benefit of every individual in the society. In this paper, we used different machine learning techniques and deep learning technique for SMS spam detection/classification. With the help of machine learning techniques such as Naive Bayes, logistic regression, and other classifiers, the spam detection model has been developed in this project. We used a dataset to train the machine learning and NLP method and Naive Bayes, Decision Tree and Voting Classifier.

Index Terms – SMS, HAM, SPAM, CLASSIFICATION, METHODOLOGY, ALGORITHM, DATASET

I. INTRODUCTION

The communicational technology is rapidly glowing. This means everybody can access or receive the information easier than the past using web browsing, text messaging and emailing. The easiest way to achieve this is to send short messages related to their intention. Individual's phone number is not only known to the people whom they wish to give their number. In terms of customer feedback, registrations, shopping, etc., a common man's mobile number is given to various sorts of organizations. Some companies try to promote their brand by sending SMS (Small Messaging Services) to the mobile numbers which they have obtained from the above listed ways. On the other hand, there are some unethical companies who buy or sell these numbers for illegal offences. These companies try to manipulate the common man's interest by sending spam messages like lot selection, cash award, lottery, fake bank messages, etc., criminal cases are recorded as cyber theft because of loss of money. A common man cheated because of the greed, lack of awareness of these kinds of theft and interest towards their luck. Hence, detection of spam SMS is very important to safe to avoid these kinds of cyber thefts. It is very important to classify which message is HAM and which message is SPAM. We used various machine learning and deep learning techniques for SMS spam detection. Machine Learning is a technology, where machines learn from previous data and made a prediction on future data. This research work focuses to detect spam SMS with the contents using machine learning and implemented in python.

1.1 Existing System

Over the years, many reviews and implementations have been done on SMS spam filtering with various algorithms and techniques. Some of the algorithms could be time-consuming, and some techniques require more lines of code. Some algorithms like K-Nearest Neighbors, which don't process the images as much, are used. The pre-processing techniques need to be updated, and therecould be an appropriate user interface to make users make use of it. Hence the NB is trained on features extracted; the training set error and test set error are close to each other. Therefore, we do not have a problem of high variance, and gathering more data may not result in much improvement in the performance of the learning algorithm. As the result, we should try reducing bias to improve this classifier. This means adding more meaningful features to the list of tokens can decrease the error rate, and is the option that is explored next.

1.1 Proposed System

In the proposed system, we will use machine learning algorithms that have higher accuracy by comparing them. We need an accurate algorithm to preprocess given text and check all the possibilities for which category it belongs, whether spam or not. So, we first get to see all classification algorithms applied to the dataset. The accuracy and precision were calculated. Rather than do this first, we tokenize the text and remove stop words and a bag of words so we can focuson the main theme of the text and check which words match with spam and which words matchwith not-spam.

Applying Naive Bayes algorithm to the dataset using extracted features with different training set sizes. The performance in learning curve is evaluated by splitting the dataset into 70% training set and 30% test set. The Naive Bayes, decision tree and voting classifier algorithm shows good overall accuracy.

Additionally, going through the misclassified samples, we notice that text messages with length below a certain threshold are usually hams, yet because of the tokens corresponding to the alphabetic words or numeric strings in the message they might be classified as spams.

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.

II. MATERIALS AND METHODS

2.1 Python Libraries

There are several libraries used in the data visualization of this operation.

NumPy is a Python library used for working with arrays. NumPy aims to provide an array object that is up to 50x faster than traditional Python lists.

Pandas is a Python library used for working with data sets. It has functions for analyzing, cleaning, exploring, and manipulating data. The name "Pandas" has a reference to both "Panel Data", and "Python Data Analysis" and was created by Wes McKinney in 2008.

Matplot library is used for creating static and amped visualizations.

Seaborn is data visualization with the Matplot library.

2.2 Data Set

The data set is needed and can be collected. The dataset consists of all spam and non-spam messages. Here we are taking over 5000 messages. The data is taken in the form of rows and columns in a table and saved in the form of a CSV file. It can contain text message examples and indicate whether it is spam or not. We categorize and pre-process them.

Table 2. Dataset preview

ID	Label	Message
0	ham	Go until Jurong point, crazy. Available only in bugs n great world la e buffet... Cine there got amore wat...
1	ham	Ok lar... Joking wif u one...
2	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
3	ham	U dun say so early hor... U c already then say...
4	ham	Nah I don't think he goes to us', he lives around here though
		.
		.
		.
5567	spam	This is the 2nd time we have tried 2 contact u. U have won the £750 Pound prize. 2 claim is easy, call 087187272008 NOW1! Only 10p per minute. BT-national-rate.
5568	ham	Will ü b going to esplanade fr home?
5569	ham	Pity, * was in mood for that. So. any other suggestions?
5570	ham	The guy did some bitching but I acted like i'd be interested in buying something else next week and he gave it to us for free
5571	ham	Rofl. Its true to its name

2.3 Models

There are various machine learning models you can use to build a spam SMS detector. Here's a simplified outline of the process:

Data Collection: Gather a dataset of SMS messages labeled as either spam or not spam (ham).

Data Preprocessing: Clean and preprocess the text data, which may involve tasks like tokenization, lowercasing, removing punctuation, Stop word removal, Bag of Words (BoW) vectorization

Feature Extraction: Convert text data into numerical features. Common techniques include TF-IDF (Term Frequency-Inverse Document Frequency) and word embeddings like Word2Vec or GloVe.

$$TF(t,d) = \frac{Number\ of\ times\ term\ t\ appears\ in\ document\ d}{Total\ number\ of\ terms\ in\ document\ d}$$

$$IDF(t,D) = log\left(\frac{Total\ number\ of\ documents\ in\ d}{Number\ of\ documents\ with\ term\ t\ in\ d}\right)$$

2.3 Model Selection

Naive Bayes: The Multinomial Naive Bayes algorithm, stemming from the family of Naive Bayes classifiers, is particularly suited for classification tasks with discrete features, such as text data represented as word vectors.

Support Vector Machines (SVM): Can be used with various kernel functions.

Random Forest: An ensemble method that can handle a wide range of features.

Neural Networks: You can use deep learning models like Recurrent Neural Networks (RNNs) or Convolutional Neural Networks (CNNs) for more complex patterns.

Model Training: Split your dataset into training and testing sets. Train your chosen model(s) on the training data.

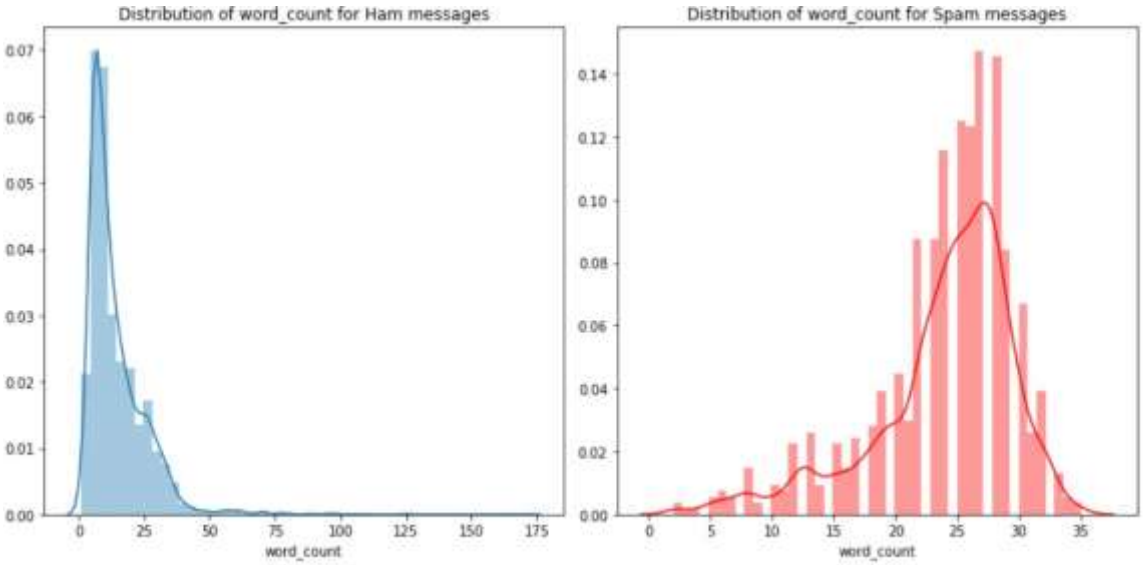
Model Evaluation: Evaluate your model's performance using metrics like accuracy, precision, recall, F1-score, and ROC-AUC on the testing data.

Deployment: Once satisfied with the performance, deploy your model to detect spam SMS messages in real-time.

Continuous Improvement: Monitor the model's performance and update it as needed to adapt to changing spam patterns..

2.3 Model Evaluation

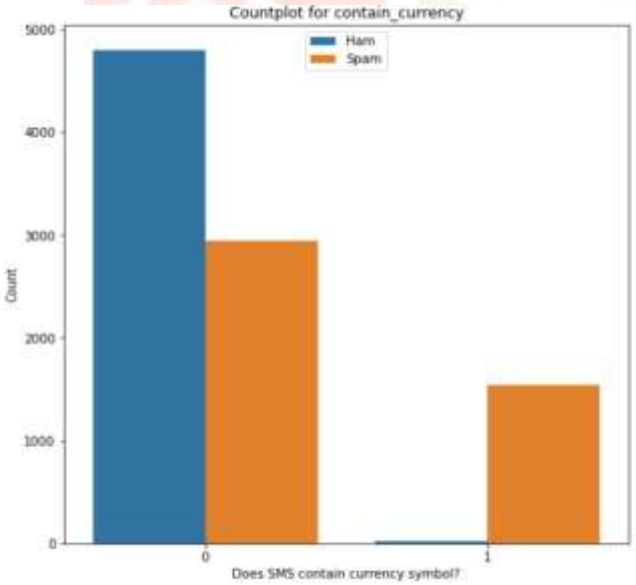
2.3.1 Word Count



***Insight: Spam messages** word_count fall in the range of **15-30 words**, whereas majority of the **Ham messages** fall in the range of **below 25 words**.*

	label	message	word_count	contains_currency_symbol
5537	1	Want explicit SEX in 30 secs? Ring 02073162414...	16	0
5540	1	ASKED 3MOBILE IF 0870 CHATLINES INCLU IN FREE ...	33	1
5547	1	Had your contract mobile 11 Mnths? Latest Moto...	28	0
5566	1	REMINDER FROM O2: To get 2.50 pounds free call...	28	0
5567	1	This is the 2nd time we have tried 2 contact u...	30	1

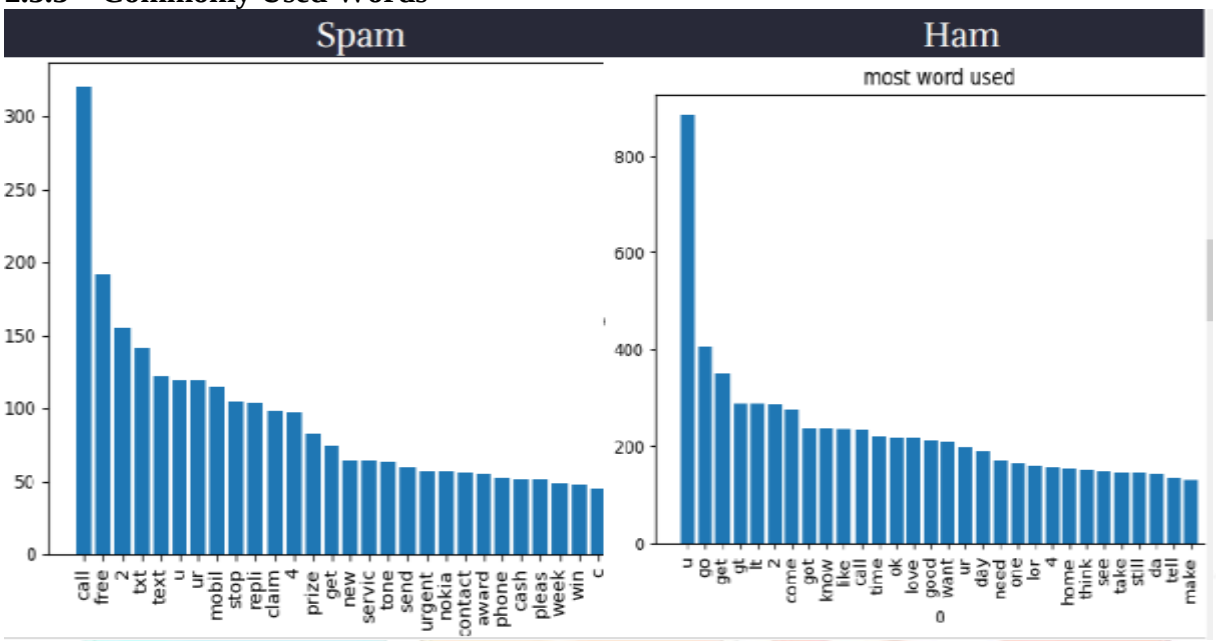
2.3.2 Currency Count plot



Insight: Almost 1/3 of Spam messages contain currency symbols**, and currency symbols are **rarely used in Ham messages.

	label	message	word_count	contains_currency_symbol	contains_number
0	0	Go until jurong point, crazy.. Available only ...	20	0	0
1	0	Ok lar... Joking wif u oni...	6	0	0
2	1	Free entry in 2 a wkly comp to win FA Cup fina...	28	0	1
3	0	U dun say so early hor... U c already then say...	11	0	0
4	0	Nah I don't think he goes to usf, he lives aro...	13	0	0

2.3.3 Commonly Used Words

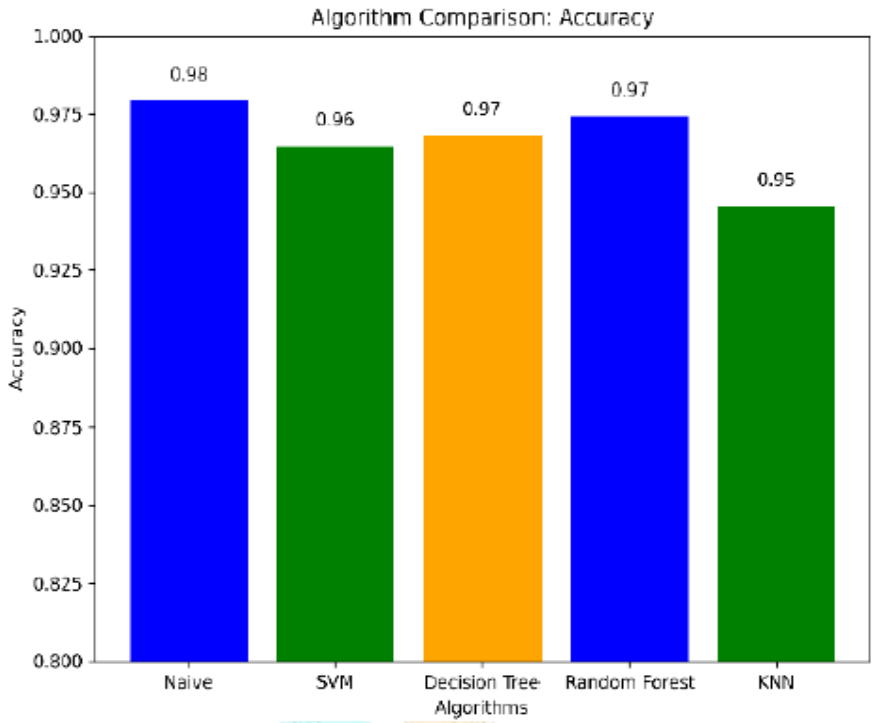


III. RESULTS

We have compared our dataset with other algorithms, namely SVM, Decision Tree, Random Forest, and KNN, and observed that Naïve Bayes performs exceptionally well, as shown in Figure 4. The Precision, Recall, and F1-Score of Naïve Bayes for Ham SMS were 98%, 99%, and 99%, respectively (Table 3). Similarly, for Spam SMS, the Precision, Recall, F1-Score, and Accuracy of Naïve Bayes were 99%, 85%, and 92%. The accuracy of the algorithm was found to be 98%, which was the highest among all.

Table 3. Performance comparison of ham-spam SMS classification

Algorithms	Precision		Recall		F1-Score		Accuracy
	Ham	Spam	Ham	Spam	Ham	Spam	
SVM	0.96	0.99	0.99	0.70	0.98	0.82	0.97
Decision Tress	0.98	0.91	0.99	0.81	0.98	0.86	0.96
Random Forest	0.97	0.99	0.99	0.79	0.99	0.88	0.97
KNN	0.94	0.99	0.99	0.54	0.97	0.70	0.95
Naïve Bayes	0.98	0.99	0.99	0.85	0.99	0.92	0.98



IV. CONCLUSION

The rise in spam messages transmitted via SMS has emerged as a prominent issue demanding immediate attention. To face this problem, a proposed system for automated SMS classification and spam analysis aims to offer a robust solution. Leveraging the power of ML techniques, this system effectively categorizes SMS messages as either spam or non-spam by analyzing extracted features inside the messages. With its accurate classification capabilities, the system helps combat the pervasive issue of spam, enhancing the overall SMS experience for users. The rise in spam messages transmitted via SMS has emerged as a prominent issue demanding immediate attention. To face this problem, a proposed system for automated SMS classification and spam analysis aims to offer a robust solution. Leveraging the power of ML techniques, this system effectively categorizes SMS messages as either spam or non-spam by analyzing extracted features from the messages. With its accurate classification capabilities, the system helps combat the pervasive issue of spam, enhancing the overall SMS experience for users.

ACKNOWLEDGEMENT

I would like to thank Dr. G. SHARMILA SUJATHA, Assistant Professor, Dept. of Computer Science & Systems Engineering, Andhra University for extending her help, support and guidance during this work.

V. REFERNCES

1. Liu, X., Lu, H., Nayak, A. (2021). A spam transformer model for SMS spam detection. IEEE Access, 9: 80253-80263. https://doi.org/10.1109/ACCESS.2021.3081479
2. 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.
3. Alzahrani, A., Rawat, D.B. (2019). Comparative study of machine learning algorithms for SMS spam detection. In 2019 SoutheastCon, untsville, AL, USA, pp. 1-6. https://doi.org/10.1109/SoutheastCon42311.2019.9020530
4. Navaney, P., Dubey, G., &Rana, A. (2018). "SMS Spam Filtering Using Supervised Machine Learning Algorithms." 2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence).
5. 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