International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

SMS Spam Detection using Tokenization and Feature Engineering

Akshay Divakar, Sitaraa Krishnakumar

Abstract— The enormous development of innovation and mobiles, the clients have been exposed to more spam messages than any other time in recent memory ever. SMS spam separating is a nearly an exceptionally ongoing answer for arrangement with such a significant issue.. This paper moves us to chip away at the assignment of separating versatile spam messages as whether it is Ham or Spam for the clients by adding messages to the worldwide accessible SMS dataset. The paper plans to break down various AI classifiers on huge corpus of SMS messages for the individuals around the globe. This paper also informs or tells the readers about the existing algorithms and it's inefficiency in filtering the ham messages from spam messages. This paper makes use of tokenization to create tokens which are then fed into the feature engineering model to extract features and then to predict the outcome.

Keywords— Mobile Phone Spam; SMS Spam; Spam

I. INTRODUCTION

The SMS, which is generally named to as "Short Message Service" is an administration for shipping messages of short length around 120 characters to various gadgets, for example, cell phones, PDAs and PDAs utilizing primary institutionalized correspondences conventions. In Today's estimation, it means that billions of SMS's are sent every day and all the time. As there is a huge rise in the number of spam messages being sent to the users from different SMS servers, there is a huge demand to solve these issues for the users and also to make filtering a more easier, efficient and pronounced process. In Today's scenario, there has been a huge rise in the number of mobile phone users and thus hackers and spammers have found a way to bombard users with spam messages and they are paid for doing the same. There has also been an immense rise in the number of applications in an android or an ios device, which makes room for these companies to send messages to the users and thus resulting in a lot of spam getting collected in their inbox. By and large the exertion and cost of sending a spam message is normally significantly more than the expense of sending spam emails or spam messages. In spite of the fact that this practice is normally restricted in western America or western nations it is still generally winning in numerous pieces of Europe, where a client sends a normal of 20 messages every day and practically the majority of the populace are owning a cell phone, truth be told, in numerous nations like China, Russia or South America, genuine clients are being Emulated by botnets and PC bot frameworks when communicating something specific through a free informing administration . Thus, by and large we can securely expect that there has been a radical abatement in the expense and adequacy of SMS spam. That is, SMS spam has turned into a productive business with great rate of profitability There has additionally been a consistent exchange on this fundamental point, and individuals have thought of positive specialized measures and calculations so as to handle this issue. The greater part of these measures and practices are inadequate in anticipating spam and ham messages. In SMS spam messages, the substance based separating for SMS spam messages is trying because of their short length size and furthermore combined with the nearness of huge number of provincial words and constrained header data when contrasted with emails and messages.

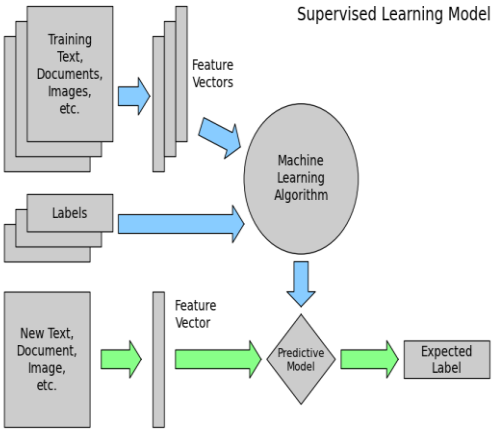

Fig 1 Supervised Learning Model

II. FEATURE ENGINEERING AND EXTRACTION METHODS

In this Natural Language enlivened methodology, we chose to sort the calculation into essential two of the most significant highlights which is the length of the message and the count vectorizer network. We have utilized the length of the message as ascribe quality measurement to decide the quality. During our investigation, which is exploratory in nature we made sense of that messages with spam have distinctive message lengths when contrasted with other ham messages.. Feature engineering is basically selecting feature words or features which is essentially a group of individual words or grouped words taken specially from each message from the given dataset which are mostly maximum occurring words in the dataset.

Manuscript published on 30 September 2019
Akshay Divakar*, Student Institute : SRM Institute of Science and Technology Chennai Tamil Nadu India
Sitara Krishnakumar, Student Institute : SRM Institute of Science and Technology Chennai Tamil Nadu

Feature engineering is essential for the text classification model to understand the problem better and for it to provide an efficient solution, it has to understand the different features existing in each paragraph of the message.

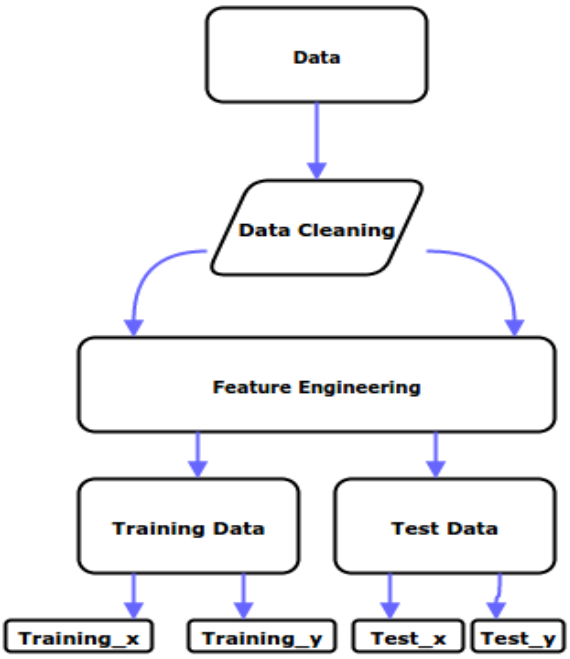

Fig 2 Flowchart of Feature Engineering

III. ALGORITHMS USED

A. Naive Bayes Classifier (NBC)

Naive Bayes is one of the best and most competent in terms of predicting neighboring and accurate values and it is helpful in inductive learning method for machine learning and natural language processing . When the process of classification is executed, it accuracy and its reasonable presentation is amazing. The purpose for better sensible introduction is the better restrictive autonomy and reliance theory on which it depends on only occasionally exists in physical and intelligent space applications. Naïve bayes algorithm is a set or group of special algorithmic classifiers which are based on the Naïve bayes theorem which is often seen in mathematical and statistical learning and analysis. Naive bayes classifier is unique because in this classifier each feature is assumed to make independent and equal contributions to the outcome.

B. Support Vector Machine (SVM)

Support Vector Machines has gained crucial and significant improvements over the presently better performing and executing approaches and performs dynamically over a multiple or various learning assignments. Also additionally support vector machines are entirely automatic, and thus eliminating the requirement or need for labor-intensive work.

C. Logistic regression:

Logistic regression is one of the most foundational algorithms present in Machine Learning and is mostly used for binary classification and binary encoding. Thus, logistic regression is a linear algorithm. The result in logistic regression is basically derived with the help of a double variable(which has the possibility of having 2 outcomes). The predictions in logistic regression helps to maximize the output values and creates a non linear graph if represented graphically..

IV. TOKENIZATION

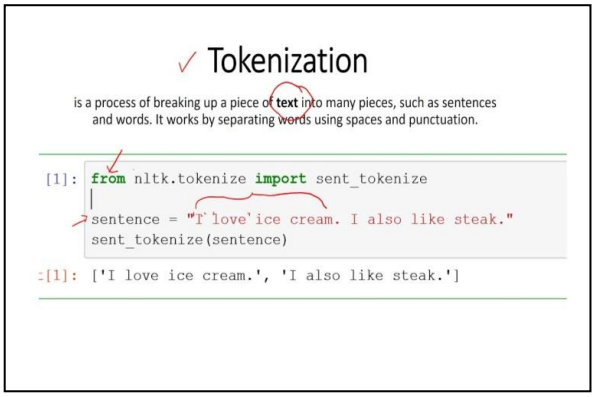

Fig 3 Tokenization

In the above image, an example of tokenization is provided with it's output. If given a sequence of characters and a derived document module, tokenization is usually termed as the process of splitting the document up into multiple meaningful pieces, or tokens . These chunks or tokens which are often generally referred to as words or terms, but it is sometimes very imperative to make a type or token separation. A *token* is mostly an object of a character sequence in some particular document that are clubbed together as a useful semantic module for processing phase. A type is referred to as the class of all tokens containing the same character sequence.

V. IMPLEMENTATION

Cyber-crime and threats to cyber security has risen over the years and now there are many techniques to deal with this Spam and Ham problem. One of the ways is to create a classifier to detect the difference between a ham and a Spam message. In our model, we have utilized feature engineering and ensemble method to detect the difference between a spam and a ham message.

A. Methodology

In this area, the system here portrays the general structure of work process of the trial which is led in an exceptionally animated condition. In this investigation Natural language Processing Tool is utilized for the examination and Machine learning libraries are utilized for the order of the dataset. In the absolute first level, the dataset is accumulated from different assets to make a pleasant dataset of spam and ham in the content configuration which is spared in a content record structure and the information is given as contribution to the Natural language preparing Model. In the second level which is a principle step before doing investigation, we changed over the information dataset which was available in content configuration to CSV group (Comma Separated Value). In the third degree of examination ,the preprocessing stage starts and it is accomplished for an increasingly attractive quality info which is finished by actualizing different information investigation and highlight extractions systems.

The dataset or the information present in the dataset is first marked before opening it in detail for sometime later. In the fourth step of the examination, numerous Machine Learning Classifiers are connected to the dataset which we have used. Subsequently the information in the dataset is prepared utilizing these numerous models. From that point onward, the way toward testing is done on the dataset to get alluring outcomes. In the fifth and last advance in our trial, a Confusion Matrix is gotten from the information , and these aftereffects of the distinctive ML classifiers are talked about and investigated utilizing ensembling strategies.

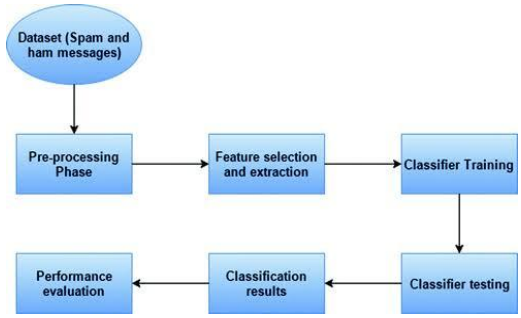

Fig 6 Methodology

VI. CODE SNIPPET

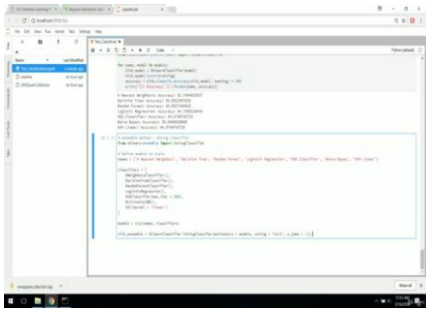

Fig 4 Code Snippet

```
In [8]: totalMails = mails['message'].shape[0]
        trainIndex, testIndex = list(), list()
        for i in range(mails.shape[0]):
            if np.random.uniform(0, 1) < 0.75:
                trainIndex += [i]
            else:
                testIndex += [i]
        trainData = mails.loc[trainIndex]
        testData = mails.loc[testIndex]

In [9]: trainData.reset_index(inplace = True)
        trainData.drop(['index'], axis = 1, inplace = True)
        trainData.head()
```

Out[9]:

	message	label
0	Go until jurong point, crazy.. Available only ...	0
1	Free entry in 2 a wkly comp to win FA Cup fina...	1
2	U dun say so early hor... U c already then say...	0
3	FreeMsg Hey there darling it's been 3 week's n...	1
4	As per your request 'Melle Melle (Oru Minnamin...	0

Fig 5. Output

A. Evaluation Metrics

The Evaluation metrics measure the accuracy and efficiency of our Model. The metrics are an essential part of any Machine learning model to predict the correct accuracy of the predicted values. The confusion matrix is mostly utilized for this purpose, for evaluating with the predicted and the actual values of our SMS Spam dataset. As suggested above, confusion matrix gives us the exact output and it also describes the complete performance of the model. We have our own classifier which predicts a particular class for a given input sample and the predicted class may vary depending on the type of input and the complexity of the input dataset.

VII. RESULTS

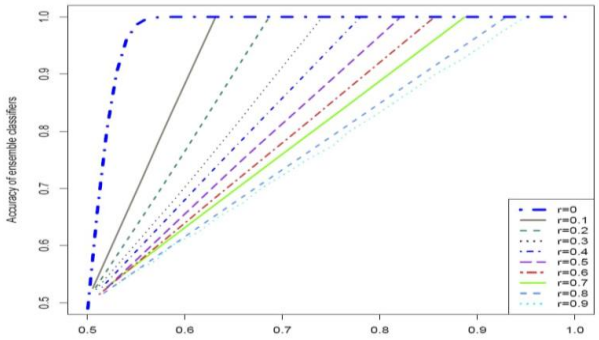

Fig 7 Accuracy of individual classifiers

VIII. CONCLUSION

As we are living in the period of quickly creating time of innovation and data, individuals are currently moving ceaselessly from the customary or standard methods for correspondence. Hence, this has cleared path for blasting sms spam messages. Different Machine Learning classifiers were connected on the dataset: Our Altered SMS Spam Collection Data Set incorporates worldwide substance just as Indian substance. In our analysis, the outcomes demonstrated that Support Vector Machine and Naive Bayes are among the better classifiers for the SMS spam recognition. The best aftereffects of University of California, Irvine SMS Spam Collection Data Set including Global substance turned out to be 98.21% of ACC%, 92.79% of SC% and 0.55 % of BH% with SVM. Future work must execute a few essential ways to deal with increment the part of the component plot. From the point of view of commonsense usage, especially for Indian portable clients, near examination of spam recognition and aversion from spam messages will bring a brilliant future for informing industry. We took inspiration from many research papers which made sms spam detection understandable and implementable to the customers and users and thus, in future an a mobile application can be developed which incorporates advanced AI or Ml algorithms to predict spam detection from non-spam or ham messages.

REFRENCES

1. SMS (Nov 2010).
2. J. M. G. Hidalgo, T. A. Almeida and A.Yamakami. "On the validity of a new SMS spam collection
3. T. A. Almeida , J. M. G. Hidalgo and A. Yamakami, "Contributions to the study of SMS spam filtering: new collection and results."
4. H. Zhang, "The optimality of naive Bayes." *AA* 1, vol. no. 2, 2004.
5. T. Joachims, "Text categorization with support vector machines: Learning with many relevant features"
6. L. Breiman, "Random forests", *Springer Berlin Heidelberg*, vol. no. 43
7. G. Ratsch, T. Onoda and K. R. Muller, "Soft margins for AdaBoost", *Springer Berlin Heidelberg*, vol. 42, issue 3
8. J. M. G.Hidalgo, G. C. Bringas, E. P. Sanz and F. C. García, "Content Based SMS Spam Filtering,"
9. G. V. Cormack, J. M. G. Hidalgo and E. P Sanz,"Spam Filtering for Short Messages".