Text Classification model using Natural Language Processing

Introduction:

Since the beginning of digital documents, automatic text classification has been a significant application and research area. Today, text classification is essential since we must deal with a huge volume of text documents every day.

Many individuals use mobile phones in the modern world, especially with the invention of the Internet. SMS is one of the most crucial services used by the client. Text messages can be sent via wireless systems using the two-way short messaging service (SMS).

SMS is a mobile phone service that enables text messaging between users. Sending undesirable content to a group of people for various reasons, such as fraud, is known as spamming. One type of spam is SMS spam, in which spammers send undesired messages to a large number of clients. To stay up with the current evolution of message services, it has therefore become important to design an SMS spam detection system.This corpus was compiled from free or open-access online sources.Three classifiers are employed in the proposed system to categorize the spam and ham messages. which include Logistic Regression, Decision Tree, and Random Forest.The project's goal is to create a system for filtering SMS spam.This system must be able to categorize datasets more securely and identify spam messages that the initial classifier mistakenly identified as ham communications. finally obtain a system that is more effective and secure.

Problem Statement:

Text messaging between users is made possible through SMS, a mobile phone service. Spamming is the act of sending unwanted content to a group of individuals for a variety of causes, such as fraud. SMS spam is one sort of spam in which spammers send unwanted messages to several clients. Designing an SMS spam detection system has consequently become crucial to keeping up with the current evolution of message services. The proposed system uses three machine learning models which are Logistic Regression, Decision Tree, and Random Forest to categorize spam and ham messages. The project's objective is to build a system for SMS spam screening.

Dataset Description:

From the Grumbletext website, 425 SMS spam messages were manually collected.3,375 SMS messages from the NUS SMS Corpus (NSC), a corpus of roughly 10,000 genuine messages gathered for research at the National University of Singapore's Department of Computer Science, were randomly selected as ham messages.450 SMS ham messages were compiled for this list from Caroline Tag's PhD thesis.And finally, 1002 unsolicited SMS messages and 322 spam texts are available to the public.

There are Two columns spam and ham in the dataset with 5572 rows which consumes around 87.2kb of memory.

Methodology:

The Natural Language Processing model is being implemented in the suggested system utilizing some machine learning algorithms. To carry out the categorization task, we implemented both white box and black box models. Decision Tree and Logistic Regression are white box models.

Although we have access to textual data, several pre-processing procedures must be performed on the data in order to convert the words into numerical features that may be used by machine learning algorithms.

- **Text Lowercase:** To cut down on the quantity of our text data's vocabulary, we lowercase the text.
- **Remove Numbers:**The numbers should be either eliminated or replaced with textual representations. The numbers are eliminated using regular expressions (re).
- **Remove Punctuations:** Since we don't have many spellings of the same term, we eliminate punctuation. Punctuation will be regarded differently if we don't delete it.
- **Removing Stopword:** Stopwords are words that do not advance the argument being made. As a result, they can be safely deleted without changing the sense of the phrase. We may use the stopwords in the NLTK (Natural Language Toolkit) library to eliminate stopwords from our text and obtain a list of word tokens as a result.

- **Stemming:** We will proceed with the procedure of obtaining a word's root form from stemming. The part to which inflexible prefixes (such as -ed, -ize, etc.) are added is known as the root or stem. We would take a word's prefix or suffix off to generate the stem words. Consequently, stemming a word might not produce genuine words.
  As an illustration: Mangoes ---> Mango
  Males ---> Male
  going —-> go
  We must translate our statements into tokens if they are not already in token form. We can translate word tokens into their root forms once we transform text strings into tokens. The Porter Stemmer, Snowball Stemmer, and Lancaster Stemmer are these. Among these, we frequently use Porter stemmers.
- **Tokenization:** It involves breaking down each speech into words or other smaller units called tokens.We can determine the number of times each token appears in the text by looking at word frequency. Types and tokens were differentiated while discussing word frequency. In a corpus, types are the unique words, whereas tokens are the words, including repeats.
  After completion of these Text Preprocessing tasks now we are ready to train the machine learning model for classification.
- **Label Encoding:**Label encoding is the process of transforming labels into a numeric form so that they may be read by machines. The operation of those labels can then be better determined by machine learning techniques. It is a significant supervised learning preprocessing step for the structured dataset.

Now,a we are training our machine learning model by using some classifiers like Decision Tree, Random Forest and Logistic Regression.For this purpose we split our dataset into train and test split where training dataset length is around 75% and the testing dataset length is of 25%.

1)After performing the Decision Tree classifier we get 95% accuracy.

2)After performing the Random Forest model we get 98%accuracy.

3)And for the Logistic Regression model we get 98% accuracy.

A voting classifier is a machine learning model that gains experience by training on a collection of several models and forecasts an output (class) based on the class with the highest likelihood of being the output.

To predict the output class based on the highest majority of votes, it merely averages the results of each classifier that was passed into the voting classifier. The concept is to build a single model that learns from these models and predicts output based on their aggregate majority of voting for each output class, rather than building separate dedicated models and determining the accuracy for each of them.

In hard voting, the predicted output class is a class with the highest majority of votes i.e the class which had the highest probability of being predicted by each of the classifiers.Suppose three classifiers predicted the output class(A, A, B), so here the majority predicted A as output. Hence A will be the final prediction.

In soft voting, the output class is the prediction based on the average of probability given to that class. Suppose given some input to three models, the prediction probability for class A =

(0.30, 0.47, 0.53) and B = (0.20, 0.32, 0.40). So the average for class A is 0.4333 and B is 0.3067, the winner is clearly class A because it had the highest probability averaged by each classifier.
Hence, We get Voting Classifier accuracy is 98%.
Here is the confusion matrix for the same.

	(Predicted Values)	
(Actual Values)	ham	spam
ham	1191	8
spam	19	175

Table 1. Confusion Matrix

Conclusion:

In the above system we have performed Text Classification by making comparative analysis using machine learning classifiers and Natural Language Processing techniques for use of a business where we can detect the message whether it is spam or ham, We have use Decision Tree , Random Forest and Logistic Regression to train the model. We have used 75% of the dataset for training and 25% dataset for the testing purpose. Finally, we use Voting classifier which is a machine learning model that gains knowledge from an ensemble of many models and forecasts an class based on the class that has the highest likelihood of being selected as the output with 98% accuracy.