TEXT MESSAGE CLASSIFICATION USING NATURAL LANGUAGE PROCESSING

Seminar (IT290) Report

Submitted in partial fulfilment of the requirements for the degree of

BACHELOR OF TECHNOLOGY

In

INFORMATION TECHNOLOGY

by

Gayathri Nisha (191IT116)

K Sakshi Thimmaiah (191IT124)

Niraj Nandish (191IT234)

DEPARTMENT OF INFORMATION TECHNOLOGY

NATIONAL INSTITUTE OF TECHNOLOGY

KARNATAKA SURATHKAL, MANGALORE - 575025

APRIL, 2021

D E C L A R A T I O N

I hereby *declare* that the *Seminar (IT290) Report* entitled **"Text message classification using Natural Language Processing"** which is being submitted to the National Institute of Technology Karnataka Surathkal, in partial fulfilment of the requirements for the award of the Degree of Bachelor of Technology in the department of Information Technology, is a ***bonafide report of the work carried out by us.*** The material contained in this seminar report has not been submitted to any University or Institution for the award of any degree.

Gayathri Nisha (191IT116)

K Sakshi Thimmaiah (191IT124)

Niraj Nandish (191IT234)

Signature of the Students

Place : NITK, Surathkal
Date : 13th April, 2021

CERTIFICATE

This is to certify that the Seminar entitled **"Text message classification using Natural Language Processing"** has been presented by Gayathri Nisha (191IT116), K Sakshi Thimmaiah (191IT124), Niraj Nandish (191IT234), the students of IV semester B.Tech. (IT), Department of Information Technology, National Institute of Technology Karnataka, Surathkal, on 13th April, 2021, during the even semester of the academic year 2019 - 2020, in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Information Technology.

Guide Name
Signature of the Guide with Date

Place:

Date:

ABSTRACT

This project describes classification of text messages using Natural Language Processing(NLP) which help in determining whether the text message a user receives is spam or not. NLP is a branch of artificial intelligence and has a wide range of uses namely, speech recognition, summarization, text searching, information grouping etc. The dataset came from the UCI Machine Learning Repository and it contains over 5000 labelled SMSs. The data was preprocessed to extract the useful information, replacement of email addresses, URLs etc. and the removal of stop words. The data was then split into two groups and the larger group was used to train our model. The accuracy of our model is over 98.5%.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 INTRODUCTION

Over the years, there has been an increase in demand for text analysis in various fields of work such as research, marketing, government-related operations etc. Text classification or text categorisation is an essential part of text analysis. It is the activity of labelling natural language texts with relevant categories from a predefined set. Or in simple words, it is the process of categorising text into organised groups.

Natural language processing is a branch of artificial intelligence with an objective to read, decipher, understand, and make sense of the human language in a manner that is valuable. It relies on machine learning for deriving the meaning from human languages. Algorithms are applied to extract the natural language rules in order to convert the unstructured data to a form that computers can understand. NLP techniques can be used to enable computers to process human language in the form of text or voice data and to 'understand' its full meaning, complete with the speaker or writer's intent and sentiment.

It goes hand in hand with text analytics, which categorizes words to extract structure and meaning from large volumes of content. It's helpful when we need the computer or program to understand and interpret human languages in order to extract information useful for that particular operation. For example, companies and/or brands need to know what their customers want and improve their product based on the results obtained. Subsequently, they would find better ways to market their product. To do this efficiently, we can analyse the activity of potential customers in social media platforms, as there is a rise in utilisation of these platforms to advertise and promote products and brands in general.

NLP can be used to perform a wide range of operations. A few of them are listed below :

1. **Searching** : This is one of the basic functions fulfilled by NLP. It can be used to find specific words in a document or to detect misspelled words, etc.

2. **Speech recognition** : Here, speech needs to be converted into text data. This can get challenging as there are different types of voices to be considered, with varying tones, accents and pronunciations

3. **Sentiment analysis** : It is often important to detect the sentiment behind written text or speech in order to understand the user/customer's opinion about a particular product or topic of discussion. This can be done by running sentiment analysis on the given text or audio data.

4. **Information grouping** : This is basically the classification of data according to specific attributes such as author, document type, etc.

5. **Summarization** : NLP can be used to summarise long texts like articles, specific documents, etc while keeping only the necessary information.

6. **Machine translation** : This involves converting or translating one natural language to another while preserving the meaning and of the given text.

We can see that NLP is a valuable technique, and it has numerous applications in quite a lot of areas. Yet, there are some challenges faced by machine and/or deep learning experts and researchers. Encoding schemas, tokenization in a few languages and understanding the context of the given data/text are a few of them.

Despite these challenges, NLP techniques today perform tasks decently well and are successful in adding value to many domains. They improve the performance of certain tasks which were otherwise performed by humans alone. One such domain is the healthcare industry, where NLP has been used to improve various purposes like disease diagnosis, care delivery and subsequently bringing down the costs involved. Some of these methods are not accurate enough and require some improvements.

NLP has been growing at a great pace, and it is likely that we reach great advancements in the years to come, and solving complex problems would become easier and can potentially be done in a shorter time.

CHAPTER 2 LITERATURE SURVEY

Natural language processing has a lot of real-world applications like, for example:

1. **Spam filtering** – Each one of us keep getting mails or SMS for things we didn't subscribe to and so it's better if they were marked as spam so that they don't mix up with the normal mails or SMS. Gmail and a lot of other mail platforms make use of NLP to identify spam emails.

2. **Speech-to-text conversion** – Translating speech into written text is important as it will help in making textual records of lectures, speeches, court hearings etc. There are a lot of apps which take note of everything we speak.

Our base paper for this project was SMS Spam Detection using Machine Learning Approach by Houshmand Shirani-Mehr. This project is basically about improving the accuracy of the model used in our base paper. The base paper's model had an accuracy of around 97.5%.

CHAPTER 3 TECHNICAL DISCUSSION

3.1 METHODOLOGY

For this particular project, we are using the dataset named “SMS Spam collection” which has a collection of text messages that can be classified into normal text messages and spam messages. This dataset comes from the UCI Machine Learning Repository. It contains over 5000 SMS labelled messages that have been collected for mobile phone spam research.

Our task is to classify this data through Natural Language Processing (by making use of machine learning techniques).

3.1.1 Preprocessing the dataset

In order to begin with the text classification, it is important that we preprocess the data present in the dataset. We start by extracting some useful information such as the column information and class distributions from the dataset. After this is done, we convert the class labels to binary values where 1 = spam, 0 = ham (normal text messages) using the LabelEncoder from sklearn. Then we replace the email addresses, URLs, phone numbers, and other symbols by using regular expressions. To ensure that all the letters in the text are of the same case, we convert all the text to lowercase using the method lower(). After this step, it is important that we remove the stop words such as “the”, “have” etc. from the text messages. The last step would be to extract the word stems (re-, un-, -ing, etc).

3.1.2 Feature Generation

Next we generate the features of sample data for the model which is the words in our case. We choose the 1500 most common words for our features. Then we shuffle the selected words and split them into two groups, one for training the model and the other is used to test the accuracy of the model. The ratio of training to test data is 3:1 .

3.1.3 Scikit-Learn Classifiers with NLTK

From SkLearn, import and use classifiers K Nearest Neighbors, Decision Tree, Random Forest, Logistic Regression, SGD Classifier, Naive Bayes, SVM Linear and VotingClassifier. The training data is fed to the classification algorithm. After training the classification algorithm, predictions are made on the testing data. The Voting Classifier used is a model that trains on an ensemble of all the 7 models mentioned above. The main benefit of using it is that it is a single model which trains by these models and predicts output based on their combined majority of voting for whether a given message is spam or not. Hard voting is used as it is more efficient than soft voting in situations where the algorithm may not be optimized. The accuracy of each of the classifiers are displayed to help figure out the best algorithm.

3.2 EXPERIMENTAL RESULTS

The “SMS Spam collection” dataset has been preprocessed.-The class labels have been converted to binary values using sklearn’s LabelEncoder. Some words have been replaced using regular expressions. Text has been converted to lowercase and finally stop words are removed and stem words are extracted.1500 of the most common words in the dataset are chosen as features. The feature sets are split into training and testing datasets consisting of 4179 and 1393 messages each. The training data is fed into various classification algorithms like the K Nearest Neighbors, Decision Tree, Random Forest, Logistic Regression, SGD Classifier, Naive Bayes, SVM Linear and VotingClassifier. The accuracies of these classifiers range from 94-98%. The K Nearest Neighbors classifier has the least accuracy.

The classification report is given below:

	precision	recall	f1-score	support
0	0.99	1.00	0.99	1212
1	0.99	0.93	0.96	181
accuracy			0.99	1393
macro avg	0.99	0.97	0.98	1393
weighted avg	0.99	0.99	0.99	1393

Table 3.1

The confusion matrix is given below:

		Predicted	
		HAM	SPAM
Actual	HAM	1211	1
	SPAM	12	169

Table 3.2

Screenshot of the output:

```
C:\Users\gayat\Desktop\IT 2nd year\IT 4TH SEM\IT290>python Text_Classification.py

----------Necessary Libraries are imported----------


----------Successfully loaded the dataset----------


----------Preprocessing is completed----------


----------Features are generated----------

K Nearest Neighbors Accuracy: 94.68772433596554
Decision Tree Accuracy: 97.98994974874373
Random Forest Accuracy: 98.77961234745155
Logistic Regression Accuracy: 98.99497487437185
SGD Classifier Accuracy: 98.63603732950466
Naive Bayes Accuracy: 98.77961234745155
SVM Linear Accuracy: 98.77961234745155
Voting Classifier Accuracy: 98.77961234745155


----------Classification report----------

              precision    recall  f1-score   support

           0       0.99      1.00      0.99      1212
           1       0.99      0.93      0.96       181

    accuracy                           0.99      1393
   macro avg       0.99      0.97      0.98      1393
weighted avg       0.99      0.99      0.99      1393


----------Confusion matrix----------

            Predicted
                 Ham Spam
Actual Ham      1211    1
       Spam       12  169
```

Figure 3.1

Hence, Out of 1393 text messages in the testing set:

- 1211 messages are correctly predicted as not spam.
- 169 messages are correctly predicted as spam.
- 12 messages are predicted to be not spam but are actually spam.
- 1 message is predicted to be spam but is not actually spam.

We can conclude that the model is able to correctly classify a large amount of text messages as spam or not. The percent of incorrectly classified text messages to the correctly classified ones is around 0.9%.

CHAPTER 4 CONCLUSIONS AND FUTURE TRENDS

A model that successfully classifies text messages into spam or not spam is designed. It has an accuracy of over 98.7%.Only a very small percent of the incorrectly classified messages are actual messages. Therefore, the model is letting through more spam which is better than missing out on actual messages. This model can be used to increase efficiency and safeguard businesses from potential risk. It can be used to help protect your devices against Viruses and Keep Hackers at Bay. Business employees do not have to go through numerous emails to decide which ones are spams, as sometimes that can be hard to decide. The time saved can be used to increase productivity. In the future, the model can be improved by taking all the words in the processed dataset as features instead of the current model that uses 1500 most common words in the set. It might take more time to train each classifier, but it can improve the overall accuracy of the model.

4.1 REFERENCES

1)https://www.researchgate.net/publication/328907962_A_Comparative_Study_of_Spam_SMS_Detection_Using_Machine_Learning_Classifiers

2)http://cs229.stanford.edu/proj2013/ShiraniMehr-SMSSpamDetectionUsingMachineLearningApproach.pdf

3) https://ieeexplore.ieee.org/document/8530469

4) https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7851079