FLIP ROBO

Project Report on

EMAIL (SMS) SPAM CLASSIFIER

Submitted by:

Pritam Sangle

ACKNOWLEDGEMENT

I would like to express my sincere thanks of gratitude to my mentors from Data Trained academy and Flip Robo Technologies Bangalore for letting me work on this project. Their suggestions and directions have helped me in the completion of this project successfully. All the required information & the dataset are provided by Flip Robo Technologies.

Finally, I would like to thank my family and friends who have helped me with their valuable suggestions and guidance and have been very helpful in various stages of project completion.

INTRODUCTION

In today's globalized world, email is a primary source of communication. This communication can vary from personal, business, corporate to government. With the rapid increase in email usage, there has also been increase in the SPAM emails. SPAM emails, also known as junk email involves nearly identical messages sent to numerous recipients by email. Apart from being annoying, spam emails can also pose a security threat to computer system. It is estimated that spam cost businesses on the order of $100 billion in 2007. In this project, we use text mining to perform automatic spam filtering to use emails effectively. We try to identify patterns using Data- mining classification algorithms to enable us classify the emails as HAM or SPAM.

DATA PREPROCESSING

The emails in the learning data are in plain text format. We need to convert the plain text into features that can represent the emails. Using these features, we can then use a learning algorithm on the emails. A number of pre- processing steps are first performed.

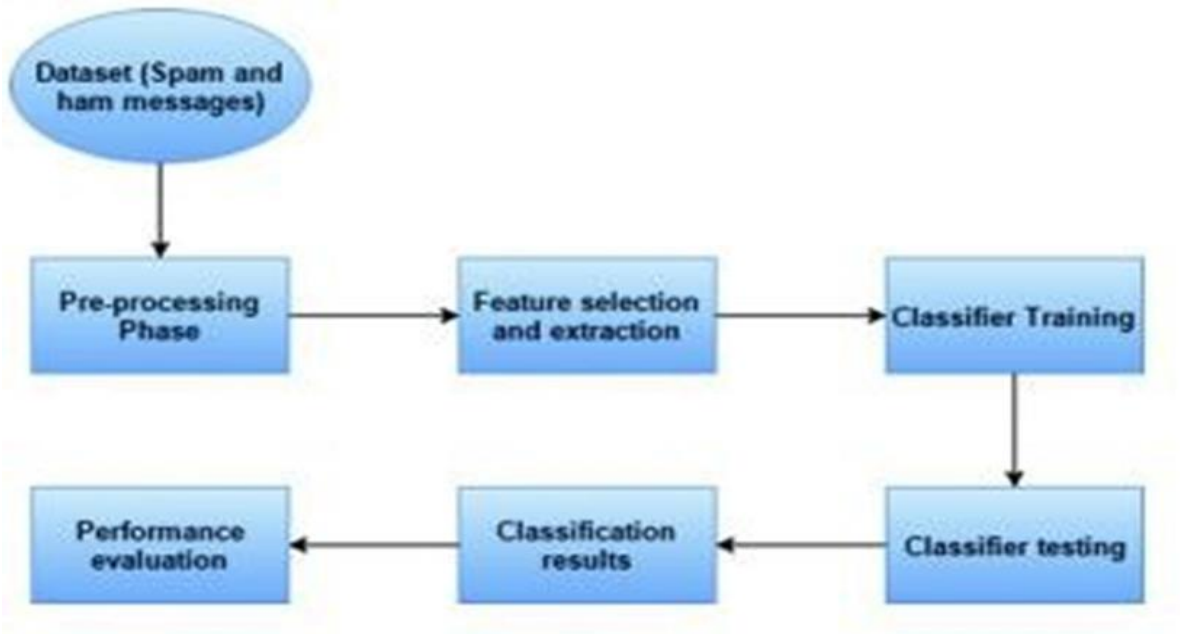
STOP WORDS

There are some English words which appear very frequently in all documents and so have no worth in representing the documents. These are called STOP WORDS and there is no harm in deleting them. Example: the, a, for etc. There are also some domain specific (in this case email) stop words such as mon, tue, email, sender, from etc. So, we delete these words from all the files using a Bourne Shell Script. These words are put in a file ‘words.txt’. The shell script takes multiple files as an argument and then deletes all the stop words mentioned in the words.txt file

STEMMING

The next step to be performed is stemming. Stemming is used to find a root of a word and thus replacing all words to their stem which reduces the number of words to be considered for representing a document.

Example: sings, singing, sing have sing as their stem. In the project, we use PYTHON implementation of Porter stemming algorithm which is slightly modified to meet our needs.



STEP FOR DATA PROCESSING

1. LOWER CASE
2. TOKENIZATION
3. REMOVING SPECIAL CHARACTER
4. REMOVING STOP WORD
5. PUNCTUATION
6. STEMMING

MODEL BUILDING

HERE WE USE THREE MODEL BUILDING ALGORITHM

1. GNB = Gaussian NB()
2. MNB = Multinomial NB()
3. BNB= Bernoulli NB()

VISUALIZATION USING WORDCLOUDS

Word Cloud is an approach to outwardly delineate the recurrence at which words show up in information. The cloud is comprised of words scattered fairly haphazardly around the figure.

Words seeming all the more regularly in the content are appeared in a bigger text style, while less normal terms are appeared in littler textual styles. This sort of figure has developed in fame as of late since it gives an approach to watch trending activities on social networking sites.

We compare the word clouds of ham and spam messages and See the difference between the frequently occurring terms in in the dataset

RESULT OF MODEL BUILDING

1. GNB = Gaussian NB()

ACCURACY SCORE-88%

PRECISION SCORE-53%

2. MNB = Multinomial NB()

ACCURACY SCORE- 96%

PRECISION SCORE -83%

3. BNB= Bernoulli NB()

ACCURACY SCORE- 97%

PRECISION SCORE -97%

CONCLUSION

The aims and objectives of the project, which achieved throughout the course, defined at the very first stage of the process. To collect all the information, the research work involved a careful study on the different filtering algorithms and existing anti-spam tools. These large-scale research papers and existing software programs are one of the sources of inspiration behind this project work. The whole project was divided into several iterations. Each iteration was completed by completing four phases: inception, where the idea of work was identified; elaboration, where architecture of the part of the system is designed; construction, where existing code is implemented; transition, where the developed part of the project is validated. However, there are still some parts that can be improved: for example, adding additional filtering techniques or changing aspects of the existing ones. The changes such as incrementing or decrementing the number of interesting words of the message and reorganizing the formula for calculating interesting rate can be done later

FUTURE SCOPE

In the future, we plan to deal with more challenging problems such as the analysis and management of report in spam SMS filters storing. Solution for this problem is another focus of work in the future.