Swathi Priya. N **Mentor: Jatin Dixit**

Capstone Project -2 SMS SPAM CLASSIFICATION MILESTONE REPORT

Problem

In this project the main goal is to predict legitimate or spam SMS messages. The SMS Spam Collection is a set of SMS tagged messages that have been collected for SMS Spam research. The files contain one message per line. Each line is composed of two columns: v1 contains the label (ham or spam) and v2 contains the raw text.

Client

Public(Mobile User)

Data Set

The dataset is taken from kaggle.

This corpus has been collected from free or free for research sources at the Internet:

-> A collection of 425 SMS spam messages was manually extracted from the Grumbletext Web site. This is a UK forum in which cell phone users make public claims about SMS spam messages, most of them without reporting the very spam message received. The identification of the text of spam messages in the claims is a very hard and time-consuming task, and it involved carefully scanning hundreds of web pages. The Grumbletext Web site is: [Web Link].

-> A subset of 3,375 SMS randomly chosen ham messages of the NUS SMS Corpus (NSC), which is a dataset of about 10,000 legitimate messages collected for research at the Department of Computer Science at the National University of Singapore. The messages largely originate from Singaporeans and mostly from students attending the University. These messages were collected from volunteers who were made aware that their contributions were going to be made publicly available. The NUS SMS Corpus is avalaible at: [Web Link].

-> A list of 450 SMS ham messages collected from Caroline Tag's PhD Thesis available at [Web Link].

-> Finally, we have incorporated the SMS Spam Corpus v.0.1 Big. It has 1,002 SMS ham messages and 322 spam messages and it is publicly available at: [Web Link].

https://www.kaggle.com/uciml/sms-spam-collection-dataset

Approach

- Loading Data
- Input and Output Data

- Applying Regular Expression
- Each word to lower case
- Splitting words to Tokenize
- Stemming with PorterStemmer handling Stop Words
- Preparing Messages with Remaining Tokens
- Preparing WordVector Corpus
- Applying Classification

Classification Steps

- Data Preparation
- Exploratory Data Analysis(EDA)
- Text Pre-processing and TF-IDF
- Model Building with Classification Algorithm

Deliverables

Trained classification model to be deployed in production to predict whether the SMS will Spam or Legitimate.

1. Data Preparation

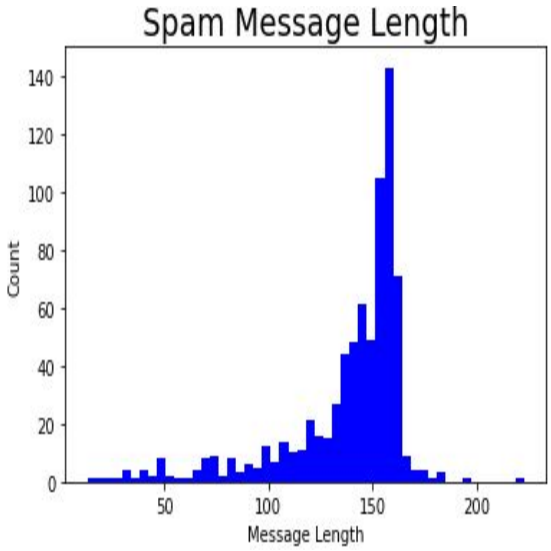
- Dropped unnecessary columns like unnamed 1, unnamed 2 and unnamed 3.
- Inserted new column called **count** to find the count the number of ham and spam messages
- Using info(), identify the null values if any and also the memory usage.
- Created corpus empty list for collective message and applied stemmer algorithm called PorterStemmer().

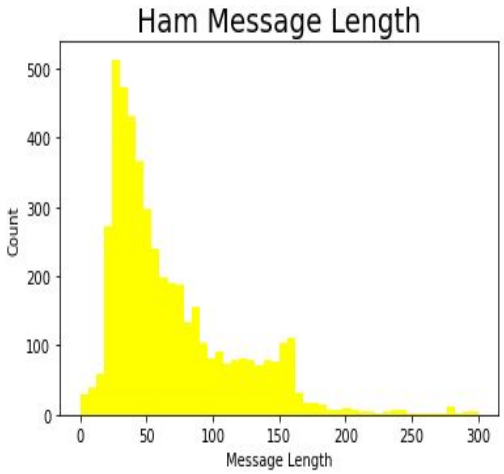
2. Exploratory Data Analysis

- Applied **groupby** to group the column information to show the ham and spam messages count and then used **describe** function to explore the statistical distribution of messages using mean , standard deviation ,etc.
- Replaced the column values and named it as label and sms instead of v1 and v2.
- Then applied count for both column label and sms which will give you the number of times a message is repeated.
- Using groupby function on column and describe : We can see the top msgs in ham and spam. Please call our customer service rep seems to be the most common spam message.
- Adding new column as sms length to find the length of the message
- **Visualize** the length of the ham and spam message by applying histogram.
- Inference of visualization is that we can see that sms with longer text tend to be spam.
- Then using word cloud we can find the repeated ham and spam word
- Inference of the word cloud code for spam words will be: we can see that sms containing words FREE,Please Call, Now , Win,Text,Call tend to be very common spam words

- Inference of the word cloud code for ham words will be: we can see the most common ham sms contain words will, know, gt (got), OK, know, Love,now.

Visualization:





Spam Words Word Cloud

Ham Words Word Cloud