SPAM MESSAGES FILTER

TUSHAR SINGH

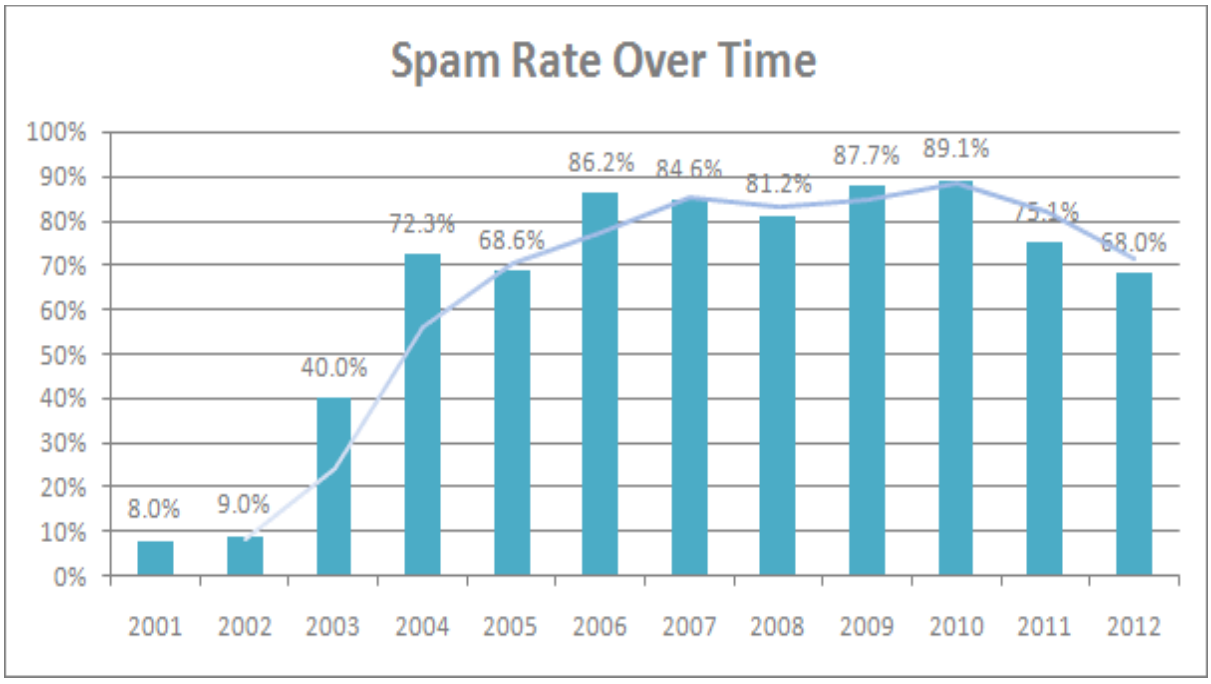
SEPTEMBER 10,2019

1.Introduction:

1.1Problem:

Nowadays as we are diving into the development in technology we faces some development as well as some of the drawbacks. Unsolicited commercial email (UCE) is the digital junk mail known as spam At the very least UCE is a nuisance, and at its worst an access point for viruses and malicious code. However you may view it, spam is an ever-present problem for most businesses and individuals.

Considering that the volume of email(more the messages) worldwide is 269 billion messages per day and that 49.7% of it is spam, you can understand the need for a good email spam filter.



We can see in this histogram that how the spam rate has changes over the years and we need a robust classifier for filtering the spam messages

1.2 Problem:

The past data of all types of spam messages would give us an insight on how different type of messages can be categorize as the spam and the normal messages .Thus this project predicts whether the Message received is either a Spam or Normal messages.

2.Data Acquisition and Cleaning:

2.1 Data sources: The data has been acquired by Kaggle dataset UCI MACHINE LEARNING DATASET . The SMS Spam Collection is a set of SMS tagged messages that have been collected for SMS Spam research. It contains one set of SMS messages in English of 5,574 messages, tagged acording being ham (legitimate) or spam.

The files contain one message per line. Each line is composed by two columns: v1 contains the label (ham or spam) and v2 contains the raw text.

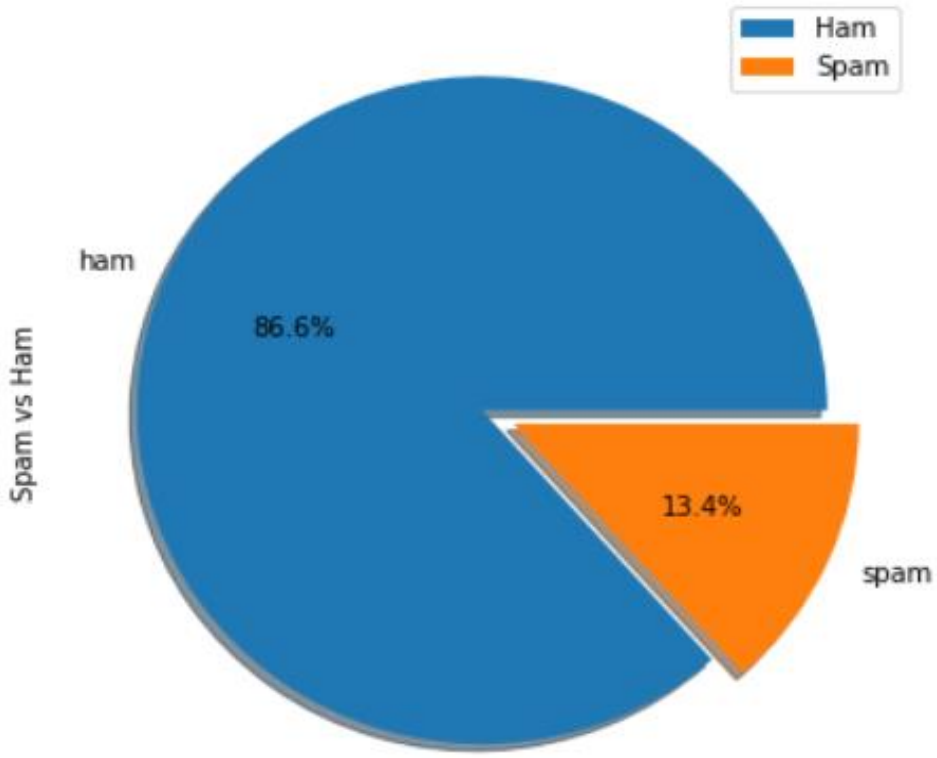
The csv file has been encoded using “latin-1” format

2.2 Data Cleaning

- I rermoved all the unimportant and outliers data/ Stopwords/ Stemming
- I have converted all the data to(Bag of words/ Tfidf Model)
- Further the columns are replaced and droped such as (Unnamed: 2, Unnamed: 3, Unnamed: 4)
- The columns V1 and V2 are replaced by Class and Text and sorted
- The Nan values are replaced using dropna() functions.
- Further a column of length is added to the loaded Dataframe so that we can use it as a feature
-

2.3 Feature Selection:

After data cleaning there were total (5572,3) rows and columns an had a size of 16716 .After detecting the correlation between the labels and text it seems that length has a strong relation with the length of text using visualization techniques through matplotlib library

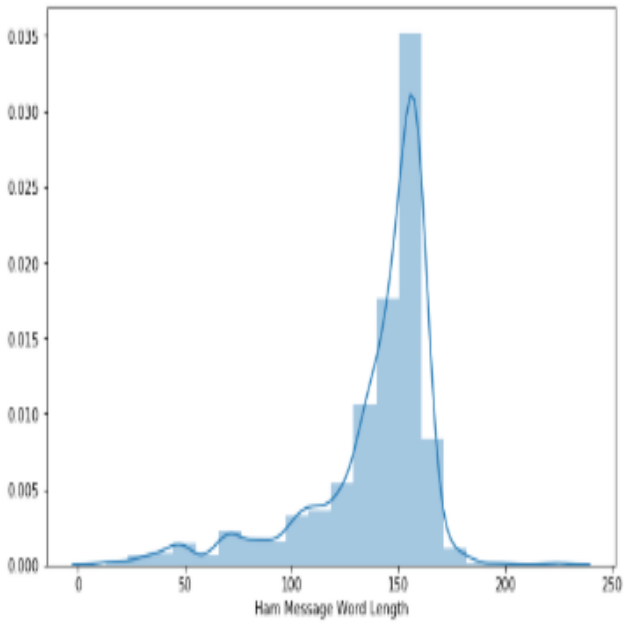


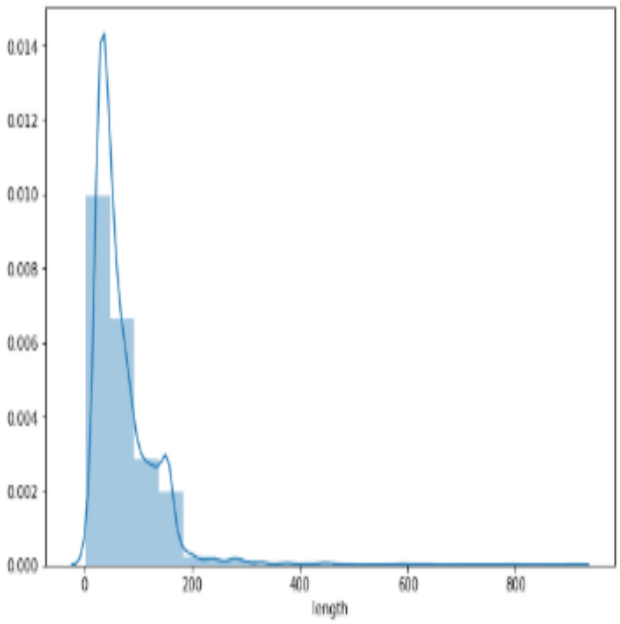
A lot of messages are actually not spam. About 86% of our dataset consists of normal messages.

- While we split our data set into train and test or when we use cross validation, we will have to use stratified sampling, otherwise we have a chance of our training model being skewed towards normal messages. If the sample we choose to train our model consists majorly of normal messages, it may end up predicting everything as ham and we might not be able to figure this out since most of the messages we get are actually ham and will have a pretty good accuracy.
- A very basic model would be a model that predicts everything as ham. It would have a decent accuracy. But then again, is that right? No. We will then have to use an accuracy metric that keeps this in mind. Goal : We don't mind if we miss the odd spam message but we surely don't want to mark a ham message as spam i.e Precision is very important.
- Further we analysed the relation between length and the class and how it affects our classification

text		
Sorry, I'll call later	30	ham
I cant pick the phone right now. Pls send a message	12	ham
Ok...	10	ham
Your opinion about me? 1. Over 2. Jada 3. Kusruthi 4. Lovable 5. Silent 6. Spl character 7. Not matured 8. Stylish 9. Simple Pls reply..	4	ham
Wen ur lovable bcums angry wid u, dnt take it seriously.. Coz being angry is d most childish n true way of showing deep affection, care n luv!.. kettoda manda... Have nice day da.	4	ham
Please call our customer service representative on FREEPHONE 0808 145 4742 between 9am-11pm as you have WON a guaranteed å£1000 cash or å£5000 prize!	4	spam
Okie	4	ham
Say this slowly.? GOD,I LOVE YOU & I NEED YOU,CLEAN MY HEART WITH YOUR BLOOD.Send this to Ten special people & u c miracle tomorrow, do it,pls,pls do it...	4	ham
7 wonders in My WORLD 7th You 6th Ur style 5th Ur smile 4th Ur Personality 3rd Ur Nature 2nd Ur SMS and 1st \Ur Lovely Friendship\"... good morning dear"	4	ham
Ok.	4	ham

Thus length acts as major part in discrimination of Spam and Ham messages
In order to apply a model, the necessary preprocessing must be completed. For text classification, usual preprocessing includes removing stop words (words that don't provide useful meaning, i.e. "and" "or"). Also the characters are converted to a single case (the below function converts to lower case). The function below then stems each word (this means that it replaces a word with the root of that word, for example "tasted" or "tasting" would become "taste").
I have visualized it as follow for better understanding:





The one on the right is the Ham messages or normal daily messages whereas the ones on the left is Spam messages we can see that the length affects the filtering a lot

The Sorted and filtered dataset is splited with using python library of train_test_split. With the test_size of 0.3 and a random_state of 111

2.3Preprocessing:

The text data column is copied, so any processing is not compelted on the original data. And then uses a TFIDF vectoriser to provide useful numerical values related to the data. TFIDF (term frequency - inverse document frequency) is a statistical method to tell how important a word is to a particular document by increasing the numerical value for an occurrence in the specific document but decreasing relative to number of occurrences in the entire corpus.

After this, a function available in the sklearn library is used to randomly assign training and test data to train and test the machine learning models.

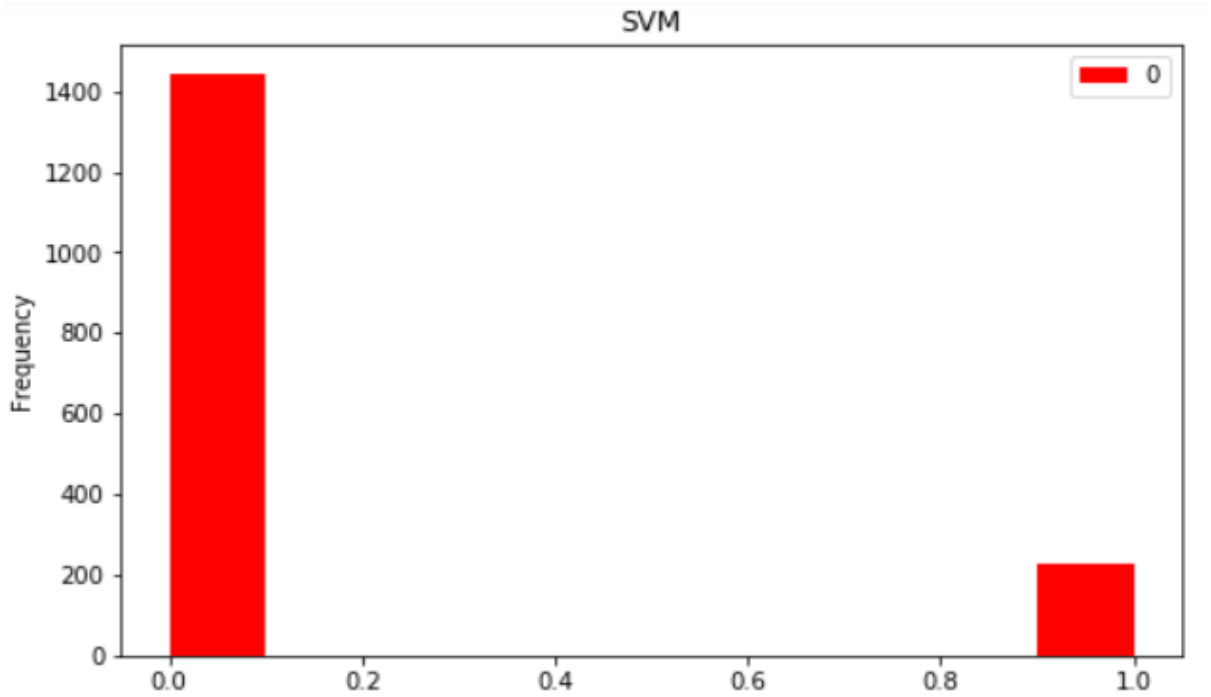
```
def pre_process(text):

    text = text.translate(str.maketrans('', '', string.punctuation))
    text = [word for word in text.split() if word.lower() not in stopwords.wor
    words = ""
    for i in text:
            stemmer = SnowballStemmer("english")
            words += (stemmer.stem(i))+" "
    return words
```

PREDICTING MODEL:

Regression Model:

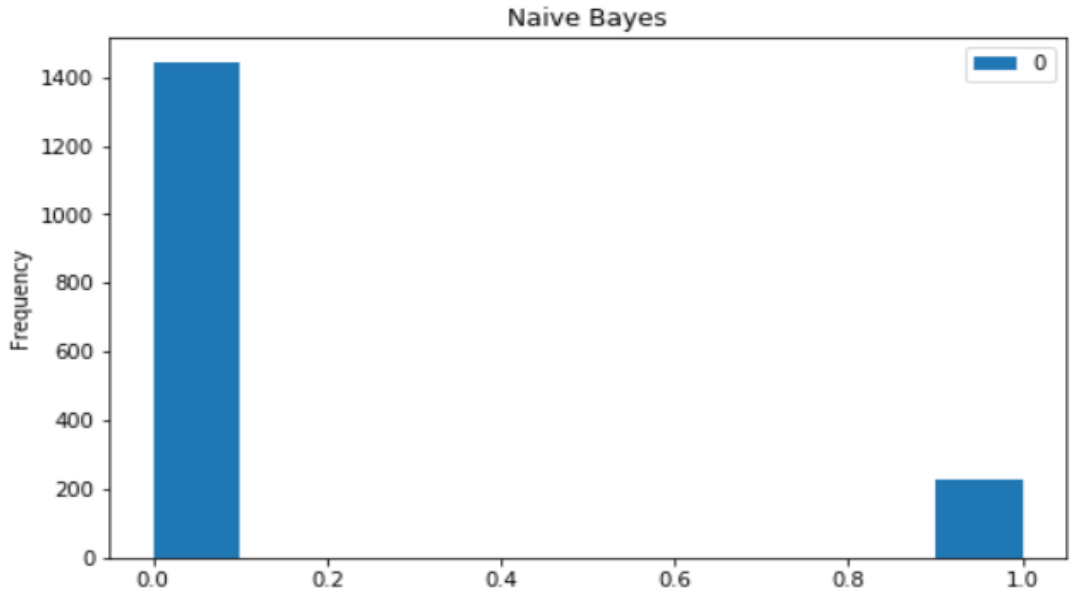
I have used a SVN model using sklearn library in python to fit my data accordingly I h ave used sigmoid function as my Kernel to fit the training dataset the training data ha s shape of `(3900, 8037)and (3900, 1)` whereas our test set has shape of `(1672, 8037)and (1672,1)` after training the data the train scores come out to be ***0.98*** and the accuracy test using accuracy_score fom sklearn metrics library it comes out to be ***0.97*** which means not only our data is fitted in optimal condition without any pr oblem of High variance (which is due to overfit)
The following code trains and tests a SVM model using sklearn, The gamma value was achieved by playing around with the model and by figure although looping or ve ctorize methond can be used for a larger dataset.
Thus the svm model suits as spam filter although a more precise model is used in s econd part i.e



The above fig represents the prediction data of our test set with **1** as the spam and **0** as Normal Messages

Naive Bayes Model:

I have used a Multinomial Naive Bayes Model using sklearn. It can be seen that is provided a slightly more accurate result than the SVM model so should therefore be used. There are many other models that may be more suitable for this dataset, however both of these model produce sufficient results .The alpha in this model has been selected as ***0.2*** after playing and modifying a lot .After training our features set we get an train accuracy for this as ***0.99*** whereas the accuracy_score comes out to be ***0.98*** which is slightly better than the SVM model although the predictions visually looks quite similar



We can summarize our model in following table:

MODEL	TRAINING ACCURACY	ACCURACY_SCORE
SVM	0.9874	0.9784
Multinomial Naive Bayes	0.9961	0.9850

Conclusions:

In this study, I analysed that how the spam filter can be formed using different models such filters can be used in many cases and has vast application working on this project gave me insight about how text data can be pre processed and can be used to classify or in case of auto fill ,Computer Vision problems(but with neural networks)

FUTURE SCOPE:

I was able to achieve almost 98% of accuracy in predicting spam filter for this project but in many cases when large data set and lots of features such as sender mobile ip device type will be added then such models would not be much efficient and then neural networks and other higher algo would come in to place so as the data and features would increase we would be using models according to have lower Bias and variance .