Email Spam Classifier Project

Submitted by:
TARKESHWER PANDEY

ACKNOWLEDGMENT

The internship opportunity I had with Flip Robo Technologies was a great chance for learning and professional development. Therefore, I consider myself as a very lucky individual as I was provided with an opportunity to be a part of it.

I would also like to appreciate Datatrained Pvt. Ltd. for provided with right trainingskills to deal with current problem statement.

INTRODUCTION

➢ Business Problem Framing

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 according being ham (legitimate) or spam.

Here we need to build an NLP Model to classify the messages asHAM or Spam.

➢ Conceptual Background of the Domain Problem

What is a Spam Filtering?

Spam Detector is used to detect unwanted, malicious and virus infected texts and helps to separate them from the nonspam texts. It uses a binary type of classification containing the labels such as '**ham'** (nonspam) and **spam**. Application of this can be seen in Google Mail (GMAIL) where it segregates the spam emails in orderto prevent them from getting into the user's inbox.

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.

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

A collection of 5572 rows 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.

➢ Review of Literature

Spam detection is one of the important Application of NLP, where it is used to detect unwanted e-mails getting to a user's inbox. NLP consists of various applications, like **speech recognition, machine translation,** and **machine text reading**. When we combine all these applications then it allows the artificial intelligence to gain knowledge of the world. Let's consider the example of AMAZON ALEXA, using this robot you can ask the question to Alexa, and it will reply to you.

NLP helps users to ask questions about any subject and get a direct response within seconds. NLP offers exact answers to the question means it does not offer unnecessary and unwanted information. NLP helps computers to communicate with humans in their languages. It is very time efficient. Most of the companies use NLP to improve the efficiency of documentation processes, accuracy of documentation, and identify the information from large databases.

Below are some of the most popular machine learning methods:

a) Naïve Bayes classifier: It is a supervised machine learning algorithm where words probabilities play the main rule here. If some words occur often in spam but not in ham, then this incoming e-mail is probably spam. Naïve bayes classifier technique has become a very popular method in mail filtering software. Bayesian filter should be trained to work effectively. Every word has certain probability of occurring in spam or ham email in its database. If the total of words probabilities exceeds a certain limit, the filter will mark the e-mail to either category.

b) Artificial Neural Networks classifier: An artificial neural network (ANN), also called simply a "Neural Network" (NN), is a computational model based on biological neural networks. It consists of an interconnected collection of artificial neurons. An artificial neural network is an adaptive system that changes its structure based on information that flows through the artificial network during a learning phase.

➢ Motivation for the Problem Undertaken

Email Spam Detection Project helps us to understands the basics of NPL model. We see some common NLP tasks that one can perform easily and how one can complete an end-to-end project. **Natural Language Processing (NLP)** is the art and science which helps us

extract information from the text and use it in our computations and algorithms.

Analytical Problem Framing

➢ Mathematical/ Analytical Modeling of the Problem

We are provided with a dataset consisting of text column and a label column. We read the data, and here is the statistical analysis of the dataset.

```
In [5]: # Let's get the information of the data
        df.info()

RangeIndex: 5572 entries, 0 to 5571
Data columns (total 5 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   v1          5572 non-null   object
 1   v2          5572 non-null   object
 2   Unnamed: 2  50 non-null     object
 3   Unnamed: 3  12 non-null     object
 4   Unnamed: 4  6 non-null      object
dtypes: object(5)
memory usage: 217.8+ KB
```

From here it is clear that the dataset consist of 5572 rows, we include object here as our columns has text data in it.

Our label column v1 consist of two classes, Spam and Ham(authenticate mail/not spam). And v2 defines the feature column

which consist of basically text data with was provided to us, mostly has all unique values.

We have total number of 4868 data with are not spam, while 704 data is spam text. Which indicate sure imbalance among the classes of label.

➢ Data Sources and their formats

We get the csv file for our model building. In that file we got two columns v1 and v2, where v1 represents the label column containing ham and spam, while v2 consist of one line message which is predicted as ham or spam.

We get to know through description that this corpus has been collected from free or free for research sources at the Internet.

```
In [2]: # load the CSV file
        df = pd.read_csv("E:\Flip Robo\Project\Spam Project\spam.csv",encoding='ISO-8859-1')

In [3]: df.head()
```

Out[3]:

	v1	v2	Unnamed: 2	Unnamed: 3	Unnamed: 4
0	ham	Go until jurong point, crazy.. Available only ...	NaN	NaN	NaN
1	ham	Ok lar... Joking wif u oni...	NaN	NaN	NaN
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...	NaN	NaN	NaN
3	ham	U dun say so early hor... U c already then say...	NaN	NaN	NaN
4	ham	Nah I don't think he goes to usf, he lives aro...	NaN	NaN	NaN

```
In [4]: df.shape
Out[4]: (5572, 5)
```

The Unnamed columns are the result of adding additional parameters while calling the dataset, so we will drop them and keep only v1 and v2. As these are the required columns for model building. The final dataset will look like this:

```
In [6]: # lets delete the Unnamed:2, Unnamed: 3 and Unnamed: 4 columns
        df.drop(columns=['Unnamed: 2','Unnamed: 3', 'Unnamed: 4'], inplace = True)

In [7]: df.sample(5)
```

Out[7]:

	v1	v2
351	ham	If you're not in my car in an hour and a half ...
4113	ham	Where are you ? What do you do ? How can you s...
3904	spam	Do you want a new video handset? 750 anytime a...
1759	ham	Do u ever get a song stuck in your head for no...
5184	ham	I'm in town now so i'll jus take mrt down later.

```
In [8]: # Replace the name of the column V1 & V2
        df.rename(columns={'v1':'target', 'v2':'text'}, inplace=True)
```

The dimension for final dataset is known by using shape function, where first element shows the number of rows while second feature shows number of columns.

```
In [4]: df.shape
Out[4]: (5572, 5)
```

➢ Data Preprocessing Done

In this step, we explore data and gain insights such as shape, accuracy, and initial patterns in the data.

First, we check the information of the given dataset and extract information about the dataset. And we see that the dataset contains 5572 entries with no NaN values in the feature message & label.

We create a new feature named text_lenght to check the length ofeach text in v2 column.

```
# finding the lenght of v2 which is our text data
data['text_lenght']=data['v2'].str.len()
data.head(2)
```

	v1	v2	text_lenght
0	ham	Go until jurong point, crazy.. Available only ...	111
1	ham	Ok lar... Joking wif u oni...	29

Besides all these steps, now we do some common tasks that can be done on every NLP project.

We have to clean the data using regex, matching patterns in the e-mail messages, and replace them with more organized counterparts or just remove them. Cleaner data leads to a more efficient model and higher accuracy. Following steps are involved in pre-processing the messages:

1. We start with importing all the necessary libraries.

2. We define a function ‘deconstructed’ to expand all the English language contractions.
3. Defining the stopwords and lemmatizer.
4. Then we pass our text data to this function, then we clean the data by remove all the hyperlinks, emails, numbers, punctuations, whitespaces, stop_words, special characters.
5. Then we perform lemmatization, and convert the entire text to lower case.
6. We perform above two steps on our text feature column and store the data into list.
7. We assigned this processed data list to the feature column v2 of our dataset.
8. Then we check the length of clean data, so that we get a clear idea about how much our data is processed.

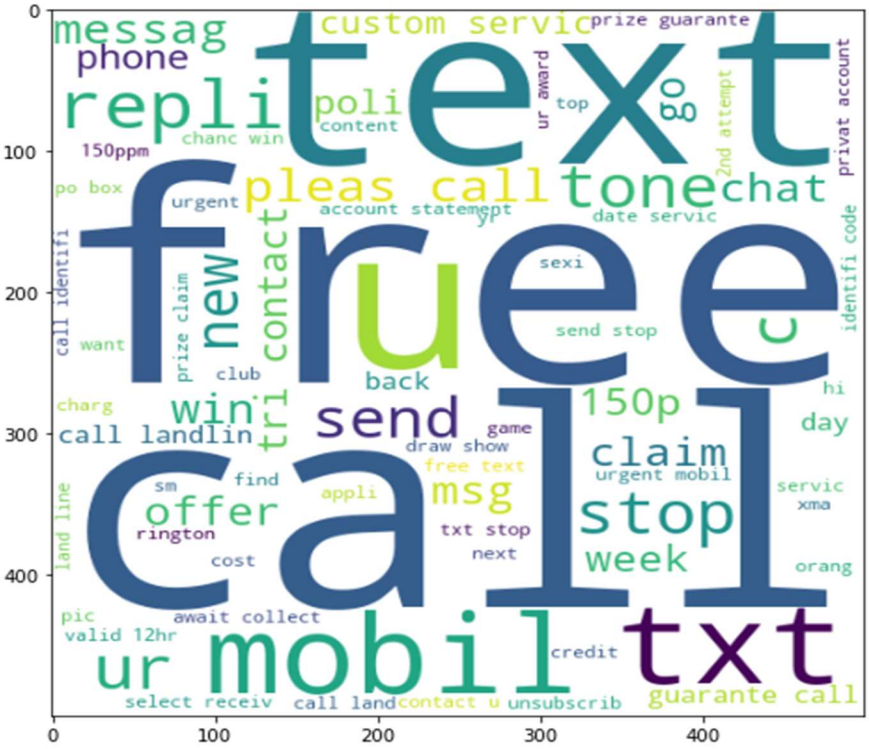
```
In [28]: df['num_characters'] = df['text'].apply(len)

In [47]: df.head()
```

Out[47]:

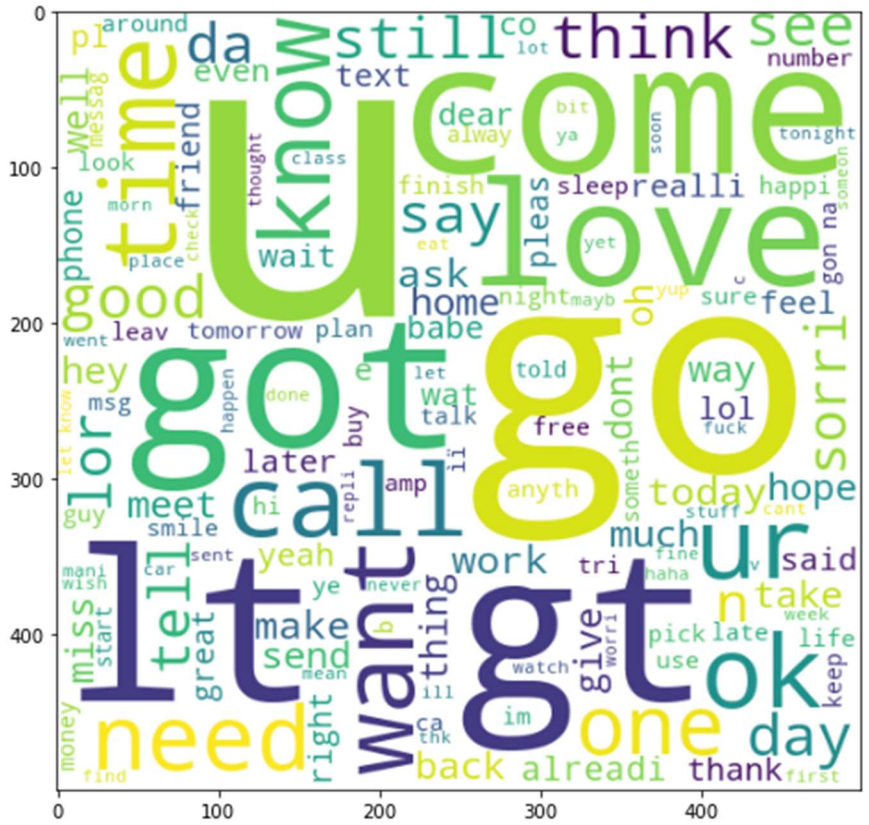
	target	text	num_characters	num_words
0	0	Go until jurong point, crazy.. Available only ...	111	24
1	0	Ok lar... Joking wif u oni...	29	8
2	1	Free entry in 2 a wkly comp to win FA Cup fina...	155	37
3	0	U dun say so early hor... U c already then say...	49	13
4	0	Nah I don't think he goes to usf, he lives aro...	61	15

➢ Data Inputs- Logic- Output Relationships

For defining the input out logic we plot word cloud for the words we are commonly seen in ham message, and a word cloud consist of common words used in spam messages.



These are top 400 words which are generally seen in a spam message. The larger font words(free, text, call and so on) are thosewhich are most occurred words in spam message.



These are set of words which generally find in a ham(not spam) messages. Here also the large fonts denotes frequently appearing words in the ham text.

- State the set of assumptions (if any) related to theproblem under consideration

We are considering that the data provided to us is collected properly, and we are building an algorithm which perfectly definesand predicts in the current dataset.

➢ Hardware and Software Requirements and Tools Used

1. Python
2. Pandas
3. Matplotlib.pyplot
4. Warnings
5. Numpy
6. Seaborn
7. Re
8. String
9. Nltk
10. Stopwords
11. Tfidfvectorizer

Here we build a classification model using classification algorithms.

Model/s Development and Evaluation

➢ Identification of possible problem-solving approaches (methods)

We perform tfidf vectorization in our feature column, to vectorizedthe data, then we use that vectorized data in our ensemble classification approach (in our case its RandomForest Classifier) to build the final model.

➢ Testing of Identified Approaches (Algorithms)

As we know that this NLP problem is a classification problem, so we follow following classification approaches to predict the best result for our dataset.

1. Logistic Regression
2. Naive Bayes (MultinomialNB)
3. Decision Tree classifier
4. RandomForest Classifier
5. Support Vector Matrix (SVC)

- Key Metrics for success in solving problem under consideration

  Our primary metrics is accuracy and log_loss for the valuation ofthe model, the model with highest accuracy and lowest log_lossseems to the best fitted model for the dataset.

  While our secondary metrics is F1 score and ROC AUC score, highestvalues for both serves as best fitted model.

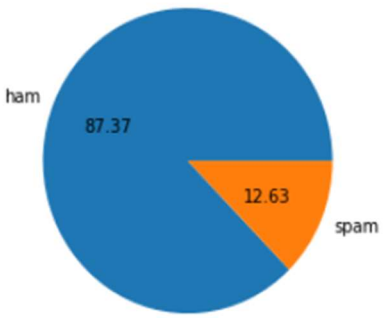
  We also perform cross validation for each model for checking the authenticity of the model.

Visualizations

Following the graphs which we plot during the EDA or visualization of data.

We can observe that we have maximum amount of data for hamthan around 4868 while for spam we have 704 messages in total. We need to take care of this while building model.

```
In [21]: # Let's create the pie chart from dataset to check the value
         plt.pie(df['target'].value_counts(), labels= ['ham', 'spam'], autopct='%0.2f')
         plt.show()
```



From here also we can see that around 87.37% of data is ham while 12.63% of data isspam. Above graph represents the percentage distribution of data among classes of labels.

Interpretation of the Results

1. Done all the pre-processing steps to make data ready for modelbuilding.
2. Removed stop words and create another feature 'clean length' for comparing cleaned and unprocessed message length.
3. Understood relationship and gain insights by using Data Visualization:
   a. Plotted count plot for checking spam and non-spam email counts.
   b. Lastly, visualize some popular terms in spam messages using the word cloud.
4. Used **Tf-idf vectorizer** to convert text into vector.
5. Found the best model as **SVC** which provides maxaccuracy of **97.58%.**
6. Found **high precision and recall score of 0.97.**
7. Confusion matrix shows **high classification accuracy** with only 16 out of 704 are incorrect.
8. **Overall model fit is good.**

CONCLUSION

- ## Key Findings and Conclusions of the Study

  With the current distribution of train data and test data, SVC seems to be the best fit for our model.

  As our dataset consist of only two columns, first is v1 which is our label column and the second is v2 which is our feature text column.

  In our case the label column is also in categorical form, so we replace the ham with '0' and spam with '1' to make it a numericalcolumn.

  For converting the text feature column to machine understandable form, we perform NLP pre-processing techniques we clean the data and then perform vectorization in other to make it machine understandable for.

  We perform TFIDF vectorization as, it vectorized the feature data or in simple term it encodes and standardized the data, so that we can take this data as it is for model building.

  The other method for vectorization is WordVec , there we encodeour data, but we need to standard scale the data before building the model.

The best fitted model for our dataset is TFIDF vectorized Random Forest Classifier. It has both highest accuracy and lowest log loss.

Even the roc_auc_score of SVC is highest of all 97.58%

- ## Learning Outcomes of the Study in respect of Data Science

Today, spamming mails is one of the biggest issues faced by everyone in the world of the Internet. In such a world, email is mostlyshared by everyone to share the information and files because of their easy way of communication and for their low cost. But such emails are mostly affecting the professionals as well as individuals bythe way of sending spam emails. Every day, the rate of spam emails and spam messages is increasing. Such spam emails are mostly sent by people to earn income or for any advertisement for their benefit. This increasing amount of spam mail causes traffic congestion and waste of time for those who are receiving that spam mail. The real cost of spam emails is very much higher than one can imagine. Sometimes, the spam emails also have some links which

have malware. And also, some people will get irritated once they see their inbox which is having more spam mails. Sometimes, the users easily get trapped into financial fraud actions, by seeing the spam mails such as job alert mails and commercial mails and offer emails.It may also cause the person to have some mental stress. To reduceall these risks, the system has proposed a machine learning model which will detect spam mail and non-spam emails, and also this system will optimize the data by removing the unwanted mails whichcontain the advertisement mails and also some useless emails and also some fraud mails. This proposed system will detect the spam mails and ham emails with the dataset consisting of spam mails.

This is one of the most popular NLP model, as it really deals with thereal time scenario.

- ## Limitations of this work and Scope for Future Work

Here we build a model which detect the spam and ham messages/emails, we can also build a system which after identifying spam mails this system will remove that spam emails and this proposed system will calculate the amount of storage before and after the removal of spam mails.
