Flip Robo Technologies

Email Spam Detection Project

Submitted by: John Tojo

Data Science Intern at Flip Robo Technologies

ACKNOWLEDGMENT

It gives me immense pleasure to deliver this report. Working on this project was a great learning experience that helped me attain in-depth knowledge on data analysis process. Flip Robo Technologies (Bangalore) provided all of the necessary information and datasets, required for the completion of the project. I express my gratitude to my SME, **Gulshana Chaudhary**, for providing the dataset and directions for carrying out the case study procedure.

INTROUDCTION

Classifying ham and spam is an important task that helps us to filter out unwanted or malicious emails. In today's digital age, electronic communication has become a crucial part of our personal and professional lives. However, this has also made us vulnerable to spam emails, which are unsolicited and often contain malicious content such as scams, phishing attacks, or viruses.

Classifying emails as either ham (legitimate) or spam (unwanted) helps us to protect ourselves from these threats by automatically directing spam emails to a separate folder or deleting them altogether. It also improves the overall efficiency of our email systems by reducing the amount of time we spend sifting through and deleting spam emails.

In this report, we will discuss the various methods for classifying ham and spam, including machine learning algorithms and the approach used to solve the problem. We will also examine the importance of accurately identifying spam emails in order to protect ourselves and our inboxes from unwanted and potentially harmful content.

Problem Statement

- The problem of spam emails is a pervasive one that affects individuals and organizations alike. Spam emails can be a nuisance, clogging up our inboxes and taking up valuable time that could be better spent on more important tasks. Moreover, spam emails can also pose a serious threat to our security and privacy, as they often contain malicious content such as scams, phishing attacks, or viruses.

- As a result, there is a need to classify ham and spam in order to protect ourselves from these threats and improve the efficiency of our email systems. However, accurately identifying spam emails can be a challenging task, as spam emails are constantly evolving and becoming more sophisticated. There is also the risk of false positives, where legitimate emails are mistakenly classified as spam.

- Given these challenges, it is important to carefully consider the various methods for classifying ham and spam and to continuously update and improve these methods as the spam landscape changes

.DATA SOURCES AND THEIR FORMATS

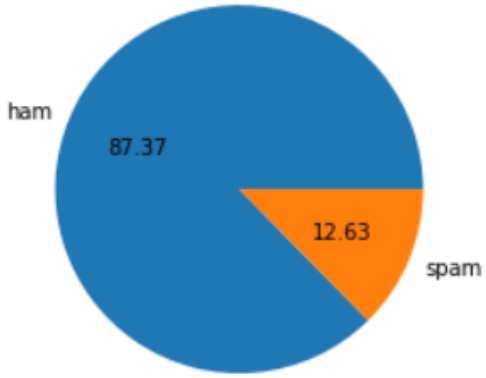
- The data was provided by FlipRobo in CSV format. The dataset had 5 columns and 5572 rows
- v1: which takes values spam or ham
- v2: text
- Unnamed: 2, Unnamed: 3, Unnamed: 4 mostly have null datas
-

Exploratory Data Analysis (EDA)

- When working with a data frame, it is important to first check for any missing or null values in the data. In this case, the columns Unnamed: 2, Unnamed: 3, and Unnamed: 4 have a high percentage of null values, which could potentially impact the accuracy of any analysis or modelling performed on the data.

- Renaming the columns in a data frame can be useful for improving the clarity and readability of your data. In this case, the columns v1 and v2 have been renamed to target and text, respectively. This can make it easier to understand the meaning and purpose of these columns when working with the data. It is also important to check for and handle duplicates in the data. Duplicate data can occur for a variety of reasons and can impact the accuracy and reliability of your analysis. In this case, it appears that duplicates accounted for 7% of the whole data, which is a significant amount. Dropping these duplicates can help to ensure that your analysis is based on unique and accurate data.

- A pie chart was used to see the distribution of spam and ham emails in the data set. It is important to carefully consider the balance of the data when performing analysis or building models. An imbalanced data set is one where the distribution of the classes is uneven, with some classes being significantly more or less represented than others. This can be a problem because it can lead to biased or inaccurate results. For example, if the data set is heavily imbalanced towards spam emails, a model trained on this data may be very good at identifying spam emails but not as effective at identifying ham emails. This could lead to false negatives, where legitimate emails are mistakenly classified as spam.



- Creating new columns in a data frame can be a useful way to extract additional information or insights from the data. In this case, new columns named num_characters, num_words, and num_sentences were created to store the number of characters, words, and sentences in the text column. This information could be useful for a variety of purposes, such as identifying trends or patterns in the data or building machine learning models.

- Cleaning and preprocessing text data is an important step in many natural language processing (NLP) tasks. In this case, the text data was cleaned by first converting it to lowercase, then removing punctuation and stop words. The text was then stemmed, which involves reducing words to their base form by removing inflections, and joined together. The resulting text was then stored in a new column called transformed_text. This process can help to improve the accuracy and effectiveness of any analysis or modeling performed on the text data. For example, removing punctuation and stop words can help to reduce noise and focus on the most important or relevant words in the text. Stemming can also help to reduce the dimensionality of the data by reducing the number of unique words. After the text data was cleaned and transformed, the target column was encoded. Encoding is the process of converting categorical data, such as spam and ham, into numerical form. This is often necessary for machine learning algorithms, which typically require numerical input. Encoding the target column in this way can allow you to more easily apply machine learning techniques to the data.

Visualization

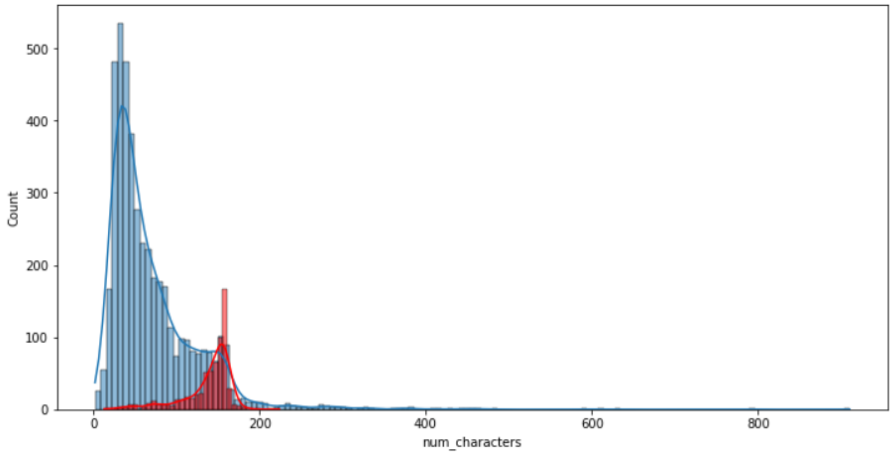


Fig1: count vs num_characters

From this plot, it appears that spam emails tend to have more characters present than ham emails.

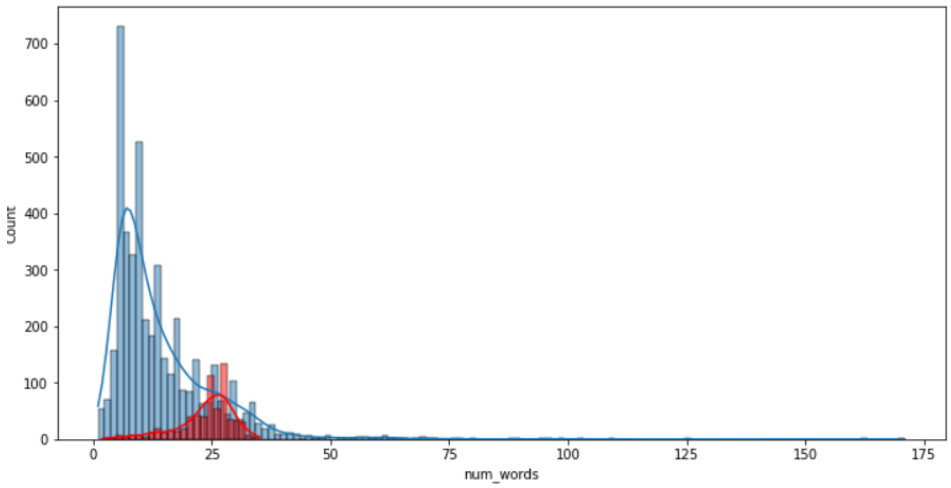


Fig2: count vs num_words

From this plot, it appears that spam emails tend to have more characters present than ham emails.

Fig3: showing word cloud for ham

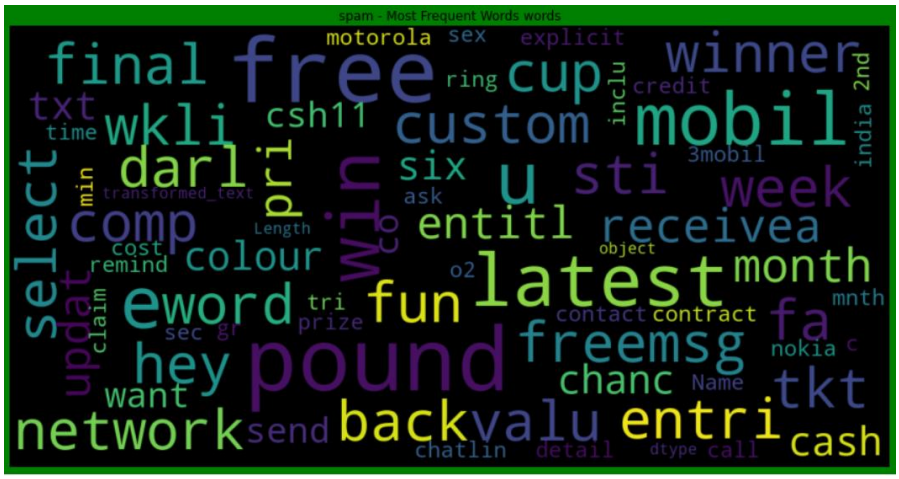


Fig4: showing word cloud for spam

Model Selection:

Vectorization is the process of converting text data into numerical form, typically by creating a vector of word counts or frequencies. In this case, the stemmed text was vectorized using a technique called TFIDF (term frequency-inverse document frequency).

TFIDF is a common method for converting text data into a numerical form that can be used in machine learning models. It takes into account the frequency of a word in a document, as well as the

overall frequency of the word across all documents. This can help to reduce the impact of common words that may not be particularly meaningful or relevant for a particular analysis or task.

The dataset being used in this case is imbalanced, meaning that the distribution of the classes is uneven. To address this issue, the dataset was balanced using a technique called SMOTE (synthetic minority oversampling technique). This involves generating new synthetic samples of the minority class in order to better balance the dataset.

The problem being addressed in this case is a classification problem, as the target variables can take on only two values: 0 or 1. The machine learning models that considered for this problem are

- SVC (support vector classifier)
- MultinomialNB (multinomial naive Bayes)
- KNeighborsClassifier (k-nearest neighbors classifier)
- RandomForestClassifier (random forest classifier)
- ExtraTreesClassifier (extra trees classifier).

To determine the best model for this task, several evaluation metrics were used, including roc_auc, f1 score, mse, mae, accuracy, and precision. These metrics provide different insights into the performance of the models and can help to identify the model that is most effective at correctly classifying spam and ham emails.

	classifier	roc_auc	f1	mse	mae	accuracy	precision
4	ETC	0.996902	0.996892	0.002914	0.002914	0.997086	1.000000
3	RF	0.993804	0.993766	0.005828	0.005828	0.994172	1.000000
0	SVC	0.989062	0.988264	0.011072	0.011072	0.988928	0.985222
1	KN	0.987887	0.987516	0.011655	0.011655	0.988345	0.994969
2	NB	0.982803	0.981527	0.017483	0.017483	0.982517	0.975520

Based on the evaluation metrics that were used, it appears that the ExtraTreesClassifier (ETC) is the best model for this task. This model achieved the highest roc_auc area, which is a measure of the model's ability to distinguish between the two classes. It also had the highest accuracy and precision and least value for FP (false positives) and FN (false negatives) which indicates that it made few mistakes when classifying the emails.

In addition, the ETC had the least error in terms of mae (mean absolute error) and mse (mean squared error). These metrics measure the difference between the predicted values and the actual values, and a lower error indicates that the model is more accurate.

Conclusion

In conclusion, classifying ham and spam is an important task that helps to protect us from unwanted and potentially harmful emails. There are various methods for classifying ham and spam, including machine learning algorithms and manual filtering techniques.

In this report, we examined the need to classify ham and spam and explored the various methods for doing so. We also looked at the importance of accurately identifying spam emails in order to protect ourselves and our inboxes from unwanted and potentially harmful content.

We found that the ExtraTreesClassifier was the best model for this task, based on the evaluation metrics that were used. This model achieved the highest roc_auc area, had the highest accuracy and precision, and made the fewest mistakes when classifying spam and ham emails.

Overall, the ability to accurately classify ham and spam is crucial for ensuring the security and efficiency of our email systems. It is important to continuously update and improve our methods for classifying ham and spam in order to stay ahead of the evolving spam landscape.