# TEXT CLASSIFICATION FOR SPAM DETECTION PROJECT REPORT

## Data Science Methodology

We have worked through each step of the data science method to complete the capstone project


## Step One: Problem Identification

After looking into the given information about the case study the problem has been identified and a problem statement is generated.

## Problem Statement

The proliferation of spam messages in communication channels poses a significant challenge, leading to user inconvenience, privacy concerns, and potential security threats. The aim of this project is to develop an efficient text classification model for spam detection, leveraging various techniques, including Bag of Words (BoW), TF-IDF, Word2Vec, and advanced BERT embeddings.

## Step Two: Data Wrangling

The second step of the data science method is data wrangling, also known as data munging or data cleaning. It involves several steps to ensure data quality and usability. Here are the steps we performed in data wrangling:

- We collected the relevant data from the Web.
- Explored the data to understand its structure and identified potential issues, such as missing values, outliers, and inconsistencies.
- Handled missing values, either by imputing them with reasonable estimates or removing them if they are too numerous or critical.
- Identified and removed any duplicate rows in the dataset.
- Examined outliers and decided whether they need to be corrected or removed.
- Converted data into a consistent format.
- Corrected any obvious errors or inconsistencies in the data.

## Step Three: Exploratory Data Analysis

The step three of DSM is Exploratory Data Analysis. In this process we summarize our findings visualize results and analyze them to find any meaningful insights. The steps we performed in this process includes:

- Summarize the key statistics and insights derived from the EDA.
- Visualize the data distributions, correlations, and patterns.
- Identify any interesting trends or relationships in the data.

## Step Four: Pre-processing and Training Data Development

This is the fourth step of the process; we pre-process and trained the data using different machine learning models and perform feature engineering to identify the best model which make appropriate predictions. This step we performed are:

- Outlined the feature engineering techniques applied to create new features or transform existing ones.
- Normalized and scaled the data using different techniques.
- Used different machine learning algorithms to test and train the data.
- Created different models such as Naive Bayes, Random Forest and a Neural Network to compare the results.
- Provide the results and performance of each model.

## Step Five: Modeling

For the step five, modeling we performed the following steps:

- Chose the best model with highest accuracy and minimum errors.
- Documented the evaluation metrics chosen to assess model performance.
- Stated the best-performing model and its evaluation metrics.
- Explained the reasoning behind choosing the winning model.
- Made suggestions for business executives on how to use the model in future.

After gathering a Dataset including 4825 Ham messages and 747 Spam messages, we imported the data into a jupyter notebook and performed some basic Data Wrangling. We analyzed the data by checking basic details like shape and information. As our data set had only two columns so it didn't require much wrangling. We encoded our text data and assigned labels to it.

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

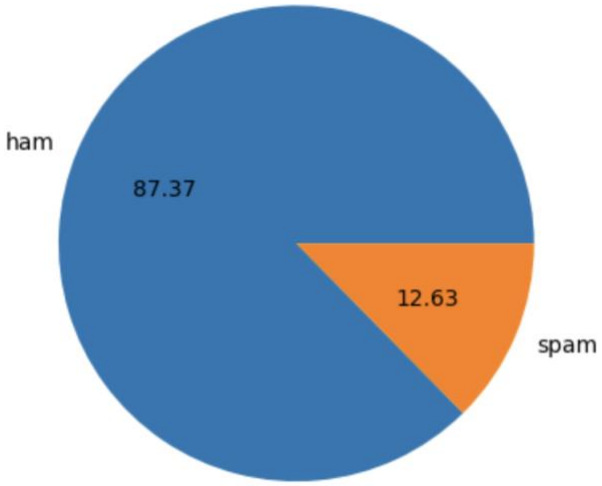
| | label | message |
|---|---|---|
| 0 | 0 | Go until jurong point, crazy.. Available only ... |
| 1 | 0 | Ok lar... Joking wif u oni... |
| 2 | 1 | Free entry in 2 a wkly comp to win FA Cup fina... |
| 3 | 0 | U dun say so early hor... U c already then say... |
| 4 | 0 | Nah I don't think he goes to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

Ham = 0 & Spam = 1

After that we further analyzed our data by looking for any null values and duplicates and find out that our data has no missing values but 403 duplicates, we then dropped the duplicates for a better analysis and to make it understandable for non-technical stakeholders. After cleaning our data, we performed a basic EDA since we have only 2 columns.

We created some visualizations to uncover some insights:

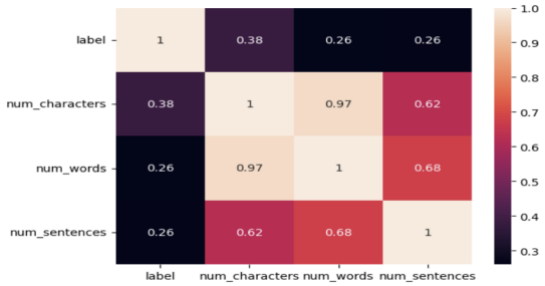
This pie chart shows that our data includes 87.37% ham and 12.63% spam messages.



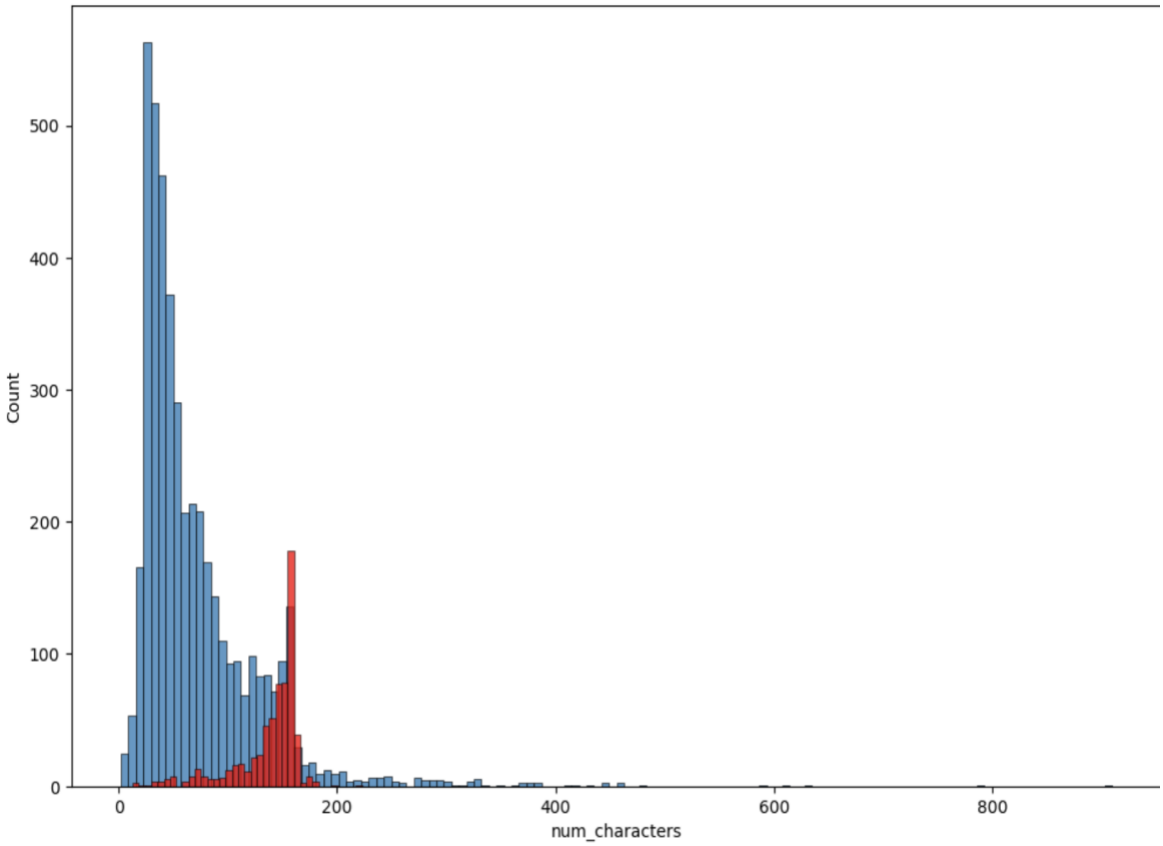

We then created three new columns including Number of Characters, Number of Words and Number of Sentences to see if we can find any further details.

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

We then investigated the stats for the new columns and created a heatmap to find any correlations.



We concluded that Number of Characters are the most correlated feature, so we plotted a histogram to check the trends in Ham and Spam messages. Ham = Blue and Spam = Red. Our Histogram shows that there are 3 times more characters in ham messages than spam.




Then we further Preprocessed our data and performed the following techniques:

- Lower case
- Tokenization
- Removing stopwords
- Stemming

After Preprocessing our data we then moved to Modelling. We created multiple **Naive Bayes** models by using **Bag of Words and Tfidf.**

The evaluation matrices for our models are:

Gaussian Naive Bayes

```
Accuracy: 0.8457399103139014
[[128  18]
 [154 815]]
Precission: 0.978391356542617
```

Multinomial Naive Bayes

```
Accuracy: 0.9883408071748879
[[128  18]
 [154 815]]
Precission: 0.9907597535934292
```

Bernoulli Naive Bayes

```
Accuracy: 0.9802690582959641
[[125  21]
 [  1 968]]
Precission: 0.9787664307381193
```

Out of these three we got the best results of 98% Accuracy and 99% Precision from Multinomial Naive Bayes model using CountVectorizer with Max_features = 3000.

We then trained a **Random Forest classifier** by creating **Word2vec** from scratch and got the following evaluation.

```
Accuracy: 0.9991007194244604
[[146   1]
 [  0 965]]
Precission: 0.9989648033126294
              precision    recall  f1-score   support

           0       1.00      0.99      1.00       147
           1       1.00      1.00      1.00       965

    accuracy                           1.00      1112
   macro avg       1.00      1.00      1.00      1112
weighted avg       1.00      1.00      1.00      1112
```

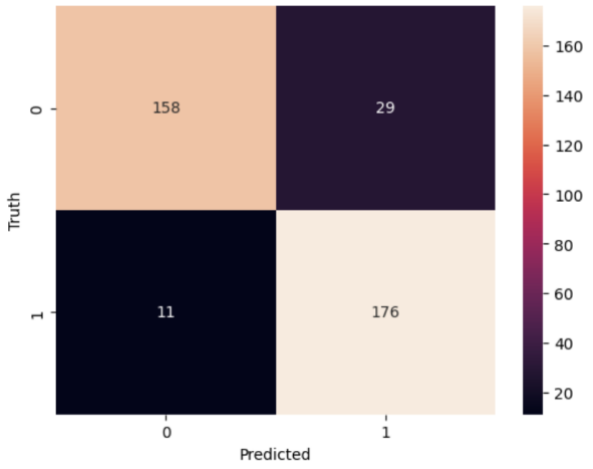
We Got some amazing results from our Random Forest model by training a word2vec from scratch. We got an accuracy of 99% and a precision of 99%.

For in depth comparison, we then created a Single Layer **Neural Network** using **Tensorflow BERT Pretrained Model**. BERT and other Transformer encoder architectures have been wildly successful on a variety of tasks in NLP. BERT models are usually pre-trained on a large corpus of text, then fine-tuned for specific tasks.

Following is the Evaluation Metric for our Neural Network using BERT:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.93 | 0.84 | 0.89 | 187 |
| 1 | 0.86 | 0.94 | 0.90 | 187 |
| accuracy | | | 0.89 | 374 |
| macro avg | 0.90 | 0.89 | 0.89 | 374 |
| weighted avg | 0.90 | 0.89 | 0.89 | 374 |

Our Neural Network is not doing very well in this case we can further fine tune our BERT pre trained model to get better results.




**Conclusion:**

In this Capstone project, Text Classification for Spam Detection we worked through a series of tasks such as Data Collection, Data Preprocessing, EDA and Modeling. We performed different NLP techniques and created three models:

- Naive Bayes We Created a Naive Bayes model using Bag of Words (BoW), Multinomial Naive Bayes model using CountVectorizer with Max_features = 3000 gave the best results with an Accuracy of 98% and Precision of 99%.
- Random Forest We Got some amazing results from our Random Forest model by training a word2vec from scratch. We got an Accuracy of 99% and a Precision of 99%.
- Neural Network We created a single layer neural network using Bert embedding ( pre-trained model ). Didn't get very good Precision score as we're looking for precision here.
-

So our best model with a Precision and Accuracy of 99% is Random Forest using Word2vec.