**COMP237 Section 001 – Introduction to AI – NLP Project**

Group No: \_\_2\_\_

List Team Members:

1. Niyanta Siddhapura -301144167

2. Vaishali Siddeshwar - 301172372

3. Shivam Verma - 301166932

4. Gagandeep Singh - 301146293

5. Divyanshu Johar - 301149021

Contents

[**1. Introduce Data** 2](#_heading=h.gjdgxs)

[**2. Data Analysis** 3](#_heading=h.30j0zll)

[**2.1 Analyzing feature- COMMENT\_ID** 3](#_heading=h.1fob9te)

[**2.2 Analyzing Feature –CLASS** 3](#_heading=h.3znysh7)

[**2.3 Analyzing Feature – AUTHOR** 4](#_heading=h.2et92p0)

[**2.4 Analyzing Feature – DATE** 4](#_heading=h.tyjcwt)

[**2.5 Analyzing Feature –CONTENT** 5](#_heading=h.3dy6vkm)

[**2.6 Analyzing length of Content Text** 5](#_heading=h.1t3h5sf)

[**3. Modelling and Training** 9](#_heading=h.4d34og8)

[**4. Model Evaluation** 9](#_heading=h.2s8eyo1)

[**5. Conclusion** 9](#_heading=h.17dp8vu)

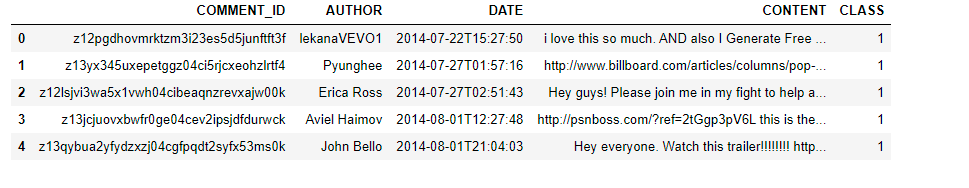
# **1. Introduce Data**

The data has **350** rows and **5** features.

The below table summarizes the data types of features.

|  |  |  |
| --- | --- | --- |
| Column Name | Data Type | # of Missing values |
| COMMENT\_ID | String | 0 |
| AUTHOR | String | 0 |
| DATE | Date | 0 |
| CONTENT | String | 0 |
| CLASS | 1. or 1 | 0 |

Below is the screenshot of the data



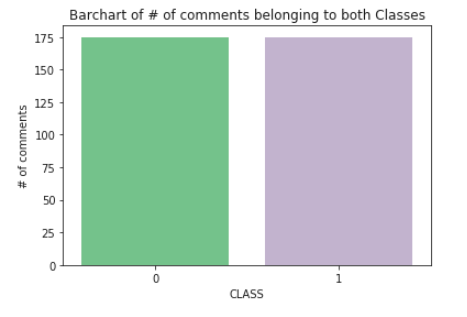
# **2. Data Analysis**

## **2.1 Analyzing feature- COMMENT\_ID**

COMMENT\_ID is a unique identifier for each row. So, no inferences could be made.

## **2.2 Analyzing Feature –CLASS**

There are 2 classes in the data – 0 that indicate NOT SPAM and 1 that indicates SPAM class. Below chart shows the distribution of the predictor variable-CLASS

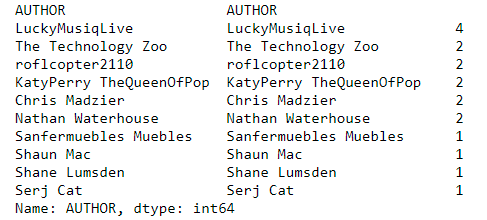


From the above chart we can see that data consists of 175 instances of SPAM class ND 175 instances of NOT SPAM class. Therefore, our dataset is “Balanced”.

## **2.3 Analyzing Feature – AUTHOR**

Most of the authors in the data are unique. There are 342 unique authors who have commented on the videos.

Next, we will group the data by Authors to get the name of authors who have commented the highest.

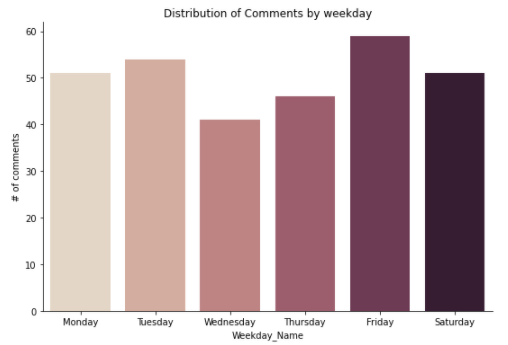


Following are our findings-

* Author with name 'LuckyMusiqLive' has commented 4 times which is the highest for this dataset.
* Only one user has given four comments and 5 users have 2 comments each. The remaining ~98% of comments have left only 1 comment.

## **2.4 Analyzing Feature – DATE**

We will now see on which weekday people commented the most for Katy Perry’s videos.



As the general norm, people commented more during Fridays and Saturday but very less on Wednesday.

## **2.5 Analyzing Feature –CONTENT**

We saw that there are 348 unique comments in the dataset. There are only 3 duplicate comments in the dataset. These 3 duplicate comments belong to class 0 and are created by different Author.

## **2.6 Analyzing length of Content Text**

Let us how the length of comments is distributed in the data.

|  |  |
| --- | --- |
|  |  |

From the above statistics we can see that 75% of the comments have a length of 116 characters. Few comments also have just 4 characters. Also, some comments are around 1200 characters long.

Now, let us see which comments are more very long and which class they belong to.



Now we can see that these long comments belong to SPAM class and they are filled with URL and special characters.

This makes us question, if there is any difference in the length of content between SPAM and NOT SPAM classes. Let us plot the histogram of their lengths together.

|  |  |
| --- | --- |
|  | Statistics for Length of SPAM Content |
| Statistics for Length of NOT SPAM Content |

**From the above screenshots, we can confirm that in our data, SPAM Class have comments that are longer(their median length is 99 characters) than the NOT SPAM Class (their median length is 58 characters).**

Here is the WordCloud for Content feature



Let us now check the most commonly occurring uni/bigrams in the CONTENT of SPAM and NON-SPAM Class.

|  |  |  |
| --- | --- | --- |
|  | |  |
|  |  | |

We can infer that CONTENT in SPAM class mostly have only URLS. That’s why, words such as http, www, amp, facebook etc are the most frequent uni/bigrams in SPAM. On the other hand, the most frequent uni/bigrams of NOT SPAM class are Katy Perry, song, love and video. (related to Katy Perry’s songs).

# **3. Modelling and Training**

**3.1 Data Pre-processing**

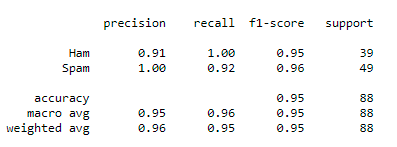
Firstly, we began with analysing the input file manually and found that out of 5 columns, two columns can be primarily used for data modelling, which are “Content” and “Class”. We identified that the “Class” column having flag 0 is HAM comment whereas “Class” column having flag 1 is SPAM comment.

Then we found that SPAM comments contained around **100 words** while HAM comments had less than **50 words**. Next, SPAM comments contained email IDs, URLs, special characters and white spaces.

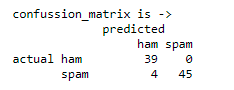
We then removed stopwords using nltk followed by the stemming of data. Finally, the remaining data was converted to lower case for further processing.

# **4. Model Evaluation**

After preprocessing the dataset, we used Naive Bias multinomial classifier model for prediction. Below mentioned is the classification report of the trained model along with the predicted result.



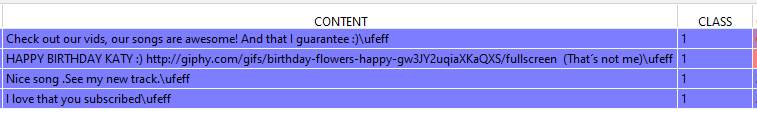
Below mentioned is the confusion matrix of the predicted output.



Below are our findings:

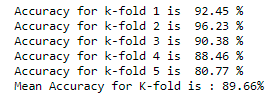
* Number of False Negatives (i.e. Ham predicted as Spam ) and False Positives(Spam predicted as Ham) are very low.
* High Precision and Recall of 0.91 and 1 respectively.
* Therefore this is a **good model.**

These are the samples of FN, which are SPAM but were predicted as 0. Because they are either missing the URL or www , and our model expected SPAM to have a URL.



Below is our hypothesis for

Below mentioned is the output of cross validation using 5-folds

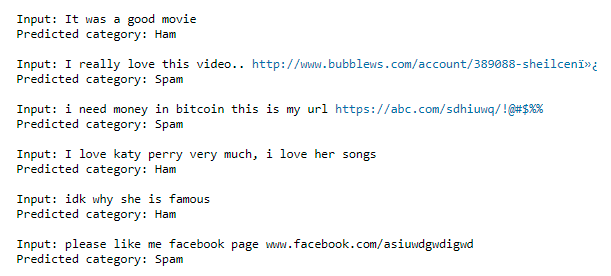


Below mentioned is the overall accuracy of the model.



The Accuracy of the model is consistent across the 5 folds. Therefore there is **no** sign of overfitting.

Here are model’s predictions on few samples.



# **5. Conclusion**

* Spam Comments are longer with URLs & special characters.
* A Naïve Bayes Model predicts Spam/Ham comments with accuracy of ~95%.