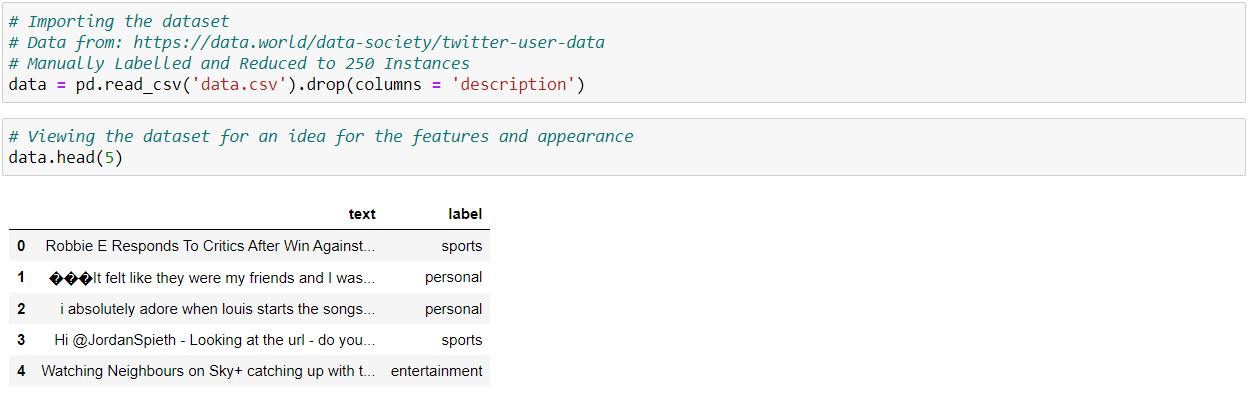
Shayma Report

# Code Progress

The code for this project is tasked with performing three different and crucial processes for the research. These processes are as follows:

## Data Pre-processing

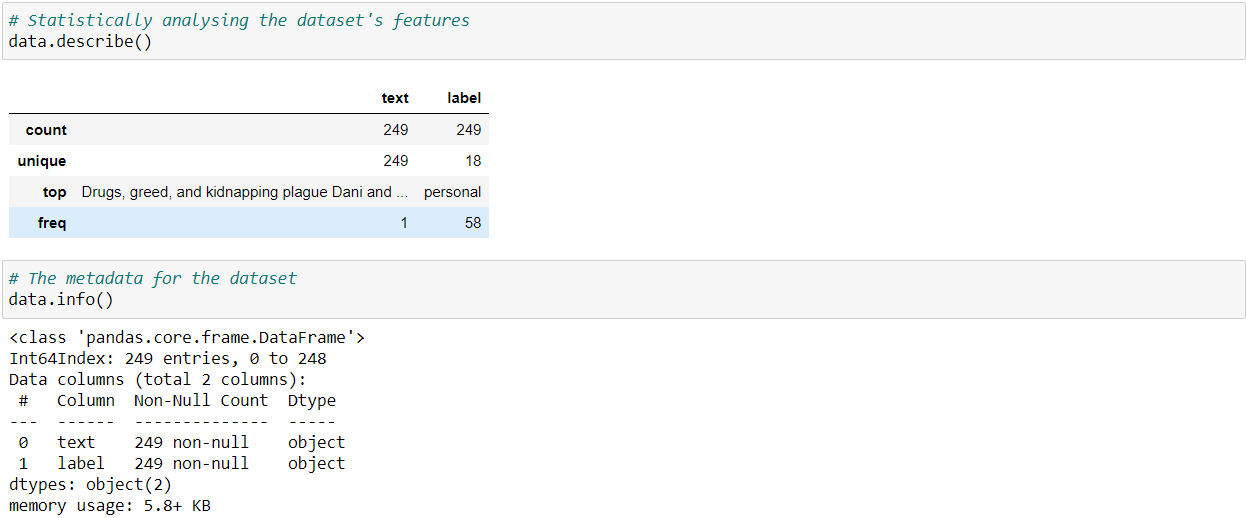
This section performs an in-depth analysis of the dataset. The in-depth analysis includes exploring the dataset’s metadata to study the properties of the dataset and the features in it. Then comes the data cleaning part where all the duplicate and null instances are removed from the dataset. Care needs to be taken here since carelessly removing instances from the data can result in the loss of the quality of prediction and the integrity of the dataset as well.



The figure above shows the importing of the dataset into the code.



The next step is described in the figure above where the dataset is rid of all the duplicate and null instances. There are no more duplicate and null values in the dataset. The next step would be to understand the dataset from a statistical point of view and also look at the metadata of the dataset.



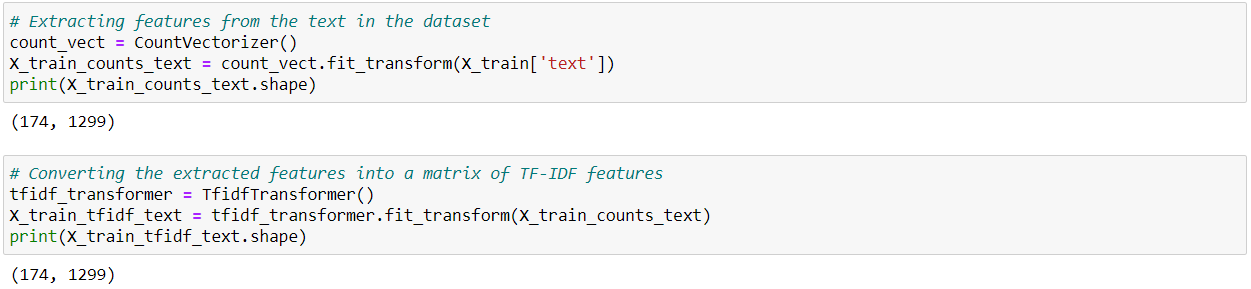
The central tendency measurements for each feature in the dataset along with the data types and the number of non-null values in each feature are seen here. They provide good insight into understanding the dataset. The next step is to split the dataset into a training and testing dataset for cross-validation. After the split, the NLP feature extraction will be applied to it in the next part.



This completes the data pre-processing part.

## Natural Language Processing

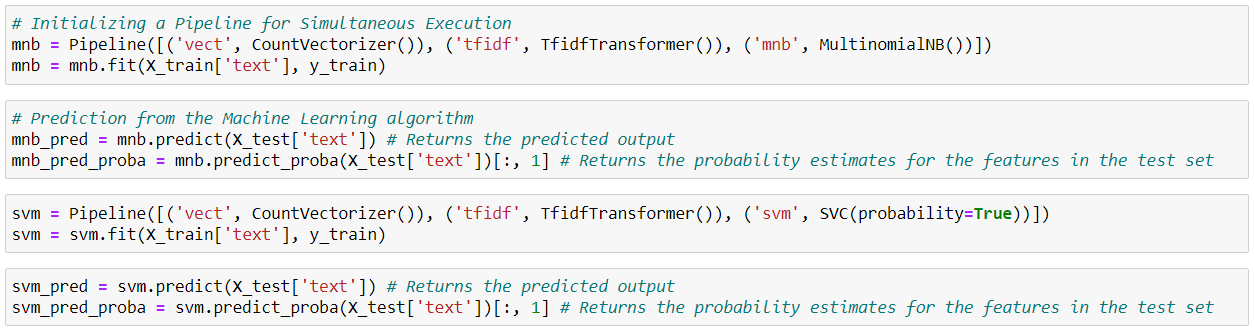
This project aims to analyze the tweets and predict the niche of the tweet. For this purpose, the dataset used here contains tweets in text form. Textual data of this kind cannot be subjected to Encoding since there will be a huge loss in the data features and the prediction results will be extremely unsatisfactory. Therefore, this research opts for applying Natural Language Processing techniques for extracting useful features from the ‘text’ and converting it into a bag of words for each output class that can then be used to train algorithms for multiclass classification. For feature extraction, the two techniques this research employs are the Count Vectorizer and the TF-IDF Transformer. The count vectorizer extracts the useful keywords from the ‘text’ and converts them into a bag of words. The tf-idf transformer converts the bag of words into a tf-idf matrix.



This completes the Natural Language Processing Part.

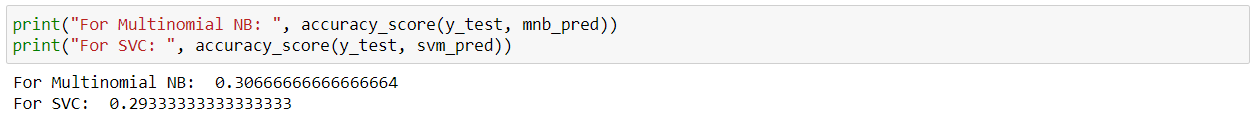
# Machine Learning

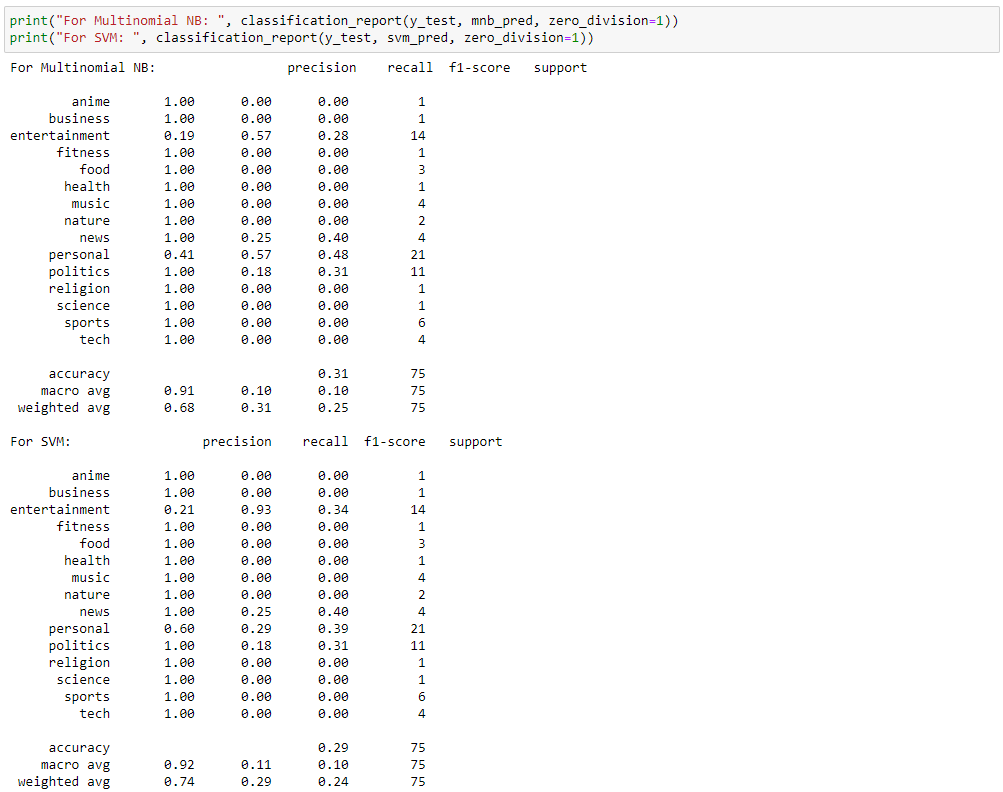
In this part of the code, two multiclass classification algorithms will be applied and their performances will be compared and pondered upon. The algorithms chosen for this research are the Multinomial NB and the SVM. A pipeline is used to apply the Natural Language Processing and the Algorithm training simultaneously.



## Performance Metrics

The accuracy of both the algorithms is measured using the performance metrics provided by scikit-learn for classification algorithms.





The accuracy for Multinomial NB is ≈ 31% and the SVM performs with an accuracy of ≈ 30%. The algorithms have trained well and have performed to the maximum of their abilities. The reason behind the low accuracy is the dataset.

Originally, the dataset contained 20,000 tweets from different users. The issue was that the dataset wasn’t labelled according to this project’s requirements. This is because the domain topic for this research is highly specific and there isn’t any dataset that can fulfil the requirements for this research topic. Therefore, we had to resort to manual labelling. Since, there is limited time for the project, only 250 of the instances could be labelled and used. Now 250 instances are a decent size for a dataset, but the real issue lies in the number of instances per output class, which is very low. As it can be seen from the classification report, the accuracy for predicting the tweets for each output class is very low due to very little data available. Therefore, the overall aggregate has also fallen. The results will be much better if the dataset provides at least 150 instances per output class.