TEAM-3

TWEET SENTIMENT ANALYSICS

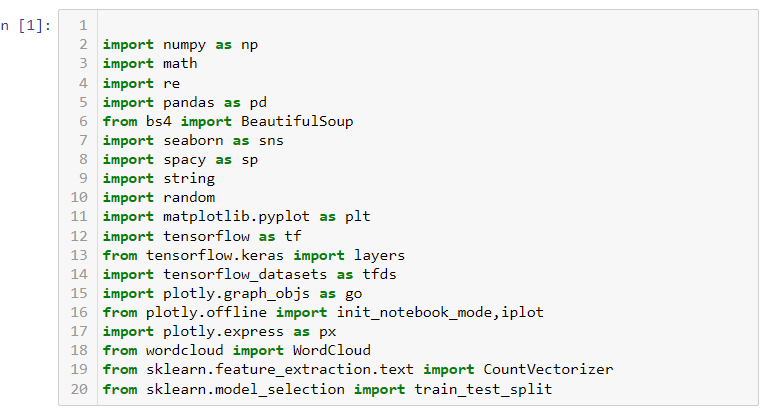
**What is business Problem we are solving :**

This is first think we should ask to ourselves. Because we can achieve our goal. With all of the tweets circulating every second it is hard to tell whether the sentiment behind a specific tweet will impact a company, or a person’s, brand for being viral (positive), or devastate profit because it strikes a negative tone. Capturing sentiment in language is important in these times where decisions and reactions are created and updated in seconds. But, which words actually lead to the sentiment description.

**Our approach to this problem:**

What we have done in this project is not to jump directly to large/complex models. we have gone from the simplest model to the complex model just to understand how Machine Learning / Deep Learning models are behaving and giving me results.

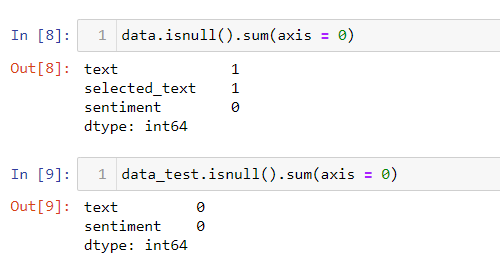
**Import from Libraries :**

****

**Uploading files :**

****

**Check duplicate and null value :**

****

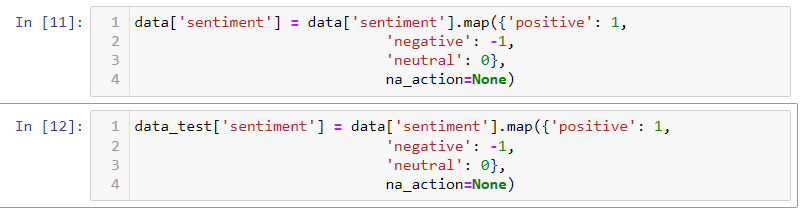
Observation :

There was one missing value present which being removed.

**Distribution of Positive, Negative and neutral (difference between text and selected text)**

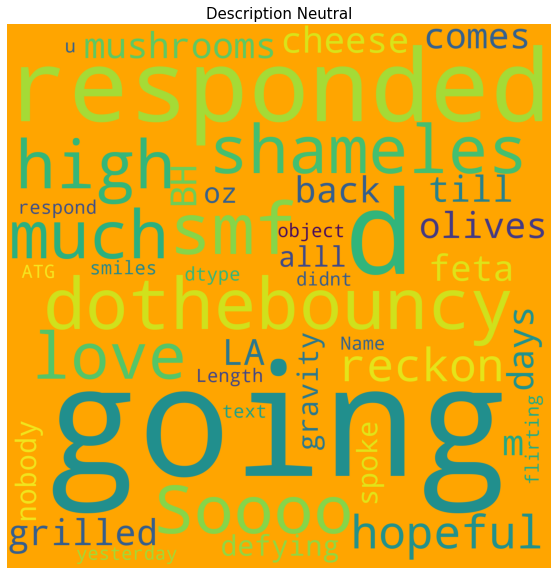
Here we observed that there are three types of sentiments, So we have labelled them as follows :

* positive = 1
* Negative = -1
* Neutral = 0



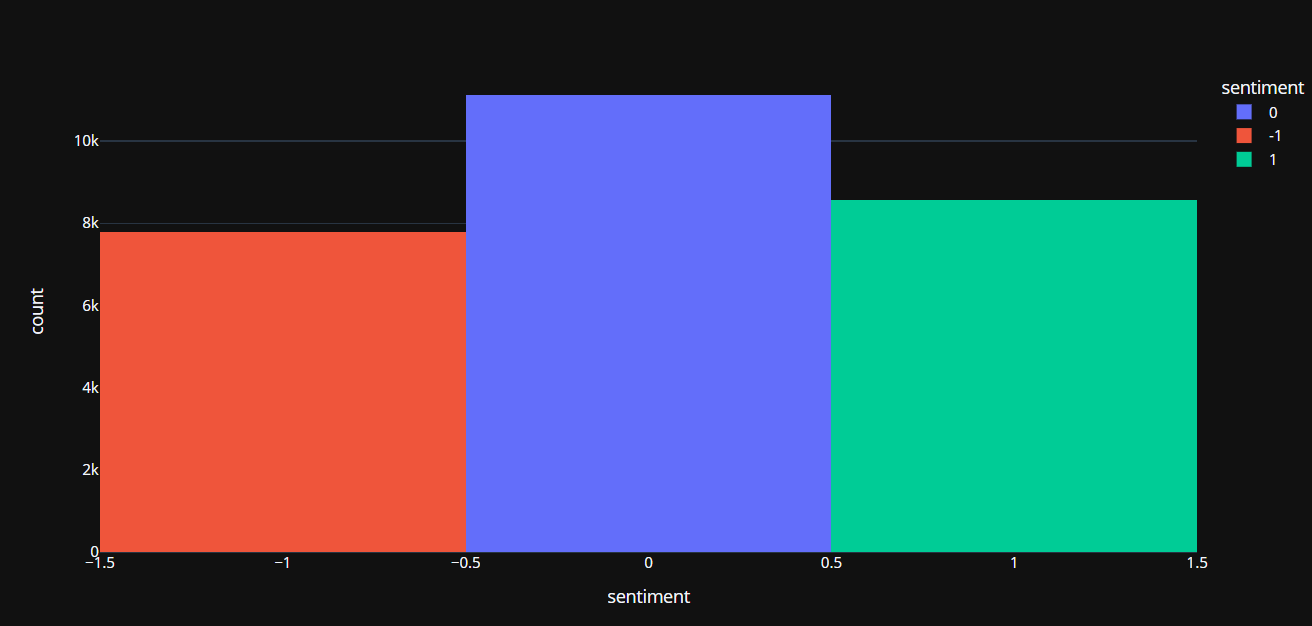
**Frequency of words (let’s build a word cloud) :**



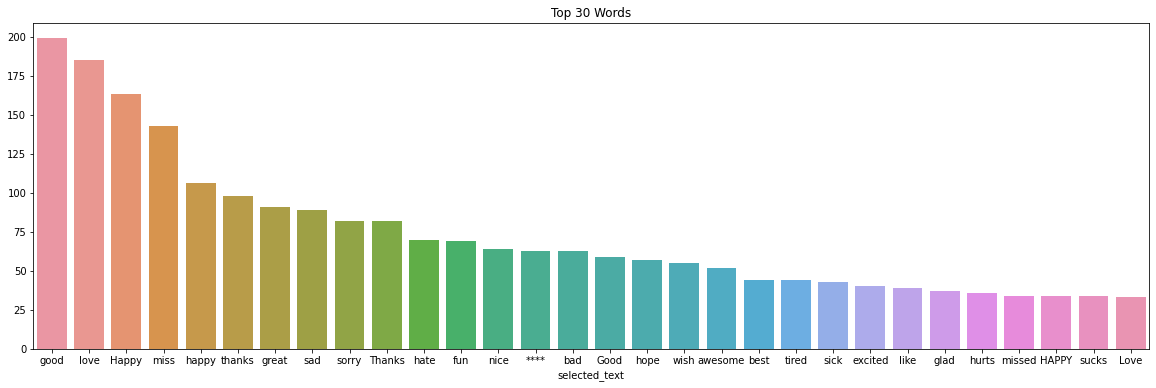


**Observation :**

If you observe the positive word cloud some words occurring more like good , fun, love , mother , great , amazing these are indication of positive tweet in negative tweet we will find don’t, can’t, miss ,sad , sick ,sucks, etc. in neutral there are mix of words which are sometime look like negative and sometime positive word but which can’t be decide based on only word , we have to look at the sentences. Let’s look at the count , that how many time these word are occurring.



**Most Frequent Word :**

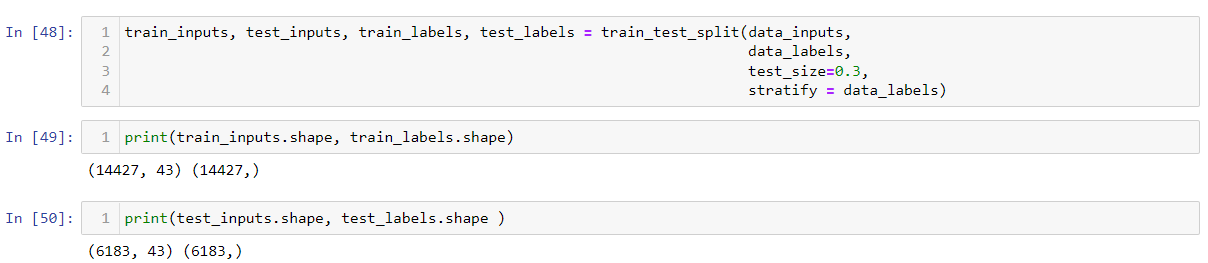


Observation :

Word like good , love, mother , happy , great ,thanks are occurring more on positive tweets not , sad , sorry, sick ,hate are occurring more on negative sentiment.

**Data split :**

In machine learning, data splitting is typically done to avoid overfitting. That is an instance where a machine learning model fits its training data too well and fails to reliably fit additional data. The original data in a machine learning model is typically taken and split into three or four sets.



**Data Cleaning :**

Data cleaning is the process of preparing data for analysis by removing or modifying data that is incorrect, incomplete, irrelevant, duplicated, or improperly formatted.

Here we are doing some pre-processing of data.

* convert to lower case
* remove text in square brackets
* remove links
* remove punctuation
* remove words containing numbers

**Tokenization:**

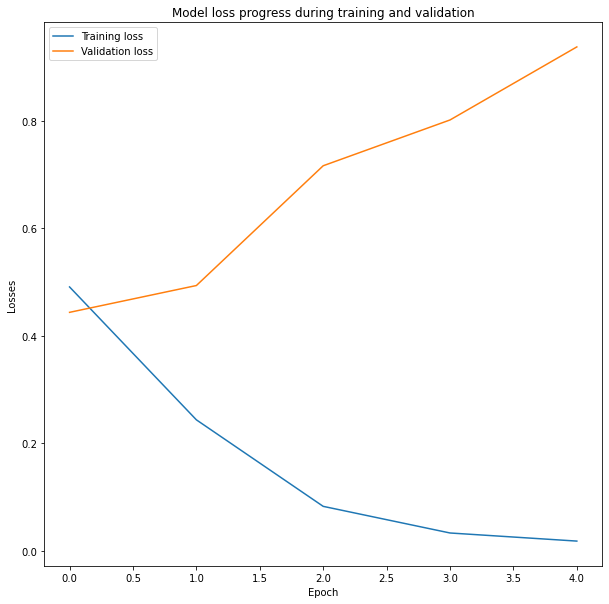
Tokenization is breaking the raw text into small chunks. Tokenization breaks the raw text into words, sentences called tokens.

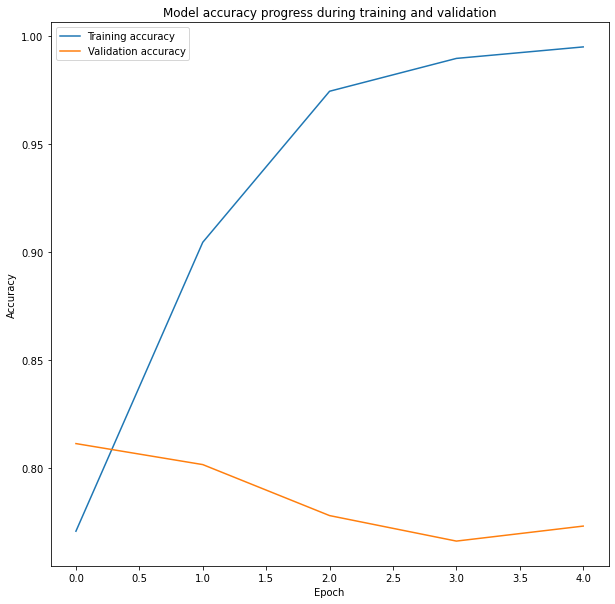


**Model building :**

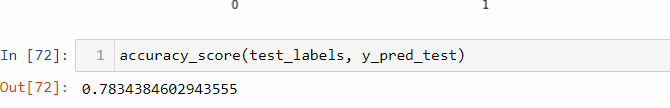
Here we have use CNN (Convolutional Neural Network) Model.

CNN is a type of artificial neural network, which is widely used for image/object recognition and classification. Deep Learning thus recognizes objects in an image by using a CNN.



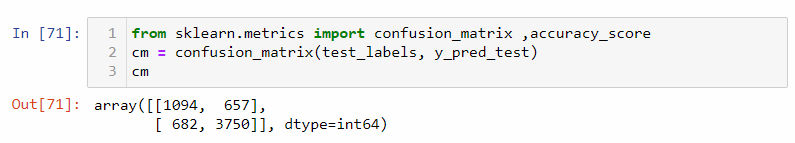


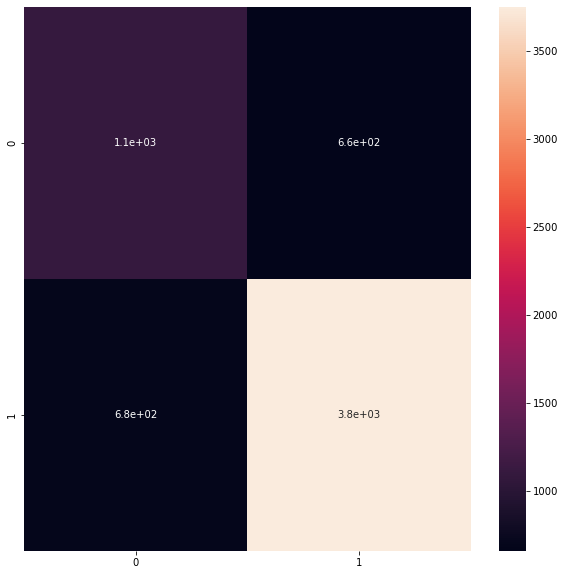
**Accuracy of the Model :**



Here we got 78.34 % accuracy.

**Confusion Metrics :**

****

****

**Conclusion :**

There is a small difficulty for the algorithm to identify the negative words, as there is a balance between the neutral and positive classes, the algorithm often understands the neutral as positive, as there is no negative word for it to relate to neutrality.

Perhaps if we had a larger database for training the algorithm we could improve it with more vocabulary words, certainly the hit margin will be much better with a broader learning of the database, even so we had a great result with some adjustments to the code.