**Project Name:**

Identify the Sentiments­­

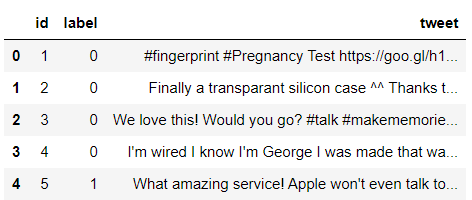
**Problem Statement:**

The problem in identify the sentiments is classifying the polarity of tweet dataset. whether the expressed opinion in a document, a sentence or an entity is positive or negative.

**Methodology:**

1. Load the Dataset
2. Data Cleaning
3. Building a Vocabulary
4. Divide the dataset
5. Vectorized our Dataset
6. Handle imbalanced Dataset
7. Training and evaluation the models
8. Save the model
9. Build a simple web application
10. **Load the Dataset**

Load the dataset which includes id, tweet, and the corresponding label



Initial dataset shape: (7920, 3)

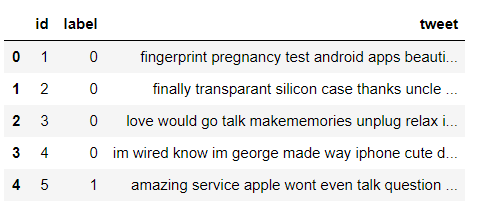
1. **Data Cleaning**

Dataset haven’t any null values and duplicate values.

Here we are working with text data. So, we can follow different text preprocessing techniques such as,

* Convert uppercase to lowercase
* Remove links
* Remove punctuation marks
* Remove numbers
* Remove stopwords
* Stemming and Lemmatization

After those preprocessing steps, dataset looks like,

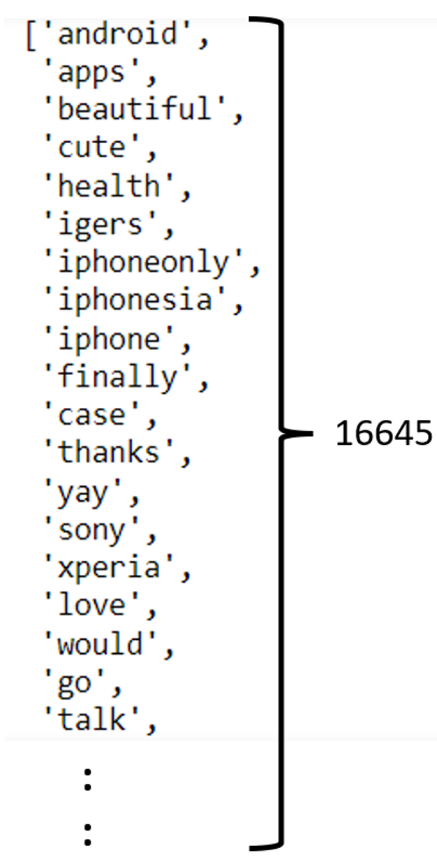


**Note:** We can see, all tweets are in lowercase. Also, there are not any punctuation marks, numbers in the resulted tweets.

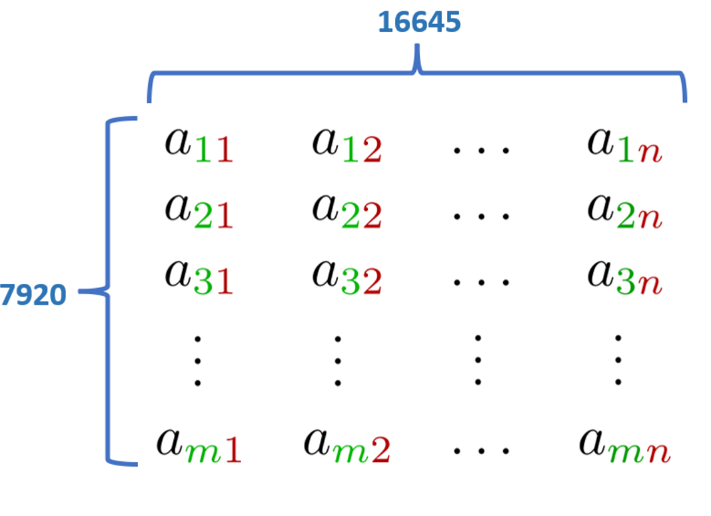
1. **Building a Vocabulary**

Before further next step, we need to build a vocabulary using all tweets that we have. To do that, we used **Counter** class which is in python collections. It gives us unique words along with frequencies.

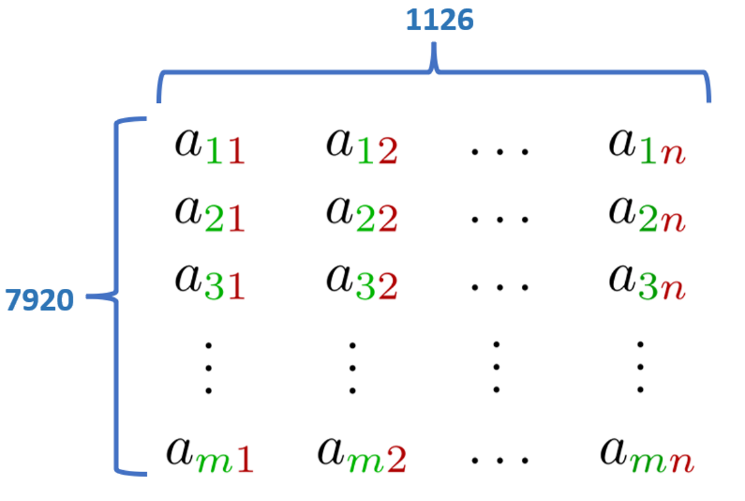
Initially our vocabulary has 16645 terms.



But we have only 7920 instances for training. Then most probably it has **overfitting** **problem** when we are training dataset like below.



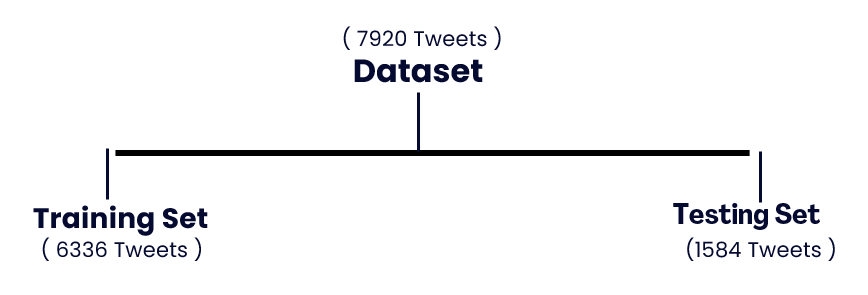
So, to reduce our vocabulary size, we consider terms that are more than 10 occurrences. That’s how we reduce the vocabulary size to 1126.



1. **Divide the dataset**

We need to divide our whole dataset into training and testing. For that, we used **train\_test\_split** method in **sklearn** library. Here we divide dataset with 5:1 ratio.

So, finally we got 6336 tweets for training purpose and 1584 tweets for testing purpose.



1. **Vectorized our Dataset**

Now we need to vectorized whole dataset using the vocabulary which we build earlier.

For example, it happened for each tweet like below.

**Sentence:** Python is a general-purpose programming language

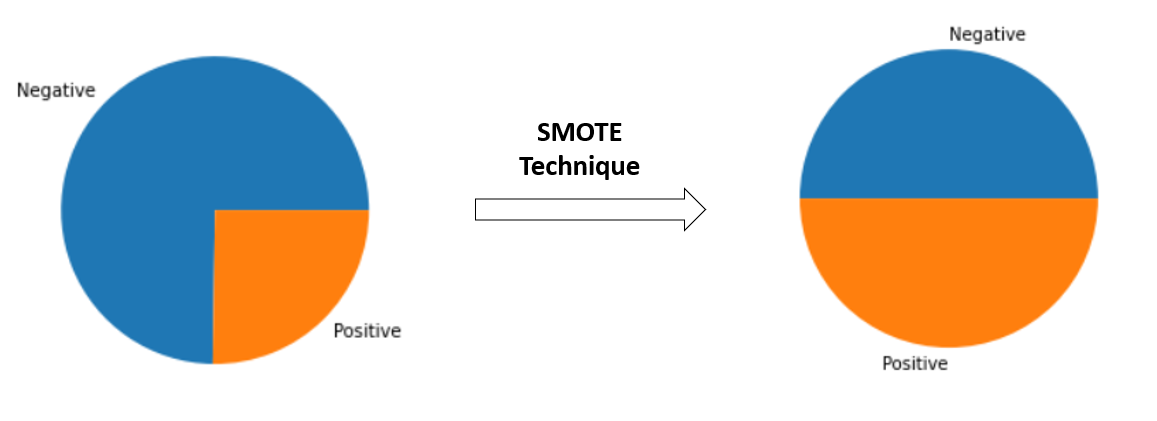
**Vocabulary:** [ I, love, Python, programming, language, is, a, popular ]

**Vectorized form:** [ 0, 0 , 1 , 1 , 1 , 1 , 1 , 0 ]

|  |  |
| --- | --- |
| **Before Vectorization** | **After Vectorization** |
|  | A picture containing scatter chart  Description automatically generated |

1. **Handle imbalanced Dataset**

Here we are working with an imbalanced dataset. Because when we considering value counts in labels, we found there are 4715 negative tweets and 1621 positive tweets. Then we try to oversample the dataset using SMOTE technique.



Then we able to get same number of negative and positive data in the training set.

1. **Training and evaluation the models**

Here we used few different machine learning algorithms to implement sentiment analysis model.

1. Logistic Regression
2. Naïve Bayes Classifier
3. Random Forest Classifier
4. Gradient Boosting Classifier
5. Support Vector Classifier

Also, we used different performance measures such as accuracy, precision, recall and f1-score for evaluate our models.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Accuracy | Precision | Recall | F1-score |
| Logistic Regression | 0.87 | 0.87 | 0.87 | 0.77 |
| Naïve Bayes | 0.86 | 0.86 | 0.86 | 0.77 |
| Random Forest | 0.87 | 0.87 | 0.87 | 0.74 |
| Gradient Boosting | 0.88 | 0.88 | 0.88 | 0.77 |
| Support Vector Machine | 0.88 | 0.88 | 0.88 | 0.78 |

Evaluate model using ROC Curve

|  |  |
| --- | --- |
| Training Phase | Testing Phase |
|  |  |

1. **Save the model**

* After checking the model performance, we saved our **Random Forest** model as a **pickle file** to further use. (“model\_pkl”)
* Also, we saved our **vocabulary** where we build using training dataset. (“vocabulary.txt”)

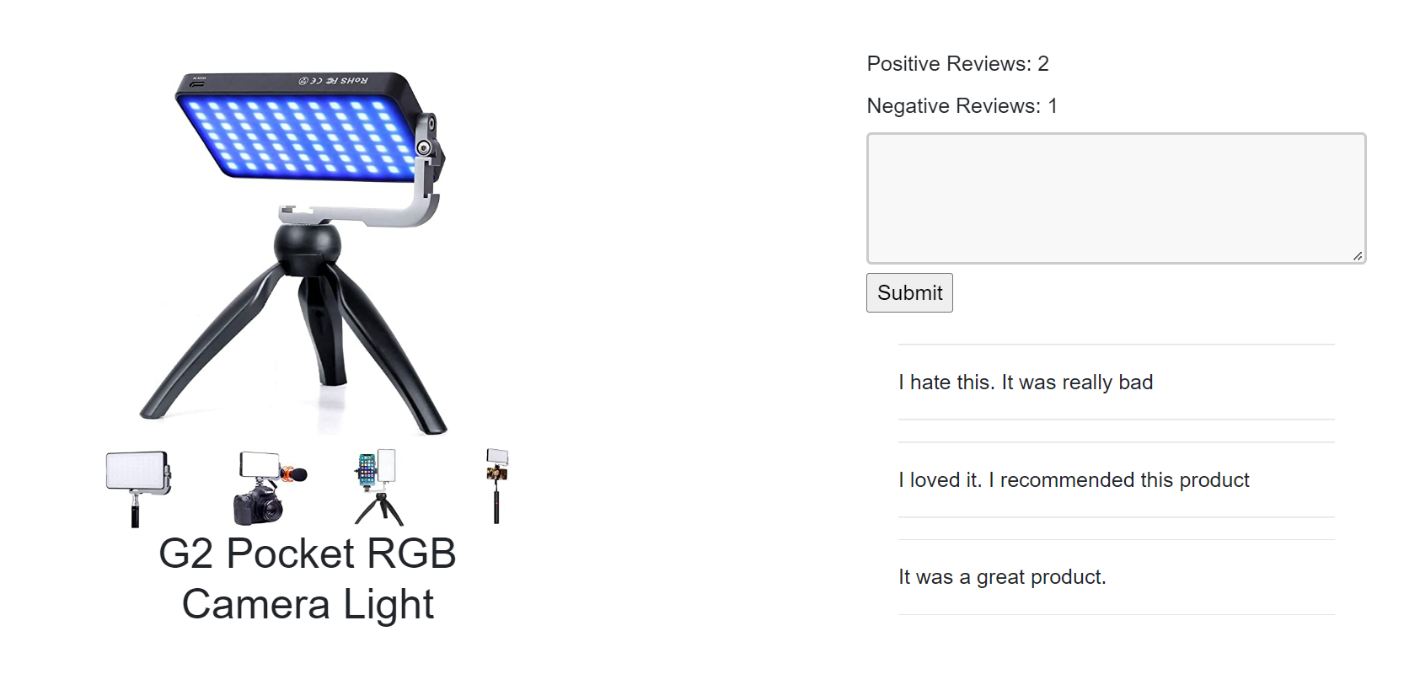
1. **Build a simple web application**

We built a web application using Flask. The application uses natural language processing techniques to classify the sentiment of the input as positive or negative.

In the Flask server, we built a simple procedure to identify sentiments with,

1. Load the model
2. Get user input
3. Text preprocessing
4. Vectorization
5. Get prediction

Then it will generate an output. So, we can demonstrate it like below,



Here we can see the number of positive and negative reviews according to our comments.

**Application link:** https://github.com/dineshpiyasamara/sentiment\_analysis\_application