SENTIMENT ANALYSIS PROJECT REPORT

USING PYTHON

(Harshitha Sakamuri - 801314903)

Contents

1. Introduction

Social media platforms are an excellent source of data, particularly for businesses and organizations interested in learning how their products or services are regarded by their customers. Twitter is one of the most popular social media sites, with millions of users and millions of tweets sent every day. Tweet sentiment analysis is a popular task in natural language processing that seeks to classify a tweet's sentiment as positive and negative. In this project, I used Python and machine learning techniques to perform sentiment analysis on tweets.

1.1. Objectives:

This project is to build a sentiment analysis model for tweets related to UVA university in the US. The project will involve the following steps:

Web scraping: Collecting 1500 unique from at least three hashtags for the university they have opted using the Tweepy library and the Twitter API.

Preprocessing: Cleaning and removing unnecessary characters such as punctuations and stop words from the raw text data.

Feature extraction: Creating a matrix of word counts for each tweet and using TF-IDF to weigh the importance of each word.

Model training: Implementing a sequential neural network model using embedding layer, LSTM, CNN, GRU, and dense output layers, and training it using binary cross-entropy loss and the Adam optimizer.

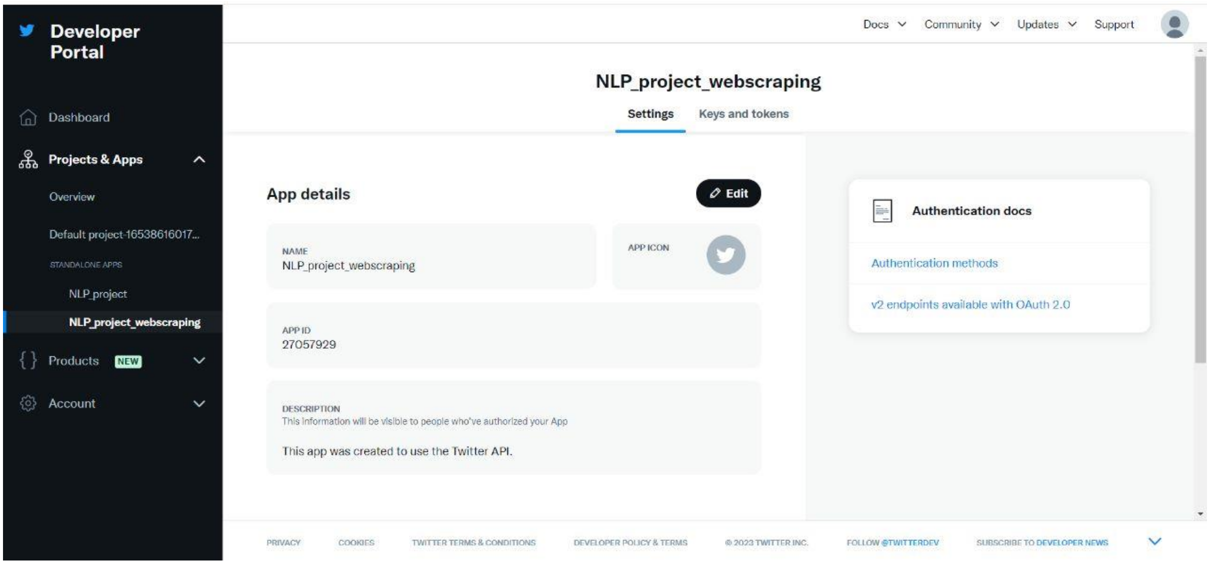
Model evaluation: Evaluating the model's performance on the testing set using metrics such as accuracy, precision, recall, and F1-score, and visualizing the prediction metrics using confusion matrix.

Predicting the sentiment of the sentence using the trained neural network model.

2. Data Collection

The data for this project was collected using Tweepy library and the Twitter API. I scraped the tweets from University of Virginia. I have collected around 2500 unique tweets from more than 6 hashtags like #UVA, #UVAAlumni, #UVALAW, #UVAHealth and so on.

To use the Twitter API, I have created the developer account, created the app and taken access token, access token secret, api key, and api secret in the developer dashboard. The Tweepy library was then used to establish a connection to the Twitter API and scrape the necessary tweets.



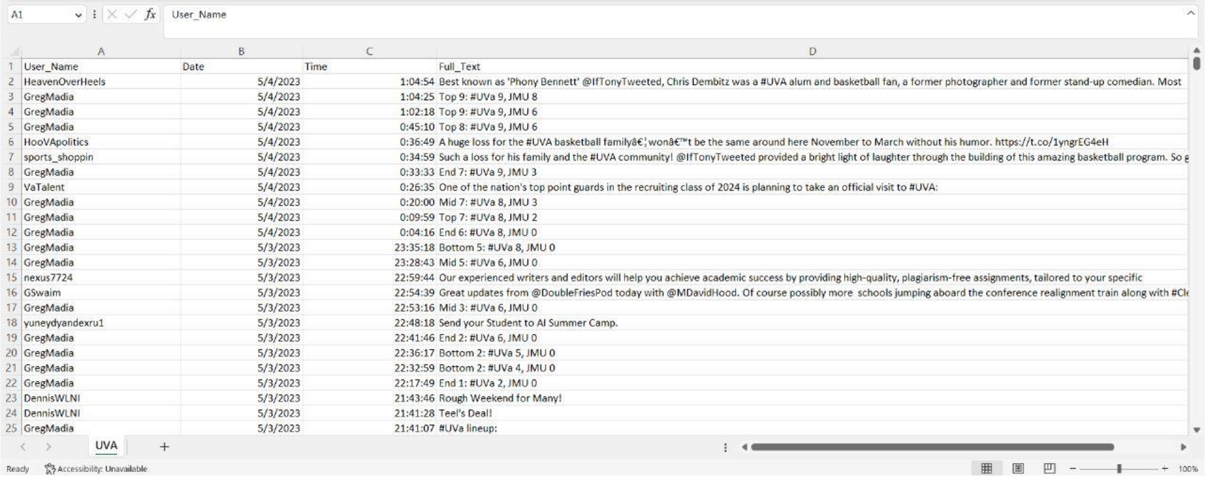
NAME
NLP_project_webscraping

APP ID
27057929

DESCRIPTION
This information will be visible to people who've authorized your App

This app was created to use the Twitter API.

Figure 2.1: Twitter API developer portal.



	A	B	C	D
1	User_Name	Date	Time	Full_Text
2	HeavenOverHeels	5/4/2023	1:04:54	Best known as 'Phony Bennett' @IfTonyTweeted, Chris Dembitz was a #UVA alum and basketball fan, a former photographer and former stand-up comedian. Most
3	GregMadia	5/4/2023	1:04:25	Top 9: #UVa 9, JMU 8
4	GregMadia	5/4/2023	1:02:18	Top 9: #UVa 9, JMU 6
5	GregMadia	5/4/2023	0:45:10	Top 8: #UVa 9, JMU 6
6	HooVApolitics	5/4/2023	0:36:49	A huge loss for the #UVA basketball familyâ€¦wonâ€™t be the same around here November to March without his humor. https://t.co/1yngrEG4eH
7	sports_shoppin	5/4/2023	0:34:59	Such a loss for his family and the #UVA community! @IfTonyTweeted provided a bright light of laughter through the building of this amazing basketball program. So g
8	GregMadia	5/4/2023	0:33:33	End 7: #UVa 9, JMU 3
9	VaTalent	5/4/2023	0:26:35	One of the nation's top point guards in the recruiting class of 2024 is planning to take an official visit to #UVA:
10	GregMadia	5/4/2023	0:20:00	Mid 7: #UVa 8, JMU 3
11	GregMadia	5/4/2023	0:09:59	Top 7: #UVa 8, JMU 2
12	GregMadia	5/4/2023	0:04:16	End 6: #UVa 8, JMU 0
13	GregMadia	5/3/2023	23:35:18	Bottom 5: #UVa 8, JMU 0
14	GregMadia	5/3/2023	23:28:43	Mid 5: #UVa 6, JMU 0
15	nexus7724	5/3/2023	22:59:44	Our experienced writers and editors will help you achieve academic success by providing high-quality, plagiarism-free assignments, tailored to your specific
16	GSwaim	5/3/2023	22:54:39	Great updates from @DoubleFriesPod today with @MDavidHood. Of course possibly more schools jumping aboard the conference realignment train along with #Cle
17	GregMadia	5/3/2023	22:53:16	Mid 3: #UVa 6, JMU 0
18	yuneydyandexru1	5/3/2023	22:48:18	Send your Student to AI Summer Camp.
19	GregMadia	5/3/2023	22:41:46	End 2: #UVa 6, JMU 0
20	GregMadia	5/3/2023	22:36:17	Bottom 2: #UVa 5, JMU 0
21	GregMadia	5/3/2023	22:32:59	Bottom 2: #UVa 4, JMU 0
22	GregMadia	5/3/2023	22:17:49	End 1: #UVa 2, JMU 0
23	DennisWLNI	5/3/2023	21:43:46	Rough Weekend for Many!
24	DennisWLNI	5/3/2023	21:41:28	Teel's Deal!
25	GregMadia	5/3/2023	21:41:07	#UVa lineup:

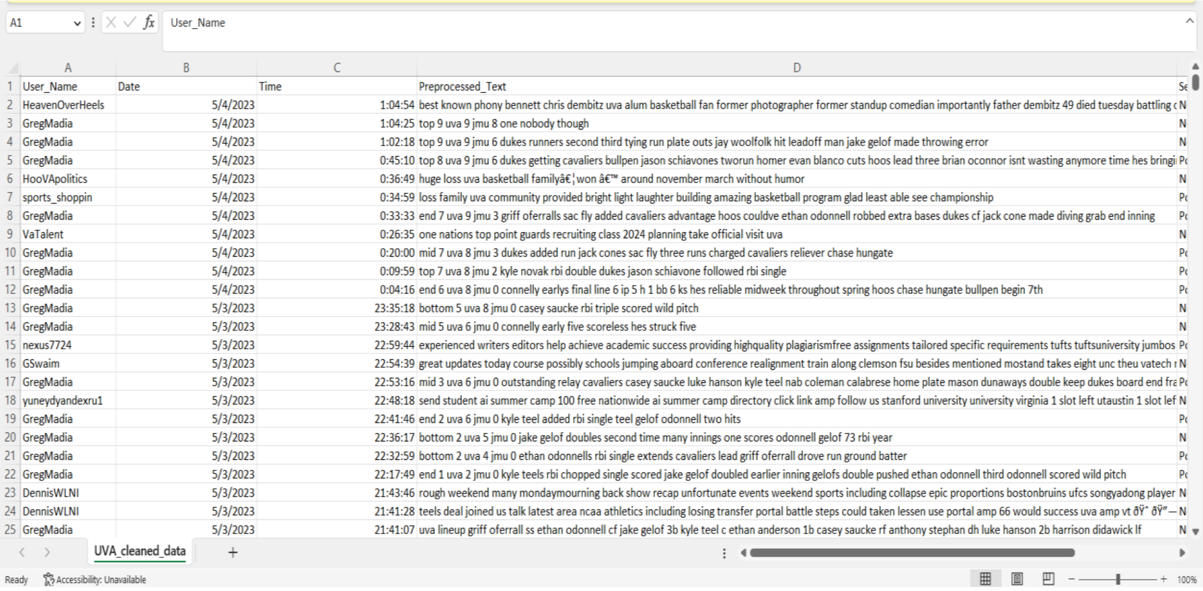
Figure 2.2 : Collected data.

3. Pre-processing

Firstly, I pre-processed the data that uses regular expressions which involves

- Removing URLs and mentions from the text,
- Replaces colons with spaces,
- Removing punctuation,
- Converting the text to lowercase,
- Removing stop words using the NLTK library

- A function clean_text is applied to the 'Full_Text' column of the 'UVA.csv' file using the Pandas apply method to create a new column called 'Preprocessed_Text' in the DataFrame 'df'.
- Duplicates are removed from the 'Preprocessed_Text' column using the Pandas drop_duplicates method, and the DataFrame is reset using the reset_index method.
- The 'Full_Text' column is dropped from the DataFrame using the drop method, and the cleaned DataFrame is saved to a new CSV file called 'UVA_cleaned_data.csv'.



	A	B	C	D	
1	User_Name	Date	Time	Preprocessed_Text	Se
2	HeavenOverHeels	5/4/2023	1:04:54	best known phony bennett chris dembitz uva alum basketball fan former photographer former standup comedian importantly father dembitz 49 died tuesday battling c	N
3	GregMadia	5/4/2023	1:04:25	top 9 uva 9 jmu 8 one nobody though	N
4	GregMadia	5/4/2023	1:02:18	top 9 uva 9 jmu 6 dukes runners second third tying run plate outs jay woolfolk hit leadoff man jake gelof made throwing error	N
5	GregMadia	5/4/2023	0:45:10	top 8 uva 9 jmu 6 dukes getting cavaliers bullpen jason schiavones tworun homer evan blanco cuts hoos lead three brian oconnor isnt wasting anymore time hes bringi	Pc
6	HooVApolitics	5/4/2023	0:36:49	huge loss uva basketball familyâ€¦won â€™ around november march without humor	N
7	sports_shoppin	5/4/2023	0:34:59	loss family uva community provided bright light laughter building amazing basketball program glad least able see championship	Pc
8	GregMadia	5/4/2023	0:33:33	end 7 uva 9 jmu 3 griff oferralls sac fly added cavaliers advantage hoos couldve ethan odonnell robbed extra bases dukes cf jack cone made diving grab end inning	Pc
9	VaTalent	5/4/2023	0:26:35	one nations top point guards recruiting class 2024 planning take official visit uva	N
10	GregMadia	5/4/2023	0:20:00	mid 7 uva 8 jmu 3 dukes added run jack cones sac fly three runs charged cavaliers reliever chase hungate	Pc
11	GregMadia	5/4/2023	0:09:59	top 7 uva 8 jmu 2 kyle novak rbi double dukes jason schiavone followed rbi single	Pc
12	GregMadia	5/4/2023	0:04:16	end 6 uva 8 jmu 0 connelly earlys final line 6 ip 5 h 1 bb 6 ks hes reliable midweek throughout spring hoos chase hungate bullpen begin 7th	Pc
13	GregMadia	5/3/2023	23:35:18	bottom 5 uva 8 jmu 0 casey saucke rbi triple scored wild pitch	N
14	GregMadia	5/3/2023	23:28:43	mid 5 uva 6 jmu 0 connelly early five scoreless hes struck five	N
15	nexus7724	5/3/2023	22:59:44	experienced writers editors help achieve academic success providing highquality plagiarismfree assignments tailored specific requirements tufts tuftsuniversity jumbos	Pc
16	GSwaim	5/3/2023	22:54:39	great updates today course possibly schools jumping aboard conference realignment train along clemson fsu besides mentioned mostand takes eight unc theu vatech r	N
17	GregMadia	5/3/2023	22:53:16	mid 3 uva 6 jmu 0 outstanding relay cavaliers casey saucke luke hanson kyle teel nab coleman calabrese home plate mason dunaways double keep dukes board end fra	Pc
18	yuneydyandexru1	5/3/2023	22:48:18	send student ai summer camp 100 free nationwide ai summer camp directory click link amp follow us stanford university university virginia 1 slot left utaustin 1 slot lef	N
19	GregMadia	5/3/2023	22:41:46	end 2 uva 6 jmu 0 kyle teel added rbi single teel gelof odonnell two hits	Pc
20	GregMadia	5/3/2023	22:36:17	bottom 2 uva 5 jmu 0 jake gelof doubles second time many innings one scores odonnell gelof 73 rbi year	N
21	GregMadia	5/3/2023	22:32:59	bottom 2 uva 4 jmu 0 ethan odonnells rbi single extends cavaliers lead griff oferrall drove run ground batter	Pc
22	GregMadia	5/3/2023	22:17:49	end 1 uva 2 jmu 0 kyle teels rbi chopped single scored jake gelof doubled earlier inning gelofs double pushed ethan odonnell third odonnell scored wild pitch	Pc
23	DennisWLNI	5/3/2023	21:43:46	rough weekend many mondaymourning back show recap unfortunate events weekend sports including collapse epic proportions bostonbruins ufcs songyadong player	N
24	DennisWLNI	5/3/2023	21:41:28	teels deal joined us talk latest area ncaa athletics including losing transfer portal battle steps could taken lessen use portal amp 66 would success uva amp vt ðŸ˜ ðŸ”—	N
25	GregMadia	5/3/2023	21:41:07	uva lineup griff oferrall ss ethan odonnell cf jake gelof 3b kyle teel c ethan anderson 1b casey saucke rf anthony stephan dh luke hanson 2b harrison didawick lf	N

Figure 3.1: Pre-processed data.

4.Feature Extraction

Feature extraction is the process of transforming raw text data into a format that can be used for machine learning models. I have used the CountVectorizer and TfidfVectorizer classes from the scikit-learn library for the feature extraction which reads the cleaned data from the 'UVA_cleaned_data.csv' file.

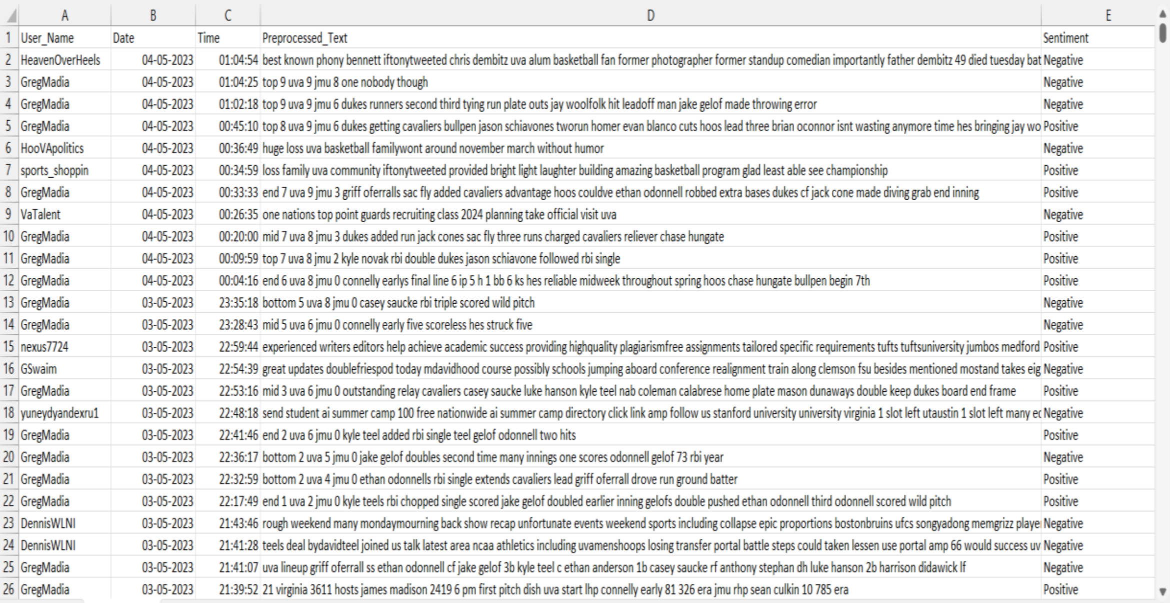
The CountVectorizer and TfidfVectorizer objects are instantiated and fitted to the 'Preprocessed_Text' column of the DataFrame using the fit_transform method.

The resulting count and tf-idf matrices are saved as 'count_matrix' and 'tfidf_matrix', respectively.

Sentiment Labelling:

In order to perform sentiment analysis on the preprocessed text data

- I have used a pre-trained model from the Transformers library.
- I used the DistilBERT model fine-tuned on the SST-2 dataset for sentiment analysis.
- The model is loaded using the pipeline function, specifying the task of sentiment analysis. The preprocessed text data is then read from a CSV file, and sentiment labels are assigned to each tweet using the loaded model.
- The sentiment labels are then appended to a list, and finally added as a new column to the dataframe.
- The sentiment labels can be either Positive, Negative or Neutral, depending on the output of the model.



	A	B	C	D	E
1	User_Name	Date	Time	Preprocessed_Text	Sentiment
2	HeavenOverHeels	04-05-2023	01:04:54	best known phony bennett iftonytweeted chris dembitz uva alum basketball fan former photographer former standup comedian importantly father dembitz 49 died tuesday bat	Negative
3	GregMadia	04-05-2023	01:04:25	top 9 uva 9 jmu 8 one nobody though	Negative
4	GregMadia	04-05-2023	01:02:18	top 9 uva 9 jmu 6 dukes runners second third tying run plate outs jay woolfolk hit leadoff man jake gelof made throwing error	Negative
5	GregMadia	04-05-2023	00:45:10	top 8 uva 9 jmu 6 dukes getting cavaliers bullpen jason schiavones tworun homer evan blanco cuts hoos lead three brian oconnor isnt wasting anymore time hes bringing jay wo	Positive
6	HooVApolitics	04-05-2023	00:36:49	huge loss uva basketball familywont around november march without humor	Negative
7	sports_shoppin	04-05-2023	00:34:59	loss family uva community iftonytweeted provided bright light laughter building amazing basketball program glad least able see championship	Positive
8	GregMadia	04-05-2023	00:33:33	end 7 uva 9 jmu 3 griff oferralls sac fly added cavaliers advantage hoos couldve ethan odonnell robbed extra bases dukes cf jack cone made diving grab end inning	Positive
9	VaTalent	04-05-2023	00:26:35	one nations top point guards recruiting class 2024 planning take official visit uva	Negative
10	GregMadia	04-05-2023	00:20:00	mid 7 uva 8 jmu 3 dukes added run jack cones sac fly three runs charged cavaliers reliever chase hungate	Positive
11	GregMadia	04-05-2023	00:09:59	top 7 uva 8 jmu 2 kyle novak rbi double dukes jason schiavone followed rbi single	Positive
12	GregMadia	04-05-2023	00:04:16	end 6 uva 8 jmu 0 connelly earlys final line 6 ip 5 h 1 bb 6 ks hes reliable midweek throughout spring hoos chase hungate bullpen begin 7th	Positive
13	GregMadia	03-05-2023	23:35:18	bottom 5 uva 8 jmu 0 casey saucke rbi triple scored wild pitch	Negative
14	GregMadia	03-05-2023	23:28:43	mid 5 uva 6 jmu 0 connelly early five scoreless hes struck five	Negative
15	nexus7724	03-05-2023	22:59:44	experienced writers editors help achieve academic success providing highquality plagiarismfree assignments tailored specific requirements tufts tuftsuniversity jumbos medford	Positive
16	GSwaim	03-05-2023	22:54:39	great updates doublefriespod today mdavidhood course possibly schools jumping aboard conference realignment train along clemson fsu besides mentioned mostand takes eig	Negative
17	GregMadia	03-05-2023	22:53:16	mid 3 uva 6 jmu 0 outstanding relay cavaliers casey saucke luke hanson kyle teel nab coleman calabrese home plate mason dunaways double keep dukes board end frame	Positive
18	yuneydyandexru1	03-05-2023	22:48:18	send student ai summer camp 100 free nationwide ai summer camp directory click link amp follow us stanford university university virginia 1 slot left utaustin 1 slot left many ec	Negative
19	GregMadia	03-05-2023	22:41:46	end 2 uva 6 jmu 0 kyle teel added rbi single teel gelof odonnell two hits	Positive
20	GregMadia	03-05-2023	22:36:17	bottom 2 uva 5 jmu 0 jake gelof doubles second time many innings one scores odonnell gelof 73 rbi year	Negative
21	GregMadia	03-05-2023	22:32:59	bottom 2 uva 4 jmu 0 ethan odonnells rbi single extends cavaliers lead griff oferrall drove run ground batter	Positive
22	GregMadia	03-05-2023	22:17:49	end 1 uva 2 jmu 0 kyle teels rbi chopped single scored jake gelof doubled earlier inning gelofs double pushed ethan odonnell third odonnell scored wild pitch	Positive
23	DennisWLNI	03-05-2023	21:43:46	rough weekend many mondaymourning back show recap unfortunate events weekend sports including collapse epic proportions bostonbruins ufcs songyadong memgrizz playe	Negative
24	DennisWLNI	03-05-2023	21:41:28	teels deal bydavidteel joined us talk latest area ncaa athletics including uvamenshoops losing transfer portal battle steps could taken lessen use portal amp 66 would success uv	Negative
25	GregMadia	03-05-2023	21:41:07	uva lineup griff oferrall ss ethan odonnell cf jake gelof 3b kyle teel c ethan anderson 1b casey saucke rf anthony stephan dh luke hanson 2b harrison didawick lf	Negative
26	GregMadia	03-05-2023	21:39:52	21 virginia 3611 hosts james madison 2419 6 pm first pitch dish uva start lhp connelly early 81 326 era jmu rhp sean culkin 10 785 era	Positive

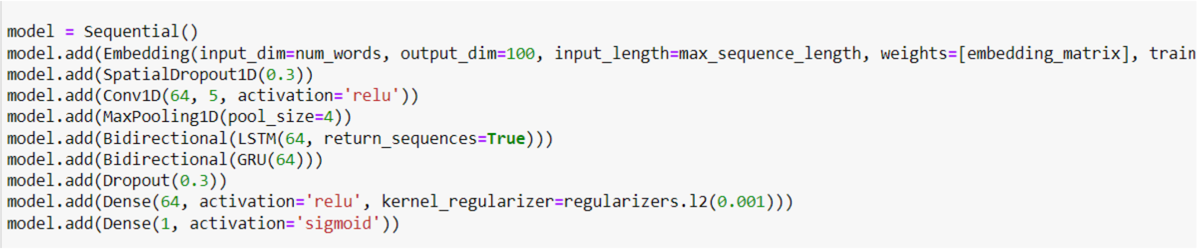
Figure 4.1: Sentiment Labelling.

5. Model Architecture

The Model Architecture includes a deep learning model for sentiment analysis using Keras. Structure of model architecture:

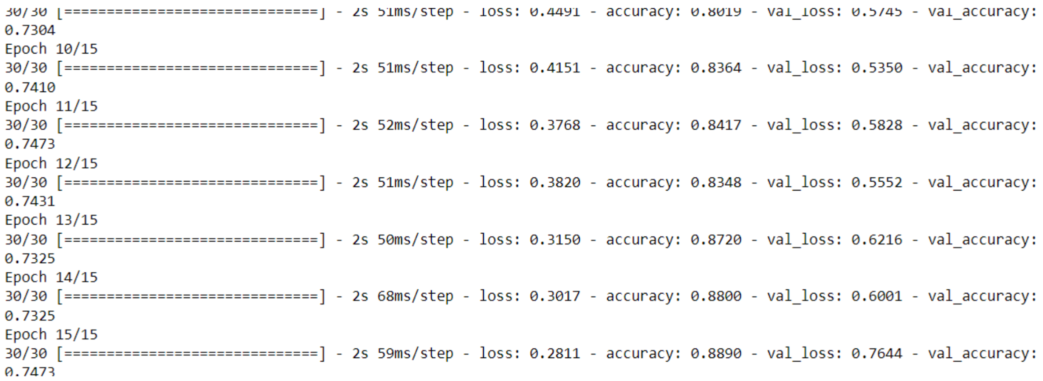
- Embedding Layer: Converts the words into vectors with 100 dimensions.
- SpatialDropout1D Layer: Randomly drops out 30% of the embedding dimensions to reduce overfitting.
- Conv1D Layer: Applies a convolutional operation with 64 filters and a kernel size of 5 to the sequence of word vectors.
- MaxPooling1D Layer: Performs max pooling operation with a pool size of 4.
- Bidirectional LSTM Layer: A layer with 64 LSTM units that reads the sequences forward and backward.
- Bidirectional GRU Layer: A layer with 64 GRU units that reads the sequences forward and backward.
- Dropout Layer: Randomly drops out 30% of the previous layer's output to further prevent overfitting.
- Dense Layer: A fully connected layer with 64 units and ReLU activation function, which introduces non-linearity.
- Output Layer: A single output unit with a sigmoid activation function that produces a probability score between 0 and 1.

I have used pre-trained GloVe word embeddings to initialize the embedding layer. The weights of the embedding layer are frozen during training to prevent overfitting, while other layers are trained and to increase the accuracy. The model is compiled with the binary cross-entropy loss function, Adam optimizer with a learning rate of 0.0005, and accuracy as the evaluation metric. Finally, the model is trained for 15 epochs with a batch size of 64, and its performance is evaluated on the test set.



```
model = Sequential()
model.add(Embedding(input_dim=num_words, output_dim=100, input_length=max_sequence_length, weights=[embedding_matrix], train
model.add(SpatialDropout1D(0.3))
model.add(Conv1D(64, 5, activation='relu'))
model.add(MaxPooling1D(pool_size=4))
model.add(Bidirectional(LSTM(64, return_sequences=True)))
model.add(Bidirectional(GRU(64)))
model.add(Dropout(0.3))
model.add(Dense(64, activation='relu', kernel_regularizer=regularizers.l2(0.001)))
model.add(Dense(1, activation='sigmoid'))
```

Figure 5.1: Model Architecture.

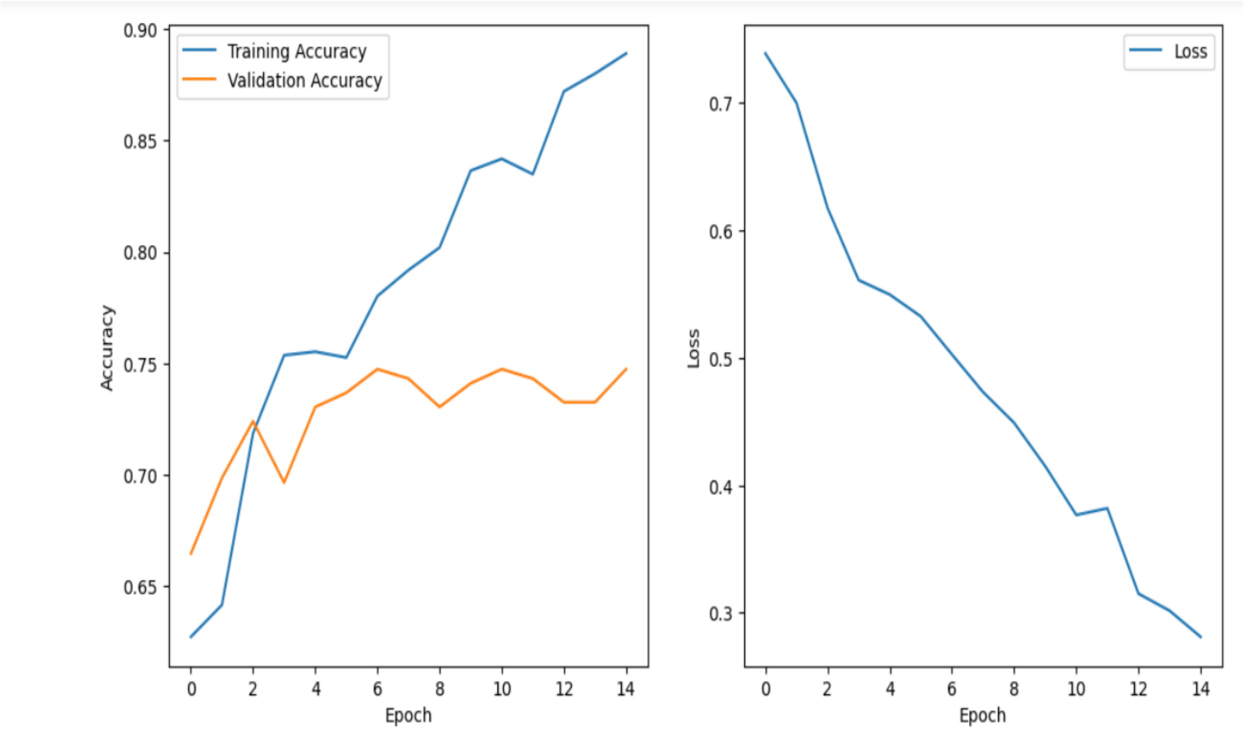


```
30/30 [==============================] - 2s 51ms/step - loss: 0.4491 - accuracy: 0.8019 - val_loss: 0.5745 - val_accuracy:
0.7304
Epoch 10/15
30/30 [==============================] - 2s 51ms/step - loss: 0.4151 - accuracy: 0.8364 - val_loss: 0.5350 - val_accuracy:
0.7410
Epoch 11/15
30/30 [==============================] - 2s 52ms/step - loss: 0.3768 - accuracy: 0.8417 - val_loss: 0.5828 - val_accuracy:
0.7473
Epoch 12/15
30/30 [==============================] - 2s 51ms/step - loss: 0.3820 - accuracy: 0.8348 - val_loss: 0.5552 - val_accuracy:
0.7431
Epoch 13/15
30/30 [==============================] - 2s 50ms/step - loss: 0.3150 - accuracy: 0.8720 - val_loss: 0.6216 - val_accuracy:
0.7325
Epoch 14/15
30/30 [==============================] - 2s 68ms/step - loss: 0.3017 - accuracy: 0.8800 - val_loss: 0.6001 - val_accuracy:
0.7325
Epoch 15/15
30/30 [==============================] - 2s 59ms/step - loss: 0.2811 - accuracy: 0.8890 - val_loss: 0.7644 - val_accuracy:
0.7473
```

I got the accuracy as 0.88

6. Results

- The model is evaluated using the following metrics:
- Accuracy: The percentage of correctly classified tweets.
- Precision: The percentage of tweets classified as positive that are actually positive.
- Recall: The percentage of positive tweets that are correctly classified as positive.
- F1-score: The harmonic mean of precision and recall.



6.1 Plots for Accuracy and Loss

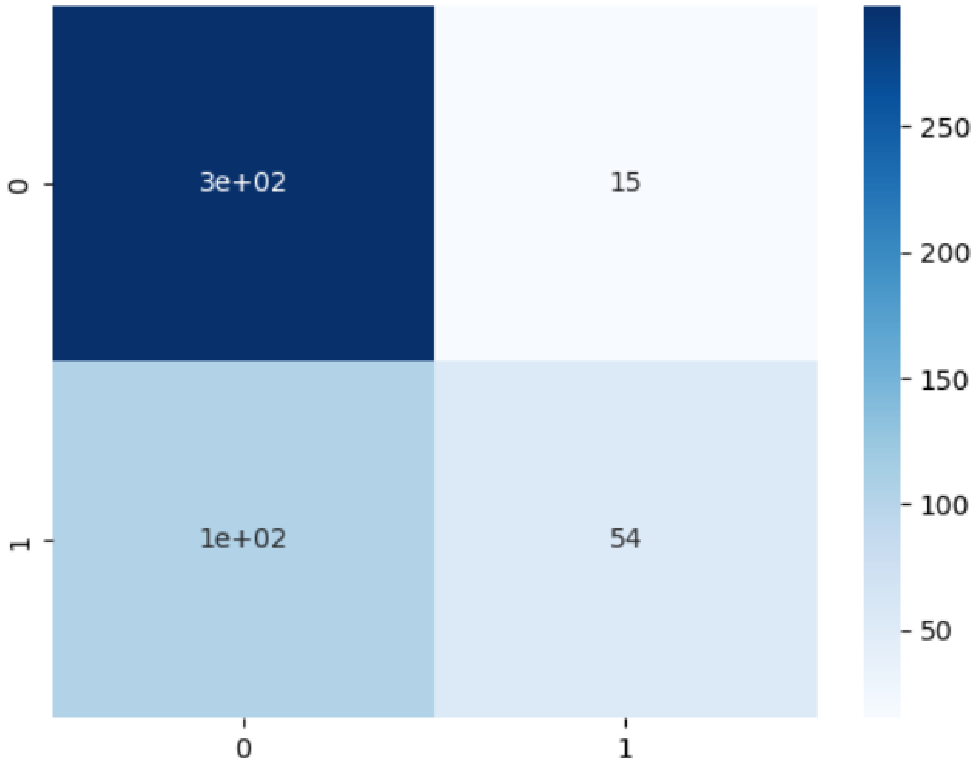
The results include subplots, one for training and validation accuracy and the other for loss. The x-axis represents the number of epochs, while the y-axis represents the accuracy and loss values.

From the plot, we can observe that the training accuracy and loss are increasing and decreasing respectively with each epoch, which means that the model is learning and fitting well to the training data. The validation accuracy and loss seem are also increasing and decreasing which says that the model is performing well.

Overall, this plot is a useful tool for evaluating the performance of the model and identifying any potential issues with overfitting or underfitting.

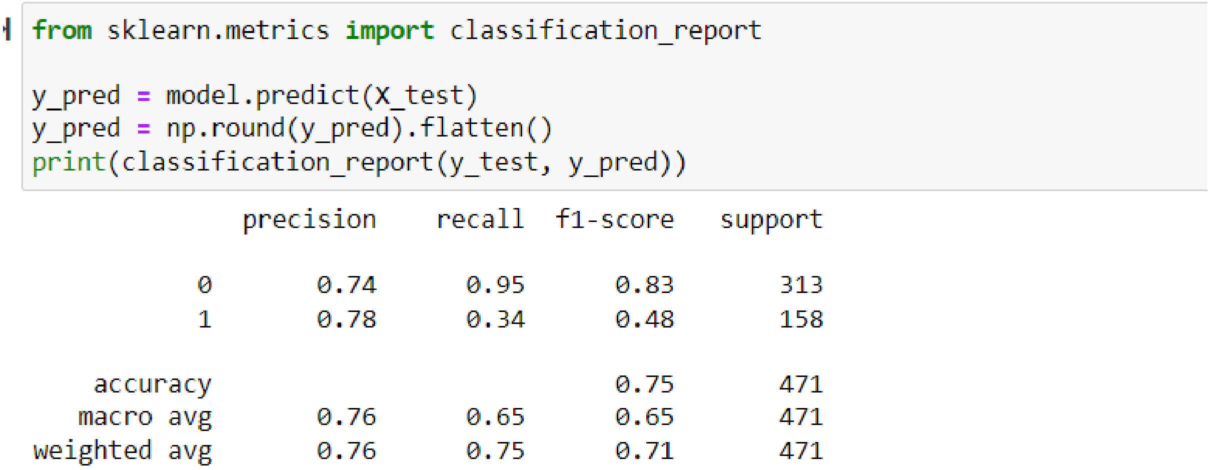
Evaluation Metrics

Confusion Matrix:



confusion matrix for the predictions is made by a model on a test dataset. The confusion matrix shows the number of true positive, false positive, true negative, and false negative predictions made by the model. The heatmap function from the seaborn library is used to visualize the confusion matrix, with each cell showing the number of predictions. The 'Blues' colormap is used to differentiate between the different values in the matrix. Overall, the confusion matrix and its visualization help in evaluating the performance of the model and identifying any misclassifications.

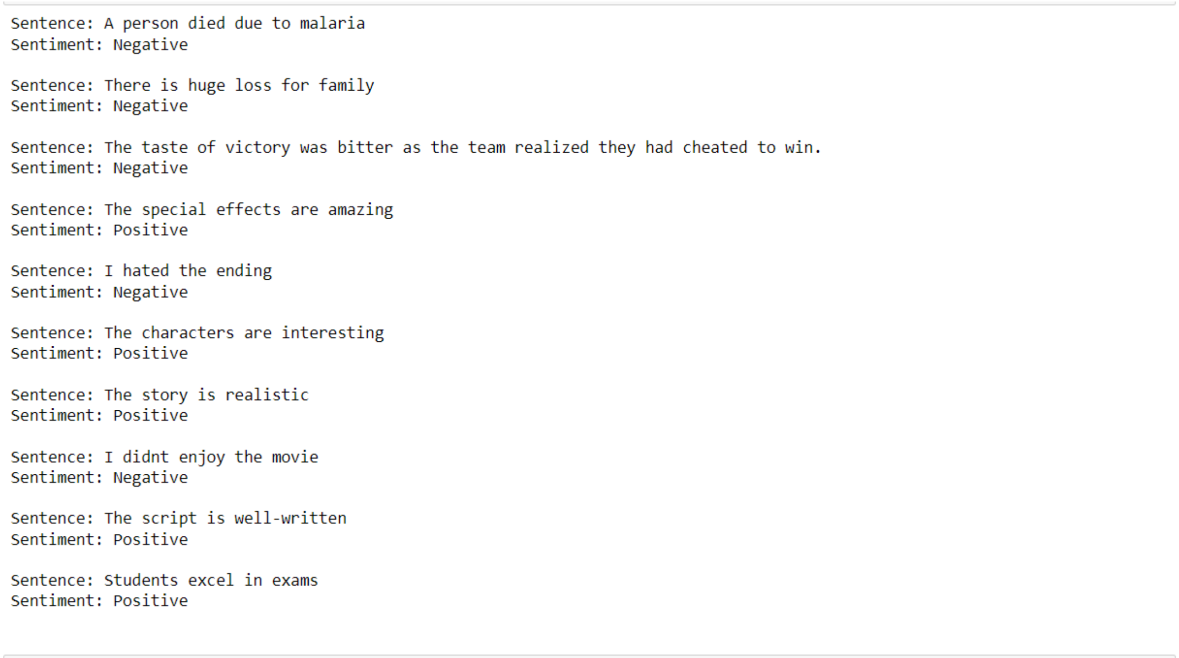
Classification Report:



```
from sklearn.metrics import classification_report

y_pred = model.predict(X_test)
y_pred = np.round(y_pred).flatten()
print(classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.74      0.95      0.83       313
           1       0.78      0.34      0.48       158

    accuracy                           0.75       471
   macro avg       0.76      0.65      0.65       471
weighted avg       0.76      0.75      0.71       471
```

The classification report function shows precision, recall, and F1-score for each class along with their support, which is the number of samples in each class. In this case, the model has an overall accuracy of 0.75 and has a higher precision for class 0 than class 1. The recall values show that the model is better at predicting class 0 than class 1. The weighted average F1-score is 0.71, which is lower than the accuracy, indicating that the model may not perform well on imbalanced datasets. The macro average F1-score is 0.65, which indicates the overall performance of the model.



```
Sentence: A person died due to malaria
Sentiment: Negative

Sentence: There is huge loss for family
Sentiment: Negative

Sentence: The taste of victory was bitter as the team realized they had cheated to win.
Sentiment: Negative

Sentence: The special effects are amazing
Sentiment: Positive

Sentence: I hated the ending
Sentiment: Negative

Sentence: The characters are interesting
Sentiment: Positive

Sentence: The story is realistic
Sentiment: Positive

Sentence: I didnt enjoy the movie
Sentiment: Negative

Sentence: The script is well-written
Sentiment: Positive

Sentence: Students excel in exams
Sentiment: Positive
```

Predicted sentences

7. Limitations

Different people may interpret the same text in different ways, leading to different sentiment scores. One of the limitations it is taking more time for training the data for the given epochs.

The model may not be able to analyse text that contains slang, typos, or misspellings, as these can significantly alter the meaning of the text and affect the sentiment analysis results. Similarly, model may struggle with analysing text that contains abbreviations or acronyms that are not well-known or commonly used.

Furthermore, sentiment analysis models may not be able to accurately analyse text that contains a mix of sentiment, such as when a review contains both positive and negative comments. In such cases, the model may struggle to accurately determine the overall sentiment of the text.

8. Conclusion

In conclusion, I developed a model for sentiment analysis of tweets using Python and machine learning techniques. This involves web scraping to collect 1500 unique tweets using Tweepy library and Twitter API, followed by preprocessing to remove unnecessary characters and and feature extraction using TF-IDF. A sequential neural network model with embedding layer, LSTM, CNN, GRU, and dense output layers will be trained using binary cross-entropy loss and the Adam optimizer. The model will be evaluated using accuracy, precision, recall, and F1-score, and a confusion matrix will be used to visualize the prediction metrics. Finally, the model will be used to predict the sentiment of a given sentence. Overall, the model performed well with best accuracy.