Sentiment Analysis of Tweets about Georgia State University

A Natural Language Processing Use-Case

David King Agbi
Joel Dassoundo
Ifeanyi Moses Uzowuru

J. Mack Robinson College of Business, Georgia State University

Introduction

- Sentiment analysis is simply identifying and classifying sentiments (which are the point of view or emotions) that are expressed in the message or text source

Introduction

- Sentiment analysis is simply identifying and classifying sentiments (which are the point of view or emotions) that are expressed in the message or text source
- A Natural Language Processing is simply preparing computers to understand and process the language we speak, either through audio or texts. Siri, Cortana, Bixby and google translator are all examples of this.

Introduction

- Sentiment analysis is simply identifying and classifying sentiments (which are the point of view or emotions) that are expressed in the message or text source
- A Natural Language Processing is simply preparing computers to understand and process the language we speak, either through audio or texts. Siri, Cortana, Bixby and google translator are all examples of this.
- Build a machine learning model that can analyze several tweets made about GSU with the ability to estimate the sentiments of those tweets.

  Thus, we would classify these tweets as positive, neutral or negative.

WHY?

Introduction

Problem Statement

Methodology

Data Cleaning

Data Transformation & Vectorization

Machine Learning & Model Evaluation

Conclusion

David King Agbi Joel Dassoundo Ifeanyi Moses Uzowuru

Steps Involved

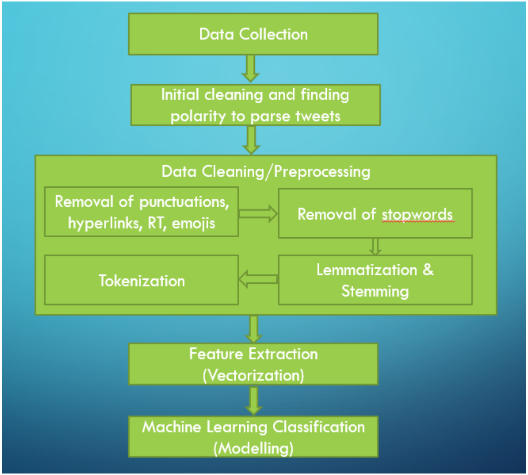
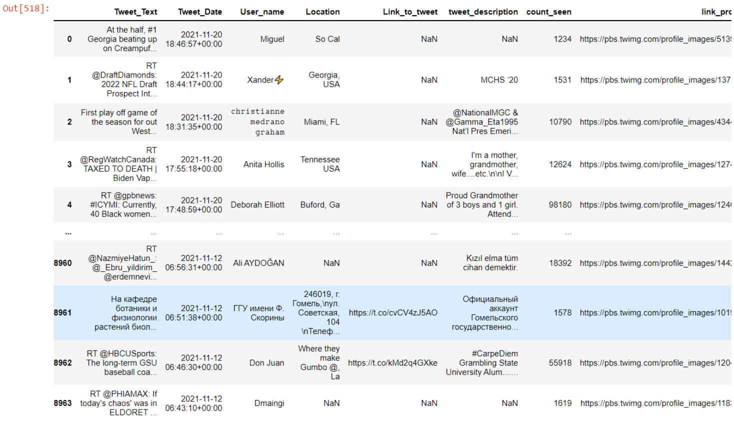

Data Collection
Initial cleaning and finding polarity to parse tweets
Data Cleaning/Preprocessing
Removal of punctuations, hyperlinks, RT, emojis
Removal of stopwords
Lemmatization & Stemming
Tokenization
Feature Extraction (Vectorization)
Machine Learning Classification (Modelling)

Figure: Process involved in the project

Task 1 - Exploratory Data Analysis & Visualization



Out[518]:

	Tweet_Text	Tweet_Date	User_name	Location	Link_to_tweet	tweet_description	count_seen	link_prc
0	At the half, #1 Georgia beating up on Creampuf...	2021-11-20 18:46:57+00:00	Miguel	So Cal	NaN	NaN	1234	https://pbs.twimg.com/profile_images/513
1	RT @DraftDiamonds: 2022 NFL Draft Prospect Int...	2021-11-20 18:44:17+00:00	Xander⚡	Georgia, USA	NaN	MCHS '20	1531	https://pbs.twimg.com/profile_images/137
2	First play off game of the season for out West...	2021-11-20 18:31:35+00:00	christianne medrano graham	Miami, FL	NaN	@NationalMGC & @Gamma_Eta1995 Nat'l Pres Emeri...	10790	https://pbs.twimg.com/profile_images/434
3	RT @RegWatchCanada: TAXED TO DEATH | Biden Vap...	2021-11-20 17:55:18+00:00	Anita Hollis	Tennessee USA	NaN	I'm a mother, grandmother, wife....etc.\n\nl V...	12624	https://pbs.twimg.com/profile_images/127
4	RT @gpbnews: #ICYMI: Currently, 40 Black women...	2021-11-20 17:48:59+00:00	Deborah Elliott	Buford, Ga	NaN	Proud Grandmother of 3 boys and 1 girl. Attend...	98180	https://pbs.twimg.com/profile_images/124
...	...	...	...	...	...	...	...	
8960	RT @NazmiyeHatun_: @_Ebru_yildirim_ @erdemnevi...	2021-11-12 06:56:31+00:00	Ali AYDOĞAN	NaN	NaN	Kızıl elma tüm cihan demektir.	18392	https://pbs.twimg.com/profile_images/144
8961	На кафедре ботаники и физиологии растений биол...	2021-11-12 06:51:38+00:00	ГГУ имени Ф. Скорины	246019, г. Гомель,\nул. Советская, 104 \nТелеф...	https://t.co/cvCV4zJ5AO	Официальный аккаунт Гомельского государственно...	1578	https://pbs.twimg.com/profile_images/101
8962	RT @HBCUSports: The long-term GSU baseball coa...	2021-11-12 06:46:30+00:00	Don Juan	Where they make Gumbo @, La	https://t.co/kMd2q4GXke	#CarpeDiem Grambling State University Alum.......	55918	https://pbs.twimg.com/profile_images/120
8963	RT @PHIAMAX: If today's chaos' was in ELDORET ...	2021-11-12 06:43:10+00:00	Dmaingi	NaN	NaN	NaN	1619	https://pbs.twimg.com/profile_images/118

Figure: Output of extracted tweets showing the various columns

Task 1 - Exploratory Data Analysis & Visualization

- We do this by first extracting the tweets using Twitter API. The figure below shows the flow diagram of the process

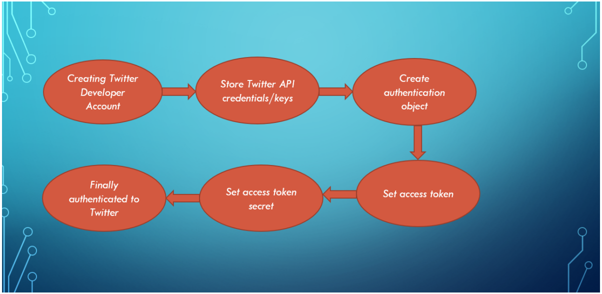

Creating Twitter Developer Account
Store Twitter API credentials/keys
Create authentication object
Set access token
Set access token secret
Finally authenticated to Twitter

Figure: Extracting Tweets using Twitter API

Task 1

- Next thing is to import all the necessary libraries we employed in the project

Task 1

- Next thing is to import all the necessary libraries we employed in the project
- We then analyse or explore the dataset such as getting the information, describing the data, shape of the data, checking for null elements and so on. We visualized the sentiments using the countplot and visualized the null elements by using the heatmap and got these results:

Task 1 - Exploratory Data Analysis & Visualization

Introduction

Problem Statement

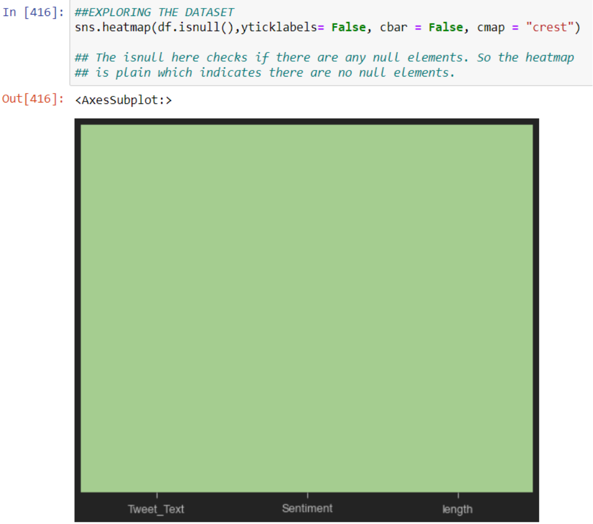
Methodology

Data Cleaning

Data Transformation & Vectorization

Machine Learning & Model Evaluation

Conclusion



In [416]:

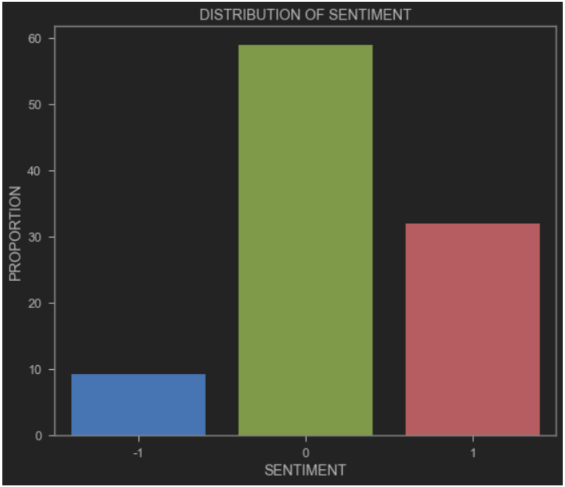
```
##EXPLORING THE DATASET
sns.heatmap(df.isnull(),yticklabels= False, cbar = False, cmap = "crest")

## The isnull here checks if there are any null elements. So the heatmap
## is plain which indicates there are no null elements.
```

Out[416]:

```
<AxesSubplot:>
```

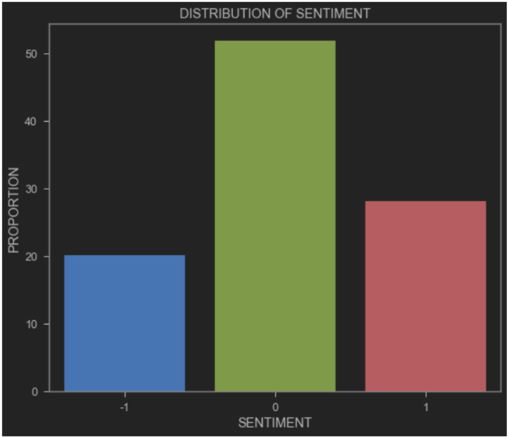
Task 1 - Exploratory Data Analysis & Visualization



As we can see, this looks not balanced

Task 1 - Exploratory Data Analysis & Visualization

- Therefore, we applied the augmentation technique to populate the negative tweets to make it more significant. This is the new plot of the distribution of sentiments



Task 2 - Text Normalizations

- The following are the processing techniques that we employed in the listed order to normalize the text:

Task 2 - Text Normalizations

- The following are the processing techniques that we employed in the listed order to normalize the text:
- Removal of punctuations.

Task 2 - Text Normalizations

- The following are the processing techniques that we employed in the listed order to normalize the text:
- Removal of punctuations.
- Removal of stop words.

Task 2 - Text Normalizations

- The following are the processing techniques that we employed in the listed order to normalize the text:
- Removal of punctuations.
- Removal of stop words.
- Tokenization

Task 2 - Text Normalizations

- The following are the processing techniques that we employed in the listed order to normalize the text:
- Removal of punctuations.
- Removal of stop words.
- Tokenization
- Stemming/Lemmatization

Task 2 - Text Normalizations

- The following are the processing techniques that we employed in the listed order to normalize the text:
- Removal of punctuations.
- Removal of stop words.
- Tokenization
- Stemming/Lemmatization
- Converting capital letters to small letters

Task 3 - Feature Extraction

- Data in the form of text is not suitable for training a machine learning model. For this reason, we had to convert the tweets into numerical features, ensuring that the inherent learnable pattern is conserved in the best way possible.

Task 3 - Feature Extraction

- Data in the form of text is not suitable for training a machine learning model. For this reason, we had to convert the tweets into numerical features, ensuring that the inherent learnable pattern is conserved in the best way possible.
- To do this, we used the count vectorizer or vectorization (Term Frequency-Inverse Document Frequency (TF-IDF)) to perform textual transformations into vectors

Task 4 - Machine Learning & Model Evaluation

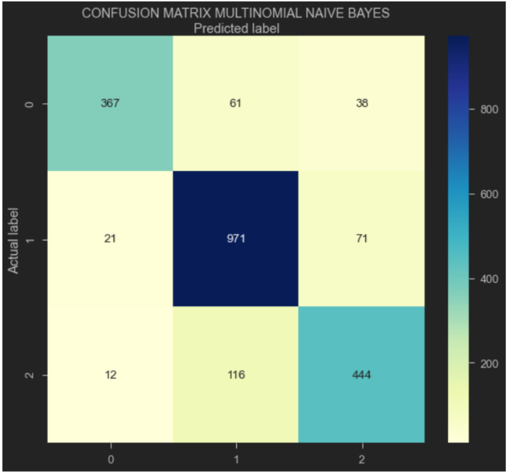
- Now that the feature extraction is done, the data is ready to be fed into a model

Task 4 - Machine Learning & Model Evaluation

- Now that the feature extraction is done, the data is ready to be fed into a model
- Finally, the data is split into training and test using stratified sampling so that the split follows the population distribution and then we applied the machine learning algorithms. Thus, we compared the performance of the:
  1. Multinomial Naive Baye's Model
  2. Multinomial Logistic regression Model
  3. The Decision Tree (Random Forest) Model

Multinomial Naive Baye’s Model

- We had the following confusion matrix and classification report after using the Multinomial Naive Baye’s Model



CONFUSION MATRIX MULTINOMIAL NAIVE BAYES

Actual label \ Predicted label	0	1	2
0	367	61	38
1	21	971	71
2	12	116	444

Multinomial Naive Baye’s model

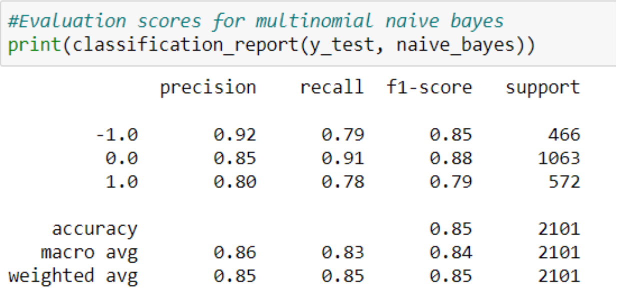
Introduction

Problem Statement

Methodology

- Data Cleaning
- Data Transformation & Vectorization
- Machine Learning & Model Evaluation

Conclusion



```
#Evaluation scores for multinomial naive bayes
print(classification_report(y_test, naive_bayes))
```

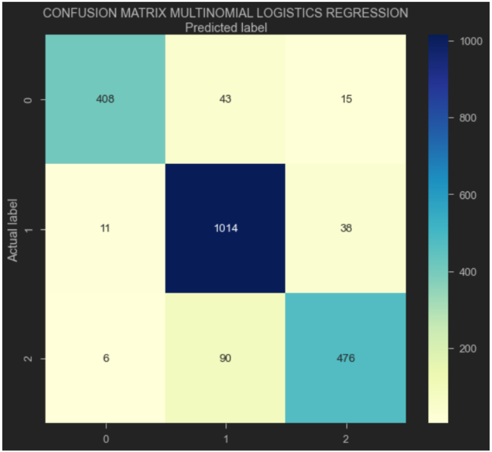
```
              precision    recall  f1-score   support

        -1.0       0.92      0.79      0.85       466
         0.0       0.85      0.91      0.88      1063
         1.0       0.80      0.78      0.79       572

    accuracy                           0.85      2101
   macro avg       0.86      0.83      0.84      2101
weighted avg       0.85      0.85      0.85      2101
```

Multinomial Logistic Regression model

- We had the following confusion matrix and classification report after using the Multinomial Logistic Regression model



Multinomial Logistic Regression model

Introduction

Problem Statement

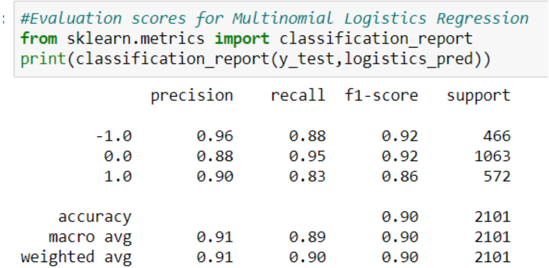
Methodology

Data Cleaning

Data Transformation & Vectorization

Machine Learning & Model Evaluation

Conclusion



```
#Evaluation scores for Multinomial Logistics Regression
from sklearn.metrics import classification_report
print(classification_report(y_test,logistics_pred))
```

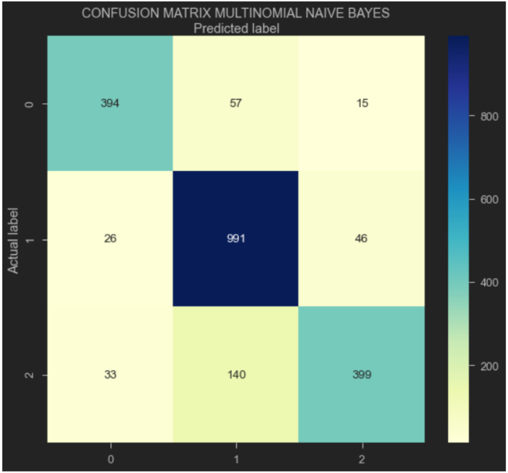
```
              precision    recall  f1-score   support

        -1.0       0.96      0.88      0.92       466
         0.0       0.88      0.95      0.92      1063
         1.0       0.90      0.83      0.86       572

    accuracy                           0.90      2101
   macro avg       0.91      0.89      0.90      2101
weighted avg       0.91      0.90      0.90      2101
```

Random Forest Model

- We had the following confusion matrix and classification report after using the Random Forest Model



Random Forest Model

Introduction

Problem Statement

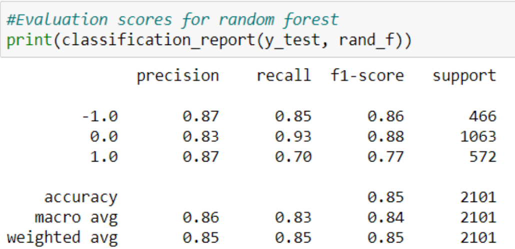
Methodology

Data Cleaning

Data Transformation & Vectorization

Machine Learning & Model Evaluation

Conclusion



```
#Evaluation scores for random forest
print(classification_report(y_test, rand_f))
```

```
              precision    recall  f1-score   support

        -1.0       0.87      0.85      0.86       466
         0.0       0.83      0.93      0.88      1063
         1.0       0.87      0.70      0.77       572

    accuracy                           0.85      2101
   macro avg       0.86      0.83      0.84      2101
weighted avg       0.85      0.85      0.85      2101
```

Conclusion

- In terms of negative prediction, Multinomial logistics Regression performed the best in classifying correctly negative tweets and in its precision power, which is why it has the best F1 score and we thereby confirm the algorithm as our model of choice.

Thank You!

Georgia State University
THE STATE WAY
GEORGIA
1
STATE
SUN BELT