Twitter Sentiment Analysis Report

M Salman Ali Khan

October 16th, 2023

Contents

Introduction:

Twitter (X) sentiment analysis is a powerful technique in the realm of natural language processing and data analytics, focused on gauging and interpreting the emotional tone and opinions expressed within the vast expanse of tweets that populate the Xverse. In essence, it involves the use of computational tools and algorithms to sift through the sea of short messages, extracting valuable insights about public sentiment, whether positive, negative, or neutral, towards a particular topic, product, event, or entity.

The significance of X sentiment analysis lies in its ability to offer real-time, large-scale, and unfiltered insights into the collective consciousness of the online community. It provides businesses with a means to understand customer feedback, helps political campaigns gauge public opinion, aids in brand management, and offers researchers a unique lens into the evolving dynamics of society's reactions to current events, making it an invaluable tool for decision-making and market research.

Executive Summary:

Project Objective:

The objective of this project is to perform entity level sentiment analysis on multilingual tweets to understand trends and insights about the Tweets. A model should also be designed to categorize the sentiments of future tweets based on similar datasets.

Key Insights:

- Negative and positive tweets collectively outnumber neutrals, with a prevalence of negative sentiment.
- Games like Assassins Creed, Borderlands, and Cyber Punk are associated with a high proportion of positive tweets.
- Sports categories such as NGL, NBA, and FIFA exhibit the highest negative tweet percentages.
- Prominent brands like Facebook, Amazon, and Google have a substantial percentage of neutral or irrelevant tweets.
- English tweets dominate, followed by German, Somali, and Afrikaans in descending frequency.
- Croatian, Hungarian, and Polish languages lead with the highest positive tweet percentages.
- Ukrainian, Albanian, and Indonesian languages show the highest negative tweet percentages, while Macedonian, Thai, and Bulgarian tweets predominantly contain neutral sentiments.

Immediate Future Actions:

- Dataset can be increased to get similar representation in tweets by language as is the case with entity.

- With a prevalence of negative sentiments, it's essential to identify and address the root causes of negative sentiment, especially in sports-related content (NGL, NBA, FIFA).
- The high positive sentiment percentages for popular games (Assassins Creed, Borderlands, Cyber Punk) can be leveraged to strengthen positive brand associations and promote these games further.
- For languages with high negative tweet percentages (Ukrainian, Albanian, Indonesian), implement language-specific strategies to understand negative sentiment more effectively.
- Engaging with audiences in languages other than English, such as German, Somali, and Afrikaans, can be means to expanding reach.
- The Multinomial Naive Bayes Algorithm is good for future sentiment analysis as shown by high accuracy, precision, recall and F1 score on test data but better model can be built on top of this model for better sentiment prediction.

About the Dataset:

The open-source dataset is taken from **Kaggle**:

https://www.kaggle.com/datasets/jp797498e/twitter-entity-sentiment-analysis

This dataset comprises sentiment analysis at the entity level on Twitter data, classifying messages into three distinct categories: Positive, Negative, and Neutral. In this classification, messages deemed irrelevant to the entity are categorized as Irrelevant. The dataset is structured with columns such as Entity, Sentiments, and Content. Specifically, 'twitter_training.csv' serves as the training set, while 'twitter_validation.csv' functions as the testing set.

Data Wrangling:

The csv data file was imported using **pandas** in **python**. The Data Wrangling operations were performed using **NumPy** and **pandas** in **python**. Some of the performed operations were:

- Missing Values replacement or removal
- Removal of Duplicates
- Filtering and Sorting Columns
- Data Verification

Data Transformation:

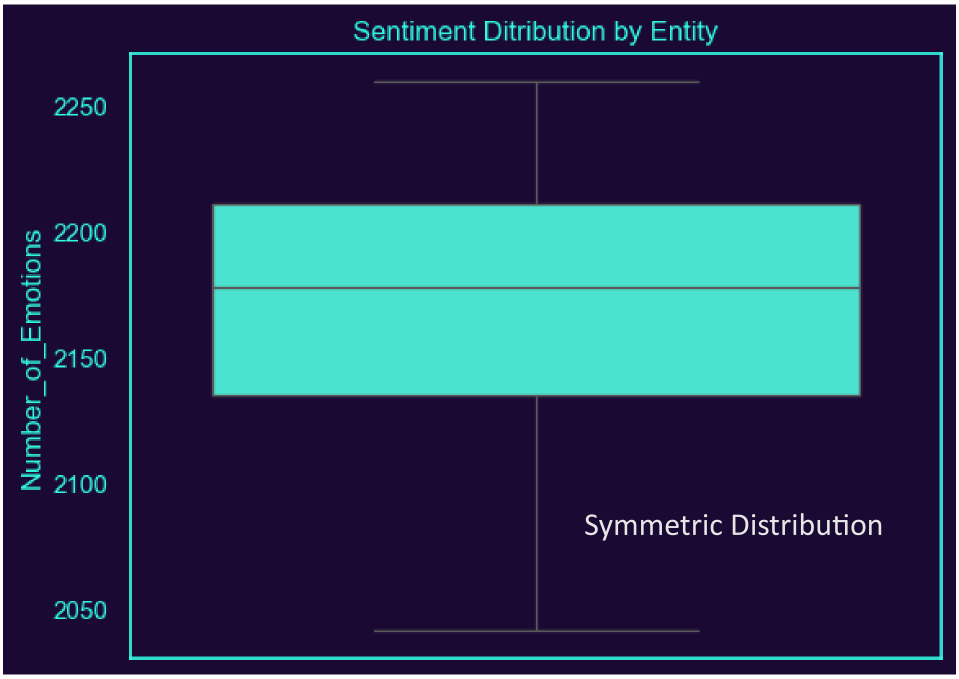
Data Transformation was done by using **pandas**, **sqlite3** and **langdetect** in **python**. Some of the performed operations were:

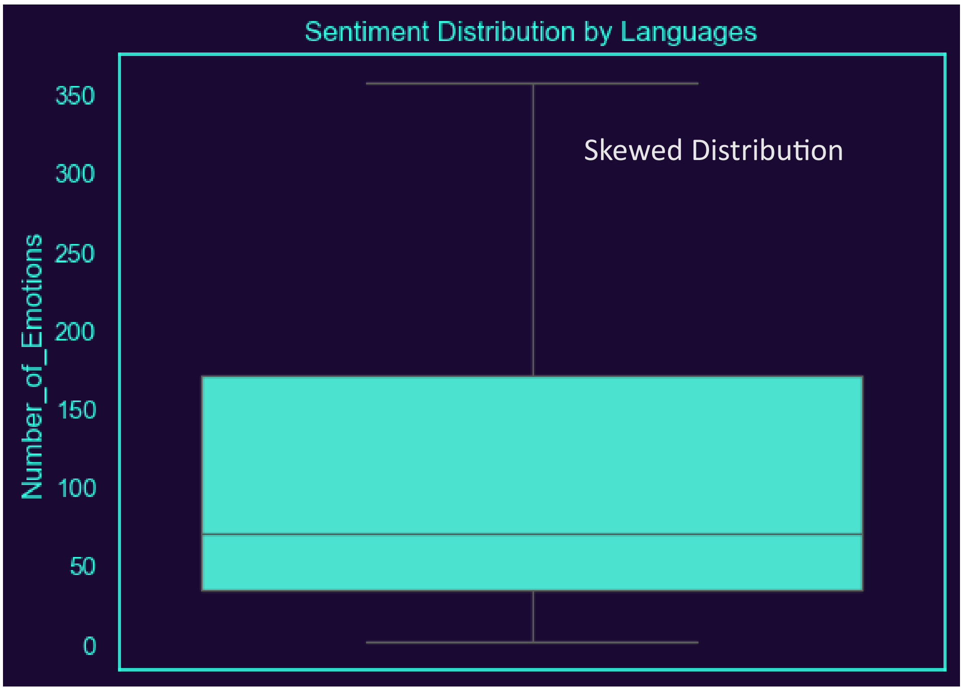
- Dealing with Outliers
- Addition of calculated columns of languages to analyze the sentiments by language
- Creation of specific pivot tables and data views for further analysis

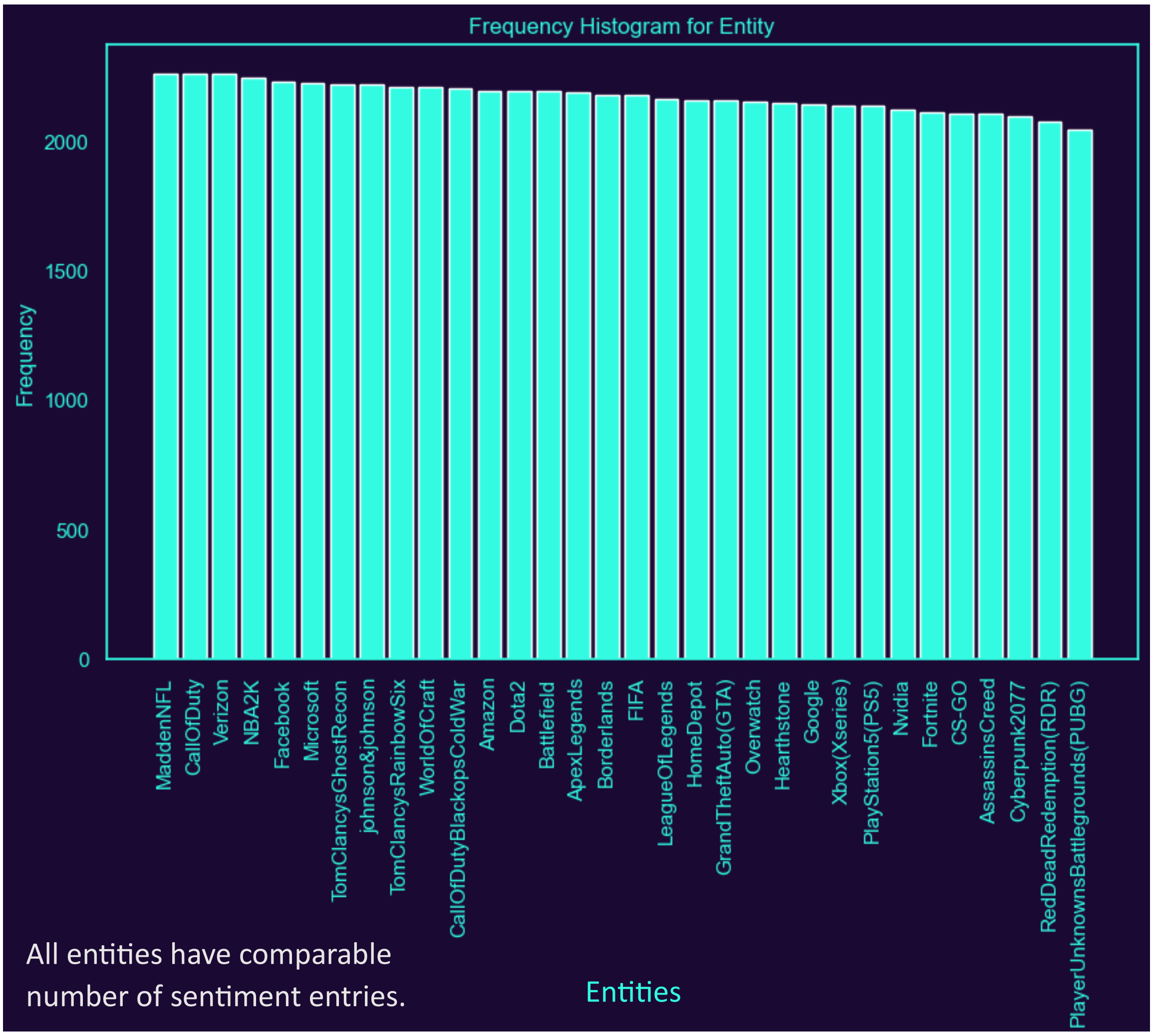
Exploratory Data Analysis:

Exploratory Data Analysis was done by using **sqlite3**, **pandas**, **matplotlib** and **seaborn** in python. Some of the performed operations were:

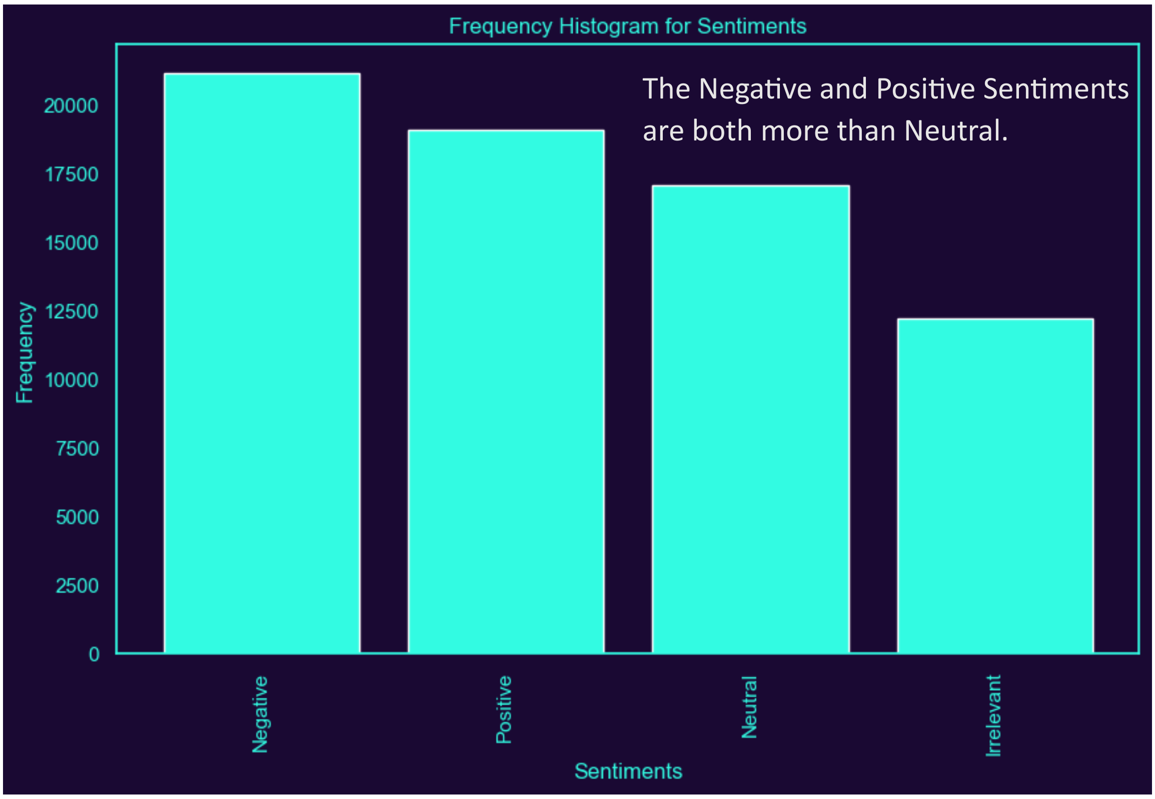
- Summary Statistics
- Grouping data for creating visualizations
- Box plot and Frequency Histogram Creation



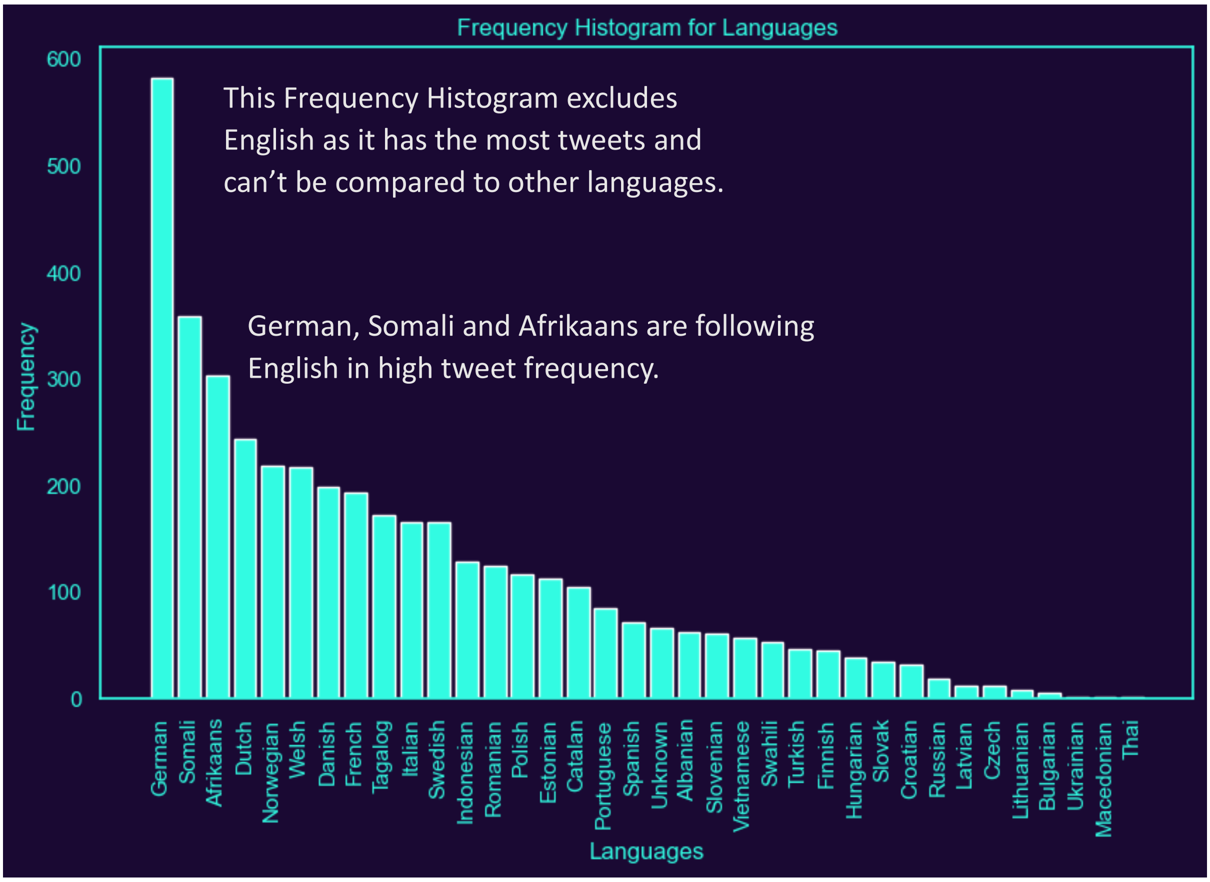




All entities have comparable number of sentiment entries.



The Negative and Positive Sentiments are both more than Neutral.



This Frequency Histogram excludes English as it has the most tweets and can’t be compared to other languages.

German, Somali and Afrikaans are following English in high tweet frequency.

Exploratory Data Analysis Results:

1. Sentiments exhibit a balanced distribution pattern when organized by entity, while they display as skewed distribution when categorized by language. This suggests that sentiment expression tends to vary consistently across different entities, but can significantly differ within the context of various languages.
2. All entities are associated with a relatively similar number of sentiment entries. This implies that each entity has a comparable representation in terms of sentiment expressions.
3. The total count of Negative and Positive Sentiment tweets surpasses the count of Neutral Sentiment tweets, with Negative Tweets being the most prevalent. This indicates that there is a noticeable prevalence of both Negative and Positive Sentiments in the dataset, with Negative Sentiments being the most dominant category.
4. English language tweets are the most abundant, followed by German, Somali, and Afrikaans in descending order of tweet frequency. The variation in tweet frequency across languages could indicate the dataset's linguistic diversity and potentially reflect the platform's user demographics.

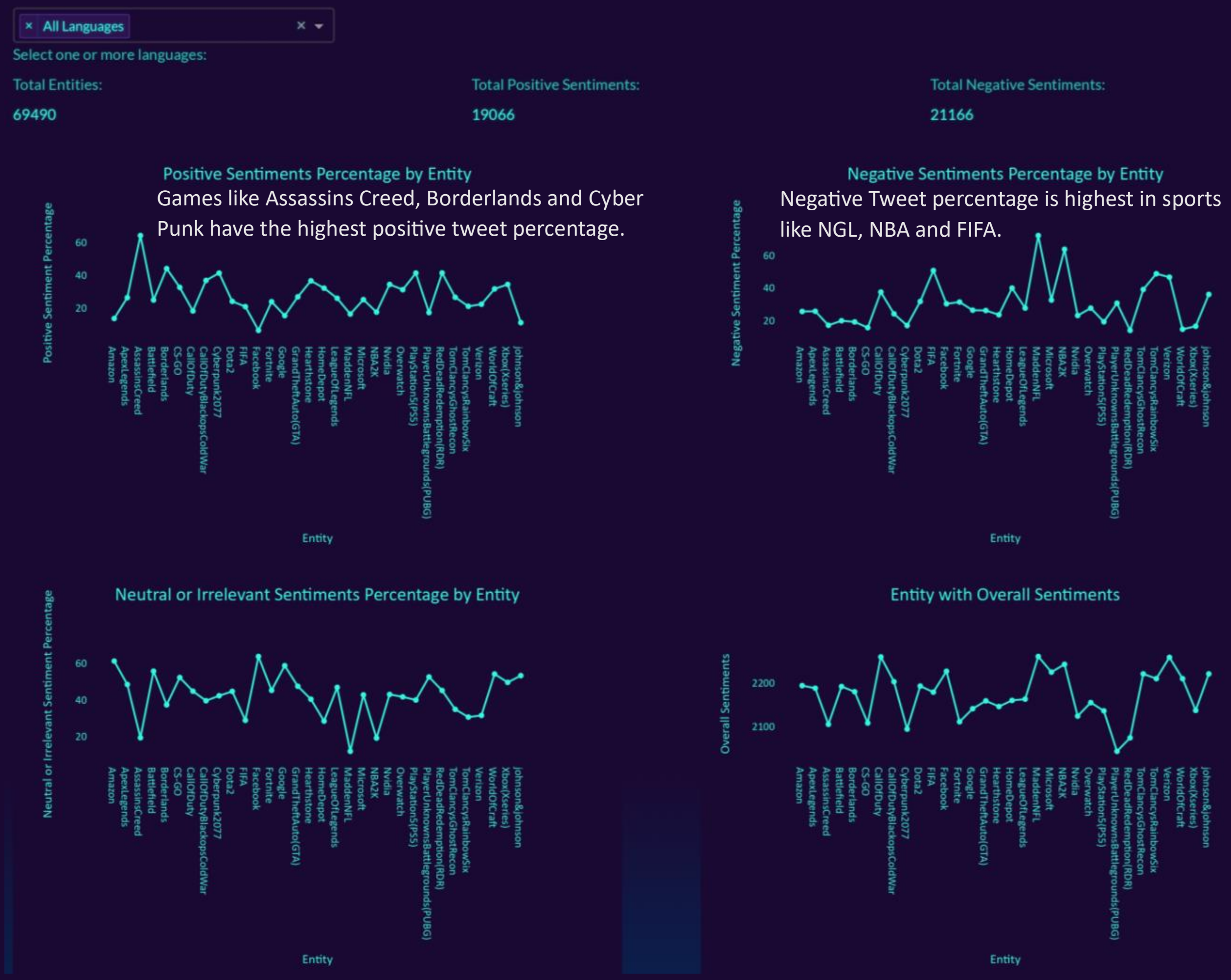
Data Visualizations and Insights:

2 dashboards with 8 specific charts are produced using **dash**, **plotly** and **pandas** in **python**. These interactive dashboards were designed to allow users to personalize their experience. Each visualization was made with a specific objective in mind. The most useful visualizations and dashboards containing the obtained insights are detailed below.

Dashboards:

1. The first dashboard visualizes the different types of sentiments against their entity while the KPIs tells they aggregate entities, positive and negative sentiments filtered by languages.

Entity Dashboard


× All Languages × ▾

Select one or more languages:

Total Entities: 69490

Total Positive Sentiments: 19066

Total Negative Sentiments: 21166

Games like Assassins Creed, Borderlands and Cyber Punk have the highest positive tweet percentage.

Negative Tweet percentage is highest in sports like NGL, NBA and FIFA.

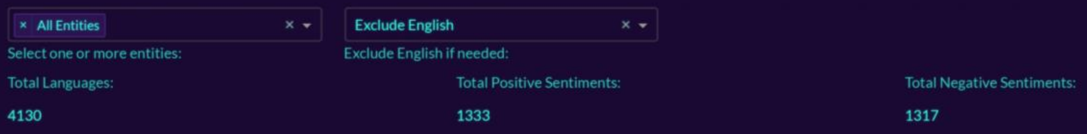
Famous Brands like Facebook, Amazon and Google have the highest neutral or Irrelevant tweet percentage.

Most tweets are about Call of Duty, NFL and Verizon.

2. This dashboard illustrates the different types of sentiments against their language while the KPIs tells they aggregate entities, positive and negative sentiments filtered by Entities.

Language Dashboard

English Language is excluded here because of its very high volume but if we analyze English Language we find that most English tweets are neutral or irrelevant. Positive and Negative Tweet Percentage in English is nearly equal.



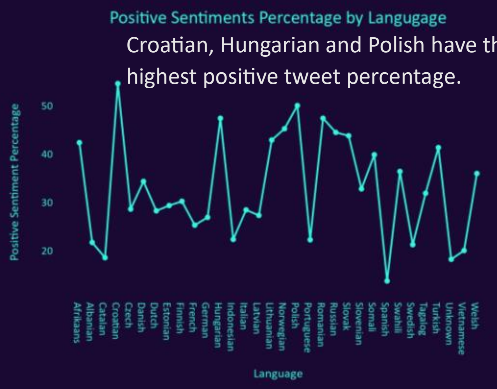
× All Entities × ▾
Select one or more entities:

Exclude English × ▾
Exclude English if needed:

Total Languages: 4130

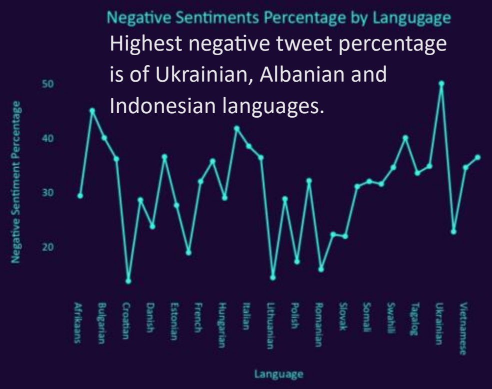
Total Positive Sentiments: 1333

Total Negative Sentiments: 1317



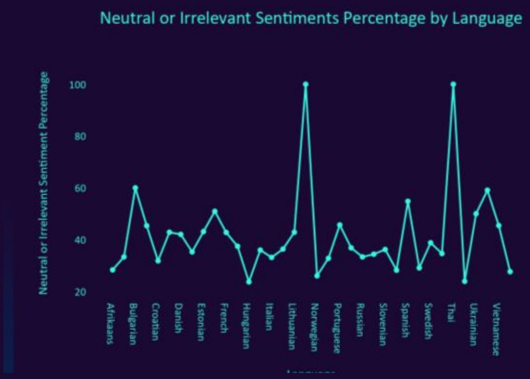
Positive Sentiments Percentage by Langugage

Croatian, Hungarian and Polish have the highest positive tweet percentage.



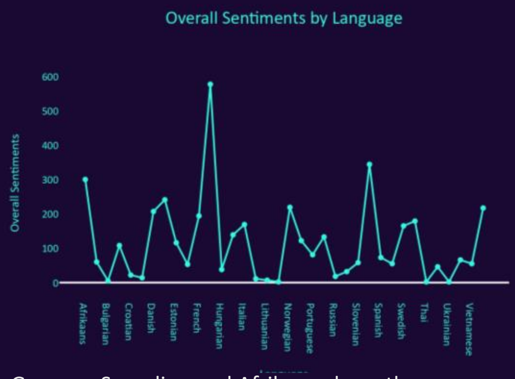
Negative Sentiments Percentage by Langugage

Highest negative tweet percentage is of Ukrainian, Albanian and Indonesian languages.



Neutral or Irrelevant Sentiments Percentage by Language

Neutral Tweet Percentage is highest in Macedonian, Thai and Bulgarian Tweets.

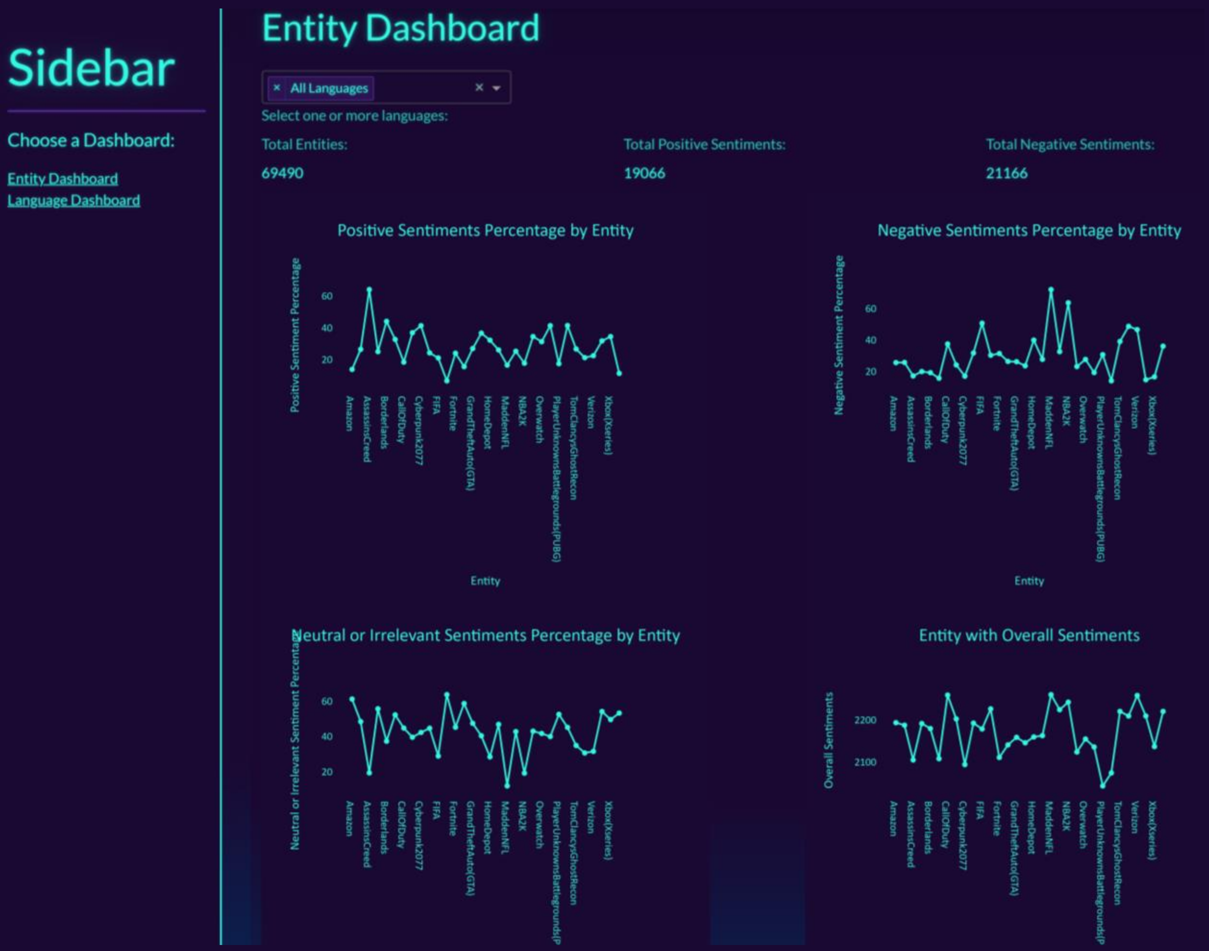


Overall Sentiments by Language

German, Somalian and Afrikaans have the most tweets after English.

Dashboard App:

A Dashboard App is created for easy maneuvering between the dashboards. Navigation is carried out using the Side Panel making it easy to study both the dashboards at once.



Sidebar

Choose a Dashboard:

Entity Dashboard
Language Dashboard

Entity Dashboard

× All Languages × ▾

Select one or more languages:

Total Entities: 69490

Total Positive Sentiments: 19066

Total Negative Sentiments: 21166

Positive Sentiments Percentage by Entity
Positive Sentiment Percentage
60
40
20
Amazon
AssassinsCreed
Borderlands
CallOfDuty
Cyberpunk2077
FIFA
Fortnite
GrandTheftAuto(GTA)
HomeDepot
MaddenNFL
NBA2K
Overwatch
PlayerUnknownsBattlegrounds(PUBG)
TomClancysGhostRecon
Verizon
Xbox(Xseries)
Entity
Negative Sentiments Percentage by Entity
Negative Sentiment Percentage
60
40
20
Entity
Neutral or Irrelevant Sentiments Percentage by Entity
Neutral or Irrelevant Sentiment Percentage
60
40
20
Entity with Overall Sentiments
Overall Sentiments
2200
2100

Key Data Insights:

1. Games such as Assassins Creed, Borderlands, and Cyber Punk exhibit the highest proportion of positive tweets. **(Dashboard 1)**
2. Sports categories like NGL, NBA, and FIFA have the highest percentage of negative tweets. **(Dashboard 1)**
3. Prominent brands like Facebook, Amazon, and Google display the highest percentage of neutral or irrelevant tweets. **(Dashboard 1)**
4. The most frequently discussed topics in tweets are related to Call of Duty, NFL, and Verizon. **(Dashboard 1)**
5. Croatian, Hungarian, and Polish languages demonstrate the highest percentage of positive tweets. **(Dashboard 2)**
6. The languages with the highest negative tweet percentages are Ukrainian, Albanian, and Indonesian. **(Dashboard 2)**
7. Macedonian, Thai, and Bulgarian tweets have the highest proportion of neutral tweets. **(Dashboard 2)**
8. Following English, German, Somali, and Afrikaans are the most frequently used languages in the tweets. **(Dashboard 2)**

Data Modelling:

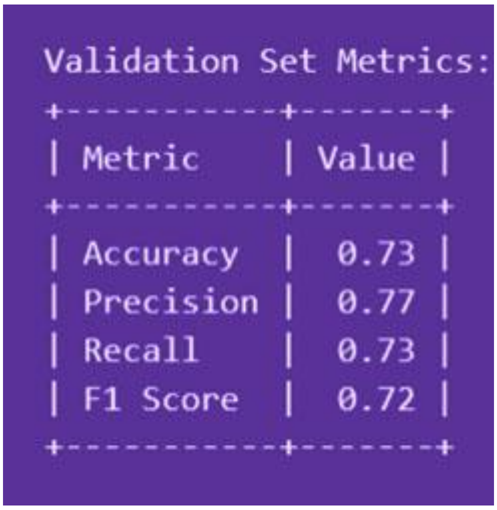
A text classification model, specifically a Multinomial Naive Bayes (NB) classifier, is used for sentiment analysis. Its purpose is to analyse text data and classify it into predefined categories, in this case, sentiment categories such as 'Positive,' 'Negative,' or possibly 'Neutral.'

The model works by converting text data into numerical features using TF-IDF (Term Frequency-Inverse Document Frequency) vectorization, which quantifies the importance of words within the text data. It then trains on the labelled training data to learn the relationships between these numerical features and the corresponding sentiment labels. Once trained, the model can make predictions on new, unlabelled text data by using the learned patterns and associations.

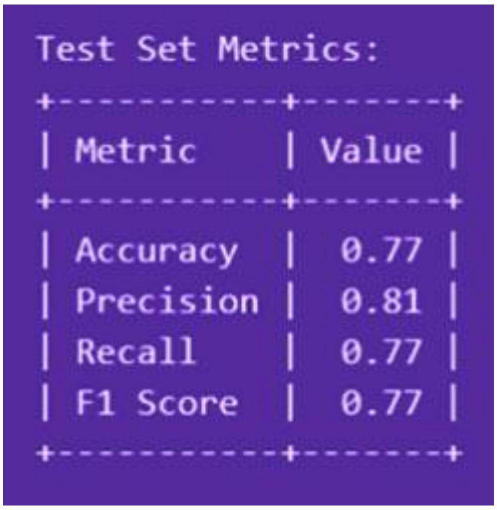
The significance of this model lies in its ability to automate sentiment analysis, which is vital in various applications, such as social media monitoring, customer feedback analysis, and product reviews. By accurately classifying text data into sentiment categories, the model can help businesses and organizations gain insights into public sentiment, identify areas for improvement, and make data-driven decisions.

The model's evaluation metrics (accuracy, precision, recall, and F1 score) provide a measure of its performance, indicating how well it generalizes from the training data to new, unseen data. Furthermore, the confusion matrix and its visualization help in understanding the model's ability to correctly classify different sentiment categories and identify any potential misclassifications. Overall, this model streamlines the sentiment analysis process, making it more efficient and scalable, and can be a valuable tool for making data-driven decisions based on textual data.

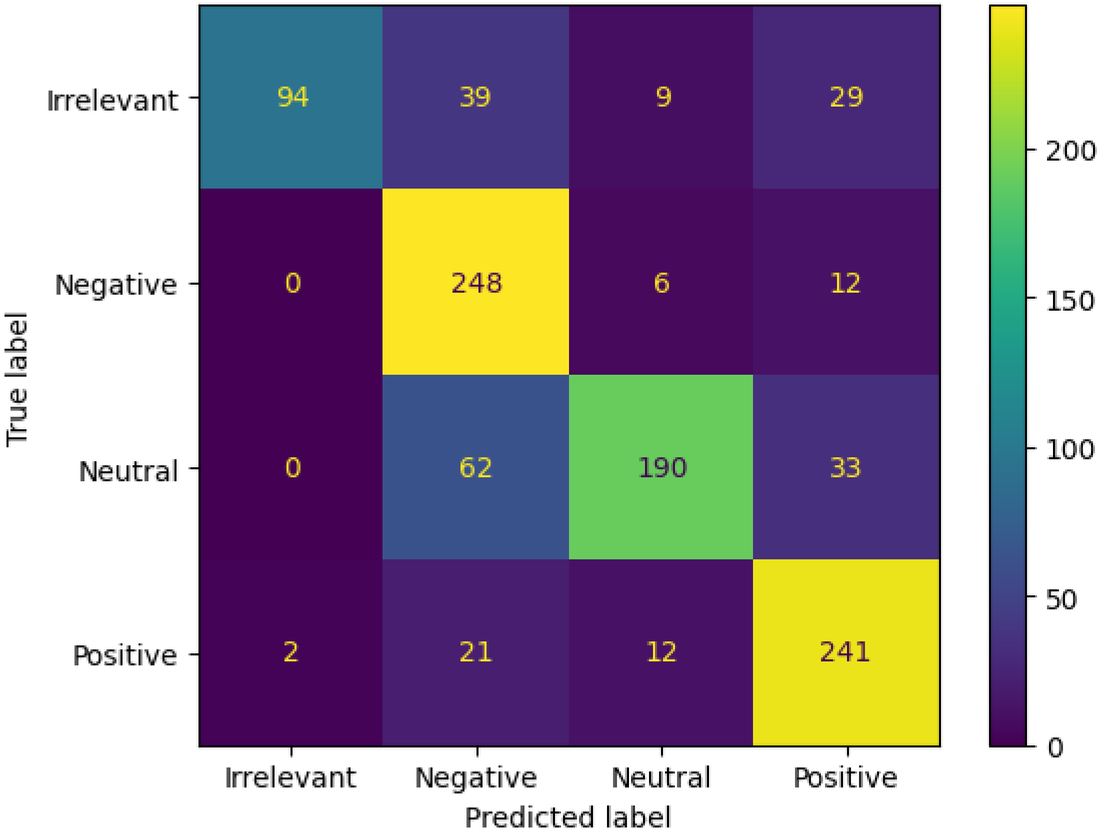
Data Modelling Results:



```
Validation Set Metrics:
+-----------+-------+
| Metric    | Value |
+-----------+-------+
| Accuracy  |  0.73 |
| Precision |  0.77 |
| Recall    |  0.73 |
| F1 Score  |  0.72 |
+-----------+-------+
```



```
Test Set Metrics:
+-----------+-------+
| Metric    | Value |
+-----------+-------+
| Accuracy  |  0.77 |
| Precision |  0.81 |
| Recall    |  0.77 |
| F1 Score  |  0.77 |
+-----------+-------+
```



1. The accuracy metric quantifies the model's ability to make correct predictions across the entire dataset. It demonstrates that the model maintains a consistently high level of accuracy and performs particularly well when dealing with new, previously unseen data.
2. Precision provides insight into the proportion of predicted positive instances that are genuinely positive. Notably, the model exhibits a high level of precision, and it is worth mentioning that the test data outperforms the training data in this regard.

3. Recall measures the model's effectiveness in correctly identifying actual positive instances. The model consistently achieves a high recall rate, indicating its capability to perform well on previously unencountered data.
4. The F1 score, which combines precision and recall through their harmonic mean, signifies the model's capacity to simultaneously maximize both precision and recall. The model consistently demonstrates a high F1 score, with the additional observation that the test data surpasses the training data's performance.
5. The confusion matrix reveals that the model exhibited stronger predictive performance for categorizing tweets as neutral or irrelevant compared to its performance with positive or negative tweets. Notably, the model demonstrated a higher degree of accuracy when classifying positive tweets as opposed to negative tweets.

Analysis Conclusions:

1. Sentiments display a balanced distribution by entity but a skewed distribution by language, indicating language-specific sentiment trends.
2. Entities maintain similar sentiment entry counts, showcasing balanced sentiment representation.
3. Negative and positive tweets collectively outnumber neutrals, with a prevalence of negative sentiment.
4. English tweets dominate, followed by German, Somali, and Afrikaans in descending frequency.
5. Games like Assassins Creed, Borderlands, and Cyber Punk are associated with a high proportion of positive tweets.
6. Sports categories such as NGL, NBA, and FIFA exhibit the highest negative tweet percentages.
7. Prominent brands like Facebook, Amazon, and Google have a substantial percentage of neutral or irrelevant tweets.
8. Frequently discussed topics include Call of Duty, NFL, and Verizon.
9. Croatian, Hungarian, and Polish languages lead with the highest positive tweet percentages.
10. Ukrainian, Albanian, and Indonesian languages show the highest negative tweet percentages, while Macedonian, Thai, and Bulgarian tweets predominantly contain neutral sentiments.

Considerations to keep in mind:

1. The dataset is not inclusive of all the details regarding tweets. As a result, the accuracy of this information is confined to the scope of the provided dataset.
2. Apart from the discussed metrics there are many other metrics that effect sentiment analysis like region.

Recommendations:

1. Dataset can be increased to get similar representation in tweets by language as is the case with entity.
2. With a prevalence of negative sentiments, it's essential to identify and address the root causes of negative sentiment, especially in sports-related content (NGL, NBA, FIFA).
3. The high positive sentiment percentages for popular games (Assassins Creed, Borderlands, Cyber Punk) can be leveraged to strengthen positive brand associations and promote these games further.
4. For languages with high negative tweet percentages (Ukrainian, Albanian, Indonesian), implement language-specific strategies to understand negative sentiment more effectively.
5. Engaging with audiences in languages other than English, such as German, Somali, and Afrikaans, can be means to expanding reach.
6. The Multinomial Naive Bayes Algorithm is good for future sentiment analysis as shown by high accuracy, precision, recall and F1 score on test data but better model can be built on top of this model for better sentiment prediction.