Data Quality and Web Text Mining

Final Project

Sentimental Analysis Using Twitter Dataset

Report

Written By

Sai Sirisha Madishetty

Date submitted: 12/13/2023

Table of Contents

Introduction:

In the rapidly evolving digital landscape, sentiment analysis has emerged as a crucial tool for understanding public opinion and emotions expressed in textual data. Among various platforms, Twitter stands out as a particularly rich source for such analysis, given its widespread use and the concise, opinion-rich nature of its content. This study aims to harness the power of Twitter to analyze sentiments expressed by the platform users. Despite the challenges posed by the platform's unique language style, including the use of slang and emojis, this analysis seeks to offer insightful perspectives on public opinion. To navigate these challenges, our study employs advanced natural language processing techniques and machine learning models. We will systematically gather and interpret Twitter data, aiming to understand perspectives from people on various topics around the globe. For example, companies can predict their product reception, and politicians can assess election prospects. Understanding Twitter's distinctive features is key to effective sentiment analysis.

Twitter, as a social media platform, allows users to express opinions in short messages or tweets, accompanied by images and videos. Interactions include likes, comments, and reposts. With over 206 million daily active users in 2022, Twitter's data is a goldmine for sentiment analysis. The challenges in sentiment analysis include data sparsity due to Twitter's word limit and diverse content. Building an efficient sentiment classifier involves real-time algorithm deployment, considering Twitter's open and dynamic nature.

A central component of our project methodology is the classification of sentiments into distinct polarities - positive, neutral, and negative. The ability to accurately categorize these polarities is fundamental to understanding the overall sentiment trends on Twitter. It allows for a more comprehensive understanding of public opinion, catering to the diverse emotional expressions found in the digital discourse. This polarity-based classification serves as a critical tool in our project, enabling us to systematically analyze and interpret the vast and varied sentiments expressed by users across the globe.

To overcome these challenges, our project uses the Naive Bayes classifier. This approach is particularly advantageous due to its speed and accuracy, essential for real-time sentiment analysis. Naive Bayes is well-equipped to handle the nuances of Twitter’s language, including slang and emojis, by converting these textual elements into quantifiable features for sentiment classification. This ensures a more robust and comprehensive understanding of public opinion trends as expressed on Twitter, making it a valuable tool in our sentiment analysis.

Problem definition

The primary objective of this project is to effectively perform sentiment analysis on Twitter data, with a focus on categorizing tweets as positive, negative, or neutral. This analysis is particularly challenging due to the unique characteristics of Twitter data, which includes a diverse range of expressions, the use of slang, and emojis, and the brevity of content due to the platform's word limit. The project aims to address these challenges through a comprehensive methodology that encompasses data collection, preprocessing, exploratory data analysis, feature extraction, model training and evaluation, and finally, the sentiment analysis of tweets containing specific keywords.

Background:

Sentiment analysis is not a new task, it has been studied since the 90s. However, in the 2000s Sentimental analysis attracted the interest of scientists due to its significance in different scientific areas also it had many unstudied research questions. Moreover, the wide availability of opinionated data pushed research in this area to a new stage.

In other words, sentiment analyses deal with the processing of opinionated text to extract and categorize opinions from certain documents. The polarity of sentiment is usually expressed in terms of positive or negative opinion (binary classification). However, it can be a multiclass classification, hence sentiment may have a neutral label or even a broadened variation of labels like very positive, positive, neutral, negative, and very negative, also labels can be associated with emotions like sad, anger, fearful, happy, etc.

Sentiment analysis is a developing area that arouses the interest of humans and especially organizations because it can be used for the decision-making process. Individuals are no longer limited to asking opinions from friends about products or services they can freely find such information on the Internet. Furthermore, organizations may save time and money by avoiding conducting surveys instead they can concentrate on processing opinions that can be obtained from the Web freely. Nevertheless, it is important to notice that sources that contain opinionated data are noisy sometimes, so it is important to extract the essential meaning from that information to use it further. Sentimental analysis uses different techniques and approaches for handling this challenging task.

Describing the dataset:

The Dataset Name “Tweets.csv”. This dataset comprises a collection of tweets, each uniquely identified and timestamped, reflecting a diverse range of user expressions on Twitter. This dataset contains 530,973 entries of tweets expressed by its users in the year 2009. The Columns of the dataset include:

Id: Unique identifier for each tweet, ensuring data integrity and enabling individual tweet analysis.

Date: Timestamp for each tweet, following the format "Mon Apr 06 22:19:45 PDT 2009", which is crucial for temporal analyses and trend identification.

Flag: A field named 'Flag', uniformly filled with 'NO_QUERY' in the sample, potentially indicating unfiltered tweet collection.

User: The username of the Twitter account, providing insights into user-level patterns and behaviours.

Text: The actual tweet content, a rich source for natural language processing, sentiment analysis, and social media studies.

NLP Methods/Algorithms Used:

(i) Data Pre-Processing:

In the realm of sentiment analysis, particularly when dealing with the dynamic and often chaotic nature of Twitter data, data preprocessing stands as a pivotal phase. This process transforms raw tweets into a clean, structured format, suitable for effective analysis. Initially, it involves the elimination of noise - URLs, mentions (@usernames), and irrelevant special characters are stripped away to declutter the text. This cleaning extends to the normalization of the dataset, where all text is converted to a uniform case, typically lowercase, to ensure consistency in interpretation.

For the Twitter dataset, the data pre-processing involves:

- **Cleaning:** Removing Noise: Tweets often contain URLs, hashtags, mentions, and special characters that may not be useful for analysis.
- **Normalization:** Converting text to a consistent format, like lowercasing all words.
- **Handling Emojis and Slang:** Tweets often use informal language and emojis, which require special handling to interpret correctly.
- **Tokenization:** Breaking down text into individual words or tokens. This is fundamental for most NLP tasks.

- **Stop Words Removal:** Removing common words that may not contribute to the overall meaning of the text.
- **Stemming and Lemmatization:** Reducing words to their base or root form. Lemmatization is more sophisticated as it considers the context of the word.

(ii) Feature Extraction:

In sentiment analysis, The Feature extraction is a critical step where raw text data is transformed into a numerical format that machine learning models can understand and process. This process involves extracting meaningful attributes or "features" from the text, which represent the essential information in a way that can be used for analysis.

The below is the overview of how we used feature extraction in sentiment analysis:

Vectorization:

The primary method in feature extraction is vectorization, where text data is converted into a vector of numbers.

TF-IDF (Term Frequency-Inverse Document Frequency): We used TF – IDF method, as this is a more advanced technique that not only counts the frequency of words but also adjusts for the fact that some words appear more frequently in general. It helps in highlighting the importance of words that are frequent in a document.

Machine Learning Algorithm:

The machine learning approaches can construct classifiers to complete sentiment classification by extracting feature vectors, which mainly includes steps including data collecting and cleaning, extracting features, training data with the classifier, and analyzing results. Here, we are using the **Naïve Bayes Classifier** for sentimental analysis.

Naïve Bayes Classifier:

The Naive Bayesian Classification represents a supervised learning method as well as a statistical method for classification. It is probabilistic model and it permit us to capture uncertainty about the model in a principled way by determining probabilities. It helps to solve diagnostic and predictive problems. This Classification is named as Naive Bayes after Thomas Bayes, who proposed the Bayes Theorem of determining probability. Bayesian classification provides useful learning algorithms and past knowledge and observed data can be combined. It helps to provide a useful perspective for understanding and evaluating many learning algorithms. This helps to determine exact probabilities for hypothesis, and it is robust to noise in input data.

The following is the Bayes theorem used,

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)}$$

With this approach, which assumes that every word in each category is independent of one another, the calculation can be simplified and can be written as

$$\mathrm{p}(\mathrm{X}\,|\mathrm{c_j})=\prod_{\mathrm{i}=1}^{\mathrm{n}}\mathrm{p}(\mathrm{w_i}|\mathrm{c_j})$$

By Using the above equation, It can also be written as

$$\mathrm{C_{MAP}argmax}\;\;\mathrm{p}\left(\mathrm{c_j}\right)\prod_{\mathrm{i}=1}^{\mathrm{n}}\mathrm{p}(\mathrm{w_i}|\mathrm{c_j})$$

When applied to sentiment analysis, the classifier considers the frequency and co-occurrence of words in documents that have been manually labelled as positive, negative, or neutral. Through training, it learns which words are most likely to appear in each category. During the classification process, it uses these probabilities to infer the sentiment of new, unlabelled pieces of text.

For instance, in a corpus of product reviews, the word "excellent" might appear more frequently in positive reviews than in negative ones. The classifier will use this statistical trend to predict that a new review containing the word "excellent" is more likely to be positive. Similarly, words frequently found in negative reviews, like "disappointing," will sway the classifier toward predicting a negative sentiment.

Using Naïve Bayes Classifier, we can determine the accuracy of classification. Generally, the algorithm the accuracy turns out to be 76%. after the modelling is done the data is given sentiments and distributed according to the polarity – positive, negative, and neutral.

Implementation:

Sequence of Operations:

1. Setup and Environment Configuration

The project's foundation was set in a Python development environment, chosen for its strong capabilities in data science and natural language processing (NLP). The primary libraries included:

1. **NLTK:** Employed for NLP tasks such as tokenization and stopword removal.
2. **scikit-learn:** Utilized for machine learning models and data preprocessing techniques.
3. **Pandas and NumPy:** Integral for data manipulation and numerical calculations.

2. Data Collection Process

The project utilized a pre-existing dataset of tweets, which contained a variety of entries relevant to the study's objectives. The dataset included essential metadata for each tweet, such as the timestamp, user ID, and the tweet text. This structured dataset was the primary source for all subsequent analyses.

3. Data Preprocessing

Preprocessing was a crucial phase, involving:

1. **Cleaning:** Utilization of regular expressions to cleanse the tweets by eradicating URLs, mentions, hashtags, and non-alphanumeric characters, along with emoji handling.
2. **Normalization**: Conversion of all text to lowercase to maintain consistency throughout the dataset.
3. **Tokenization:** Deployment of the NLTK library for breaking down tweets into individual words or tokens.
4. **Stop Word Removal:** Elimination of frequently occurring words that provide minimal unique information.
5. **Stemming and Lemmatization:** Application of NLTK’s Porter Stemmer and WordNetLemmatizer to condense words to their root forms.

4. Exploratory Data Analysis

This stage focused on extracting initial insights from the dataset:

- **Tweet Characteristics:** Analysis included examining tweet lengths, word counts, and common words.

- **Visualization:** Creation of word clouds and bar charts to visually represent significant terms and dataset attributes.

5. Feature Extraction

This step transformed textual data into a numerical format for machine learning:

- **TF-IDF Vectorization:** Implementation with scikit-learn to convert tweets into a matrix of TF-IDF features.
- **Dimensionality Reduction:** Consideration of techniques like Principal Component Analysis (PCA) for feature space reduction while retaining critical information.

6. Model Development

Various machine learning models were explored:

- **Naive Bayes:** Selected for its baseline efficiency in text classification tasks.
- Model parameters were finely tuned using grid search and cross-validation methods.

7. Evaluation and Testing

The models underwent thorough evaluation using:

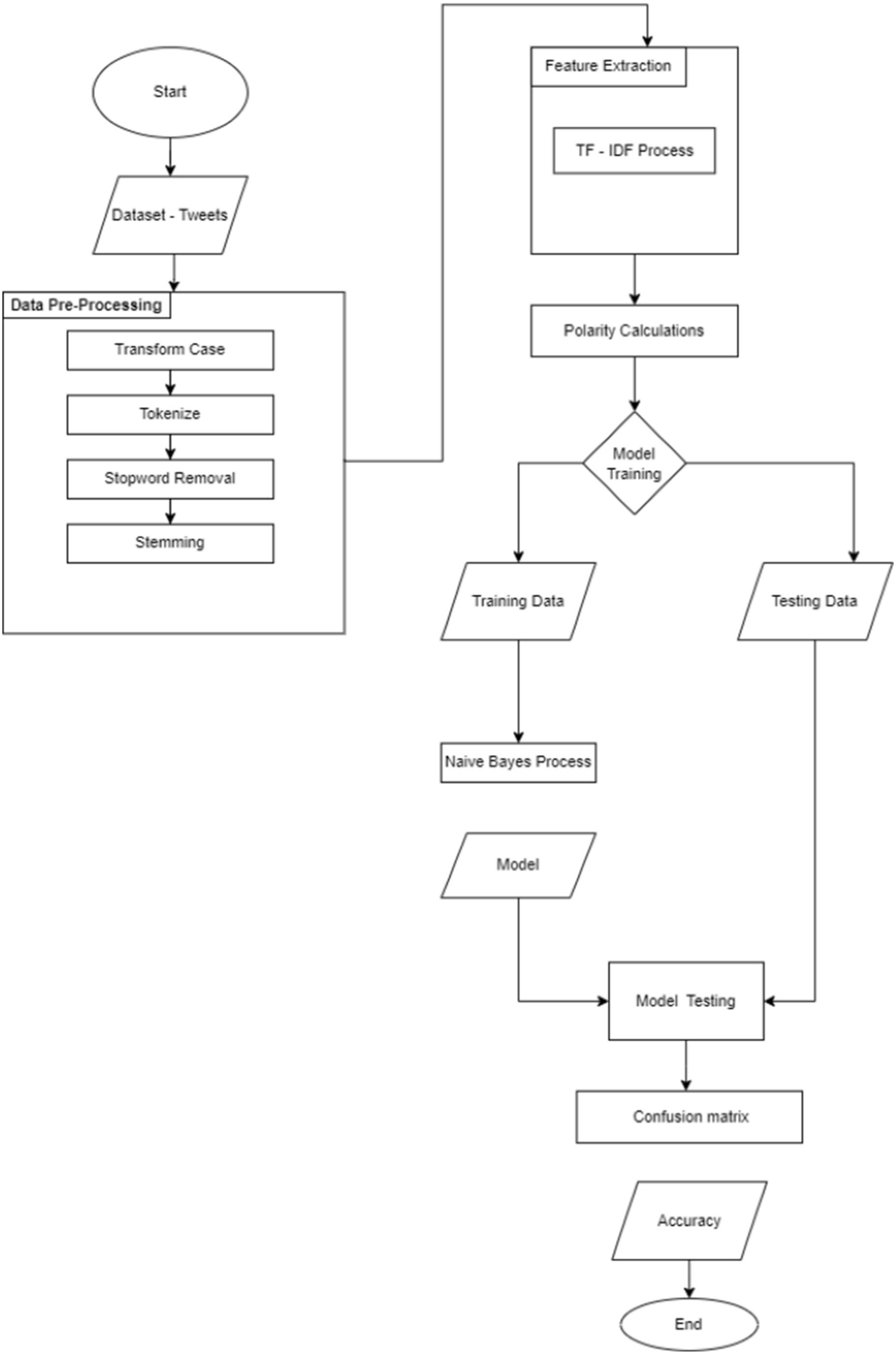
- **Accuracy, Precision, Recall, and F1-Score:** These metrics gauged model performance comprehensively.
- **Confusion Matrix:** Helped in identifying misclassifications and biases.

8. Keyword-Specific Analysis

A feature was developed to analyze sentiments in tweets containing specific keywords from the dataset:

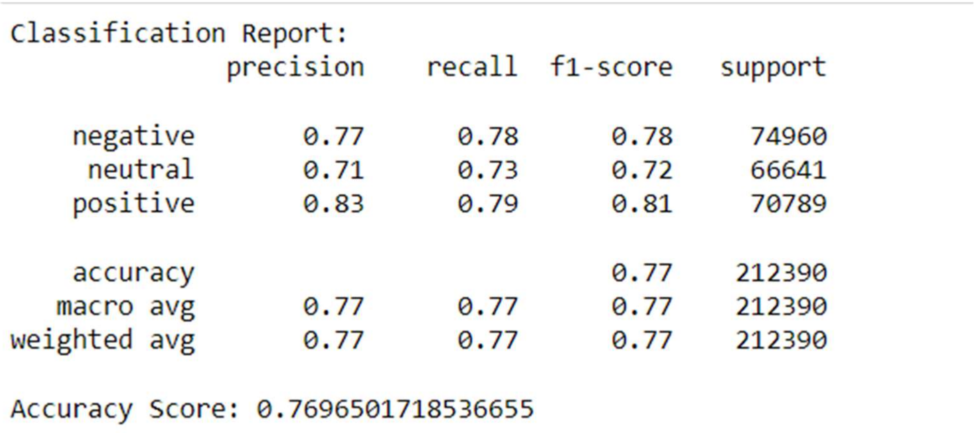
- **Filtering by Keyword**: The dataset was queried for tweets containing user-specified keywords.
- **Sentiment Prediction:** The trained models were applied to these filtered tweets for sentiment analysis.

Flow Chart:


Start
Dataset - Tweets
Data Pre-Processing
Transform Case
Tokenize
Stopword Removal
Stemming
Feature Extraction
TF - IDF Process
Polarity Calculations
Model Training
Training Data
Testing Data
Naive Bayes Process
Model
Model Testing
Confusion matrix
Accuracy
End

Results:

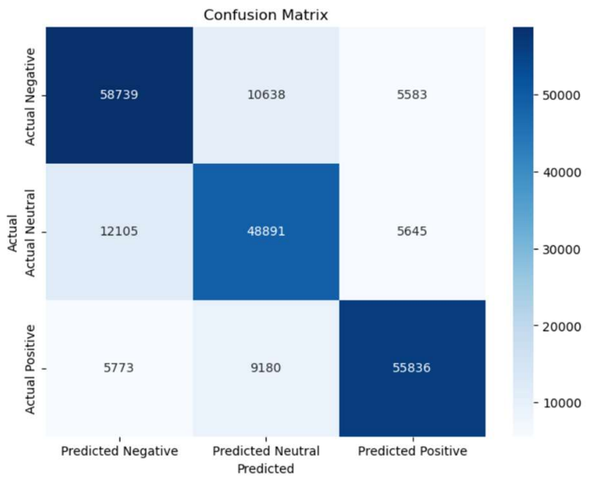
Based on the above Implementation, the results of the Naïve Bayes Classification are shown below,



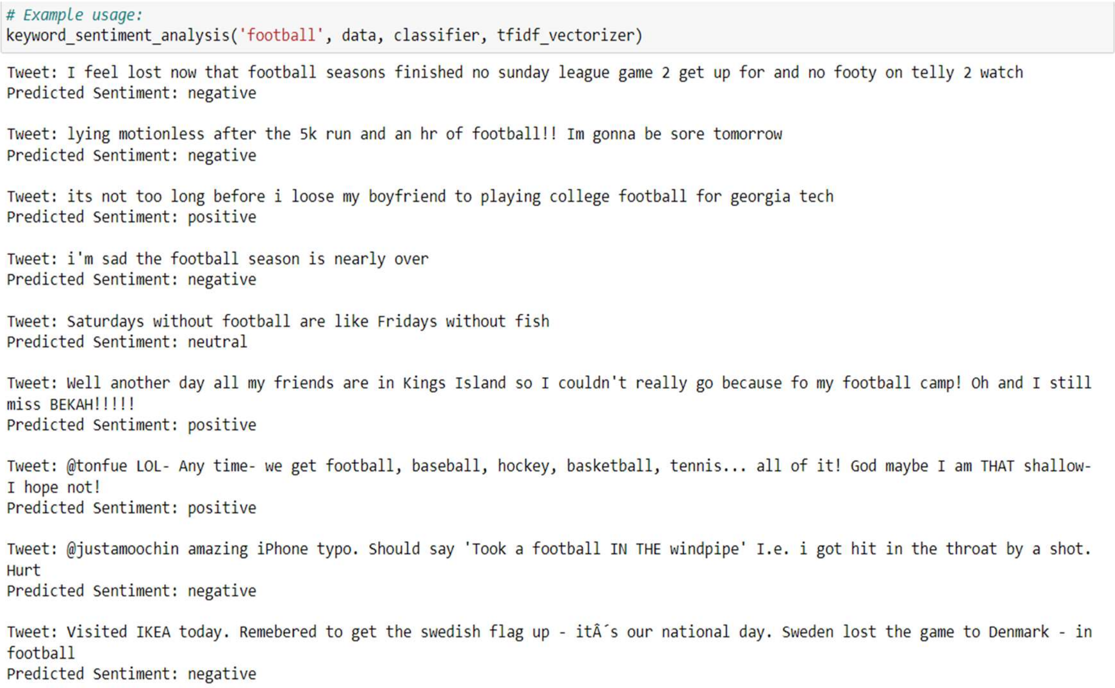
```
Classification Report:
              precision    recall  f1-score   support

    negative       0.77      0.78      0.78     74960
     neutral       0.71      0.73      0.72     66641
    positive       0.83      0.79      0.81     70789

    accuracy                           0.77    212390
   macro avg       0.77      0.77      0.77    212390
weighted avg       0.77      0.77      0.77    212390

Accuracy Score: 0.7696501718536655
```

From the above picture, the classification report for the sentiment analysis model reveals a balanced performance across three sentiment categories: negative, neutral, and positive. The model demonstrates the highest efficacy in identifying positive tweets with an F1-score of 0.81, and an overall precision and recall of 0.77. Neutral sentiments have slightly lower precision and recall, suggesting a potential area for improvement. The model achieves an overall accuracy of 77%, indicating a reliable performance in sentiment classification. The support values show a relatively even distribution of sentiment classes within the dataset, confirming the model's consistent application across different sentiments.

Here, we also generated the confusion matrix to visualizes the performance of the sentiment analysis model. The matrix shows that the model is most accurate with negative sentiment predictions, followed closely by positive and then neutral. The off-diagonal numbers represent classification errors, with the model confusing neutral with both negative and positive sentiments to a significant extent, which could suggest areas for improvement in distinguishing neutral sentiments.



In the project, there's an input cell on the ipynb file, where users can enter any keyword. The system then automatically picks 10 random tweets related to that keyword and determines their sentiment as either positive, negative, or neutral. For instance, when I used "Football" as a keyword, the machine generated the following output. Based on the analysis of this output, it's evident that the machine demonstrates high accuracy in sentiment prediction.



```
# Example usage:
keyword_sentiment_analysis('football', data, classifier, tfidf_vectorizer)
```

```
Tweet: I feel lost now that football seasons finished no sunday league game 2 get up for and no footy on telly 2 watch
Predicted Sentiment: negative

Tweet: lying motionless after the 5k run and an hr of football!! Im gonna be sore tomorrow
Predicted Sentiment: negative

Tweet: its not too long before i loose my boyfriend to playing college football for georgia tech
Predicted Sentiment: positive

Tweet: i'm sad the football season is nearly over
Predicted Sentiment: negative

Tweet: Saturdays without football are like Fridays without fish
Predicted Sentiment: neutral

Tweet: Well another day all my friends are in Kings Island so I couldn't really go because fo my football camp! Oh and I still
miss BEKAH!!!!!
Predicted Sentiment: positive

Tweet: @tonfue LOL- Any time- we get football, baseball, hockey, basketball, tennis... all of it! God maybe I am THAT shallow-
I hope not!
Predicted Sentiment: positive

Tweet: @justamoochin amazing iPhone typo. Should say 'Took a football IN THE windpipe' I.e. i got hit in the throat by a shot.
Hurt
Predicted Sentiment: negative

Tweet: Visited IKEA today. Remebered to get the swedish flag up - itÂ´s our national day. Sweden lost the game to Denmark - in
football
Predicted Sentiment: negative
```

Conclusion:

In Conclusion, the use of the Naive Bayes Classifier in sentiment analysis of Twitter data has yielded commendable results, demonstrating the classifier's adaptability and efficiency in handling complex, unstructured text data. As a fundamental part of our project, the Naive Bayes Classifier effectively categorized tweets into positive, negative, and neutral sentiments, showcasing its capability in navigating the nuances of Twitter's concise and often informal language. The classifier's performance was notable, achieving an overall accuracy of 77%, a significant accomplishment given the inherent challenges of sentiment analysis in social media contexts. To improve the evaluation accuracy, we need something to take the context and references into consideration. I will try to build an LSTM network, and benchmark its results compared to this NLTK machine learning implementation.

Future Research:

In Sentimental Analysis, there is a significant potential that can substantially elevate the depth and accuracy of our insights. The exciting development in sentiment analysis is the inclusion of audio data. This would involve the use of advanced speech-to-text conversion technologies to transform audio content from videos or voice-based tweets into analyzable text. Moreover, the implementation of emotion recognition capabilities in speech processing can delve deeper, capturing the emotional nuances and intonations inherent in spoken language, which are often lost in textual conversion.

The analysis of video content presents yet another frontier. By employing facial recognition technology, we can analyze facial expressions in video content, offering powerful insights into emotional states. Facial expressions often convey sentiments more powerfully than words alone. Additionally, contextual analysis of video content considering the background, actions, and interactions within a video can provide a more nuanced understanding of the sentiments being expressed.

Perhaps the most transformative enhancement would be the development of a multimodal sentiment analysis framework. This approach would integrate data from text, audio, and video, offering a holistic view of sentiments. By combining the strengths of each modality, such a framework could achieve a level of accuracy and depth in sentiment analysis that is currently unattainable with single-mode analyses. This comprehensive approach would not only increase the accuracy of sentiment detection but also provide a richer, more dimensional understanding of public opinion and emotional trends across various media formats.

References:

1. Musto, C., Semeraro, G., Polignano, M. (2014). A comparison of lexicon-based approaches for sentiment analysis of microblog posts. Information Filtering and Retrieval, 59
2. 19. Ding, X., Liu, B., Yu, P. S. (2008, February). A holistic lexicon-based approach to opinion mining. In Proceedings of the 2008 international conference.
3. Widiyaningtyas, Triyanna & Zaeni, Ilham & Al Farisi, Riswanda. (2019). Sentiment Analysis Of Hotel Review Using N-Gram And Naive Bayes Methods. 1-5. 10.1109/ICIC47613.2019.8985946.
4. Uğuz, H. (2011). A Two-Stage Feature Selection Method for Text Categorization by using Information Gain, Principal Component Analysis

and Genetic Algorithm. KnowledgeBased Systems, 24(7), 1024-1032. doi:10.1016/j.knosys.2011.04.014

5. Rana, S., & Singh, A. (2017). Comparative Analysis of Sentiment Orientation using SVM and Naïve Bayes Techniques. International Conference on Next Generation Computing Technologies (NGCT), 106-111. doi:10.1109/NGCT.2016.7877399

6. Muthia, D. A. (2014). Sentiment Analysis of Hotel Review Using Naïve Bayes Algorithm and Integration of Information Gain and Genetic Algorithm as Feature Selection. International Seminar on Scientific Issues and Trends (ISSIT), 25-30. Retrieved from http://issit.bsi.ac.id/proceedings/index.php/issit2014/article/view/30/30

7. H. Parveen and S. Pandey, "Sentiment analysis on Twitter Data-set using Naive Bayes algorithm," 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), Bangalore, India, 2016, pp. 416-419, doi: 10.1109/ICATCCT.2016.7912034.

8. F. Neri, C. Aliprandi, F. Capeci, M. Cuadros and T. By, "Sentiment Analysis on Social Media," 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Istanbul, Turkey, 2012, pp. 919-926, doi: 10.1109/ASONAM.2012.164.

9. Chandra Prakash Gupta, V. V. Ravi Kumar, "Sentiment Analysis and its Application in Analysing Consumer Behaviour", *2023 International Conference on Emerging Techniques in Computational Intelligence (ICETCI)*, pp.332-337, 2023.

10. Govindasamy, Shobana. (2018). Twitter Sentimental Analysis. 7. 343-346.