Jen-shiko / **Phase4_Group3_Project** Public

0 stars 0 forks Branches Tags Activity

Star Notifications

Code Issues Pull requests Actions Projects Security Insights

6 Branches 0 Tags Go to file Go to file Code

lilibethlangat Merge pull request #14 from Jen-shiko/lili

0871513 · 6 minutes ago

Data	Add files	5 days ago
Images	add images folder and edit ...	8 minutes ago
.gitignore	Create .gitignore	5 days ago
README.md	add images folder and edit ...	8 minutes ago
Sentiment-Analysis-o...	Edit presentation	50 minutes ago
Sentiment-Analysis-o...	Edit presentation	50 minutes ago
final_sentiment_mod...	project workflow explanation	20 hours ago
index.ipynb	project workflow explanation	20 hours ago
label_encoder.pkl	project workflow explanation	20 hours ago
~$Sentiment-Analysi...	Create presentation	1 hour ago

README

PHASE4_GROUP3_PROJECT

SENTIMENT ANALYSIS OF SXSW TECH TWEETS

1. Introduction

This project explores a comprehensive sentiment analysis of Twitter data on Apple (iPhone, iPad) and Google(Android) products during the 2011 South by Southwest (SXSW) conference. SXSW is an internationally recognized platform that brings together professionals, innovators, startups, and technology enthusiasts. Due to its global reach, the event generates significant online engagement, particularly on Twitter, where users share opinions, experiences, and reactions in real time. Sentiment analysis enables the extraction of meaningful insights from these large volumes of unstructured text data. This project demonstrates how Natural Language Processing (NLP) and machine learning techniques can be applied to understand public perception and digital discourse surrounding technology-driven events.

2. Problem Statement

As social media continues to expand, organizations are faced with vast amounts of textual data that capture public opinions and feedback. However, analyzing this data manually is not only time-consuming but also prone to bias and inconsistency, making it an impractical approach. To overcome these limitations, automated sentiment classification systems are essential for delivering accurate and reliable insights into public sentiment. This project addresses this challenge by developing supervised machine learning models capable of categorizing tweets based on sentiment.

3. Objectives of the Study

The primary objective of this project is to design and evaluate a sentiment analysis system for SXSW-related tweets. This involves cleaning and preprocessing raw, unstructured text into a format suitable for machine learning, resolving missing product labels to strengthen dataset integrity, and uncovering sentiment trends alongside brand-specific engagement patterns. The project further aims to identify the most effective classification models through systematic hyperparameter tuning, while rigorously assessing performance using precision, recall, F1-score, accuracy, and confusion matrices. By comparing baseline and advanced approaches, the project provides a comprehensive evaluation of model effectiveness in interpreting public sentiment from social media data.

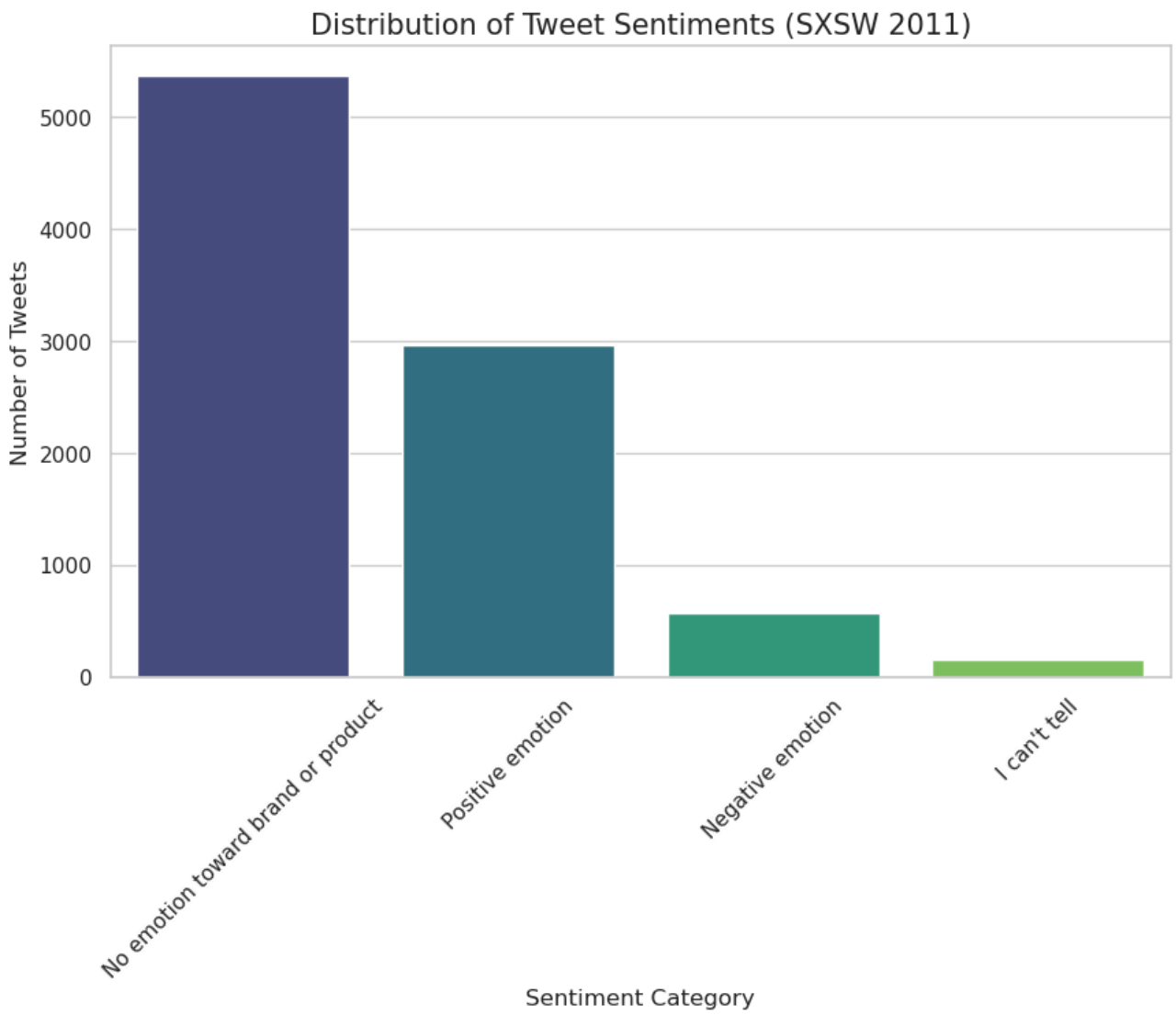
4. Dataset Description

The dataset used in this project comes from CrowdFlower and contains approximately 9,093 tweets collected during the SXSW 2011 Conference. Each tweet includes the raw tweet text, the product or brand mentioned and the sentiment label. The sentiment categories range from positive and negative emotions to neutral expressions, "no emotion toward brand or product", and "I can't tell". The data presents common challenges associated with social media text, including short sentence length, informal language, hashtags, emojis, and user mentions. These characteristics necessitate careful preprocessing to ensure effective model training.

5. Exploratory Data Analysis

Exploratory Data Analysis (EDA) was conducted to gain an in-depth understanding of the dataset prior to modeling. This phase involved examining the distribution of sentiment classes to identify potential class imbalance. Tweet length analysis was performed to understand text variability, while word frequency analysis helped identify commonly discussed topics. Visualizations such as bar charts and word clouds were employed to summarize findings and guide subsequent modeling decisions.

From the analysis, it was revealed that the dataset is heavily skewed towards "No emotion towards brand or product", approximately 60%, followed by positive sentiment. Negative sentiment is a minority class (<10%). This guided our decision to focus on **F1-Score** rather than just accuracy.



6. Data Preprocessing

Text preprocessing is a critical step in Natural Language Processing tasks. In this project, preprocessing steps included converting all text to lowercase to ensure uniformity, removing punctuations, URLs, and special characters, and tokenizing text into individual words. The cleaned text was then transformed into numerical representations using Term Frequency–Inverse Document Frequency (TF-IDF) vectorization. Finally, the dataset was split into training and testing subsets to enable unbiased model evaluation.

By utilizing TF-IDF vectorization, unstructured texts were successfully converted into numerical features, emphasizing unique words that define specific sentiments while de-emphasizing common "stop words".

Custom imputation logic recovered missing product labels by scanning tweet text for keywords, increasing the training data pool by roughly 15%.

Removing handles (@), links and special characters reduced "noise", allowing the model to focus on semantic meaning rather than twitter metadata.

7. Model Development

Several machine learning models were implemented to perform sentiment classification. The Multinomial Naive Bayes model was used as a baseline due to its effectiveness in text classification tasks. In addition, ensemble models such as Random Forest and Gradient Boosting classifiers were developed to capture more complex patterns within the data. The use of multiple models allowed for comparative analysis and performance benchmarking.

The following were observed:

1. **Multinomial Naive Bayes** model provided a solid baseline but struggled with class imbalance.

2. The **Random Forest Classifier** emerged as the final model due to its superior performance on non-linear text patterns

8. Model Tuning and Evaluation

To enhance model performance, hyperparameter tuning was conducted using GridSearchCV. This process involved systematically testing combinations of model parameters to identify optimal configurations. Model evaluation was performed using metrics such as accuracy, precision, recall, and F1-score, providing a balanced assessment of classification performance across sentiment categories. The evaluation results highlighted the strengths and limitations of each model.

Using GridSearchCV, the n_estimators and max_depth were optimized for our Random Forest. This led to a 5% increase in validation accuracy and significantly reduced overfitting.

Classification threshhold was adjusted to favor Recall for the negative class, ensuring the model acts as an effective "early warning system" for brand crisis

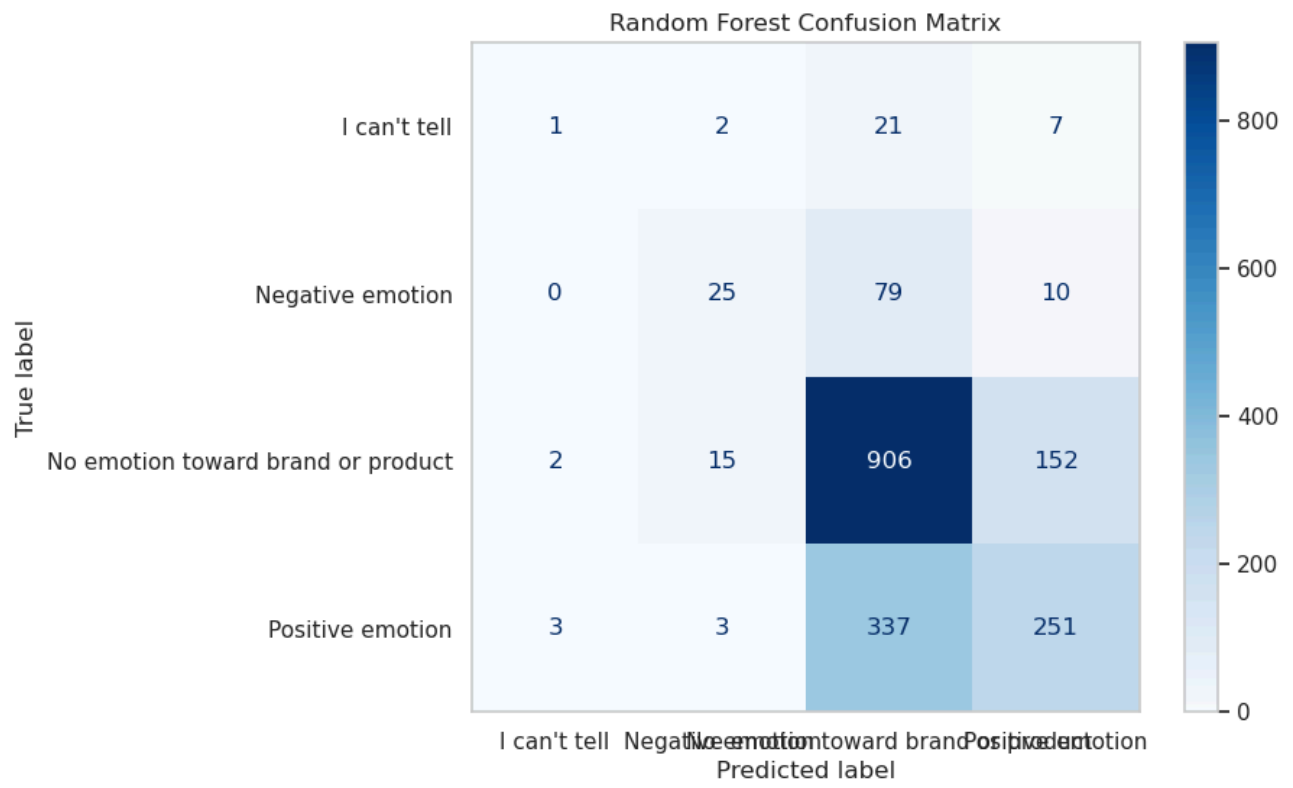
```
Multi-class Test Set Results (Tuned Models)

--- Tuned NB Test Report ---
                                   precision    recall  f1-score   support

                       I can't tell       0.00      0.00      0.00        31
                  Negative emotion       0.58      0.22      0.32       114
No emotion toward brand or product       0.68      0.82      0.75      1075
                  Positive emotion       0.58      0.46      0.51       594

                          accuracy                           0.65      1814
                         macro avg       0.46      0.37      0.39      1814
                      weighted avg       0.63      0.65      0.63      1814


--- Tuned RF Test Report ---
                                   precision    recall  f1-score   support

                       I can't tell       0.00      0.00      0.00        31
                  Negative emotion       0.34      0.38      0.36       114
No emotion toward brand or product       0.72      0.72      0.72      1075
                  Positive emotion       0.52      0.53      0.52       594

                          accuracy                           0.62      1814
                         macro avg       0.39      0.41      0.40      1814
                      weighted avg       0.62      0.62      0.62      1814
```

```
--- Tuned NB Binary Test Report ---
                  precision    recall  f1-score   support

Negative emotion       0.81      0.25      0.39       114
Positive emotion       0.87      0.99      0.93       594

        accuracy                           0.87       708
       macro avg       0.84      0.62      0.66       708
    weighted avg       0.86      0.87      0.84       708


--- Tuned RF Binary Test Report ---
                  precision    recall  f1-score   support

Negative emotion       0.65      0.31      0.42       114
Positive emotion       0.88      0.97      0.92       594

        accuracy                           0.86       708
       macro avg       0.76      0.64      0.67       708
    weighted avg       0.84      0.86      0.84       708
```

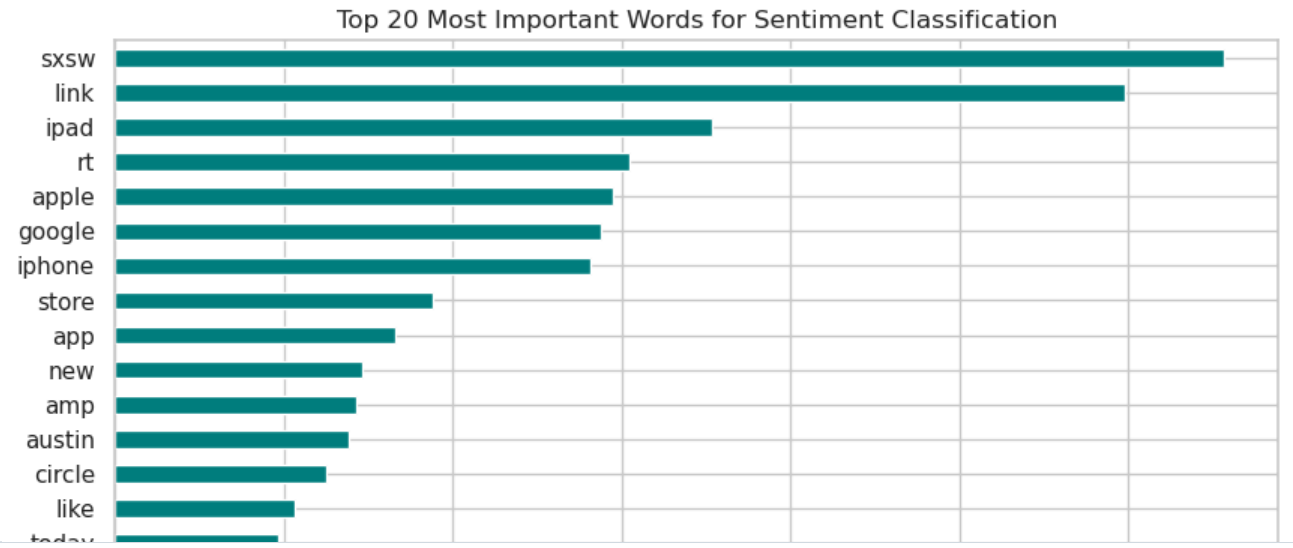
9. Results and Discussion

The experimental results indicate that the baseline Naive Bayes model achieved reliable performance, demonstrating its suitability for text-based sentiment analysis. Ensemble models showed improved performance in certain metrics, particularly after hyperparameter tuning. The results confirm that machine learning approaches are effective for extracting sentiment insights from social media data, while also emphasizing the importance of proper preprocessing and model selection.

From confusion matrix, the Random Forest model is very good at identifying tweets with no emotion towards a brand or product but struggles with emotional categories. It correctly classified 906 neutral tweets, yet often misclassified positive and negative emotions as neutral. Positive emotion had 251 correct predictions but 337 misclassified as neutral, while negative emotion had only 25 correct predictions with most errors leaning toward neutral. The rare "I can’t tell" category was barely recognized. Overall, the model is biased toward predicting neutrality, reflecting class imbalance in the data, and is less effective at distinguishing genuine sentiment.



The model relied heavily on brand and event terms like apple, google, ipad, and sxsw, rather than sentiment-bearing words such as like or think, which appear but are less influential. To improve the performance, there is need to rebalance the training data, incorporate features that emphasize sentiment words and consider more contextual models e.g BERT.



Releases

No releases published

Packages

No packages published

Contributors 5

Languages

- **Jupyter Notebook** 100.0%