Twitter Sentiment Analysis for Product review

Subject: Data Science Laboratory

Name:	Mis:
Kishor Murade	112003091
Pratik Sarode	112003124
Rushikesh Pingle	112003113

Project Guide: Dr. Y. V. Haribhakta

Contents

1 | Introduction

In this project, we will be using data science techniques to perform sentimental analysis on tweets. Twitter is one of the most popular social media platforms where users express their opinions and feelings on a wide range of topics.Our goal in this project is to build a predictive model that can accurately classify tweets as positive, negative or neutral. We will be using text data from tweets tweets are often useful in generating a vast amount of sentiment data upon analysis. These data are useful in understanding the opinion of the people about a variety of topics. Therefore, we need to develop an Automated Machine Learning Sentiment Analysis Model in-order to compute the customer perception.

2 | Objectives

The prime objective of the Twitter Sentiment Analysis for product review are:

- Understand customers' opinions and perceptions about a product.
- Identify areas for product improvement.
- Improve customer service.

3 | Requirements

This project is completed using python3, we use several python libraries like seaborn, numpy, pandas, textblob, sklearn, scipy etc.

4 | Methodology

Below is the step-by-step methodology. And methodology flowchart (Figure 4.1).

- Data Scraping : Scrap the necessary data from twitter using twitter search scraper library, including id, content, username, date reply, like and retweet.
- Data Preprocessing: Preprocess the collected data by removing punctuations, stop words, hashtags, formating the data to confirm that it is ready for analysis.
- Data Exploration and Visualization: Data exploration is the process of analyzing and understanding the structure and characteristics of a dataset. It involves examining the dataset to identify patterns, relationships.
- Model Selection: Choose an appropriate machine learning model for the predict result. Consider different types of models, including support vector machine(SVM), linear regression, naïve bayes, logistic regression.
- Model Training: Fit the chosen model to a training dataset using the features selected during the feature engineering phase.
- Model Evaluation: Evaluate the performance of the trained model on a testing dataset. Use appropriate evaluation metrics, such as accuracy, precision, recall, and F1 score, to assess the model's performance.
- Prediction: Use the best accuracy model to make prediction about product review .
- Conclusion: Evaluate the prediction outcomes and derive conclusions regarding the performance of the model. Offer suggestions for enhancing the model and its application in the future based on the findings.

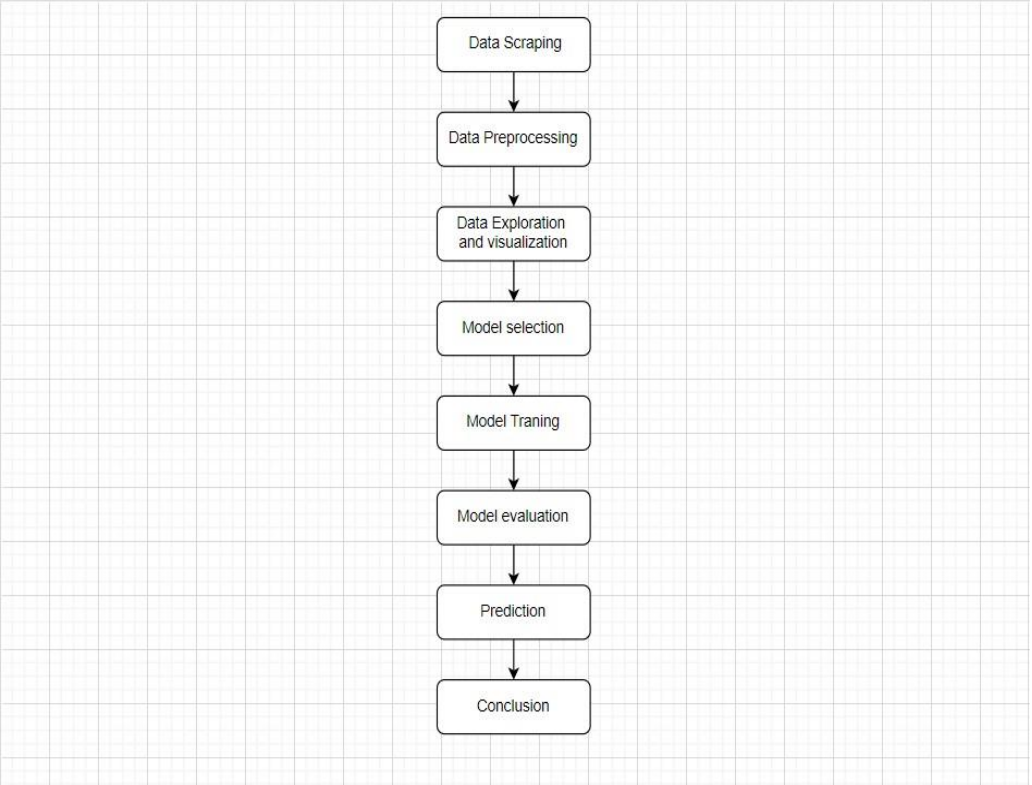
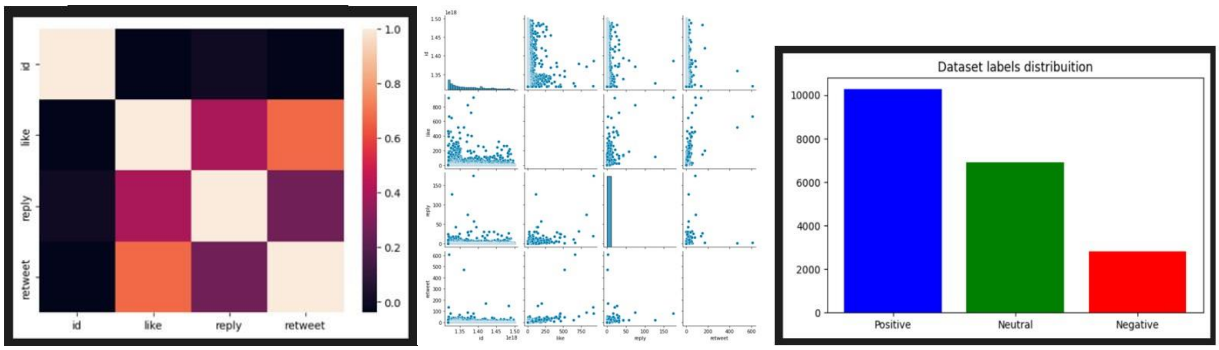


Figure 4.1: Project flow.

5 | Procedure

- Data Collection: Determine the relevant sources of data for the project, such as tweets related to the product or brand being analyzed. Collect the data from reliable and diverse sources, such as Twitter's API and other social media monitoring tools. Store the collected data in a structured format, such as a CSV file or a database, with appropriate labeling and metadata to ensure easy retrieval and analysis.
- Data Cleaning and Preprocessing: Eliminate duplicates and address missing values in the collected data. Convert the data to a uniform format to ensure that it is suitable for analysis. Standardize or normalize the data to ensure that all features have the same scale. Additionally, preprocess the text data by removing stop words, punctuation, and special characters, and perform tokenization and stemming or lemmatization as needed for the analysis.
- Data Exploration and Visualization: Explore and visualize the data to uncover insights into the underlying sentiments and themes in the data. Employ various techniques such as sentiment analysis and topic modeling to detect patterns and relationships between tweets. Utilize visualization tools such as word clouds, bar charts, and heatmaps to depict the data in a meaningful and interpretable way. Additionally, perform statistical analysis to identify any significant differences in sentiment between different categories or time periods.
- Model Selection: Select a suitable machine learning algorithm for the sentiment analysis task, such as Naive Bayes, Support Vector Machines (SVM), Random Forests, or Convolutional Neural Networks (CNN). Evaluate the strengths and weaknesses of each algorithm and choose the one that is best suited to the data and the analysis task. Additionally, consider the performance of the algorithm in terms of accuracy, precision, and recall, and tune the hyperparameters of the selected algorithm to optimize its performance

- Model Training: Partition the data into training and testing sets. Train the chosen algorithm on the training data, using the extracted features from the pre-processing phase. Utilize appropriate evaluation metrics, such as accuracy, precision, recall, and F1 score, to assess the performance of the model. Additionally, employ techniques such as cross-validation to ensure that the model generalizes well to new and unseen data.
- Prediction: Utilize the optimized and trained model to make predictions about sentiment in upcoming tweets. Compare the predictions to the actual sentiments of the tweets. Employ various techniques such as confusion matrices and ROC curves to evaluate the performance of the model
- Results and Conclusion: Analyze the results of the sentiment analysis and draw conclusions regarding the effectiveness of the model. Evaluate the accuracy of the model in terms of its ability to correctly predict the sentiment of tweets. Additionally, explore the limitations of the model and identify areas for improvement



(a) Heatmap **(b)** Pair plot **(c)** Bar plot

Figure 5.1: Data Analysis

6 | Results

Prediction Results

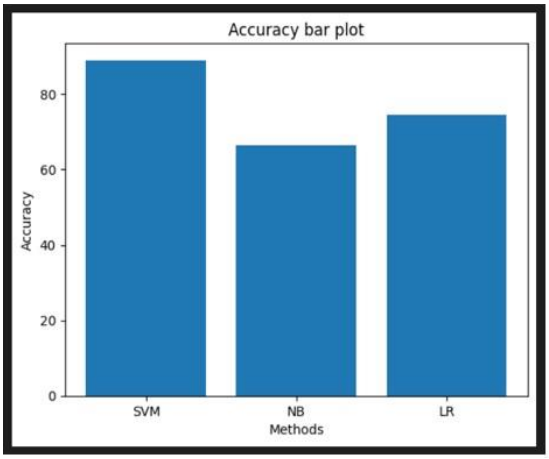


Figure 6.1

7 | Limitations

- Limited message length: Tweets are limited to 280 characters, which may result in incomplete or ambiguous messages that are difficult to accurately interpret.
- Contextual understanding: The context in which a tweet is written can greatly impact its sentiment, but context can be challenging to identify automatically. For instance, a tweet expressing negative sentiment about a sports team may be intended as playful banter rather than a genuine expression of dislike
- Biases in data: Twitter users are not a representative sample of the general population, which may lead to biases in the sentiment analysis results. For example, the opinions expressed on Twitter may not be reflective of the opinions of the broader population.

.

8 | Conclusion

In conclusion, the Twitter Sentiment Analysis project is a fascinating and challenging data science project that can provide valuable insights into public opinion and perception. By following a structured and systematic approach, it is possible to develop a powerful predictive model that can accurately classify the sentiment of tweets related to a particular topic or brand.However, it is important to keep in mind the limitations of the project, such as data quality, bias, noise, language nuances, changing trends, overfitting, and the lack of context. By acknowledging these limitations, it is possible to use the model's predictions in conjunction with other sources of information to make informed decisions.Overall, the Twitter Sentiment Analysis project is an excellent opportunity to apply data science techniques and gain insights into public sentiment and perception. With the right approach and a good understanding of the data and the domain, it is possible to develop a highly accurate and effective predictive model that can help classify the sentiment of tweets and improve decision-making.

.