Sentiment Analysis on Twitter Datasets

Team members

Jaydeep Patil - 122959431
Gautham Vijayaraj - 1229599464
Rahul Babu - 1229832450

1 Abstract

Objective

- To develop a sentiment analysis model using different machine learning techniques to classify the sentiment of text as positive, negative, or neutral.
- To achieve an accuracy more than that of existing models and provide a comparative analysis of different machine learning techniques.
- To provide a visual representation of the most used and significant words in the text using Word Cloud.

Motivation

Sentiment Analysis plays a crucial role in understanding the sentiments of public opinion regarding different products, services, etc. Twitter is a platform where users express their opinions and sentiments about various topics in real time. Analyzing sentiment on Twitter will allow researchers or businesses to gain immediate insights into public opinion, trends, and reactions to events or products. Using Sentiment Analysis on Twitter, the Government has the ability to research and evaluate public sentiment as well as the social effects of political decisions.

Expected Outcome

We aim to conduct a comprehensive study of existing sentiment analysis models and try to achieve higher accuracy using different machine learning techniques. Higher accuracy will help in better decision-making by providing actionable insights that align with customer needs and current market trends. We plan to provide a comparative analysis of different machine learning algorithms based on their accuracy, and precision, and identify the most effective approaches for sentiment analysis.

2 Execution Plan

Steps to Follow

- Data Collection which involves gathering a diverse dataset of text labeled with sentiment.
- Data Pre-processing to clean the data (removing stopwords, punctuations, etc) before model training.
- Feature Extraction to extract relevant features from the pre-processed data.
- Model Training which involves selecting the machine learning model and training it using the preprocessed data.
- Model Evaluation which includes evaluating the results obtained from the model using performance metrics such as Confusion Matrix.

Figure 1: General Flow of the Project
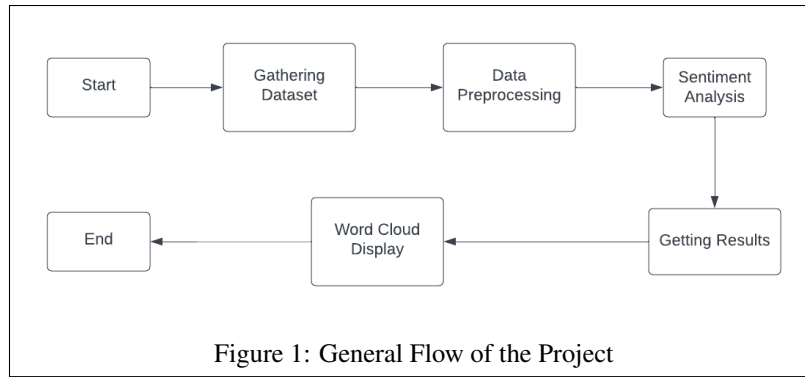
- Different tools which we will be using include NumPy, matplotlib, scikit-learn, pandas, wordcloud, etc.

Workload Distribution

| Jaydeep Patil	Gautham Vijayaraj	Rahul Babu
Finding Reference papers	Finding Reference papers	Finding Reference papers
Initiate Project Proposal Report	Data Collection	Perform Data Preprocessing
Work on training models	Work on training models	Work on training models
Perform Performance Analysis	Perform Comparative Analysis	Initiate Final Report
Perform Data Visualization	Initiate Milestone Report	Work on Project Demo

Table 1: Workload Distribution

Time Table	Plan
Week 1 - Feb 20	Forming the project team and suggesting project ideas
Week 2 - Feb 27	Worked on project ideas and looked for reference papers
Week 3 - Mar 5	Finalized project idea and worked on project proposal report
Week 4 - Mar 12	Project plan/proposal submission
Week 5 - Mar 19	Perform data preprocessing and feature extractions
Week 6 - Mar 26	Model training and performance analysis on the models
Week 7 - Apr 2	Work on the project milestone report
Week 8 - Apr 9	Project Milestone Submission
Week 9 - Apr 16	Train more models, and perform comparative analysis and performance analysis on all models
Week 10 - Apr 23	Work on the final project report
Week 11 - Apr 30	Project Report Submission

Table 2: Time Table

Expected Challenges and Solutions

- Data Collection : Collecting data and labeling it would be a big challenge and time consuming so we would be using publicly available dataset (from kaggle) for the purpose of training the model.

- Achieving optimal performance : To overcome this challenge we will try experimenting with different machine learning algorithms like Naive Bayes, SVM, Logistic Regression, Convolutional Neural Networks (CNNs).

Figure 2: Basic Workflow of Sentiment Analysis
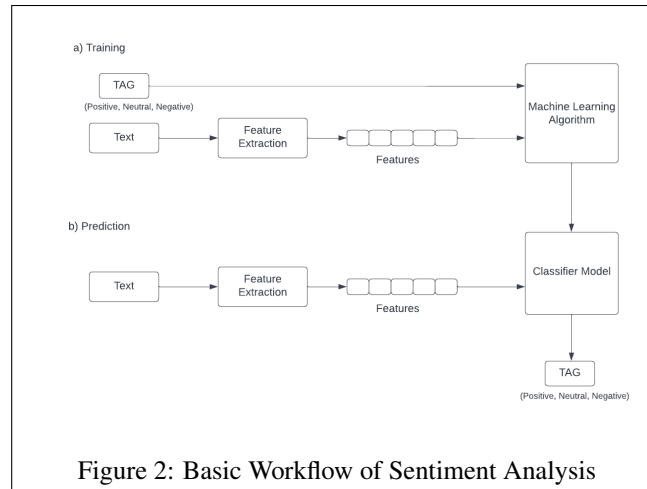
3   Evaluation Plan

To evaluate outcome of the project

We will use Confusion Matrix to evaluate the performance of our model using different machine learning techniques. These metrics include Accuracy, Precision, Recall value. Based on these parameters we can evaluate performance of our sentiment analysis model which involves using the machine learning algorithm which gives the highest accuracy.

We plan to implement Receiver Operating Characteristics Curve (ROC) plots which will help in evaluating performance of the model at different threshold settings. Performance of the model can be evaluated by comparing model predictions with ground truth labels.

How to evaluate your performance + peer review

Each Team members performance can be evaluated based on the assigned task and contributions made by each of us towards this project. Assigned tasks and contributions involve defining metrics, obtaining labeled data, pre-processing text, extracting features, selecting models, training, and validating them.

Interpretation of results provides actionable insights from peer reviews on Twitter, completing the evaluation process with meaningful findings for analysis.

Statistical plots and graphs from previous analysis

Figure 3: Comparison between Negative, Positive, and Neutral Emotions
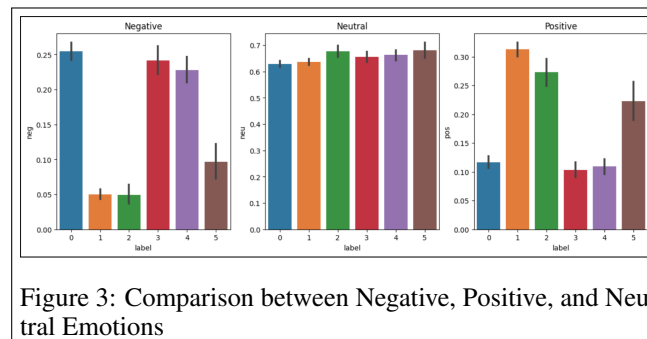
References

1. Dataset: Emotions (kaggle.com).

2. Varsha Sahayak, Vijaya Shete, Apashabi Pathan "Sentiment Analysis on Twitter Data" International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163 Issue 1, Volume 2 (January 2015).

3. Hanif Bhuiyan,Jinat Ara,Dr .MD Rashedul Islam "Sentiment analysis algorithms and applications: A survey" Journal Volume, December 2014, Pages 1093-1113.

4. Y. M. Wazery, H. S. Mohammed and E. H. Houssein, "Twitter Sentiment Analysis using Deep Neural Network," 2018 14th International Computer Engineering Conference (ICENCO), Cairo, Egypt, 2018, pp. 177-182, doi: 10.1109/ICENCO.2018.8636119.

5. M. Mertiya and A. Singh, "Combining naive bayes and adjective analysis for sentiment detection on Twitter," 2016 International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India, 2016, pp. 1-6, doi: 10.1109/INVENTIVE.2016.7824847.

6. S. Tiwari, A. Verma, P. Garg and D. Bansal, "Social Media Sentiment Analysis On Twitter Datasets," 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2020, pp. 925-927, doi: 10.1109/ICACCS48705.2020.9074208.

7. ChatGPT Prompts: Chat GPT.

8. ChatGPT Prompts: Chat GPT.