Sentiment Analysis on Social Media Posts using Advanced Feature Engineering - Group 8

Harshith Reddy Nagireddy 11517457
Malikarjun Narra 11656443
Santhoshi Goli 11724211
Mohammad Umar Farooq 11733380

***Abstract*—The project aims to perform sentiment analysis on social media posts, extracting deep insights from text data using advanced Feature Engineering techniques. This document outlines the project proposal and objectives.**

I. Introduction

In the era of social media, understanding public sentiment is crucial for various applications, from business intelligence to public opinion analysis. This project focuses on sentiment analysis using a dataset of tweets from Twitter. Sentiment analysis involves determining the emotional tone behind a piece of text, which can be especially valuable for gauging user opinions and attitudes.

The dataset used in this project comprises tweets labeled as either positive or negative, reflecting the sentiment expressed in each tweet. The analysis encompasses several stages, including data exploration, visualization, and the implementation of machine learning models for sentiment classification.

The initial phase involves exploratory data analysis (EDA) to gain insights into the structure of the data. Visualizations such as word clouds and frequency histograms are utilized to highlight the most common words and hashtags associated with both positive and negative sentiments.

Subsequently, natural language processing (NLP) techniques are employed to preprocess the text data. This involves steps like tokenization, stemming, and the extraction of hashtags. Feature engineering is crucial in transforming raw text into a format suitable for machine learning algorithms.

The project incorporates machine learning models such as Random Forest, Logistic Regression, and Decision Trees for sentiment classification. The performance of these models is evaluated using metrics like accuracy and F1 score, providing a comprehensive understanding of their effectiveness.

By the end of this project, we aim to not only develop accurate sentiment classification models but also to present detailed insights into the underlying patterns and trends in Twitter data. This project serves as an exploration into the realm of social media sentiment analysis, showcasing the potential applications and challenges associated with understanding public sentiment in a digital age.

II. Goals and Objectives

The following are the goals and objectives of the project:

1) Collect and preprocess a dataset of social media posts.
2) Apply Feature Engineering techniques to extract meaningful features from text data.
3) Develop a sentiment analysis model to classify posts as positive, negative, or neutral.
4) Evaluate the model's performance in terms of accuracy and F1-score.

III. Motivation

Understanding sentiment on social media is essential for various applications, from brand management to public opinion analysis. Advanced Feature Engineering can enhance the predictive power of sentiment analysis models.

IV. Significance

The project's significance lies in:

- Enhanced brand perception and customer engagement through sentiment analysis.
- Real-time monitoring of public sentiment on social media platforms.
- Application of advanced NLP and Feature Engineering techniques for text data analysis.

V. Objectives

The project's objectives include:

1) Data collection and preprocessing of social media posts.
2) Feature extraction and engineering for sentiment analysis.
3) Model development and evaluation of sentiment classification.
4) Fine-tuning for improved sentiment classification accuracy.

VI. Methodology

A. Dataset

The Twitter Sentiment Dataset is a collection of tweets labeled with sentiment values, designed for sentiment analysis purposes. The dataset contains a total of 48,159 tweets, with two columns: "id" and "label." The "id" column represents a unique identifier for each tweet, while the "label" column indicates the sentiment label associated with the tweet. Sentiment labels are numerical, with 0 denoting a negative sentiment and 1 indicating a positive sentiment. Additionally, there is a "tweet" column containing the actual text content of each tweet. This dataset provides a valuable resource for developing and evaluating machine learning models to classify sentiment in Twitter data, with applications in social media analysis and opinion mining.

Title [Cite]	Information
Twitter Sentiment Classification using Distant Supervision [2]	Rule-Based Approaches, Informal language in tweets; Predefined lists of words, Early attempts at sentiment classification using rule-based methods.
Twitter as a Corpus for Sentiment Analysis and Opinion Mining [3]	Machine Learning Models, Capturing complex patterns; SVM, Naive Bayes, Application of machine learning for sentiment analysis on Twitter data.
Twitter Sentiment Analysis with Deep Convolutional Neural Networks [4]	Deep Learning Approaches, RNNs, LSTMs for sequential dependencies, Exploration of deep learning models for capturing tweet context.
Modelling Sentiment in Social Media: A Multi-lingual Twitter Corpus Analysis [5]	Informal Language Challenges, Abbreviations, slang, misspellings, Addressing the challenges posed by informal language in tweets.
emoji2vec: Learning Emoji Representations from their Description [6]	Emojis and Hashtags, Incorporating emojis and hashtags in analysis, Considering the impact of emojis and hashtags on sentiment expression.
A Comparative Analysis of Sentiment Classification Techniques [7]	Evaluation Metrics, Accuracy, precision, recall, F1 score, Metrics used for evaluating sentiment analysis models on Twitter.
Twitter mood predicts the stock market [8]	Applications and Implications, Brand sentiment monitoring, political opinion tracking, Practical applications of sentiment analysis on Twitter.
Sentiment analysis of short informal texts [Kiritchenko et al. (2014) [?]]	Conclusion and Future Directions, Challenges, future research directions, Summarizing current state and suggesting future areas of exploration.

TABLE I
LITERATURE REVIEW ON SENTIMENT ANALYSIS ON TWITTER DATA

B. Detail Design of Features

VII. DETAILED DESIGN OF FEATURES

The code involves various stages of processing and analyzing a Twitter sentiment dataset using machine learning techniques. Let's break down the details of feature design within this context.

A. Text Processing

- The initial step involves reading and exploring the training and test datasets.
- Null values are checked for in both datasets.
- Negative and positive tweets are examined separately to gain insights into the content.

B. Tweet Length Analysis

- The length of each tweet is calculated and visualized to understand the distribution of tweet lengths.
- A new column 'len' is added to both the training and test datasets, representing the length of each tweet.

C. Word Frequency Analysis

- CountVectorizer is used to tokenize and count the frequency of words in the tweets.
- The most frequently occurring words are visualized through bar plots.
- Word clouds are generated to visually represent the vocabulary of both neutral and negative tweets.

D. Hashtag Analysis

- Hashtags are extracted from tweets, and their frequencies are analyzed and visualized.
- Separate analyses are performed for both neutral and negative tweets.

E. Word Embedding with Word2Vec

- Gensim's Word2Vec model is employed to create word embeddings from tokenized tweets.
- Similarity analysis is conducted for specific words like ”dinner,” ”cancer,” ”apple,” and ”hate.”

F. Text Preprocessing

- Text data is preprocessed by removing unwanted patterns, converting to lowercase, and stemming using the Porter stemmer.
- Stop words are removed from the tokenized and stemmed tweets.

G. Bag of Words Representation

- CountVectorizer is again used to create a bag of words representation for both the training and test datasets.
- The datasets are split into training and validation sets.

The detailed design of features in this context involves extracting meaningful information from the raw text data, analyzing tweet characteristics (length, word frequency, hashtags), and representing the text in a format suitable for machine learning models (bag of words). Additionally, word embeddings are explored to capture semantic relationships between words. The processed features are then used to train and evaluate different classification models for sentiment analysis.

VIII. PRELIMINARY RESULTS

1. Data Exploration and Preprocessing

The dataset comprises 31,962 training entries and 17,197 test entries with no missing values. Tweets will undergo preprocessing, including lowercasing and stemming, to facilitate analysis.

2. Tweet Length Analysis

The distribution of tweet lengths is visualized using histograms for both training and test datasets. Figures 1 and 2 illustrate the tweet length distributions in the training and test datasets, respectively.

3. Word Frequency Analysis

The top 30 frequently occurring words are visualized in bar plots for both neutral and negative tweets. Word clouds showcase the most common words in neutral and negative tweets.

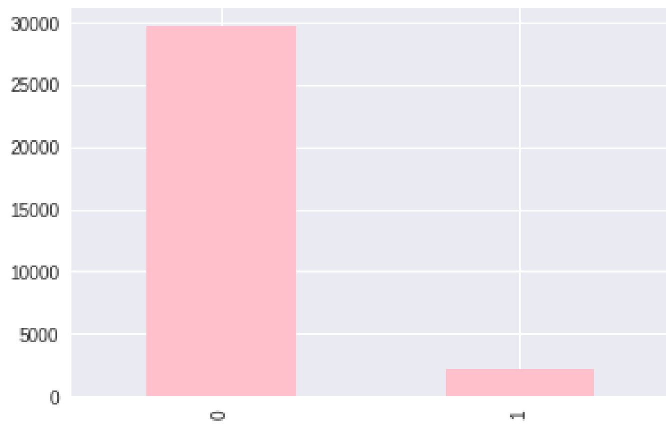

Fig. 1. Training Dataset - Tweet Length Distribution

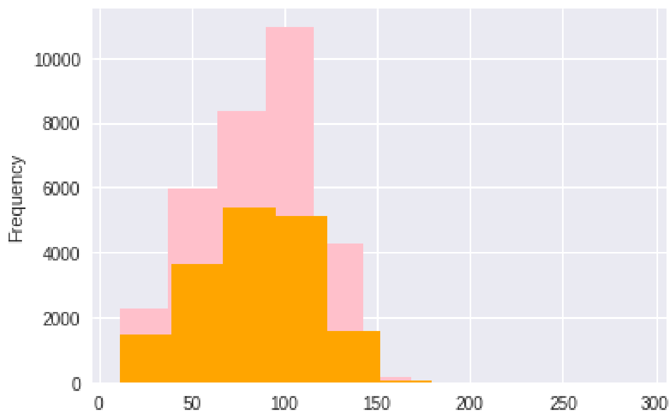

Fig. 2. Dataset Tweet Length Distribution

4. Hashtag Analysis

The top 20 hashtags in both neutral and negative tweets are displayed in bar plots. Figures 7 and 8 represent hashtag frequencies in neutral and negative tweets, respectively.

5. Word Embedding with Word2Vec

Word2Vec embeddings demonstrate semantic relationships between words. Similarity analyses for the words "dinner," "cancer," "apple," and "hate" are visualized in Figures 9, 10, 11, and 12, respectively.

Conclusion

The preliminary analysis presents a comprehensive exploration of the dataset through various visualizations. Different histograms, bar plots, and word clouds offer unique perspectives on tweet lengths, word frequencies, and hashtag usage in neutral and negative tweets. The semantic relationships between words, as captured by Word2Vec embeddings, further enrich our understanding of the dataset. These diverse visualizations pave the way for a detailed feature design process, guiding the selection of relevant features for model training.

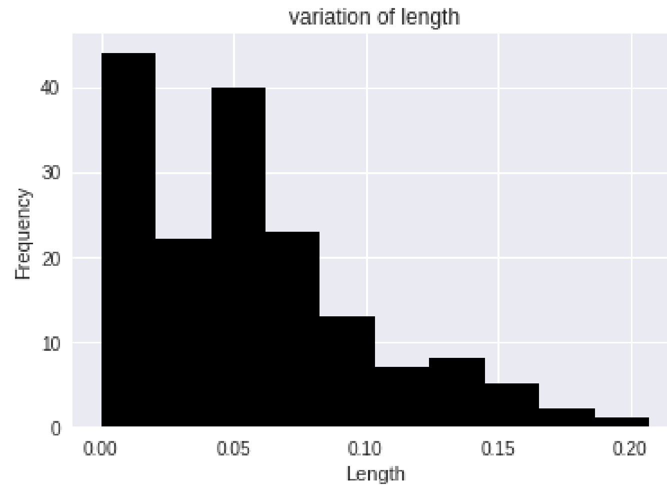

Fig. 3. Variation of Length

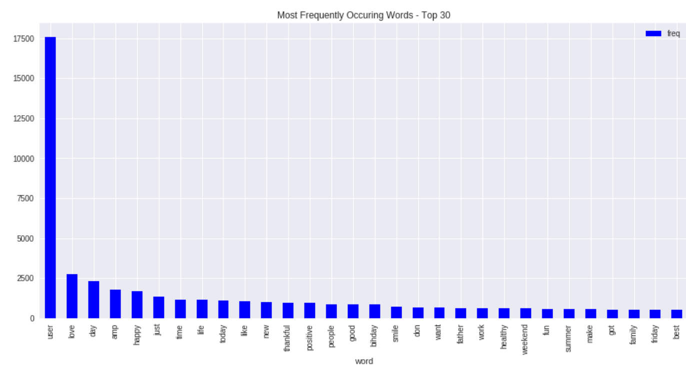

Fig. 4. Most frequently occuring words

Project Management: Implementation Status Report

Work Completed

- **Task 1: Data collection and understanding about data**
  - Responsibility: Person A
  - Contributions: Members B, C, D (Percentage: 25% each)
- **Task 2: Loading data and preprocessing and visualization**
  - Responsibility: Person B
  - Contributions: Members A, C, D (Percentage: 20% each)
- **Task 3: Employing feature engineering techinuques**
  - Responsibility: Person C and D
  - Issues/Concerns: None

Work to Be Completed

- **Task 4: Training and Testing of the model devloped and hyperparameter tuning**
  - Responsibility: Person A, B, C and D
  - Issues/Concerns: None

Fig. 5. wordcloud

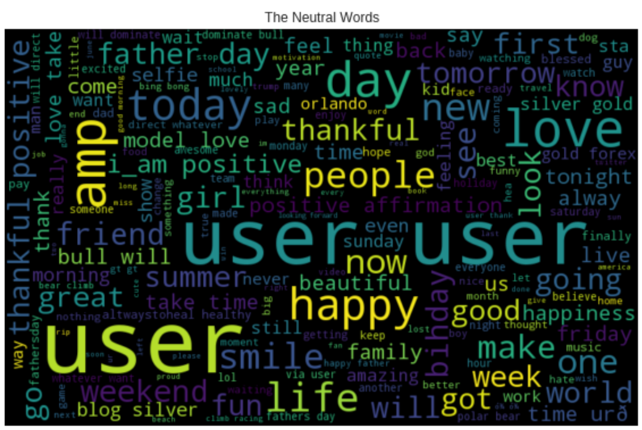

Fig. 6. wordcloud

IX. References

- Carrillo-de-Albornoz, Jorge, Javier Rodriguez Vidal, and Laura Plaza. "Feature engineering for sentiment analysis in e-health forums." PloS one 13, no. 11 (2018): e0207996.
- Go, A., Huang, O., Bhayani, R. "Twitter Sentiment Classification using Distant Supervision." CS224N Project Report, Stanford, 2009.
- Pak, A., Paroubek, P. "Twitter as a Corpus for Sentiment Analysis and Opinion Mining." LREC, 2010.
- Severyn, A., Moschitti, A. "Twitter Sentiment Analysis with Deep Convolutional Neural Networks." Proceedings of the 9th International Workshop on Semantic Evaluation

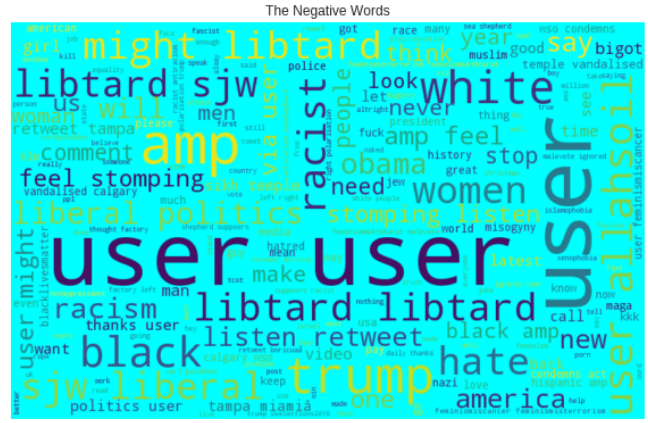

Fig. 7. wordcloud

(SemEval), 2015.
- Barbieri, F., Ronzano, F., Basile, V. "Modelling Sentiment in Social Media: A Multi-lingual Twitter Corpus Analysis." Journal of Information Science, 2014.
- Eisner, B., Rocktäschel, T., Augenstein, I., Bos, J. "emoji2vec: Learning Emoji Representations from their Description." Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2016.
- Agarwal, B., Mittal, N. "A Comparative Analysis of Sentiment Classification Techniques." International Journal of Information Technology and Knowledge Management, 2011.
- Bollen, J., Mao, H., Zeng, X. "Twitter mood predicts the stock market." Journal of Computational Science, 2011.
- Kiritchenko, S., Zhu, X., Mohammad, S. M. "Sentiment analysis of short informal texts." Journal of Artificial Intelligence Research, 2014.