Sentiment Analysis of Twitter Data Using Machine Learning

Lakshmi Sowjanya Gangumolu
lgangumo@depaul.edu

November 19, 2024

Abstract

This project explores sentiment analysis on Twitter data using the Sentiment140 dataset, containing 1.6 million labeled tweets. The goal is to classify tweet sentiments as positive, negative, or neutral using various machine learning models, including Logistic Regression, Random Forest, Long Short-Term Memory (LSTM), and BERT. BERT emerged as the top performer, achieving an accuracy of 82.47% with balanced precision, recall, and F1-scores. The integration of engineered features further enhanced the model's performance, making BERT the most effective choice for sentiment analysis tasks. The results provide valuable insights into public sentiment, benefiting businesses and researchers.

1 Introduction

In the age of social media, platforms like Twitter serve as a major source of public opinion, allowing users to express their thoughts, experiences, and sentiments freely. Analyzing these sentiments offers valuable insights that can aid businesses, researchers, and policymakers in understanding consumer behavior, public mood, and emerging trends. Sentiment analysis has become a crucial tool for market research, public opinion analysis, and trend forecasting. This project focuses on developing machine learning models to classify tweet sentiments using the Sentiment140 dataset, which contains over 1.6 million labeled tweets. By employing models such as Logistic Regression, Random Forest, LSTM, and BERT, this study aims to uncover deeper insights into public opinion dynamics, validate model performance, and identify strengths and weaknesses of different approaches.

2 Related Works

Several prior studies have focused on sentiment analysis and text classification in the context of social media. These studies provide valuable insights into the methodologies and challenges of analyzing textual data, especially from platforms like Twitter.

- Pang and Lee (2008): Their seminal work in sentiment analysis employed traditional classification techniques such as Support Vector Machines (SVM) and Naïve Bayes to classify movie reviews into positive and negative sentiments.

Their findings highlighted the challenges of working with subjective and sparse data [1].

- Aliza Sarlan and Chayanit Nadam (2015): This study developed a system that classified customer perspectives through tweets into positive and negative categories. They enhanced their model's performance by incorporating features such as VADER sentiment scores, tweet length, and frequency of specific keywords [3].
- Adyan Marendra Ramadhani and Hong Soon Goo (2020): This research applied deep learning models, specifically Long Short-Term Memory (LSTM) networks and Dynamic Convolutional Neural Networks (DCNN), to classify sentiment in Twitter data. [2].
- Peerapon Vateekul and Thanabhat Koomsubha (2021): In their study, the authors explored the use of deep learning techniques, focusing on LSTM and DCNN, to classify sentiment in Thai Twitter data. [4].

3 Preliminary/Background

The Sentiment140 dataset used in this project consists of labeled tweets (positive, negative, neutral) which provide the necessary training and validation data for building sentiment classification models. This dataset contains the following fields:

- target: The polarity of the tweet (0 = negative, 2 = neutral, 4 = positive)
- ids: A unique identifier for each tweet
- date: Timestamp when the tweet was posted
- flag: The query used to extract the tweet
- user: The username of the individual posting the tweet
- text: The actual content of the tweet

Dataset link: `https://www.kaggle.com/datasets/kazanova/sentiment140/data`

3.1 Key Concepts

- **Sentiment Analysis**: The task of determining the emotional tone behind a body of text. It involves classifying text into categories such as positive, negative, or neutral based on the sentiment conveyed.
- **Machine Learning Models**: Algorithms used to make predictions or decisions without explicit programming. In this project, several models are used to perform sentiment classification, including Logistic Regression, Random Forest, LSTM, and BERT.
- **Tokenization and Lemmatization**: Techniques for processing text data. Tokenization splits text into smaller units (tokens).

- **TF-IDF (Term Frequency-Inverse Document Frequency)**: A statistical method used to evaluate how important a word is to a document in a collection of documents. It is used as a feature extraction technique for text classification.
- **Word Embeddings (GloVe, BERT)**: A technique to represent words as continuous vectors that capture semantic meaning. GloVe is a pre-trained model for word embeddings, and BERT is a transformer-based model for deep contextual understanding of text.

4 Methodology

4.1 Data Collection

The Sentiment140 dataset was selected due to its large size (1.6 million tweets) and sentiment-labeled data. It is publicly available and provides a comprehensive set of labeled tweet data for sentiment analysis tasks.

4.2 Data Preprocessing

- **Text Cleaning**: The raw text data was cleaned by removing URLs, user mentions, hashtags, and special characters to focus on the relevant text.
- **Tokenization**: Each tweet was tokenized into individual words using Python's NLTK library.
- **Feature Extraction**: Text features were extracted using TF-IDF vectorization for traditional models and pre-trained word embeddings (GloVe, BERT) for deep learning approaches. VADER sentiment scores were additionally incorporated into the Random Forest model to capture tweet sentiment information.

4.3 Model Selection

Four models were chosen based on their suitability for text classification tasks:

- **Logistic Regression**: A linear model used for binary classification tasks, effective for smaller datasets and simple feature spaces.
- **Random Forest**: An ensemble learning method that uses multiple decision trees to improve prediction accuracy and handle more complex feature sets.
- **LSTM (Long Short-Term Memory)**: A deep learning model used to capture sequential patterns and contextual relationships in text data, particularly suitable for NLP tasks.
- **BERT (Bidirectional Encoder Representations from Transformers)**: A pre-trained transformer-based model that has demonstrated state-of-the-art performance in a wide range of NLP tasks.

4.4 Model Training and Hyperparameter Tuning

Each model was trained with default hyperparameters initially. For models like LSTM and BERT, fine-tuning was performed to improve performance:

- **Logistic Regression**: No hyperparameter tuning, used as a baseline.
- **Random Forest**: Default parameters, with feature importance based on TF-IDF and engineered features.
- **LSTM**: Hyperparameters were tuned and training was done for 10 epochs.
- **BERT**: Fine-tuned with a learning rate of 2e-5 for 3 epochs using the Hugging Face transformers library.

4.5 Evaluation Metrics:

Models were evaluated on accuracy, precision, recall, F1-score, and MAP, emphasizing recall.

5 Numerical Experiments

5.1 Overall Performance

BERT achieved the highest accuracy at 82.77%, followed by Logistic Regression (77.23%), LSTM (76.72%), and Random Forest (75.01%). The detailed performance metrics are shown below:

BERT (Best Performer)

- Accuracy: 82.77%
- Consistent 0.83 across all metrics (precision, recall, F1-score)
- Balanced performance between positive and negative classes

Logistic Regression

- Accuracy: 77.23%
- Stronger positive class recall (0.79) vs negative (0.75)
- F1-scores: 0.76 (negative) and 0.77 (positive)

LSTM

- Accuracy: 76.72%
- Lower metric scores around 0.50-0.51
- Similar performance across positive and negative classes

Random Forest

- Accuracy: 75.01%
- Uniform 0.75 across all metrics
- Most balanced performance between classes

6 Conclusion

This study demonstrated the effectiveness of various machine learning models for sentiment analysis on Twitter data using the Sentiment140 dataset. While all models achieved acceptable performance, BERT emerged as the most accurate and reliable model. Future improvements could focus on fine-tuning BERT further, incorporating more advanced text preprocessing techniques, and addressing class imbalance in the dataset. Sentiment analysis on Twitter data can provide invaluable insights for businesses, marketers, and researchers.

References

[1] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1):1–135, 2008.

[2] A. M. Ramadhani and H. S. Goo. Twitter sentiment analysis using deep learning methods. In *7th International Annual Engineering Seminar (InAES)*, pages 1–4, 2017.

[3] A. Sarlan, C. Nadam, and S. Basri. Twitter sentiment analysis. In *Proceedings of the 6th International Conference on Information Technology and Multimedia*, pages 212–216, 2014.

[4] P. Vateekul and T. Koomsubha. A study of sentiment analysis using deep learning techniques on thai twitter data. In *13th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, pages 1–6, 2016.