Big Data Major Project

Twitter and Reddit Sentiment Analysis using Spark

Submitted By:

Aanjaney Kumar Verma (2019UCS0085)

Anshuman Mishra (2019UCH0019)

Github Repository Link: https://github.com/not-anshuman/big_data

Research Paper Link: A_Topic_based_Approach_for_Sentiment_Analysis_on_Twitter_Data

1. Introduction:

In the field of data analysis, social media platforms provide a vast amount of valuable information that can be harnessed to understand public sentiment and make informed decisions. Among these platforms, Twitter and Reddit are prominent sources for real-time data analysis and sentiment research. Analyzing data from both Twitter and Reddit allows us to gain comprehensive insights into the opinions and emotions expressed by users, enabling us to uncover trending topics and sentiments.

This project focuses on the analysis of a combined dataset comprising tweets from a random city and Reddit data. The tweet dataset is stored in a CSV file, while the Reddit data encompasses a broader scope of discussions. The primary objective is to perform sentiment analysis on the combined dataset using machine learning techniques and natural language processing (NLP) methods.

To process the large volume of data efficiently, we will leverage Apache Spark, a distributed data processing framework. Spark provides a scalable and high-performance environment for data processing and analysis tasks. We will utilize Spark's capabilities to handle the combined datasets and extract meaningful insights.

The project involves several stages, including data preprocessing, feature extraction, and sentiment classification. During data preprocessing, we will clean the text by removing punctuation, hyperlinks, mentions, and non-essential characters. Stop words will be eliminated, abbreviations will be fixed, and part-of-speech tagging will be applied to identify relevant words.

For feature extraction, we will employ techniques such as tokenization and counting vectors to represent the text data numerically. These features will serve as input to machine learning models for sentiment classification. Logistic regression, naive Bayes, decision trees, and random forests are among the algorithms that will be explored and evaluated for sentiment analysis.

The project's success will be assessed based on the accuracy of the sentiment analysis models and the identification of significant trending topics within the combined dataset. By analyzing the sentiment and identifying prevalent topics, businesses, governments, and organizations can gain valuable insights to guide decision-making and strategy development.

Through this project, we aim to demonstrate the application of Spark and machine learning techniques for sentiment analysis on combined Twitter and Reddit data. By leveraging NLP methods and scalable data processing, we can extract meaningful information from the datasets and provide actionable insights for various domains, including market research, public opinion analysis, and brand reputation management.

2. Project Details

2.1 Background

In the era of digital communication, social media platforms like Twitter have become a significant source of public opinion and sentiment towards various topics. Analyzing these sentiments can provide valuable insights for businesses, governments, and other entities. However, due to the massive volume of data, manual analysis becomes practically impossible, and automated methods are needed. Machine learning, particularly Natural Language Processing (NLP), has proven effective in analyzing and extracting meaningful information from text data at scale.

2.2 Problem Statement of the Project

The project aims to build a sentiment analysis model to classify tweets into different sentiment categories. The model will be trained on a preprocessed dataset of tweets, which are labeled with their respective sentiment categories. The main challenge lies in processing the data efficiently and accurately, given the language variations, slang, and abbreviations common in tweets.

The project also focuses on assessing the effectiveness of different machine learning models for the sentiment analysis task. The models' performance will be evaluated based on their accuracy in predicting the sentiment category of unseen tweets in the test dataset.

2.3 Methodology Adopted for the Project

The project adopts a mix of NLP and machine learning techniques for data preprocessing, feature extraction, model training, and evaluation.

1. **Data Preprocessing:** The text data in tweets is first lemmatized and then tokenized using a regular expression tokenizer. Stop words are removed, and the text is transformed into count vectors (bag-of-words).
2. **Topic Modeling:** An LDA (Latent Dirichlet Allocation) model is used to assign topics to tweets. The LDA model is generated using Gensim.
3. **Model Training:** For each topic discovered by the LDA model, a separate logistic regression model is trained using tweets associated with that topic. A function `estimate_sentiment` is defined to estimate the sentiment of a given tweet based on these trained models.
4. **Cross-Validation:** To tune the hyperparameters of the logistic regression model, cross-validation is performed. This process involves training multiple models with different combinations of hyperparameters and selecting the best one based on its performance on a validation set.
5. **Model Evaluation:** The performance of the models is evaluated using a multi-class classification evaluator, which measures the accuracy of the predictions.

6. **Comparison of Models:** The project also explores the use of other machine learning models for sentiment analysis, including Naive Bayes, Decision Tree, and Random Forest classifiers. These models are trained and evaluated in the same manner as the logistic regression model.

The project is implemented using PySpark, given its ability to handle large datasets, and its robust MLlib library for machine learning tasks. NLTK is used for text preprocessing, and Gensim is used for the LDA model.

3. Results

The project successfully implemented and evaluated several machine learning models for sentiment analysis on Twitter data. The following are the results obtained from the experiments:

1. **Subset Partitioning Results**:
   - For N = 1: The logistic regression model trained on the subset of tweets related to a single topic achieved an accuracy of approximately 75.49% in predicting the sentiment category of test tweets.
   - For N = 3: When training the logistic regression model on subsets related to three topics, the accuracy improved to around 78.93%.
2. **Logistic Regression using TF-IDF Features**:
   - The logistic regression model trained on TF-IDF (Term Frequency-Inverse Document Frequency) features achieved an accuracy of approximately 69.38%.
3. **Cross-Validation Results**:
   - The logistic regression model trained with cross-validation, considering different combinations of hyperparameters, achieved an accuracy of about 75.22% on the validation set.
4. **Naive Bayes**:
   - The Naive Bayes classifier achieved an accuracy of approximately 70.02% on the test data.
5. **Decision Tree**:
   - The decision tree classifier yielded an accuracy of around 27.27% on the test data.
6. **Random Forest**:
   - The random forest classifier also had a low accuracy of approximately 27.21% on the test data.

4. Discussion

The results indicate that the logistic regression model trained on subsets of tweets related to specific topics showed promising performance, with accuracies exceeding 75%. This finding suggests that focusing on specific topics can lead to more accurate sentiment predictions compared to considering the entire dataset.

The logistic regression model using TF-IDF features performed relatively well, achieving an accuracy of approximately 69.38%. TF-IDF captures the importance of words in a document by considering their frequency and rarity across the entire dataset. This feature representation likely helped the model in understanding the distinguishing characteristics of different sentiment categories.

The cross-validated logistic regression model achieved an accuracy of around 75.22%, suggesting that the selected hyperparameters provided a good balance between model complexity and generalization performance.

The Naive Bayes classifier achieved a moderate accuracy of approximately 70.02%. Naive Bayes is known for its simplicity and efficiency but assumes that the features are conditionally independent, which might not hold true for text data. Despite this limitation, the classifier still achieved reasonable performance.

In contrast, the decision tree and random forest classifiers performed poorly, with accuracies of approximately 27.27% and 27.21%, respectively. It is likely that these models struggled to capture the complex relationships between text features and sentiment categories effectively.

Overall, the logistic regression model trained on subsets of tweets related to specific topics and using TF-IDF features showed the most promising results, outperforming the other models. The project findings suggest that focusing on specific topics and using advanced feature representations can enhance the accuracy of sentiment analysis on Twitter data.

Further improvements can be made by considering additional features, such as sentiment lexicons or contextual information, to capture more nuanced sentiment patterns. Additionally, exploring deep learning approaches, such as recurrent neural networks or transformer models, could potentially yield even better performance on sentiment analysis tasks.

The findings of this project have practical implications for businesses, government bodies, and organizations that can leverage sentiment analysis to gain insights into public opinion on Twitter and make informed decisions accordingly.

5. Scope of Improvements

While the project achieved promising results in sentiment analysis on Twitter data, there are several areas that could be further improved:

1. **Fine-Grained Sentiment Analysis**: The current project focused on classifying tweets into positive, neutral, and negative categories. However, sentiment analysis can be more nuanced, with different degrees of positivity or negativity. Implementing a fine-grained sentiment analysis approach, such as predicting sentiment scores or utilizing sentiment intensity modifiers, could provide more detailed insights.
2. **Handling Imbalanced Data**: The dataset used in this project might suffer from class imbalance, where one sentiment category has significantly more instances than others.

Addressing this issue can involve techniques like oversampling the minority class or undersampling the majority class to balance the data and improve model performance.

3. **Leveraging Deep Learning**: Deep learning models, such as recurrent neural networks (RNNs) or transformer-based architectures like BERT, have shown great success in natural language processing tasks. Exploring these advanced models could potentially yield better accuracy and capture more intricate sentiment patterns in tweets.
4. **Feature Engineering**: The project utilized TF-IDF features, but other feature engineering techniques could be explored. For instance, word embeddings like Word2Vec or GloVe can capture semantic relationships between words, allowing the model to better understand the context and sentiment of the text.
5. **Larger and Diverse Dataset**: Expanding the dataset by including more diverse and representative tweets from different regions, demographics, and topics can improve the generalizability of the sentiment analysis models. A larger and more diverse dataset can help capture a wider range of sentiments and ensure robustness.

6. Conclusion

The project successfully implemented sentiment analysis on Twitter data using various machine learning models and techniques. Through preprocessing, topic modeling, and model training, the project aimed to accurately classify tweets into sentiment categories.

The logistic regression model trained on subsets of tweets related to specific topics showed promising results, outperforming other models. The model using TF-IDF features also yielded reasonable accuracy. Cross-validation was used to fine-tune the logistic regression model's hyperparameters and enhance its performance.

Although the decision tree and random forest models achieved lower accuracies, the project highlighted the importance of feature engineering, focusing on specific topics, and leveraging advanced techniques like TF-IDF for sentiment analysis on Twitter data.

The findings of this project have practical implications for organizations seeking to understand public sentiment on Twitter and make informed decisions based on those insights. By analyzing sentiment, businesses, governments, and other entities can adapt their strategies, improve customer satisfaction, and address public concerns more effectively.

To further enhance sentiment analysis, future work could explore fine-grained sentiment analysis, address class imbalance, leverage deep learning models, experiment with different feature engineering approaches, and expand the dataset's size and diversity. These improvements would contribute to more accurate sentiment analysis and enable organizations to gain deeper insights into public opinion on social media platforms.