

Project Report: Sentiment Analysis of Twitter Data

Project Overview:

This project focused on conducting sentiment analysis on Twitter data to gain insights into public opinions and emotions regarding a particular topic, brand, or event. Sentiment analysis, also known as opinion mining, involves using natural language processing and machine learning techniques to determine the sentiment expressed in textual data.

Steps for Building a model:

- **Data Collection:** Gather a diverse and representative dataset of tweets related to the chosen topic.
- **Preprocessing:** Clean and preprocess the raw Twitter data to prepare it for sentiment analysis.
- **Model Development:** Build and train a sentiment analysis model using machine learning or deep learning techniques.
- **Evaluation:** Assess the model's performance in terms of accuracy, precision, recall, and F1 score.
- **Insights Generation:** Extract meaningful insights from the sentiment analysis results to understand public sentiment.

Methodology

1. Data Collection

1.1 Selection of Topic

The project focused on sentiment analysis related to Tweets on Twitter.

1.2 Data Gathering

Twitter data was collected using the Kaggle. The dataset included a mix of user-generated content, retweets, and replies.

Data set link: <https://www.kaggle.com/datasets/kazanova/sentiment140>

2. Data Preprocessing

2.1 Text Cleaning

Raw tweets were preprocessed to remove noise, including special characters, URLs, and user mentions. Text normalization techniques such as stemming were applied to standardize the text.

2.2 Tokenization

The cleaned text was tokenized into words to prepare it for further analysis.

3. Model Development

3.1 Model Selection

A Logistic regression and a Naive bayes model was chosen for sentiment analysis due to its proven effectiveness in similar tasks.

3.2 Training

The model was trained on a labeled dataset containing examples of tweets with corresponding sentiment labels (positive, negative). The dataset was split into training and validation sets.

4. Evaluation

4.1 Metrics

The model's performance was evaluated using standard metrics, including accuracy, precision, recall, and F1 score, to assess its ability to correctly classify sentiment.

Results and Discussion

1. Model Performance

The sentiment analysis model achieved 0.78 accuracy score, demonstrating its effectiveness in classifying sentiment in Twitter data.

2. Limitations

2.1 Data Bias

The analysis may be influenced by biases in the dataset, such as over-representation of certain demographics or perspectives.

2.2 Model Limitations

The model's performance may vary in different contexts, and it may struggle with certain types of language or sarcasm.

Conclusion

The sentiment analysis of Twitter data provided valuable insights into public opinions on the chosen topic. The project successfully developed a robust model and generated meaningful insights that can inform decision-making processes related to the subject matter.