Deep Learning-Based Sentiment Analysis of Social Media Posts: A Comparative Analysis of Reddit and Twitter Data

Farah Walid Abdelsalam

Software Engineer

Uneeq interns Machine learning task 2

***Abstract*— This project aims to analyze sentiment in social media posts using Natural Language Processing (NLP) and Deep Learning techniques. Here I employ a neural network model to classify sentiments into three categories: positive, negative, and neutral. Two datasets, Twitter and Reddit data, are used for training and testing the model. The preprocessing steps include text cleaning, tokenization, and TF-IDF vectorization. The performance of the model is evaluated using metrics such as accuracy, precision, recall, and F1 score.**

Keywords— NLP, Neural Network, Deep learning, Sentiment analysis, twitter, reddit, dataset, classification, models, preprocessing, positive, negative, neutral.

I. INTRODUCTION

Determining the sentiment or opinion expressed in text data is a crucial task in Natural Language Processing (NLP) called sentiment analysis. The need to comprehend the opinions expressed by users in their contributions is expanding as social media sites like Reddit and Twitter gain greater influence on our generation. This project aims to create a deep learning model that can reliably categorize social media messages into three sentiment categories: positive, negative, and neutral. Businesses can easily enhance their products, services, and marketing tactics by gaining insights into client opinions through the analysis of such sentiments.

II. DATA DESCRIPTION AND PREPROCESSING STEPS

Datasets-

- Twitter Data:

The Twitter dataset consists of tweets labeled into three sentiment categories: positive, negative, and neutral.
The dataset contains a (clean_text) column, which is preprocessed and used for model training.

- Reddit Data:

Similar to the Twitter dataset, the Reddit dataset comprises user comments labeled with sentiment categories.
The (clean_comment) column is used for training the model.

Preprocessing Steps-

- Text Cleaning:
- Convert text to lowercase.

- Remove special characters and punctuation.

- Tokenize the text into words.

- Stopwords Removal and Stemming:
- Remove common stopwords using NLTK's predefined list.

- Apply the PorterStemmer to reduce words to their base form.

- TF-IDF Vectorization:
- Convert the preprocessed text into numerical features using TF-IDF (Term Frequency-Inverse Document Frequency).

- Limit the number of features to 5000 for efficient model training.

- Handling Missing Values:
- Remove rows with missing data in the clean_text and category columns.

- Ensure that there are no NaN values in the training and testing sets.

III. USING METHODOLOGY/APPROACH

Neural Network Model-

A Sequential Neural Network model is implemented using TensorFlow and Keras. The model architecture consists of the following layers:

- Input Layer:
- A Dense layer with 128 units and ReLU activation, taking the TF-IDF vectorized features as input.

- Hidden Layers:
- Two additional Dense layers with 64 and 32 units, each followed by a Dropout layer to prevent overfitting.
- Both layers use ReLU activation.

- Output Layer:
- A Dense layer with 3 units (corresponding to the three sentiment categories) and a softmax activation function for multi-class classification.

- Model Training
- The model is compiled using the Adam optimizer and categorical cross-entropy loss function.

- It is trained on the TF-IDF vectorized features for 15 epochs for the Twitter dataset and 20 epochs for the Reddit dataset, with a validation split of 20% and 30%, respectively.
- The batch size is set to 32.

IV. Results

The results obtained from both datasets are as follows:

Dataset Comparison		
	Twitter	Reddit
Accuracy	77.51%	78.70%
Recall	77.51%	78.70%
Precision	77.50%	78.66%
F1-Score	77.28%	78.60%

These results indicate that the neural network model performs well in classifying sentiments in both datasets. The use of TF-IDF vectorization and dropout layers in the model helps in achieving robust performance.

V. Conclusion

This study effectively illustrates how deep learning methods may be used to analyse sentiment in social media data. In order to get the data ready for model training, preprocessing procedures including text cleaning, stopword removal, and TF-IDF vectorisation are essential. On both the Twitter and Reddit datasets, the neural network model achieves good accuracy, precision, recall, and F1 scores because to its properly thought-out architecture. Future research might investigate more intricate models, such as CNNs or RNNs, and broaden the analysis to include more datasets in an effort to enhance generalisation and performance even more.