Proposal: Sentiment Analysis Model for Tweets Dataset

Aya Ragaa	7793
Nabeel Mohemed	7547

Introduction

In today's digital world, social media platforms like Twitter have become essential sources for gauging public opinion on various topics. Our project focuses on building a **Sentiment Analysis Model** using a **Tweets dataset**. Sentiment analysis helps classify text into categories such as positive, negative, or neutral sentiments, making it useful for brands, governments, and researchers to monitor trends, public opinions, or feedback in real-time.

The main objective of our project is to **analyze the emotional polarity** of tweets, providing meaningful insights into public sentiment towards specific topics or events. This project will combine **natural language processing (NLP)** techniques with machine learning models to classify sentiments accurately.

Literature Review

Sentiment analysis has been an active research area with numerous approaches to text classification. Previous studies have shown that:

1. **Pre-trained embeddings (Word2Vec, GloVe)** significantly improve text classification accuracy by capturing semantic meanings of words.
2. **Recurrent Neural Networks (RNNs)** and **Long Short-Term Memory (LSTM)** models handle the sequential nature of text better compared to traditional machine learning algorithms like Naïve Bayes.
3. **Transformers (e.g., BERT)** have recently outperformed other models in sentiment analysis by capturing context better across sentences.

The literature suggests that a combination of advanced preprocessing and deep learning models will yield high performance. An excellent reference is from **Dr. Walid's publications** on using deep learning for NLP tasks, which emphasizes the importance of model generalization for robust results.

Dataset to be Used

We will utilize the **Tweets Dataset** for this project. This dataset contains tweets along with their labeled sentiments (positive, negative, or neutral), making it suitable for supervised learning tasks. The dataset will be divided into **training, validation, and test sets** to evaluate the model's performance effectively.

Proposed Methodology & Approach

1. Preprocessing:

- Tokenization and stopword removal.
- Lowercasing, punctuation removal, and handling special characters (e.g., hashtags and mentions).
- Applying **lemmatization** or stemming to reduce words to their base forms.

2. Model Building:

- We plan to experiment with several models:
  - **Naïve Bayes**: As a baseline model.
  - **LSTM**: To capture sequential dependencies in text.
  - **BERT**: A transformer-based model for better contextual understanding.

3. Compilation:

- Models will be compiled using suitable optimizers such as **Adam**.
- Loss functions like **Categorical Crossentropy** will be used for multi-class classification.

4. Evaluation:

- We will use metrics such as **accuracy**, **F1-score**, **precision**, and **recall**.
- **Confusion matrices** will be used to analyze misclassifications.

Expected Results & Evaluation

We expect our sentiment analysis model to achieve high accuracy in classifying tweets into positive, negative, and neutral categories. Using **BERT** or **LSTM** models, we aim for a **classification accuracy of over 85%**. The success of the project will be evaluated based on:

- **Prediction accuracy** on unseen data.
- **Interpretability** of the results through visualizations (e.g., word clouds and confusion matrices).
- **Comparative performance** between different models.

Team Members

- **Aya Ragaa** – ID: 7793
- **Nabeel Mohamed** – ID: 7547