Deep Learning-Based Sentiment Analysis: A Comparative Study of Classical Machine Learning, Deep Learning and Transformer Models

1. Introduction

Sentiment analysis is a crucial task in natural language processing (NLP), used to classify textual data into positive, negative, or neutral sentiments. This project explores and compares various sentiment analysis approaches, ranging from classical machine learning methods to modern deep learning architectures, including Transformer-based models. By implementing and evaluating models such as Naive Bayes, Logistic Regression, CNN, LSTM, and **DistilBERT**, I will analyze the strengths and limitations of each approach for sentiment classification, with a focus on real-world data from Twitter.

2. Objective

The objective of this project is to evaluate and compare the performance of different machine learning and deep learning models in sentiment analysis, specifically focusing on Naive Bayes, Logistic Regression, CNN, LSTM, and **DistilBERT**. This project will emphasize the efficiency of Transformer-based models, particularly **DistilBERT**, in handling complex text data and producing accurate sentiment predictions.

3. Dataset

For this project, I will use the **Sentiment140 dataset**, which consists of **1.6 million tweets** labeled as **positive**, **negative**, or **neutral**. The dataset provides real-world data from Twitter, making it ideal for multi-class sentiment classification tasks.

- **Description**: The Sentiment140 dataset contains 1.6 million tweets labeled as positive, negative, and neutral. It's widely used for sentiment analysis on Twitter and ideal for working with real-world short-text data.
- **Link to Dataset**: https://www.kaggle.com/datasets/kazanova/sentiment140

4. Methodology

1. **Data Preprocessing**:
    - Clean the dataset by removing special characters, URLs, and stop words.
    - Tokenize the text data using Suitable **tokenizer** for classical models and **DistilBERT tokenizer** for Transformer-based model.
2. **Model Implementation**:
    - **Classical Machine Learning Models**:
        - **Naive Bayes** and **Logistic Regression** will serve as baseline models to evaluate traditional approaches for sentiment analysis.
    - **Deep Learning Models**:
        - **CNN** (Convolutional Neural Networks): A CNN model will be built to capture local patterns in the text, useful for sentiment classification.

        - **LSTM** (Long Short-Term Memory): LSTM will be implemented to capture sequential dependencies and long-term contextual information in the text.
    - **Transformer-Based Model**:
        - **DistilBERT**: A lighter, faster version of BERT, **DistilBERT** will be fine-tuned on the Sentiment140 dataset to capture contextual relationships in the data and generate accurate sentiment predictions.
3. **Evaluation Metrics**:
    - I will use **accuracy**, **precision**, **recall**, and **F1-score** to evaluate and compare the performance of all models. These metrics will provide a comprehensive assessment of the models' ability to classify sentiments accurately.

5. Main Paper

- Zhang et al., Deep Learning Based Text Classification: A Comprehensive Review (2020).

6. References

- Sanh et al., *DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter* (2019).
- Vaswani et al., *Attention is All You Need* (2017).

Submitted by:

Sudip Pokhrel
001261230
Masters Of Computer Science
University Of Lethbridge