Sentiment Analysis Project Report

1. Executive Summary

This report outlines the progress and outcomes of the Sentiment Analysis Project, aimed at developing robust machine learning models to classify sentiments in Twitter data from the **Sentiment140 dataset**. Three models—Logistic Regression with TF-IDF, LSTM with Word2Vec, and DistilBERT—were developed and evaluated. The project achieved promising results, with DistilBERT outperforming others in accuracy and F1-score. Despite challenges such as computational resource constraints and data preprocessing complexities, the team successfully delivered deployable models, with the DistilBERT model uploaded to HuggingFace for public access. Next steps include model optimization and real-world deployment.

2. About Dataset

Context

The Sentiment140 dataset contains 1,600,000 tweets extracted using the Twitter API. The tweets are annotated for sentiment polarity (0 = negative, 4 = positive) and are designed for sentiment detection tasks.

Content

The dataset includes the following 6 fields:

- **target**: The polarity of the tweet (0 = negative, 4 = positive).
- **ids**: The ID of the tweet (e.g., 2087).
- **date**: The date of the tweet (e.g., Sat May 16 23:58:44 UTC 2009).
- **flag**: The query (e.g., lyx). If there is no query, the value is NO_QUERY.
- **user**: The user who tweeted (e.g., robotickilldozr).
- **text**: The text of the tweet (e.g., "Lyx is cool").

Acknowledgements

The dataset's official resources are available here. The methodology is detailed in the paper: Go, A., Bhayani, R., and Huang, L., 2009. *Twitter sentiment classification using distant supervision*. CS224N Project Report, Stanford, 1(2009), p.12.

Inspiration

The dataset is ideal for detecting sentiment severity from tweets, enabling applications in social media monitoring and opinion analysis.

3. Project Objectives

The primary objectives of the project were:

- To preprocess and analyze the Sentiment140 dataset for sentiment classification (positive/negative).
- To develop and compare three machine learning models: Logistic Regression with TF-IDF, LSTM with Word2Vec embeddings, and DistilBERT.
- To achieve at least 80% accuracy on the test set for the best-performing model.
- To deploy the best model for public use via HuggingFace.
- To document the process, challenges, and outcomes in a comprehensive report.

4. Progress Report

The project progressed through several key phases:

- **Data Acquisition and Preprocessing**: The Sentiment140 dataset (1.6 million tweets) was downloaded via Kaggle. Preprocessing included lowercasing, removing URLs, mentions, hashtags, emojis, numbers, and punctuation, and handling contractions. The dataset was split into 80% training and 20% testing sets.
- **Model Development**:
    - **Logistic Regression with TF-IDF**: Utilized TF-IDF vectorization with n-grams (1,2) and achieved an accuracy of approximately 81% on the test set. A pipeline was created and saved using joblib for deployment.
    - **LSTM with Word2Vec**: Employed Word2Vec for embeddings and an LSTM architecture, achieving around 83% accuracy. The model was trained for 5 epochs with a batch size of 128.
    - **DistilBERT**: Fine-tuned the distilbert-base-uncased model for 4 epochs, achieving the highest accuracy of 85% and an F1-score of 0.86 on the test set. The model was saved and uploaded to HuggingFace.
- **Evaluation**: All models were evaluated using accuracy, F1-score, and confusion matrices. DistilBERT outperformed others due to its ability to capture contextual nuances.
- **Deployment**: The DistilBERT model was successfully uploaded to HuggingFace under the repository mahmoudelbahy33/distilbert-base-uncased-finetuned-sentiment-2.

5. Budget and Financials

The project was conducted using free-tier resources, primarily Google Colab for model training and Kaggle for data access. Additional costs included:

- **HuggingFace Subscription**: Free tier used for model hosting.
- **Software Licenses**: Open-source libraries (e.g., TensorFlow, PyTorch, scikit-learn) incurred no costs.
  Total expenditure remained under $50, well within the allocated budget.
  Future scaling may require investment in premium cloud computing resources.

6. Challenges and Risks

Several challenges were encountered:

- **Data Preprocessing**: Handling emojis, contractions, and noisy text required extensive cleaning, which was time-consuming. Risk of information loss was mitigated by testing multiple preprocessing strategies.
- **Computational Resources**: Training DistilBERT and LSTM models on Colab's free tier led to memory crashes and slow training times. This was addressed by optimizing batch sizes and using smaller subsets for initial testing.
- **Model Performance**: Logistic Regression and LSTM underperformed compared to DistilBERT due to limited feature representation. This was mitigated by combining TF-IDF and Word2Vec in a hybrid model, though results were still inferior to DistilBERT.
- **Risk of Overfitting**: Addressed by incorporating dropout layers in LSTM and weight decay in DistilBERT.
  Future risks include scalability issues for real-time sentiment analysis and potential biases in the dataset.

7. Team and Resources

The project team consisted of:

- **Data Scientist**: Responsible for data preprocessing, model development, and evaluation.
- **Machine Learning Engineer**: Focused on model training, optimization, and deployment to HuggingFace.
- **Project Manager**: Oversaw timelines, resource allocation, and stakeholder communication.
  Resources utilized:
- **Hardware**: Google Colab (free tier with GPU access).
- **Software**: Python, TensorFlow, PyTorch, scikit-learn, HuggingFace Transformers, NLTK, and Kaggle.
- **Data**: Sentiment140 dataset (1.6 million tweets).

8. Timeline and Milestones

The project spanned 8 weeks, with the following milestones:

- **Week 1-2**: Data acquisition, exploratory data analysis, and preprocessing.
- **Week 3-4**: Development and evaluation of Logistic Regression and LSTM models.
- **Week 5-6**: Fine-tuning and evaluation of DistilBERT model.
- **Week 7**: Model comparison, selection, and deployment to HuggingFace.
- **Week 8**: Documentation and final report preparation.
  All milestones were met on schedule, with minor delays in DistilBERT training due to resource constraints.

9. Stakeholder Updates

Key updates shared with stakeholders:

- **Mid-Project Review**: Presented initial results from Logistic Regression (81% accuracy) and LSTM (83% accuracy), highlighting preprocessing challenges and resource needs.
- **Model Selection**: Informed stakeholders of DistilBERT's superior performance (85% accuracy) and planned deployment.
- **Deployment Confirmation**: Shared the HuggingFace repository link (mahmoudelbahy33/distilbert-base-uncased-finetuned-sentiment-2) for public access.

Stakeholders expressed satisfaction with the project's progress and outcomes.

10. Next Steps

The following actions are planned:

- **Model Optimization**: Explore hyperparameter tuning and larger batch sizes for DistilBERT to improve performance.
- **Real-World Deployment**: Integrate the DistilBERT model into a web application for real-time sentiment analysis.
- **Scalability Testing**: Evaluate model performance on larger datasets and real-time Twitter streams.
- **Bias Analysis**: Investigate potential biases in the Sentiment140 dataset and apply debiasing techniques.

11. Conclusion and Recommendations

The Sentiment Analysis Project successfully developed and deployed three machine learning models, with DistilBERT achieving the highest accuracy (85%) and F1-score (0.86). The project met its objectives within budget and timeline constraints, despite challenges with computational resources and data preprocessing.
Recommendations include:

- Investing in premium cloud computing resources for faster training and scalability.
- Conducting a follow-up project to integrate the model into a production environment.
- Exploring ensemble methods to combine strengths of all three models for improved performance.
  The deployed DistilBERT model is accessible at https://huggingface.co/mahmoudelbahy33/distilbert-base-uncased-finetuned-sentiment-2 and ready for further development.