Project Proposal Template: NLP-Based Sentiment Analysis

Your Name(s)

April 8, 2025

1 Introduction

[Briefly introduce the problem you are tackling.] Sentiment analysis is a crucial task in Natural Language Processing (NLP), enabling automatic classification of text into sentiment categories. This project aims to explore different machine learning models and pre-processing techniques to improve sentiment classification on social media texts.

1.1 Research Questions

[Clearly state your research questions. Example:]

- **RQ1:** Does the model architecture significantly affect sentiment analysis accuracy?
- **RQ2:** Do pre-processing techniques (e.g., stop-word removal, stemming) improve model performance?

2 Methodology

2.1 Technical Approach

[Describe the step-by-step approach for the project. Example steps:]

1. **Data Preprocessing:** Cleaning data by removing URLs, punctuation, and emojis.
2. **Feature Extraction:** Using TF-IDF for logistic regression and BERT embeddings for the pretrained model.
3. **Model Training:** Training both models and fine-tuning the last layer of BERT.
4. **Evaluation:** Comparing models using cross-validation.

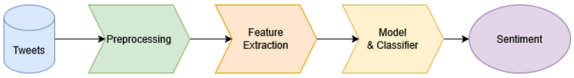

Figure 1: [Insert an architecture diagram here. You can use tools like draw.io to create a simple diagram.]

2.2 NLP Techniques

[List and explain the techniques you will use. Example:]

- **Logistic Regression:** A baseline model using TF-IDF.
- **BERT:** A pre-trained transformer model for sentiment analysis.
- **Preprocessing Techniques:** Tokenization, stop-word removal, and lemmatization.

3 Team Contributions

[Define who will do what. Example:]

3.1 Shared Responsibilities

All Members: Data gathering, pre-processing, train-test split creation, and final project poster preparation.
Sample Deliverables:

- A git repository with all materials to verify the experiments.
- A poster for the poster session.

3.2 Individual Responsibilities

3.2.1 Student 1

Role: Implement and train the logistic regression model.
Deliverables: Trained logistic regression model.

3.2.2 Student 2

Role: Implement and train the BERT model.
Deliverables: Trained BERT model.

3.2.3 Student 3

Role: Develop the evaluation framework and calculate performance metrics.
Deliverables: Detailed evaluation comparing both models.

4 Evaluation and Dataset

4.1 Dataset Description

[Specify which dataset(s) you will use, consider sources such as:]

- Hugging Face Datasets
- Kaggle Datasets
- Papers with Code Datasets

[Example dataset details:] We are using the Sentiment140[1] dataset containing 1.4 million tweets.

- **target:** Sentiment class (0 = negative, 2 = neutral, 4 = positive).
- **id:** Unique sample identifier.
- **date:** Timestamp of the tweet.
- **user:** Username of the author.
- **text:** Tweet content.

Target	ID	Date	User	Text
0	1467811184	Mon Apr 06 22:19:57 PDT 2009	ElleCTF	My whole body feels itchy and like it's on fire.

Figure 2: A sample row from a sentiment dataset.

4.2 Experimental Setup

[Define how you will evaluate your models. Example metrics:]
We will evaluate using the following metrics on our train / validation / test split. For this we once evaluate BERT, then logistic regression and then a combination of different pre-processing steps.

- Accuracy
- F1 Score
- ROC-AUC Score

References

[1] A. Go, R. Bhayani, and L. Huang, “Twitter sentiment classification using distant supervision,” *CS224N Project Report, Stanford*, vol. 1, p. 12, 2009. Accessed: 2025-03-27.