Project Proposal: Sentiment Analysis on Social Media Reviews

Krish Chaudhary

COE 379L – Machine Learning Applications

April 17, 2025

Introduction and Problem Statement

In this project, I aim to explore the task of sentiment analysis by focusing on social media reviews as the primary source of data. The goal is to build a system that can automatically classify the sentiment behind user-generated posts—determining whether they express a positive, negative, or neutral opinion. Social media platforms like Twitter and Reddit are filled with informal language, slang, emojis, and sarcasm, which make them uniquely challenging for traditional natural language processing techniques. However, they are also rich sources of real-world, emotionally charged data that reflect genuine user reactions. By training and evaluating various machine learning models on this kind of data, I hope to gain a deeper understanding of how well these models can handle noisy, short-form text, and identify the limitations that arise in this domain.

Data Sources

For this project, I will primarily use the Sentiment140 dataset, which contains 1.6 million tweets that have been automatically labeled as positive, negative, or neutral based on the presence of emoticons. This dataset is widely used for benchmarking sentiment classification tasks and provides a substantial amount of real-world data that has already been cleaned to some extent. If necessary, I may also explore additional datasets such as Amazon product reviews or Reddit comment sentiment datasets, particularly if I need to experiment with longer-form text or additional domains. All datasets used will be publicly available and sourced through platforms like Kaggle.

Methods, Techniques, and Technologies

To approach the problem, I will begin with classical machine learning techniques that we have covered in class, such as Naive Bayes and logistic regression. These models will be combined with feature extraction techniques like TF-IDF or CountVectorizer to convert raw text into numerical representations. These classical models will serve as baselines for comparison. In the second phase of the project, I plan to implement a more advanced deep learning approach, specifically using an LSTM (Long Short-Term Memory) network with pre-trained word embeddings such as GloVe. Finally, I will evaluate a transformer-based model, likely DistilBERT, using the Hugging Face Transformers library. This will allow me to compare how different types of models handle informal, real-world text.

For preprocessing, I will apply standard NLP techniques such as tokenization, stop-word removal, lemmatization, and emoji handling. Because social media text often includes sarcasm and abbreviations, part of the analysis will involve observing how different models perform on these more ambiguous samples. I will use tools such as Python, scikit-learn, TensorFlow/Keras, Hugging Face Transformers, and either NLTK or spaCy for preprocessing tasks.

Products to Be Delivered

At the end of the project, I will deliver a fully documented Jupyter Notebook that includes all stages of the project—from data preprocessing and model training to evaluation and visualization. The notebook will include both classical and deep learning models, along with detailed performance analysis using accuracy, precision, recall, F1-score, and confusion matrices. Additionally, I will submit a final written report that explains the problem, describes the dataset, outlines the methodology used, and discusses the results and their significance. If time allows, I also plan to develop a lightweight web interface or REST API to showcase the model's predictions in real time. This optional feature would demonstrate how the model could be deployed in a real-world application.