Fall 2024
Project Proposal

Edward Wang

1 Project Description

The objective of this project is to identify the most effective models for Natural Language Processing (NLP), with a particular focus on sentiment analysis of online text. In particular, various tweets from 𝕏, formerly known as Twitter will be used where the contents of each post will be classified into three main categories: negative, neutral, and positive.

2 Data Collection

The data source for this project is the Stanford Sentiment140 Dataset[3], which contains 1.6 million tweets extracted from 𝕏 (formerly Twitter) API. This dataset was collected in 2009, when posts were limited to 140 characters, even though the current limit has been increased to 280 characters. Unfortunately, the original dataset is no longer hosted by Stanford. It is now accessible through Kaggle[5] and Hugging Face.

3 Proposed Methodology

3.1 Data Cleaning

In this dataset, tweets often exhibits typical attributes found in online social media including user mentions, hyperlinks, emojis, abbrevations, and slang. To prevent specific usernames from influencing sentiment analysis, it is essential to eliminate mentions (e.g., @User123) and hyperlinks. For instance, a user like @PositiveQuoteDaily, who consistently shares positive content, could inadvertently bias the model towards associating this username with positive sentiment.

Furthermore, to enhance the quality of the data, it is crucial to remove stopwords—those commonly occurring words in any natural language that contribute little semantic value. Stopwords includes articles, conjunctions, prepositions, pronouns, and frequently used verbs.

Since words can take various forms based on tense, employing techniques such as stemming or lemmatization is necessary to reduce words to their root forms[4]. Identifying the most effective method for this purpose will be a key aspect of this analysis.

3.2 Modeling

The objective is to conduct sentiment analysis on online text, for which a neural network model is most suitable. Specifically, the Recurrent Neural Network (RNN). RNN is a type of deep learning architecture designed to handle sequential data, making it particularly effective for natural language processing (NLP) and speech recognition tasks[1].

The goal would be to explore the various RNN models currently available. For example, the Long Short-Term Memory would be a specific RNN that is good at capturing the dependencies in text sequences. Additionally, Bidirectional LSTMs allows processing input data in both forward and backward directions which provides improved performance for context retrieval in both future and past states. Additionally, exploring state of the art models such as Bert[2] created by Google that have significant improvements in the field of NLP.

References

[1] Shervine Amidi. Recurrent neural networks cheatsheet, 2021. Accessed: 2024-10-20. URL: https://stanford.edu/~shervine/teaching/cs-230/cheatsheet-recurrent-neural-networks.

[2] Hugging Face. Bert model documentation, 2024. Accessed: 2024-10-20. URL: https://huggingface.co/docs/transformers/en/model_doc/bert.

[3] A. Go, R. Bhayani, and L. Huang. Twitter sentiment classification using distant supervision. In *CS224N Project Report*, volume 1, page 12. Stanford University, 2009.

[4] Stanford NLP Group. Stemming and lemmatization, n.d. Accessed: 2024-10-20. URL: https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html.

[5] Kazanova. Sentiment140 dataset, 2017. Accessed: 2024-10-20. URL: https://www.kaggle.com/datasets/kazanova/sentiment140/data.