Project 2: Proposal and Data Selection

Mithil Patel

Bellevue University

DSC680-T302 Applied Data Science

Prof. Catherine Williams

April 16th, 2023

Describe and name your project in 1-2 sentences max

The text-based emotion recognition model is designed to identify people's emotions, attitudes, or sentiments toward a particular goal, such as an individual, an organization, a topic or a product.

Describe the business problem your project is trying to solve and/or the research questions you will explore

Sentiment analysis involves analyzing text to determine a particular topic's overall sentiment (positive, negative, or neutral). Meanwhile, emotion detection is a subset of sentiment analysis that involves identifying specific emotions in a text. Many researchers have previously worked on emotion identification of facial and speech expressions; however, text-based emotion detection is underdeveloped as it is a laborious task due to missing cues such as tone or facial expressions in speech. Therefore, the project will focus on creating a model capable of detecting specific emotions based on text using optimal machine learning (ML) and deep learning (DL) algorithms.

Where are you getting your data? Describe the data that you will use to solve the problem

To develop a machine learning model, a good dataset will be vital to accurate predictions. The text-based emotion recognition model will use datasets from ISEAR and WASSA. The International Survey on Emotion Antecedents and Reactions (ISEAR) dataset, which was created by psychologists who conducted surveys in 37 countries using text and emotional stimuli, includes data on 173 emotional experiences that have been categorized into seven types of emotions: sadness, fear, shame, joy, anger, surprise, and disgust. The WASSA-2017 dataset comprises a collection of tweets annotated with emotion labels, with each tweet accompanied by an emotion intensity score ranging from 0 to 1.

What analysis methods will you use to complete this project?

To enhance the efficiency of the model, the raw dataset will undergo a series of preprocessing steps. First, unnecessary symbols and text will be eliminated. Then, techniques such as tokenization, removal of stop words, stemming, and lemmatization will be employed. In addition, feature extraction will be performed using CountVectorizer and TF-IDF Transformer. The preprocessed data will then be used to train and evaluate various machine learning classification algorithms, including Naïve Bayes, Decision Tree, and Random Forest. Deep learning algorithms such as Convolutional Neural Network (CNN) will also be used to predict emotions. Through these processes, we aim to achieve accurate predictions and insights that can inform decision-making.

What are some potential ethical concerns of this topic or analyzing the data?

While text-based emotion detection may have potential benefits in various industries, there are several ethical concerns to consider in ensuring the technology is used in a responsible manner. The most critical ethical concern to consider is privacy. People may feel an invasion of privacy since the model is trained on written communication; therefore, there is a risk of exposing sensitive information, such as political opinions, mental health issues, or personal beliefs. Technology can also be used to manipulate people by generating targeted messages intended to influence their behavior or actions. The model's accuracy and reliability can also be questionable, especially when the model fails to account for complex human emotions such as sarcasm or irony.

What are some issues and challenges do you think you might face?

One may encounter several issues when creating a text-based emotion recognition model. The most challenging aspect of the following project is the preprocessing stage. The survey was conducted in 37 countries; therefore, the dataset may contain text in different languages. Also, handling large datasets with noisy text, such as emojis, typos, and misspellings, can be challenging. If the dataset is presented in a different format, feeding the data through various preprocessing techniques, such as lemmatization, stemming, tokenization, and stop-word, can be challenging and may compromise the model's accuracy. Once the preprocessing is complete, a minor challenge I may face could be determining the ideal machine learning and deep learning algorithms to apply and how to combine both algorithms to create a hybrid model.

What sources will you use to validate your results and support your project topic?

To ensure the reliability of our findings, a portion of the dataset will be reserved as the validation set, which will not be used for training the model. Additionally, various metrics, such as accuracy score, F1 score, AUC-ROC score, and precision score, are used to evaluate the model's performance. To further validate the model's capabilities, we will ask it to predict emotions from manually inputted text containing diverse elements, such as sarcasm, misspellings, emojis, and hashtags. By subjecting the model to such varied inputs, we can ensure that it performs well across a range of real-world scenarios.

References

- Bharti, Santosh Kumar, et al. “Text-Based Emotion Recognition Using Deep Learning Approach.” *Computational Intelligence and Neuroscience*, U.S. National Library of Medicine, 23 Aug. 2022, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9427219/.