ISyE 6740 – Spring 2024
Project Proposal

Team Member Names: Mahshid Jafar Pour, Mehdi Sadeghi (Project Team 159)

Project Title: **Extracting Emotions from Twitter Posts: A Multi-class Classification NLP Problem**

Problem Statement:

Sentiment analysis is a classic NLP (Natural Language Processing) problem where the goal is to find the sentiment expressed in the text. Often, it is run on social media posts, product reviews or comment sections of a news article. This is crucial for businesses to understand public opinion on a particular topic or product. Companies often use Sentiment analysis to track customer satisfactions, monitor social media, perform market research or analyze employees survey [1].

Although sentiment analysis is widely used, it has its own challenges: detecting sarcasm in text is hard; emojis are used in text more than ever which cause difficulties in accurate classification; language biases; dealing with large volume of data; use of words that have multiple meanings which can be deceiving to the model; negation does not always mean a negative sentiment, etc. [2]. The trick is to train the model on a large enough dataset and perform the pre-processing stages cautiously.

In this project, we are aiming to run a similar problem by focusing on the content shared on X (former Twitter). However, instead of positive, negative and neutral labeling for the text, we are assigning emotions to them such as love, hate, anger, etc. In other words, our approach involves utilizing various multi-class classification algorithms to precisely categorize Twitter posts based on their emotional content.

Data Source

We will use "Emotions" dataset, which is shared on Kaggle [3] for our project. Contrary to traditional positive/negative sentiment analysis, this dataset has six distinct emotion categories: sadness (0), joy (1), love (2), anger (3), fear (4), and surprise (5). It serves as a valuable resource for understanding and analyzing the diverse spectrum of emotions expressed in short-form text on social media.

This dataset has 416,810 messages shared on Twitter, each entry consists of a text feature accompanied by a corresponding emotion label. It's important to note that the dataset is unbalanced, with a greater percentage of texts expressing sadness and joy compared to other emotions. This imbalance requires careful consideration during model development and evaluation to ensure robust performance across all emotion categories.

Methodology

Python is the main language used for this project. The steps involved will include:

1) Data Pre-processing

As the initial step, we will perform data pre-processing to clean the data, and ensure optimal data quality and consistency. This involves a series of steps including converting all text to lowercase to eliminate case sensitivity and ensure uniformity. Additionally, whitespace removal is necessary to eliminate unnecessary whitespace characters to improve readability, tokenization for further analysis and feature extraction, removal of emojis to focus only on linguistic content, removal of English stopwords ("is", "a", "the", "are", etc), removal of punctuations, replacing chatword abbreviations with their full words (e.g. "you" instead of "U") to ensure clarity and consistency in our dataset. Next, we will perform stemming or lemmatization, which is an NLP technique that reduces words to their root. Finally, spell checking will be done to correct potential misspelled words.

2) EDA

Once the data are cleaned and pre-processed, we can perform Explanatory Data Analysis (EDA) to have a better understanding of data distribution. As our data is unbalanced, we will perform some undersampling or oversampling techniques to overcome this challenge and ensure the final model learns effectively and improve its performance on classifying messages in correct emotions.

3) Feature Extraction

The next step is feature extraction, we will use bag of words and will explore other word embedding methods to convert text to numerical vectors. Data are then partitioned into train, validation and test datasets for model training, model selection, and accuracy measurements.

4) Model Training

We are planning to use different classical machine learning classifiers (such as Naïve Bayes, support vector machine, logistic regression, etc.) as well as neural network to train the model with the data. Additionally, we plan to also explore open source pre-trained Large Language Models (LLMs) for the purpose of transfer learning and potentially achieving a better result leveraging LLMs to classify our data.

Different models' results will be compared by various evaluation metrics to identify the best performing model.

Evaluation and Final Results

Different metrics will be used to measure the performance of the models and choose the best one. We will use confusion matrix, precision, recall and F1 score as well as overall accuracy to select the best model. If we have class imbalance, it is important to get a better class accuracy for the smallest class, rather than overall accuracy.

References

1. 8 applications of sentiment analysis, https://monkeylearn.com/blog/sentiment-analysis-applications/#:~:text=Some%20popular%20sentiment%20analysis%20applications,text%20by%20emotion%20and%20opinion.
2. Begüm Yılmaz, Top 5 Sentiment Analysis Challenges in 2024, https://research.aimultiple.com/sentiment-analysis-challenges/.
3. https://www.kaggle.com/datasets/nelgiriyewithana/emotions/data