Tweet Sentiment Analysis

Team Members

1. Bradley Payne - A01935193 (Grad Student):
2. Nithya Alavala – A02381733 (Grad Student):

Description of the Project

Sentiment analysis is an important task for many domains including for product feedback and how the general public feels about a given company, product or topic. Many people express their feelings and concerns, movie analysis, and others on twitter through tweets. The goal of this project is to take texts from twitter data and classify it as negative, neutral, or positive. There are many things to take into consideration with language data such as sarcasm that may be difficult for some tweets to be correctly classified.

One method that has been used in conjunction with classical text analysis is the use of emoji data and how that impacts the sentiment of tweet text. We would like to use both classical text preprocessing analysis for classification as well as emoji. There is a prelabeled dataset on kaggle that we can analyze for this task. There is also a dataset on kaggle that is a compilation of emojis and the associated sentiment of the tweets.

An important task in the classification of tweet data is preprocessing to be compatible with machine learning methods. We will need to use Tokenization and text preprocessing methods to clean the data such as feature extraction, working with misspelled words, and unwanted or unuseful data. These considerations will likely be the most difficult part of the project.

To accomplish the classification task there are many methods used in the literature. We would like to explore the following algorithms: Random Forest, Decision Trees, SVM, RNN, Naive Bayes, and KNN. These algorithms are available in many popular Python learning packages such as Sci-Kit Learn, PyTorch, and others. For this project, the implementation of the algorithms will be using the packages rather than from scratch.

We all also select the best model as the method for use on the test dataset that would be used in a sentiment analysis application. To compare our method with state-of-the-art methods we can use a pretrained model such as BERT to determine if our method is valid.

Dataset

The initial datasets we have chosen for analysis is a tweet dataset that has been prelabeled and contains 1.6 million tweets. This is a publicly available dataset located on the Kaggle website and is free to use. The dataset can be found and obtained: https://www.kaggle.com/kazanova/sentiment140 The benefit of using this dataset is that this is larger than could be obtained and manually labeled by us during the scope of this project. This data is represented as a csv file and has the sentiment, user id, date, query, username, and tweet text. The output variable the models will aim to predict is the sentiment. The dataset has labeled negative, neutral, and positive sentiments as (0 = negative, 2 = neutral, 4 = positive). For classification this may be better represented as a one hot encoded vector. For training and testing we will use a Train/Test split of 70%/30%.

The other dataset that may serve as a starting point for use of emojis in sentiment analysis is a compilation of emojis and the associated sentiment of tweets that contained them from a previous data mining task. This data can also be found on the Kaggle website here: https://www.kaggle.com/thomasseleck/emoji-sentiment-data . This dataset is represented as a csv file and contains the emoji character, the unicode representation, number of occurrences, the normalized position it was in the tweet, and the occurrences in negative, neutral, and positive tweets, respectively. This dataset also contains a text summary of the emoji. This can be a useful starting point for classification.

Use of Data Analytics

For this project we are targeting text preprocessing, text and emoji representation for machine learning models, and machine learning for the classification task. All of these are relevant to data analysis because they can be used by companies and researchers to understand how a group of people feel about a topic and can then be applied to improving satisfaction or predicting consumer behavior. Knowing that a topic or item has a general negative feedback can be used to improve the product to overcome and provide near real-time response to how a new business plan is working.

Expected Outcomes/Results

The outcome of this project will be to determine which machine learning model can perform the best on sentiment classification for our dataset. We expect that with proper preprocessing and model hyperparameter tuning, our model will be relatively accurate in determining text sentiment. We do not, however, expect to perform better than the state of the art method. Our results could then be used to start with the best model for data mining with sentiment analysis.

References

http://davidliedtka.com/docs/cs224u.pdf

https://arxiv.org/pdf/2101.00430.pdf

https://www-cs.stanford.edu/people/alecmgo/papers/TwitterDistantSupervision09.pdf

https://towardsdatascience.com/sentiment-analysis-in-10-minutes-with-bert-and-hugging-face-294e8a04b671