CALIFORNIA STATE SCIENCE FAIR
2015 PROJECT SUMMARY

Name(s)

Manan A. Shah

Project Number

35761

Project Title

Improving the Accuracy of Sentiment Classification: A Novel Synthesis of Computational and Analytical Methods

Abstract

Objectives/Goals

The determination of individuals' mood in a review of a restaurant or the public sentiment of a political campaign are incredibly important statistics that cannot be accurately estimated manually or with simple computational techniques.

The purpose of this project is to a) make use of supervised and weakly supervised machine learning algorithms to accurately classify the sentiment of a string of text, and b) incorporate negation handling, word n-grams, feature selection by mutual information, subjectivity classification, and polarity determination to improve classification of sentences. This project is unique as it is the first in the field to holistically explore a novel combination of both supervised and weakly-supervised machine learning models.

Methods/Materials

The two approaches studied were tested against corpora of data from IMDb, Amazon reviews, and Twitter for accuracy, precision, and recall. For each specified dataset, numerous iterations were run with different sample sizes ranging from 15 to 100. The primary analysis involved the use of IMDb pre-classified polar movie reviews. Every review was split into sentences, which were preprocessed, classified for subjectivity and polarity, and stored for future predictions.

Results

After training, feature selection by mutual information, and further textual analysis, the supervised model yielded an average accuracy of 88.7%. The weakly supervised model predictions continually increased in accuracy and were able to consistently predict results with an accuracy of greater than 83% after only 600 iterations. The weakly supervised model was more adept at making predictions on novel data due to its use of pattern matching and objectivity classification, whereas the supervised model prevailed at classifying sentences similar to its training set.

Conclusions/Discussion

The weakly supervised model improved on the foundations of the supervised model. The addition of subjectivity and polarity classification as well as feature selection vastly improved accuracies as only highly subjective sentences were included in overall calculations. The primary difference between the supervised and weakly supervised model was the analysis of linguistic patterns in sentences, which allowed for better classification of unseen cases.

Summary Statement

This project compared and improved weakly supervised and supervised machine learning models using linguistic analysis, polarity and subjectivity classification, and negation handling to effectively classify the sentiment of provided text.

Help Received

Parents helped with the board assembly. Computer Science teacher and mentor Dr. Eric Nelson helped with algorithm testing.