Comparison of sentiment prediction techniques on a common movie review data set.

Matthew Smith

***Abstract*—In language, sentiment is a major component of overall expression. Sentiment analysis and classification uses traditional machine learning techniques but learns to predict what the text is trying to express. In this paper we will survey, discuss, and compare various techniques for sentiment classification. Such techniques include older methods such as bag of words classification and more modern techniques such as neural net classification. Some of these methods were implemented and tested on the IMDB movie review dataset. This dataset expresses the star review system as a binary value, where greater than seven is a 1, and less than five is a 0 and ¡find out what happens in between¿.These tests showed ¡results go here¿**

I. Introduction

When implementing machine intelligence techniques in practical applications, we often wish to predict the feeling of the author through text. Rather than predicting the subject of the text, sentiment analysis aims to determine how the text is speaking about a subject. Sentiment is crucial in language and communication, and therefore important to consider in text analysis. In this paper, we will be aiming to compare common machine learning models to determine which model succeeds in arbitrary length paragraph sentiment classification.

Were currently living in a world saturated with textual reviews, discussions, and opinions. Sentiment is a crucial datapoint for being able to take these pieces of text and extract a useful understanding of the authors intention.

-talk about what separates sentiment analysis from subject analysis
-talk about what kind of data is used for sentiment analysis
-talk about the IMDB dataset and the limitations and benefits of using it
-talk about what kind of experiments will be done in this paper

II. Discussion on the dataset

In this section we will be discussing the IMDB dataset and its format. Discussion on the data split, the labeled, and unlabeled, train and test data obtained from a kaggle competition page.

Also discussed will be the processing of the data, such as removing any html tags, removing any unwanted punctuation, and splitting the data on whitespace to create an array from which to learn on. Additionally, discussion on what punctuation should be kept and what should be ignored

-origins of the dataset
-kaggle
-format of the data set
-separation of dataset, labeled and unlabeled, test and train
-processing the dataset to make it more useful
-gram models
-including/excluding emojis and punctuation

III. Survey and comparison of methods

Here we will be an overview of methods to solve the problem of sentiment analysis as well as a comparison of their strengths and weakness.

-bag of words
-svm?
-neural net
-deep learning

IV. Method of Test

In this section, we will be discussing the methods that will be chosen for comparison. Also to be discussed is why each was chosen.

V. Results

VI. Conclusion

Acknowledgment

References

[1] Sentiment Prediction Using Collaborative Filtering, Jihie Kim, Jaebong Yoo, Ho Lim, Huida Qiu, Zornitsa Kozareva, Aram Galstyan
[2] A Sentimental Education: Sentiment Analysis Using Subjectivity: Summarization Based on Minimum Cuts Bo Pang and Lillian Lee
[3] Thumbs up? Sentiment Classification using Machine Learning Techniques Bo Pang and Lillian Lee, Shivakumar Vaithyanathan