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.

Sentiment analysis is distinct from subject analysis in the inference phase of training. For both types of analysis, vocabularies are built for the model to have an understanding of surrounding words. These methods split during training for classification and differ based on classification ground truth labels. For this paper, we will be using the Stanford IMDB dataset. This dataset consists of 100k individual reviews, with 50k labeled with either a positive or negative sentiment. This will allow for a large vocabulary to be available for training producing a diverse model to be used on unknown reviews. This dataset will be beneficial due to the medium to large sized paragraphs available, which will be allow for more robust models. In this paper, we will be implementing several models beneficial in sentiment analysis, including Bag of Words, Word vectors, and paragraph vectors. These models will be compared using identical classification trees in a random forest configuration.

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

The Large Movie Review Dataset (ref here) released by Stanford University, was built with sentiment analysis in mind. This dataset consists of 100k movie reviews from the Internet Movie DataBase (IMDB). Half of the reviews in the dataset were labeled using Amazons Mechanical Turk. For each sample, if a review gave greater than or equal to seven stars, the review was given a positive label. Likewise, a review with less than or equal to four, it received a negative label. This binary classification labels allow for the task of classification to be easier later on, as it avoids vague moderate reviews and their effect on the learning.

In raw form, the data was separated into 5 directories, each containing many text files of individual reviews. Test and train data were separated into separate directories, and within those, positive and negative reviews were separated as well. In order to make use of these reviews, each directory in the lowest level was concatenated into a single TSV file, with the filename as the ID, the sentiment, and the full review, each separated by tab characters.

Next, the data must be further processed and cleaned. For each review, punctuation, capital letters, and HTML tags were stripped and the clean reviews were written to a new TSV file with the same format. This process was done a second time, but this time removing stop-words from the reviews. Stop words are words which connect together sentences and often do not provide much meaning. These are removed to avoid noise in the models.

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 @InProceedingsmaas-EtAl:2011:ACL-HLT2011, author = Maas, Andrew L. and Daly, Raymond E. and Pham, Peter T. and Huang, Dan and Ng, Andrew Y. and Potts, Christopher, title = Learning Word Vectors for Sentiment Analysis, booktitle = Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, month = June, year = 2011, address = Portland, Oregon, USA, publisher = Association for Computational Linguistics, pages = 142–150, url = http://www.aclweb.org/anthology/P11-1015