Sentiment Analysis Using Semi-Supervised Recursive Autoencoders

Vinay Kumar

March 17, 2015

1 Introduction

Sentiment analysis is a growing field targeting to extract insights or subjective conclusions from the sources like text or large amount of data. Identifying the sentiment from online text gives businesses and organizations insights into what are main issues hitting the public opinion and what what they should focus on improving or changing. Getting significant attention from both business and research communities, sentiment analysis has many potential applications like summarizing user-reviews, brand-management and public relations management of business organizations and governments. Most of the past work has been focused on classifying the data in two classes: positive and negative. In my project, I aim to classify the data in five classes: very positive, positive, neutral, negative and very negative.

2 Algorithm: Semi-Supervised Recursive Autoencoders

I wish to use the Semi-Supervised Recursive Autoencoders algorithm from [1]. The algorithm consists of an unsupervised part and a supervised part. The unsupervised part is a recursive auto-encoder that creates an N-dimensional vector that represents the phrase or simply called as the 'code'. In the second part after obtaining an N-dimensional vector or 'code' for each phrase from the unsupervised recursive auto-encoder, supervised learning is to be used to classify this vector into a particular class. Standard supervised learning algorithms such as support vector machines or naïve bayes could be used for the classification. I wish to use softmax regression or one-vs-all logistic regression for multi-class classification. Apart from using the randomized N-dimensional vectors, I will use google word2vec tool [4] which maps words with similar meaning to similar positions in the N-dimensional vector space.

3 Dataset

I will use sentiment analysis dataset from Kaggle [3], which contains phrases and sentences from Rotten Tomatoes movie reviews. This dataset [3] consists of 8544 sentences which is converted to 156060 English phrases from movie reviews.

I will do the standard 70-30 percentage split from this dataset for the training set and the test set respectively.

References

[1] R. Socher, J. Pennington, E. H. Huang, A. Y. Ng, and C. D. Manning. 2011b. SemiSupervised Recursive Autoencoders for Predicting Sentiment Distributions. In EMNLP Url: http://ai.stanford.edu/ ang/papers/emnlp11-RecursiveAutoencodersSentimentDistributions.pdf

[2] Bahareh Ghiyasian and Yun Fei Guo. 2014. Sentiment Analysis Using Semi-Supervised Recursive Autoencoders and Support Vector Machines. Stanford.edu

[3] Kaggle. 2014. Sentiment Analysis on Movie Reviews. https://www.kaggle.com/c/sentimentanalysis-on-movie-reviews

[4] Google. 2014. word2vec. https://code.google.com/p/word2vec/