Sentimental Analysis with Amazon Review Data

Mingxiang Chen *ming1993@stanford.edu*

Yi Sun *ysun4@stanford.edu*

Introduction

Sentimental analysis often refers to using a combination of techniques like natural language processing and text analysis to identify positive or negative opinion, emotions or evaluations accurately. It could be hugely instrumental for us to get an overview of a paragraph. For example, it is widely used in social media monitoring. Many researchers use this as a tool to investigate problems in the field of social science, such as predicting stock price, and political preference. Some even use it as a tool to predict president election.

In this project, we compared different classification method using the Amazon review dataset to see which one works better under what situation.

Binary Classification

A sentence can be either “happy” or “not happy”. So far, there are already a bunch of natural language processing tool kits on the Internet. Here, we are comparing our Naïve Bayes model with the one called “Textblob”.

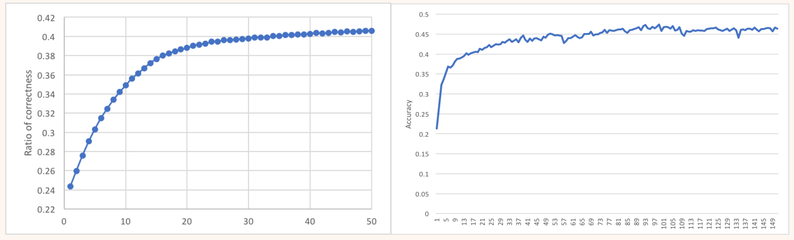
The Naïve Bayes model is trained on 8.8 million Amazon book reviews, and calculated from the probabilities of how likely a single word can lead to a good review (4-star or 5-star). Then multiply the probabilities for all the words in the sentence, and we can get the probability of “happy” for the sentence. The test set is 100 thousand book reviews. Since the Naïve Bayes is trained on a similar dataset, it achieves a higher accuracy at 78.78%. The ”Textblob” model achieves an accuracy at 63.34%.

Five-Category Classification

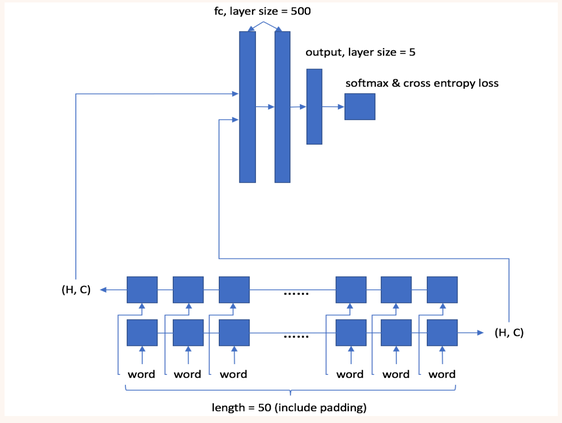
We compared 3 different methods in this multi-category classification: K-nearest neighbor (KNN), support vector machine (SVM), and neural networks.

To obtain the sentence vectors, we simply used the mean of pretrained GloVe (Global Vectors for Word Representation, https://nlp.stanford.edu/projects/glove/) word vectors for all the words appeared in the sentence.

The k values for the KNN models are 1, 5, and 10. The simple neural network model has 2 hidden layers with 200 units and the activation functions for each layer is Relu. Using the softmax function, we can get the probabilities. We used the ADAM optimizer to minimize the loss. Since the dataset is huge, we set a batch size of 100.



Training history for the simple NN model and the LSTM model



This is the structure of the LSTM model. Since we do not have a GPU, The network size is relatively small.

Result

Binary Classification	
Method	Accuracy
“TextBlob”	63.34%
Naïve Bayes	78.78%

Multiple Classification	
Method	Accuracy
KNN-1	25.52%
KNN-5	27.05%
KNN-10	27.92%
SVM (linear, one to all)	37.94%
Simple neural network	40.53%
LSTM recurrent neural network	46.66%

Discussion

The following are some example of misclassification:

Have purchased many in the past like in the 1970'safter I received one as a gift. I recommend to all ages.

Overall: 1 Prediction: 5

A must in everyone's library for graceful poetry, inpirational reading and lessons to live by and to even to let go. A new printing would be a welcome as the \"used\" ones in good condition are quite expensive.

Overall: 4 Prediction: 2

This work still holds up decades later. Every reflective, thoughtful person should have The Prophet in the home library. It can be read one chapter at a time. The chapter on love is as good as anything I've ever read.

Overall: 4 Prediction: 5