Capstone Proposal

Xiaodong TAN
21 March 2017

https://review.udacity.com/#!/reviews/414661

Domain Background

Sentiment analysis, sometimes known as opinion mining, is one of the main applications of natural language processing techniques. Its main aim is to determine if an individual's attitude/reaction towards a certain topic or event is positive or negative by analyzing his or her writings, voices or something else. Sentiment analysis is widely used in market research and customer analysis, but can also be used in other areas such as analyzing patients' feelings in the clinic settings and analyzing public opinion.

The classic way of performing sentiment analysis relied on representing text with a fixed-length of numbers, mostly using either clustering or a bag-or-words model. Bo and Lee (2008) provide a comprehensive review of sentiment analysis techniques in their classic paper.[1]. With the development of deep learning techniques, recent research has tried to compute embeddings that capture the semantics of word sequences (phrases, sentences, and paragraphs), with methods ranging from the word vectors to sophisticated architectures such as convolutional neural networks and recurrent neural networks (e.g., Maas etc., 2011,[2] Socker etc., 2013,[3] Kiros etc., 2015[4])

The project is inspired by Kaggle competition "Bag of words meets bags of popcorn". Besides, sentiment analysis on text and voices can also play an important role in my profession in fraud detection.

Problem Statement

[1] Bo Pang and Lillian Lee. 2008. Opinion mining and sentimentanalysis. Found. Trends Inf. Retr., 2(1-2):1-135.
[2] Andrew L Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, pp. 142–150. Association for Computational Linguistics, 2011.
[3] Richard Socher, Alex Perelygin, Jean YWu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the conference on empirical methods in natural language processing (EMNLP), 2013.
[4] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In Advances in neural information processing systems, 2015.

The aim of this project is to design a model to determine if the sentiment of a written review is positive or negative.

Datasets and Inputs

In this project, I will use dataset with labeled movies reviews from IMBD, used in Learning Word Vectors for Sentiment Analysis (Mass et al, 2011).[5] The dataset contains 50,000 reviews split evenly into 25k train and 25k test sets and there are no overlap between the two sets. The overall distribution of labels is balanced (25k positive reviews and 25k negative reviews). In the entire collection, no more than 30 reviews are allowed for each movie because reviews for the same movie tend to have correlated ratings. An additional 50,000 unlabeled documents, if necessary, might be used to train the word2vec model as the model can deal with unlabeled data.

In the labeled train/test sets, a negative review has a score <= 4 out of 10, and a positive review has a score >= 7 out of 10. Thus reviews with more neutral ratings are not included in the train/test sets. In the unsupervised set, reviews of any rating are included and there are an even number of reviews > 5 and <= 5.

Solution Statement

In this project, I will first convert tests (words) to numerical values (vectors) using some word embedding method. Then I will use the vectors as the input, consider the orders and put them into a Recurrent Neural Network (RNN) with Long Short term Memory (LSTM) architecture.

Benchmark Model

My benchmark models will include combinations of different methods to convert words to numerical features (bag of words methods/word2vec) and different classifiers. Examples of these benchmark models include bag of words + random forest/SVM/logistic regression and words to vectors + random forest/SVM/logistic regression.

Evaluation Metrics

[5] http://ai.stanford.edu/~amaas/data/sentiment/

Accuracy will be used as the main evaluation matrix. In this project accuracy is defined the proportion of correctly classified reviews in all the reviews in the test set. This is the evaluation matrix used in most of the other research on sentiment analysis (e.g., Socker etc., 2013[6]).

[6] Richard Socher, Alex Perelygin, Jean YWu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the conference on empirical methods in natural language processing (EMNLP), 2013.

Project Design

First I will clean up and preprocess the data so they can be fit into the models later. Issues I will need to deal with include but are not limited to html markup from the website, non-letters, stopwords, and sentence splitting. The data will be preprocessed differently depending on the converting methods used later.

Second, I will convert the cleaned and preprocessed text to different types of numerical features.

- I will use bag-of-words method to learn a vocabulary from all of labeled training set, then models each document by counting the number of times each word appears. The vocabulary size will be limited as the vocabulary might be very large.

- I will also use word2vec method to convert words to vectors by counting the co-occurance of words in the same sentence and creating a vector representation of each word. Depending on the computational resources needed, I might train a model myself using the method described by Kaggle or use a pre-trained model as suggested by http://www.volodenkov.com/post/keras-lstm-sentiment-p2/

Third, I will put features generated by different methods into different classifiers. Features generated by the word-of-bags method will be used as the input for classifiers such as random forest, SVM and logistic regression. Features generated by the word2vec method will be used as the input for the three classifiers above as well as LSTM network.

Last but not the least, the results from different models will be compared.

The design of the project can be summarized as follows.