Attention-Based Sentiment Analysis on Movie Reviews

Li-Hen Chen (UIN:928003907), Yun He (UIN:326005850),
Qifan Li (UIN:127004551), Fan Yang (UIN:525004621),
Yining Zhou (UIN:927009507)

March 25, 2019

Problem Statement: In this project, we propose to classify the reviews from movie website into positive and negative comments, using different machine learning models as well as deep learning models. Besides, we also aim to provide interpretations to the sentiment classification task with attention-based method, which could possibly help developers further debug the system.

Problem Motivation: In the web era, sentiment classification has been becoming a significant problem nowadays, due to the tons of online available textual information. To better facilitate human in doing online textual analysis for data mining, machine learning and deep learning techniques, such as SVM and neural networks, have been utilized in many real-world applications. However, most of those applied learning systems typically lacks interpretability, which largely hinders developers to know whether their deployed models are reasonable or not. To guarantee that the deployed model does not make predictions based on those bias and artifacts of data, developers need further interpretations of the generated predictions. To this end, in this project, we aim to provide some interpretation to the sentiment classification task beyond the prediction results, so as to help developers better understand the system they deploy.

Proposed Solution: We plan to use an interpretable attentional model for sentiment classification. First, we will apply word2vec to initialize the word embedding. Then, a RNN-based model is applied to capture the contextual information of each word. After that, an attention layer is trained to model the weight of each word. Specifically, the higher the weight is, the more important of the word is in terms of the information of sentiment of movie reviews. Moreover, attention weights can be used to interpret our model. For example, a list of words along with their attention weights can be returned to users as the reason why our model give such a classification result. Besides, we will also investigate the attention from the post-hoc manner, which would be agnostic to different RNN architectures.

Data Description: We will use SST-2(Standard Sentiment Treebank), which is composed of sentences extracted from movie reviews and human annotations(positive or negative) of sentiment. This dataset consists of training dataset, development dataset and test dataset. The number of instances for each split are 67K, 872 and 1.8K, respectively.

Expected Outcomes: We expect our outcomes to achieve state-of-the-art classification accuracy. Also, we would compare the performance of different models including logistic regression, SVM, decision tree, LSTM+RNN and CNN. Besides, in order to generate interpretations, we will show the corresponding attention scores from both intrinsic manner and post-hoc manner. By doing so, we are able to visualize the deployed model with heatmaps for sentiment analysis on movie reviews.