CISC 882 - Natural Language Processing

Final Project Proposal
Ye Fan, Lu Zhan, Mengdie Tao, Chen Ling
11/26/2018

Sentiment Analysis on Movie Reviews

BACKGROUND

Sentiment Analysis is also known as *Opinion Mining* is a field within Natural Language Processing (NLP) that builds systems that try to identify and extract opinions within the text (Bo and Lillian, 2008). Currently, sentiment analysis has become a topic with great interests and developments since it has many practical applications, including social media monitoring, marketing analysis, customer service, and product analytics. With the advantage of growing data size in most areas, a large number of texts expressing opinions and attitudes are available in review sites, blogs, and social media.

PROJECT DETAILS

The objective of this project is to apply different machine learning and deep learning methods in the task of sentiment analysis of movie reviews. Specifically, there are multiple movie review websites such as Rotten Tomatoes, IMDB, and Flixster, where users can rate based on their own feelings. Our sentiment analysis project aims at using raw movie review text with associated labels from Rotten Tomatoes to classify phrases on a scale of five classes: negative, somewhat negative, neutral, somewhat positive, positive (We use numerical value 0, 1, 2, 3, 4 to represent them respectively). The challenging part of the task is dealing with obstacles like sentence negation, abbreviation, language ambiguity, and metaphors.

The first part of the project is data-preprocessing. Since the original dataset contains only raw text with its label, we need to transform the shape of natural language and remove noises in order to better fit our model. Therefore, the data preprocessing contains the following steps: tokenization (Segmentation), Noise Removal (Remove stop words) and Normalization. Besides, for the purpose of contrast with traditional NLP classification methods, it is also essential to vectorize phrases (TF-IDF, Word Embedding) in order to learn the correlation between texts rather than treating words as discrete symbols like the *bag of words* model.

The essence of this project is to apply multi-class classification on sentiment analysis, which is also a classical machine learning project. To accomplish a better performance (Precision, Recall, and F1 Score), our team proposes to take advantage of ensemble learning. Particularly, each group member will take charge of a machine learning method for the classification task. Not only can we compare different models' performance, but also aggregate the scores of each model to fulfill a voting mechanism. Our baseline model will be Naive Bayes, which is straightforward and easy to understand. Other than that, traditional machine learning models like kNN, SVM, Random Forest as well as deep learning models like TextCNN and LSTM will be considered.

The last part of the project is the data visualization. We plan to draw several word clouds with the most representative words for each of the five classes in our training set. Besides, histogram and pie graph can also help us to understand the dataset better.

DATASET

The Rotten Tomatoes movie review dataset has been taken from Kaggle.com’s competition, which is a corpus of movie reviews used for sentiment analysis, originally collected by Pang and Lee (Pang and Lee, 2005). The feature is raw text divided by lines with its associated sentimental labels. Labels are like the following:

0 - negative

1 - somewhat negative

2 - neutral

3 - somewhat positive

4 - positive

Users can rate a movie’s quality by their feelings in the scale of 0 to 4 (from negative to positive), which is also our labels of the training set. The output of our model will also be transformed into the numerical format.

MILESTONES

1. Prepare the dataset
2. Finish the proposal and determine the basic role for each group members
3. Discuss machine learning and deep learning algorithms that we expect to use in this task
4. Finish the coding part and test model’s results
5. Aggregate our models and come up with a final resemble model for the sentiment analysis
6. Data visualization
7. Generate final report

REFERENCES

Bo Pang and Lillian Lee (2008), "Opinion Mining and Sentiment Analysis", *Foundations and Trends® in Information Retrieval*: Vol. 2: No. 1–2, pp 1-135. http://dx.doi.org/10.1561/1500000011

Pang, Bo, and Lillian Lee. "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales." *Proceedings of the 43rd annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2005.