CIS 519 Project Write-Up

Fall 2018

Dong Ku (David) Im, Sun Jae (Jasmine) Lee, Nina Chang

Objective

To create a model that accurately performs sentiment analysis of reviews. Given a user's product review in text, our model will predict how many stars (out of 5) that the user would give the product. This model will be used for predictions on not only product reviews (like those of Amazon), but also movie reviews (on IMDb and Rotten Tomatoes).

As we saw in Homework 2, a model trained for the NER task on news data does not perform well on email data because it fails to generalize. With this in mind, instead of building a domain-specific model, we want to build a model that can perform relatively accurate predictions on reviews in multiple domains. This would require that we choose features that accurately capture sentiment, but are general enough so that they are informative about sentiment in any domain.

Approach

1. Read papers on work that has already been done in sentiment analysis
2. Identify features that reflect sentiment but are not specific to a particular domain
3. Experiment with different learning algorithms on the data
4. Include the possibility of noise in labels as the rating datasets are subjectively built
5. Fine-tune parameters
6. Analyze the experimental data to find the best learning algorithm
7. Discuss further applications of the algorithm and enhancement of the project

Reading List

Mohammadamir Kavousi, Sepehr Saadatmand. *Estimating the Rating of Reviewers Based on the Text*. (https://arxiv.org/pdf/1805.08415.pdf)

Mingxiang Chen, Yi Sun. *Sentimental Analysis with Amazon Review Data.* (http://cs229.stanford.edu/proj2017/final-reports/5163147.pdf)

Jason Jong. *Predicting Rating with Sentiment Analysis*. (http://cs229.stanford.edu/proj2011/Jong-%20PredictingRatingwithSentimentAnalysis.pdf)

Raheesa Safrin, K.R.Sharmila, T.S.Shri Subangi, E.A.Vimal. Sentiment Analysis on Online Product Review. (https://www.irjet.net/archives/V4/i4/IRJET-V4I4598.pdf)

What We Have Already Done

We have researched the different datasets available for this task, and plan on using a dataset of IMDb reviews and one of Amazon product reviews. We also discussed algorithms that we could potentially use as our baselines.