W266 Project Proposal/Abstract

Yi Jin, Ashley Levato, Umber Singh

Text Reviews Rating Prediction

We intend to use customer text reviews to predict a review's star ratings. Review rating prediction could be used to recommend items, detect suspicious or fake online reviews, better understand customers' opinions, improve the default rating system, etc.

The work will be focused on the available yelp challenge dataset comprised of 2.7M reviews from 687K users for 85K businesses. The review file is a json file in the following format:

```
{
    'type': 'review',
    'business_id': (encrypted business id),
    'user_id': (encrypted user id),
    'stars': (star rating, rounded to half-stars),
    'text': (review text),
    'date': (date, formatted like '2012-03-14'),
    'votes': {(vote type): (count)},
}
```

Although extensive work has been done on portions of this dataset, very few work utilized the deep learning algorithms. We intend to implement a CNN text classification for sentiment scale/polarity analysis. The challenge of this project is to apply the NLP models and algorithms to improve the prediction accuracy.

The baseline model is bags of words model and logistic regression. We are going to use accuracy and cross-entropy loss to measure the results.

References/Resources

1. https://www.yelp.com/dataset_challenge
2. https://arxiv.org/abs/1408.5882
3. http://www.wildml.com/2015/12/implementing-a-cnn-for-text-classification-in-tensorflow/
4. https://kevin11h.github.io/YelpDatasetChallengeDataScienceAndMachineLearningUCSD/
5. http://cseweb.ucsd.edu/~jmcauley/cse255/reports/fa15/031.pdf
6. http://cs229.stanford.edu/proj2014/Chen%20Li,%20Jin%20Zhang,%20Prediction%20of%20Yelp%20Review%20Star%20Rating%20using%20Sentiment%20Analysis.pdf