Will Crawford

CECS 590 – Deep Learning

Review 1

I will be using the following database: <https://www.kaggle.com/yelp-dataset/yelp-dataset> to solve the problem of sentiment analysis. The dataset contains information about businesses spanning multiple states and cities, the user reviews, the check ins by users and the user info. I will be using the business dataset and the corresponding reviews to find restaurants with the better reviews.

State of the art algorithms that are popular in this approach are n-grams, bag of words and TF-IDF, these algorithms however are used in more common approaches[[1]](#footnote-1). There is, however, a study done that showed CNN (Convolutional Neural Network) and LSTM (Long Short-Term Memory) models proved to do sentiment analysis on the yelp database[[2]](#footnote-2). The following study does also show that a CNN-LTSM model which uses both CNN and LTSM with a pooling layer between in order to perform sentiment analysis with a high degree of accuracy[[3]](#footnote-3).

Case Study:

The yelp dataset contains many useful reviews but also many bad ones that are not entirely representative of certain businesses. The proposed model will be able to discern between the helpful reviews and the ones that are not. This will help businesses discern what users thought of their visit. They will be able to accurately determine what is a useful review, either negative or positive which will in turn show what the businesses’ pain points are and solve them.

This can also be used as a recommendation engine. Users will be able to find other restaurants that are most like the ones they already like based off of user reviews.

Dataset:



This is the first 500,000 reviews of the dataset with the standard English stopwords removed and bag of words as the input to the word cloud generator. My first thoughts are that more data trimming is required to remove strange formatting in the dataset, such as (, `, $, !. These formatting issues are from the fact that the reviews are written similar to tweets. So, further data cleaning is required so that plain English words are only in the data. Also, to a human, the helpfulness of the review, or the star rating, is pretty easy to recognize however, with the help of the CNN-LSTM models, the relationship will come apparent.

Finally, for the reviews, the statistics for the first 500,000 star reviews are:

count 500000.000000

mean 3.729382

std 1.455030

min 1.000000

25% 3.000000

50% 4.000000

75% 5.000000

max 5.000000

The mean of the data is around 4 which means that the reviews are typically helpful to other users. This however, does not represent if the actual sentiment of the review is either negative or positive.

Activity Calendar:

Week 1 Further cleanup of data, continuing to remove punctuation from the data.

Week 2 Implement CNN/LSTM from Bag of Words or TF-IDF from the data

Week 3 Analyze output and tune hyperparameters of the models

Week 4 Impelement CNN-LSTM

Week 5 Analyze output and tune hyperparameters of the models

Week 6 Implement LDA for the model

Week 7 Analyze output and tune hyperparameters of the models

Week 8 Summarize results into final presentation

1. Ay Karakuş, B, Talo, M, Hallaç, İR, Aydin, G. Evaluating deep learning models for sentiment classification. *Concurrency Computat Pract Exper*. 2018; 30:e4783. <https://doi.org/10.1002/cpe.4783> [↑](#footnote-ref-1)
2. Tang D, Qin B, Liu T. Document modeling with gated recurrent neural network for sentiment classification. Paper presented at: Conference on Empirical Methods in Natural Language Processing; 2015; Lisbon, Portugal. [↑](#footnote-ref-2)
3. Ay Karakuş, B, Talo, M, Hallaç, İR, Aydin, G. Evaluating deep learning models for sentiment classification. *Concurrency Computat Pract Exper*. 2018; 30:e4783. <https://doi.org/10.1002/cpe.4783> [↑](#footnote-ref-3)