Capstone Project 2 Proposal:
Understanding Yelp Text and Star Reviews for Business Success

Gurdeep Sullan
4/6/2020

Background:
User reviews can provide useful feedback for business growth and establishment. With the ease of online review apps like Yelp, consumers have added an ever growing amount of text and “star” review data for businesses they visit. Machine learning techniques in NLP can be used to decipher the importance of text features of a business for its success. Specifically, NLP can be used to analyze the actual text of the reviews for the data and perform sentiment analysis. This sentiment analysis can be powerful in gauging public tone. It can be used to understand and respond to customer criticisms of business marketing and products in real time.

Example Yelp star average star rating with business attributes

Problem:
There is a large amount of text review data that is connected to businesses on Yelp. Can we build a model to accurately predict sentiment of reviews (using the star rating as labels)? Can we find meaningful words in reviews that correlate highly with success in star rating?

Client:
The clients for this project would be businesses that are listed on Yelp. An analysis of their text reviews can give businesses a perspective of how they are doing and overall levels of customer satisfaction. On top of the numerical star reviews, the important words from this analysis can point businesses to potential key areas of improvement that will have a greater impact on reviews (and thus performance)

Data:
The data that will be used comes from the Yelp dataset on Kaggle:
https://www.kaggle.com/yelp-dataset/yelp-dataset
The project specifically uses two of the json files in this dataset:

- yelp_academic_dataset_business.json
- Yelp_academic_dataset_review.json

The dataset contains data from businesses in 11 metropolitan areas located in 4 countries. There are 5,200,000 user reviews and information on 174,000 businesses.

Additionally, a recurrent neural network will be trained in tensorflow.keras framework. This model will use byte-pair encoded data that has been generated and is available through tensorflow datasets:
https://www.tensorflow.org/datasets/catalog/yelp_polarity_reviews#yelp_polarity_reviewsplain_text_default_config
This dataset has 560,000 highly polar yelp reviews in its training set, and 38,000 highly polar yelp reviews in its test set. Though this dataset is also from Yelp, it is not the exact same dataset and therefore a 1:1 comparison between the RNN model trained on this dataset and the other ML models trained on the Kaggle dataset cannot be performed. However, it is a useful model to demonstrate and quantify the performance of a RNN model for sentiment analysis.

Approach:

- ➔ The data will be loaded in pandas dataframes using the read_json method. For the reviews dataset, due to the large size of the dataset, the data will be read in as chunks, and a random small sample (1%) of each chunk will be read into a dataframe, which will be concatenated into a larger dataframe after reading through all of the chunks.
- ➔ Data exploration and descriptive statistics will be completed on the data using pandas, scipy, spacy and matplotlib.
- ➔ The spacy NLP package will be used to engineer numerical features for the text data, both in the business dataset and in the review dataset. Due to the large size of the sampled review data, a generator will be used to generate lemmas from the text data.
- ➔ The data will be split into training and test data to be run on a variety of models. The reviews dataset labels will be generated, splitting the data into “positive” and “negative” classes based on the star_rating feature. A positive review is defined as having a star rating that is greater than 3. A negative review is defined as having a star rating that is less than or equal to 3.
- ➔ Models will be trained based on three approaches: (1) a bag of words approach using TfidfVectorizer from sklearn, (2) a BOW approach + SVD using TruncatedSVD from sklearn, and (3) word embeddings using the builtin spacy doc.vector attribute.
- ➔ Additionally, a RNN (recurrent neural network) model using keras in tensorflow will be trained to perform sentiment analysis on the raw text data.
- ➔ Hyperparameters and model performance will be recorded in order to keep track of optimal hyperparameters and models that perform well.

Deliverables:

The output of this project will be a machine learning model on the reviews dataset that predicts sentiment (positive vs negative) based on the text of the review. This model will return the top words that appear in positive reviews as well as negative reviews. There will be an associated hyperparameter table with this model as well.

All of the notebooks for data analysis, statistical analysis, and machine learning model implementation will be put on Github. Additionally, all reports, presentations, and data will be put on Github.