Project Proposal

Topic: Unsupervised Sentiment Analysis

Jennifer Caceres, Jenifer Vivar, Ali Salem, Zoya Shafique

Introduction

The aim of this project is to create an unsupervised model for sentiment analysis. To do so, we plan to use the Yelp Dataset which provides nearly 7 million reviews across approximately 150,000 businesses. However, these reviews do not have a "positive" or "negative" label attached to them and labeling them manually would take a copious amount of time. Therefore, this problem is a prime candidate for unsupervised learning.

More generally, being able to classify the reviews as positive or negative would help provide a more standardized rating system than using a 5 star rating for businesses. One person's scale of what constitutes 3 out of 5 stars may not be the same as another person's, making the 5 star scale very subjective. For instance, someone may have loved the food at a restaurant but have rated it 3 stars because of the ambience but someone else may not care about the ambiance and rate that same restaurant as 5 stars. Based on sentiment analysis of reviews, we could provide another metric to supplement the 5 star rating and provide a more standardized outlook on the business performance among customers for people searching on Yelp.

To achieve this goal, we plan on implementing approaches based on LSTM networks and a small scale autoencoder. As our baseline, we plan on using either a distance based metric to measure the distance between words in a review and words labeled as positive in an embedding space or a machine learning model such as KNN. Since we are attempting an unsupervised problem, we will not have ground truth labels with which to calculate accuracy or other evaluation metric. We plan to use cosine similarity between the word embeddings to show that the output of our models are relatively close to each other for positive reviews and negative reviews and that the two spaces are distinct from each other. We would also like to interpret our model results to see why reviews were classified as positive or negative.

Objective:

1. To develop an unsupervised method for sentiment analysis
2. To use deep learning algorithms for NLP to analyze the Yelp dataset. Namely, we plan on studying unsupervised LSTMs and autoencoders to classify reviews as "positive" or "negative"
3. To evaluate and ensure accurate prediction even without ground truth labels
4. Time permitting, we aim to answer two basic questions for every business: What is the business doing well? What does the business need to improve on? We plan to do so by interpreting the model results to see why it classified a review as positive or negative.

Methodology:

The methodology involves following steps

1. EDA to understand the structure of the data and what the reviews look like/what processing needs to be done
2. Data preprocessing to get rid of punctuation, standardizing and embedding the text, possibly using word2vec, and possibly dimensionality reduction
3. Computing baseline models such as k-Nearest Neighbors
4. Building a LSTM and/or unsupervised autoencoder to predict if a review is positive or negative
5. Interpreting the LSTM/autoencoder neurons to see why reviews were classified as positive or negative.

Expected outcome:

The expectation is that the model will be able to not only classify the reviews as positive or negative but also be able to provide useful feedback to the business based on the reviews. Specifically, in addition to relaying the overall sentiment of customers towards the business, the model should be able to convey why the customers have a largely positive or largely negative opinion about their operations.

Timeline:

The timeline for this project will be as follows:

- EDA (0.5 weeks)
- implement baseline (0.5 weeks)
- Work on main model (1.5 weeks)
- Fine tune model and work on model interpretation (1 week)
- Write paper and make presentation (1 week)

Deliverables:

The deliverables for this project include:

1. A deep unsupervised model for sentiment analysis
2. An interpretation of the model results of why the review was classified as positive or negative to give businesses feedback about customer opinion
3. A written report on the development and evaluation of the unsupervised sentiment analysis model and the Yelp Dataset
4. A presentation summarizing the main findings of the project