Natural Language Processing
CS 585
Milestone Report

Text Mining and Sentiment Analysis on Yelp Dataset Challenge

AUTHOR	ARNAB MUKHOPADHYAY, GEORGE MATHEW, SMRITI RAJ
CWID	A20353463, A20352131, A20364719
DATE	03/30/2017

Table of Contents

1. Abstract

- Yelp is a user service online which provides information and latest insights on local businesses
- Yelp allows registered users located across the world to rate and review business
- Yelp generates revenue by selling ads and sponsored listings to small businesses.
- We are trying to solve certain aspects of the yelp data set challenge which contains

  4.1M reviews and **947K** tips by **1M** users for **144K** businesses
  1.1M business attributes, e.g., hours, parking availability, ambience.
  Aggregated check-ins over time for each of the **125K** businesses

2. Introduction

- For our study, we are interested in the only restaurant data
- We have considered out only those businesses that are categorized as restaurants.
- We have filtered out restaurants that have ratings below 3 stars.
- Try to understand why the ratings are so low.Example: “The service is awful”,

“The ambience was not clean”

- This feedback helps business owners to improve upon certain areas
- This feedback also enables business owners to get an edge over their competition
- The yelp data set has 48485 restaurants
- Number of Reviews 528134 below 3 stars
- These 528134 reviews were given to 42009 restaurants

3.Method

Data clean up

- We load the json data and create a CSV of the data relevant to us
- The code extracts only those value we need ['business_id','type','stars','text']
- We then filter those reviews with 1 and 2 stars
- We use a tokenizing function that removes hyperlinks and stop words

Libraries Used:

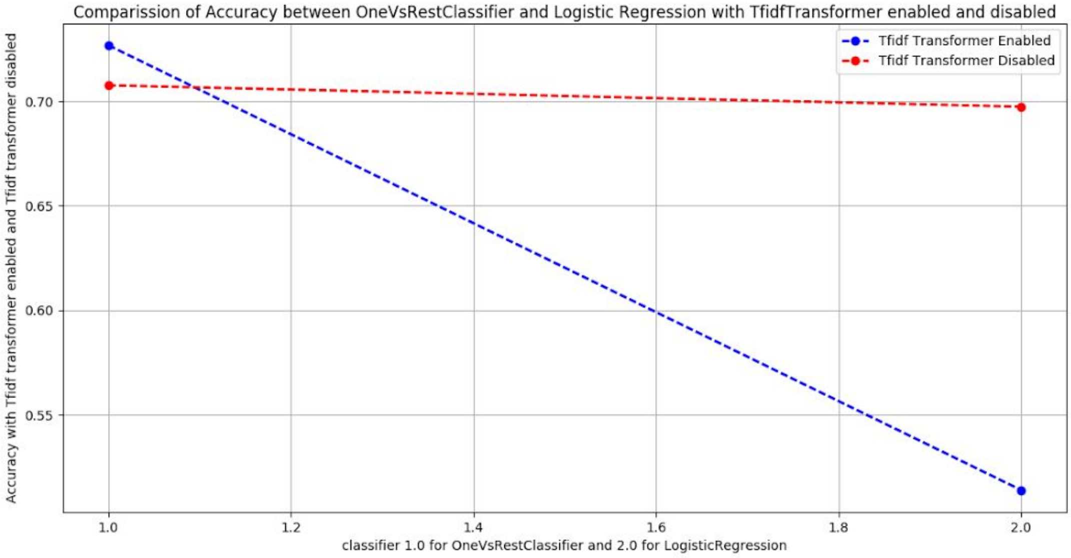
- Scikitlearn
- Numpy
- Pandas
- Itertools
- NLTK

3. Approach

- The first part of our code revolves around cleaning up the data and getting the bad reviews for Restaurant and food category
- Once we have our data we need to decide on the sentences relevant for our decision making.
- For example, sentences such as “my friend suggestion doesn’t help us understand" the reason behind a lower rating
- Sentences such as “bland difficult” to eat give us a more clear understanding behind the low rating
- We had to manually sift through the sentences and identify factors influencing the rating keywords such as Food, Ambience, Service
- The real challenge was to identify the keywords, as there are so many synonyms used
- To identify sarcasm in a statement and get around it
- Once the attributes are identified, we build a classifier to segregate relevant sentences from non-relevant ones by building a classifier
- Using the relevant sentences we are going to build our training data set
- Once we build the training data set we are going to identify the reasons behind why a restaurant is bad

4. Experiment

- We built one Logistic Regression and One v/s All classifier and to classify sentences as relevant or non-relevant
- We experimented with TFID enabled and disabled and found the when it is enabled leads to higher accuracy for One v/s All classifier
- We sifted through a number of reviews and picked the most common ones or the ones that have most common attributes
- We adapted different combinations of min_df, max_df, binary in our classifiers the highest accuracy is 0.746794871795 using a One v/s Rest Classifier with min_df, max_df, binary as 1, 2, True respectively



5. Conclusion

	Logistic Regression Classifier	One v/s Rest Classifier
TfidfTransformer-enabled	0.513974358974	0.726794871795
TfidfTransformer-disable d	0.697435897436	0.707692307692
Highest Accuracy min_df-1 max_df-2 binary -true		0.746794871795

We adapted different combinations of min_df, max_df, binary in our classifiers the highest accuracy is 0.746794871795 using a One v/s Rest Classifier with min_df, max_df, binary as 1, 2, True respectively

6. Related Work

- https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1928601
    a. In this paper three main findings are presented about the impact of consumer reviews on the restaurant industry:
    b. An increase by one star in Yelp rating leads to a 5-9 percent increase in revenue, this effect noticeable in independent restaurants not a chain
    c. restaurant chains have had a declined market share as Yelp penetration increased. This suggests that online consumer reviews have a more powerful effect than more traditional forms of reputation.

- https://arxiv.org/abs/1401.0864
    a. There are an overwhelming number of reviews for businesses especially restaurants, it is very hard for users to go through all reviews and find the related information
    b. This rating is mostly subjective and biased toward users' personality. In this paper, a business rating is predicted based on user-generated reviews texts.
    c. They use a combination of three feature generation methods and four machine learning models to find the best prediction. This approach creates a bag of words from the top frequent words in all raw text reviews, or top frequent words/adjectives from results of Part-of-Speech analysis

- http://dl.acm.org/citation.cfm?id=2631784
    a. This paper analyzes restaurant reviews to extract food words and create food recommendation. We used this paper to start analyzing our reviews and to understand the general nature of restaurant reviews.
    b. Our project tries to analyze restaurant reviews and understand the reasons behind a lower review. Though our review analyzis was inspired from this paper, our objectives are completely different

- https://pdfs.semanticscholar.org/8b2b/ada22181916196116f1711d456ea212f2b3b.pdf
    a. This paper discusses a food recommendation model based on the yelp data. K-nearest algorithm is used to build the model to recommend food to users. This helped us in understanding yelp data. We decided not to use the k-nearest algorithm and devised our own methods to build a model that understands the reasons behind lower reviews.
- http://varianceexplained.org/r/yelp-sentiment/
    a. In this project they used sentiment analysis methods to classify words as “positive” or “negative”, then to average the values of each word to categorize the entire document.Then they predicted a customer’s rating based on their written opinion.Finally they compared the predicted ratings with the original ratings.In our project we are taking the original ratings to filter out the data which has ratings above 3.Then we we worked on those reviews with lower rating to classify the sentences and to analyse the reason behind the bad ratings.

7. Who did what

- ○ Data Clean up: George Mathew
- ○ One v/s All Classifier: Smriti Raj
- ○ Logistic Regression Classifier: Arnab Mukhopadhyay

8. Timeline

- ○ We were able to build our classifier by March 30th
- ○ We will be completing our language model by April 13th
- ○ We will be performing various experiments and enhancing our language model for accuracy and efficiency by April 20th
- ○ We will be completing our report by April 25th
- ○ We will submit the complete project by April 27th