Sentiment Analysis of Yelp's Reviews

Anwar Alharbi
College of Computing & Informatics
Drexel University Philadelphia, PA 19104, USA
Email: asa393@drexel.edu

1 INTRODUCTION

Millions of users visit Yelp platform to search for businesses, posts and read reviews. Yelp was founded by Jeremy Stoppelman and Russel Simmons In July 2004. The platform connects people with great local businesses. It also offers consumers a one-stop local platform where they can discover, interact with, and transact with local businesses of all sizes. According to Yelp statistics, there are 244M reviews; this massive amount of data has great value and can bring helpful insight to the business owners to enhance their business. [1]

Therefore, this project aims to use sentiment analysis by utilizing Apache Spark to classify Yelp reviews as positive or negative. The ultimate objective of this project is to use Spark NLP to create a sentiment analysis model for the Yelp dataset's customer reviews, which will then be evaluated using classification methods.

2 METHODOLOGY

2.1 DATA DESCRIPTION

The data was obtained from the Yelp Open source dataset. It has more than 7 million reviews for more than 150,346 businesses across eight metropolitan areas in the USA and Canada.

The primary dataset includes a subset of businesses, reviews, and user data, but this project focuses mainly on two datasets: business and Reviews. [2] The Business dataset has 117,618 samples after the missing values were removed, and 7 of the 13 features were selected. In addition, the Reviews dataset contained 6,990,280 samples and 2 of 9 features chosen for analysis.

2.2 Data Features

The dataset is composed of different features. The following list is the features in the Business dataset:

- business id: unique identifier for each business.
- categories: list of different business categories.
- city: location.
- business Name: complete business name.
- review count: number of reviews.
- rating: star rating, rounded to half-stars.
- state: 2 character state code, if applicable.

While the Reviews dataset have the following list of features:

- review id: 22 characters unique review id.
- business id: 22 character business id, maps to business in business table.
- date: date formatted YYYY-MM-DD.
- text: the review itself.

The vast dataset size was too much for Google Colab to handle. Therefore reviews were filtered only to restaurants in Pennsylvania state. Additionally, only reviews with stars 1,2,4 and 5 were considered in the analysis. The final dataset contains information about 26,317 restaurants with 544,840 reviews. For labeling, Restaurant reviews with three stars or more were rated positively. Otherwise, they were negative.

2.3 EDA

It's important to visualize the data to get useful insight and understand it more clearly. The Yelp dataset for Pennsylvania contains 26,317 businesses, 5,252 of which are restaurants. Figure 1, shows the Rating Distribution of restaurants reviewed in Pennsylvania, 409,790 restaurants have received 4 stars. Following that, restaurants with three stars with a

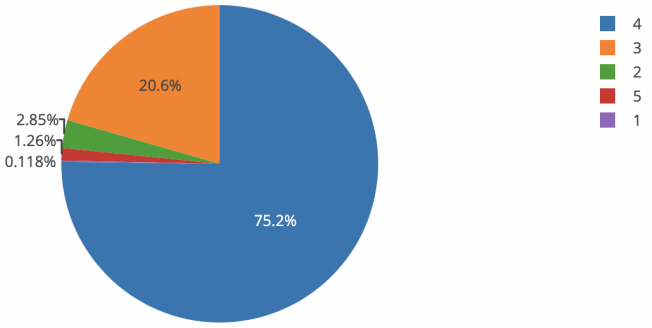

Fig. 1: **Rating Distribution of Restaurants Reviewed in the state of PA.**

total of 112,008. And the minority of ratings were in restaurants with 1,2 and 5 stars with a total of 643 , 15514, 6885 respectively. Another finding obtained during the analysis is that most ratings were in restaurants in Philadelphia with a total of 2400.

3 Experiments and Results

The sentiment analysis pipeline was built using Spark NLP once the dataset has been explored and various pre-processing techniques have been used on the reviews.

Data pre-processing is the first phase and it generally focuses on text cleaning by removing null values, removing steps words, tokenization, and normalization. The word cloud of the 181 eliminated steps words from the reviews is shown in Figure 2. According to the analysis, there were 528,683 positive sentiments overall and 16,157 negative ones. Additionally, positive and negative cloud words were obtained, as shown in Figure 3 and Figure 4, respectively.

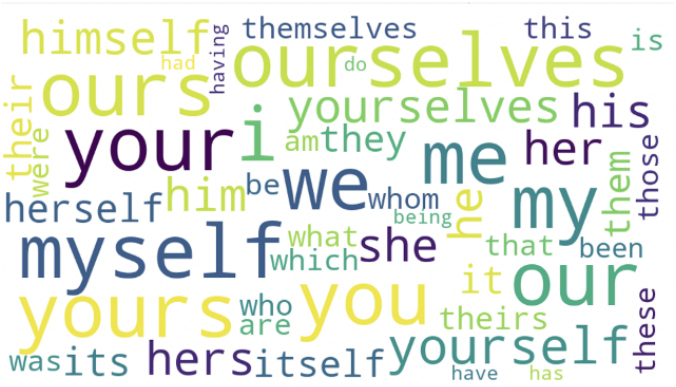

Fig. 2: **Word cloud illustration for positive words.**

Fig. 3: **Word cloud illustration for the eliminated steps words.**

The process of calculating TF, or the frequency of each term in a review was done in the following step using CountVectorizer. Hence, the data is now prepared for text classification after the sentiment analysis pipeline has been constructed.The data was Split into 80% for training a total of 435,848 and 20% for testing with a total of 108,992.And to handle Class Imbalance in the training dataset's minority classes, which is the negative label, weights in the Loss Function approach were applied. [3]

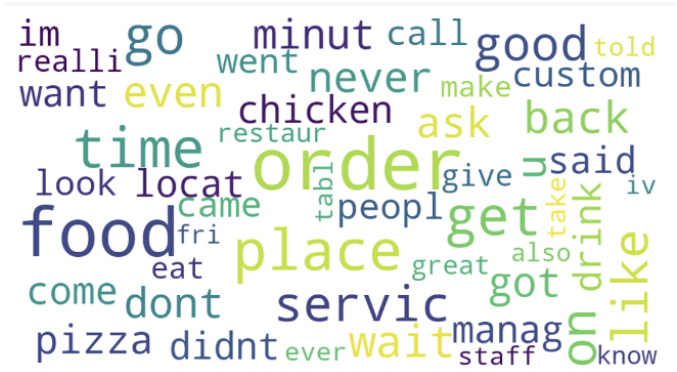

Fig. 4: **Word cloud illustration for negative words.**

4 Conclusion

In this project, Supervised machine learning algorithms were applied to train the model including Logistic Regression, Random Forest and Support Vector Machines. As a result, we examined the accuracy of several methods to find that Support Vector Machines had the highest classification accuracy 97% and Logistic Regression had the lowest classification accuracy 85%. While the classification accuracy of Random Forest was 95%.

References

[1] Yelp facts. Yelp. (n.d.). Retrieved August 24, 2022, from https://www.yelp-press.com/company/fast-facts/default.aspx.
[2] Yelp open dataset. Yelp Dataset. (n.d.). Retrieved August 24, 2022, from https://www.yelp.com/dataset.
[3] Handling Class Imbalance Using Weights. (n.d.). Retrieved from http://blog.madhukaraphatak.com/spark-3-introduction-part-4/.