Sentiment Analysis of the Yelp Reviews Using NLP & Topic Modeling

Leena AlQasem
Leenabdulh@gmail.com

Randa Mohammed
randa1414@gmail.com

Abstract:

Analyzing the massive text is very time-consuming. Therefore, Natural Language Processing (NLP) was used in this project because it helps resolve ambiguity in language and adds useful numeric structure to the data. Moreover, the machine learning technique that we used is topic modeling, it extracts useful information from any text and analyzes it. It also allows us to decrease the dimension of data using matrix decomposition in order to categorize the text into handful topics for easier analysis.

Design:

After gathering data, we have analyzed reviews through exploratory data analysis using some visualization techniques such as histograms and word cloud. EDA helped us to know how the reviews are distributed between positives and negatives reviews. Then, the text of reviews were pre processed through cleaning and tokenizing the text. Stop words, punctuations, urls, numbers, capitalization, non-alphabeticals characters were all removed in the cleaning process. In addition, lemmatization was applied in order to shorten the words into its base forms. TF-IDF technique was used for tokenizing because it is a better approach than count vectorizer. Finally, three topic modeling techniques were applied to the data which are SVD, LDA, and NMF.

Data:

The Yelp dataset was obtained from Kaggle as a public source Here. It includes *business_id, date, review_id, stars, text, etc.*

- **Scope:** The dataset we used contains 6911 records with 10 columns, we have selected all 6911 rows from the original dataset for the purpose of training the model on as much data as we can. Thus, the model can have high performance. In addition, we took only the text to be analysed through NLP techniques.

Algorithms:

First, we divided the dataset into two different datasets which are positive and negative reviews. We treated each dataset separately in order to explore each one alone. We divided them based on the rating stars where the positive dataset contains all reviews that have 3 or more stars. On the other hand, the negative dataset contains all reviews with 2 or less rating stars. Then we split both dataset into 90% training and 10% test.

1. NLP: Cleaning data → Remove Stop words, punctuations, urls, numbers, capitalization, non-alphabeticals characters

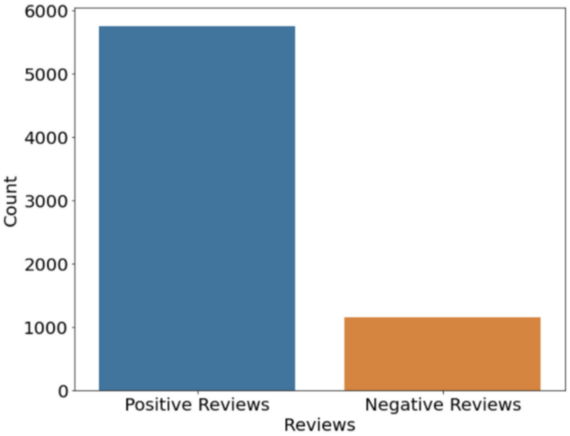
   Tokenize → using TF-IDF

2. Topic Modeling: We have tried multiple models in order to see which one makes more sense with our dataset. LDA, LSA, and NMF models were all trained on both positive and negative datasets. NMF turned to make most sense on providing meaningful topics. Thus it was used on test data. To analyze the performance of the model on test data 10 documents of each dataset were manually analyzed.

Tools:

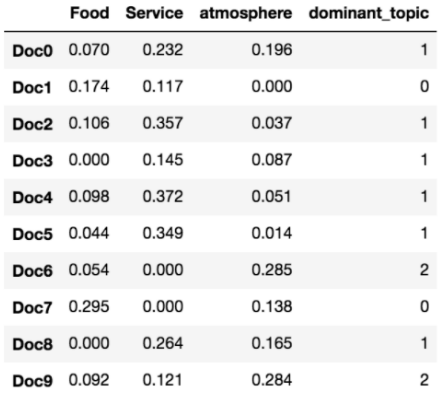
- **Technologies:** Jupyter Notebook, Python.
- **Libraries:** Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn, and NLTK

Communication:



Visualize the Number of both the Positive and Negative Reviews.

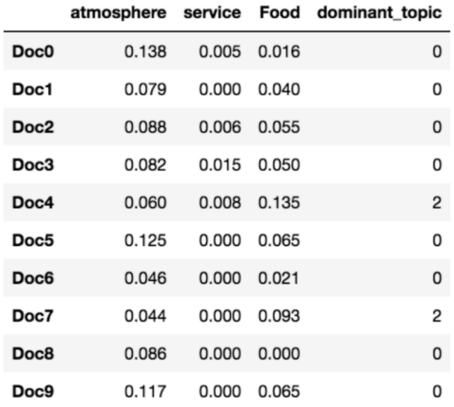
The stars with 3 and above are considered as positive reviews AND stars with 2 and 1 are considered as negative.



	Food	Service	atmosphere	dominant_topic
Doc0	0.070	0.232	0.196	1
Doc1	0.174	0.117	0.000	0
Doc2	0.106	0.357	0.037	1
Doc3	0.000	0.145	0.087	1
Doc4	0.098	0.372	0.051	1
Doc5	0.044	0.349	0.014	1
Doc6	0.054	0.000	0.285	2
Doc7	0.295	0.000	0.138	0
Doc8	0.000	0.264	0.165	1
Doc9	0.092	0.121	0.284	2

Experiments the NMF model on the Negative Reviews (Testing)

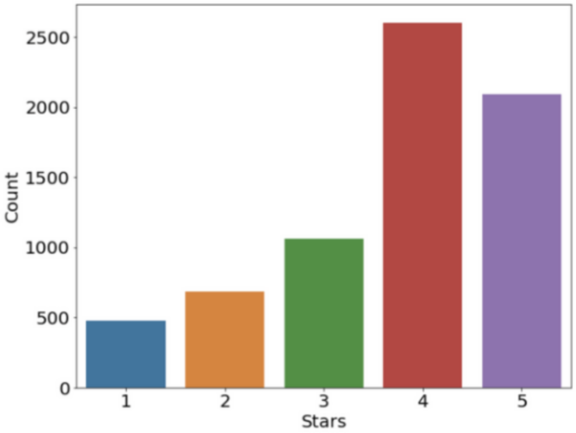
Comparing the topics with the documents. Found out that our model predicts 80% of topics right (8 out of 10)



	atmosphere	service	Food	dominant_topic
Doc0	0.138	0.005	0.016	0
Doc1	0.079	0.000	0.040	0
Doc2	0.088	0.006	0.055	0
Doc3	0.082	0.015	0.050	0
Doc4	0.060	0.008	0.135	2
Doc5	0.125	0.000	0.065	0
Doc6	0.046	0.000	0.021	0
Doc7	0.044	0.000	0.093	2
Doc8	0.086	0.000	0.000	0
Doc9	0.117	0.000	0.065	0

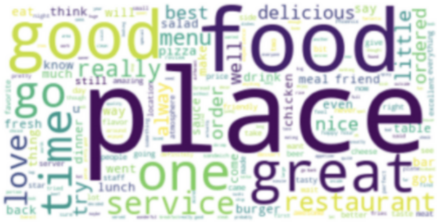
Experiments the NMF model on the Positive Reviews (Testing)

Comparing the topics with the documents. Found out that our model predicts 70% of topics right (7 out of 10)

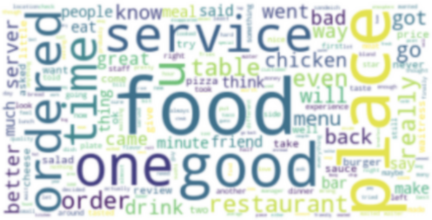


Visualize the Number of both the Positive and Negative Reviews.

The stars with 3 and above are considered as positive reviews AND stars with 2 and 1 are considered as negative.



Visualize all positive words



Visualize all negative words