Open in app

Pooja Prasannan

Lists About

Exploratory Data Analysis and Data Mining on Yelp Restaurant Review

Pooja Prasannan 1 hour ago · 6 min read

We all depend on reviews to make decisions in wide range of situations starting from which doctor to consult, which restaurant to go, which food to order etc. Not only the customers benefit from reviews, but businesses too rely on customer reviews to understand what went well, what did not and what could be done better. A research shows that one star rise in Yelp 's ranking resulted in a 5–9% increase in restaurant sales. But the real challenge lies in how to understand and make better use of this enormous amount of data in the Internet. Through this article we will understand how the authors of this research paper[1] have used Exploratory Data Analysis(EDA) and Data Mining towards it. We will also see the application of the Term Frequency method (Bag-of-Words) on reviews to get frequent words and phrases in each review class, whether positive, negative, or neutral.

Before digging further lets first understand how EDA and Data Mining helps through different use cases.

1. In a particular study, researchers analyzed Twitter data of four different airlines to find out the most common words used for compliments and slanderous words by the people. They represented the positive and negative tweets graphically in both spatial and temporal context. This visualization helped them to find out the words.[4]

2. In another study, researchers tried to identify the most common topics discussed in YELP along with the their sentiment using EDA.[2]

Before diving deep into YELP dataset, lets discuss a little about YELP. YELP provides crowd-based local business feedback. YELP had a monthly average of 61.8 million unique visitors. YELP reported that it had 192 million ratings by 2019. The platform has sections for specific businesses such as restaurants, clinics, hotels, beauty salons etc. It allows users to submit a one to five-star rating for a businesses' goods or service along with a comment or post as a textual review.

Dataset:

The dataset is from the Kaggle. There are five CSV files, among those two are used here yelp_business and yelp_review. There are 174,567 records of different businesses in yelp_business dataset. The yelp_review dataset contains 5,261,668 documents and nine descriptive features.

Exploratory Data Analysis:

Exploratory Data Analysis (EDA) is a data analysis technique mostly accompanied with visual methods to understand the data characteristics.
Visualizations helps to get insights about data beyond formal models or testing hypotheses. The National Institute of Standards and Technology (NIST) defines EDA as a data analysis approach/philosophy using a range of techniques for optimizing insights into the data set; disclosing the underlying structure; capturing significant variables; identification of outliers and anomalies; checking underlying assumptions; designing promising models; and determining optimum factor settings.

Bag-of-Words:

The Bag-of-Words model is a technique to extract textual features from texts. It describes the occurrence of words in a text corpus. It is called a bag because the order or meaning of the words doesn't matter. It is concerned with how many recognized terms are there in the document and their frequency. The model's output is a vector containing words and their count in a document by applying a sum function on the one-hot encoding representation.

EDA on Yelp Dataset:

The authors have plotted the graphs using Plotly and Cufflinks library of Python. First, the most relevant category of business is plotted as a bar chart. Below Fig:1 shows that restaurant is the top category. So in further analysis restaurant dataset has been filtered out to understand what are the best restaurants, which category is more famous etc.

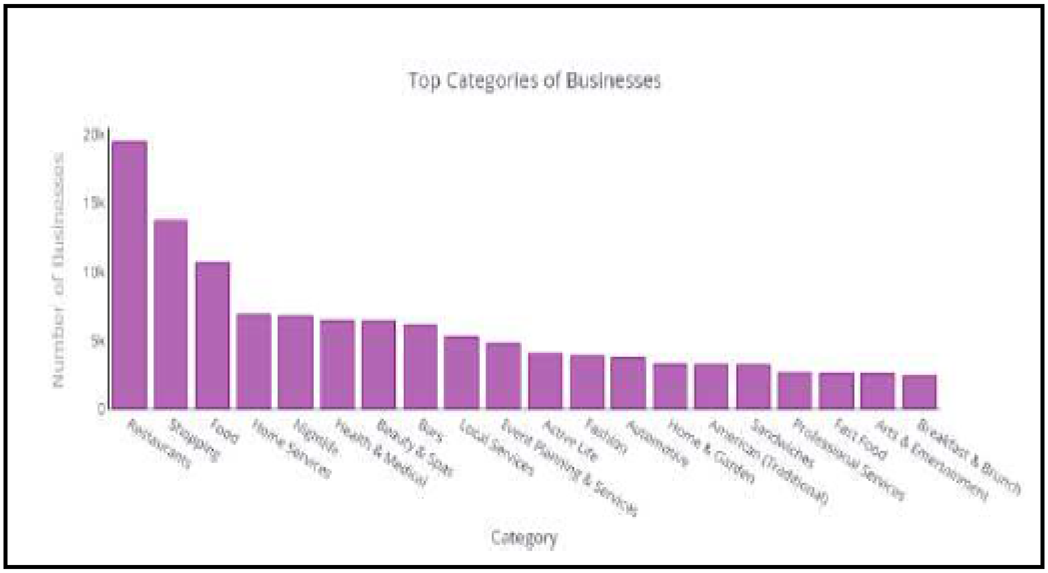

Fig-1 : Bar chart for top business categories (src : https://ieeexplore-ieee-org.libaccess.sjlibrary.org/document/9428850)

Classifying the restaurants, in Fig-2: we can see the most famous categories. Fast food restaurants, followed by sandwich restaurants, then American restaurants, are the most prevalent restaurants.

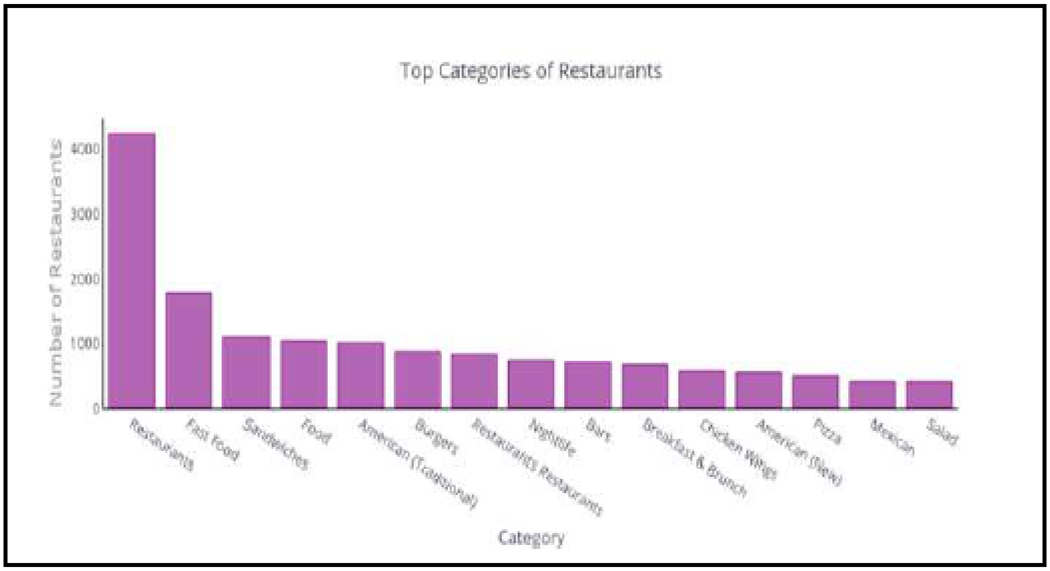

Fig-2 : Top categories among restaurants (src : https://ieeexplore-ieee-org.libaccess.sjlibrary.org/document/9428850)

To understand the ratings of these restaurants, the authors has classified the ratings to positive, negative and neutral. The 4-stars and above corresponds to positive, 2-stars and below means it is negative and 2–4stars represents neutral. Fig-3 shows these classes. From the figure we can see positive reviews are the highest, while neutral is the least.

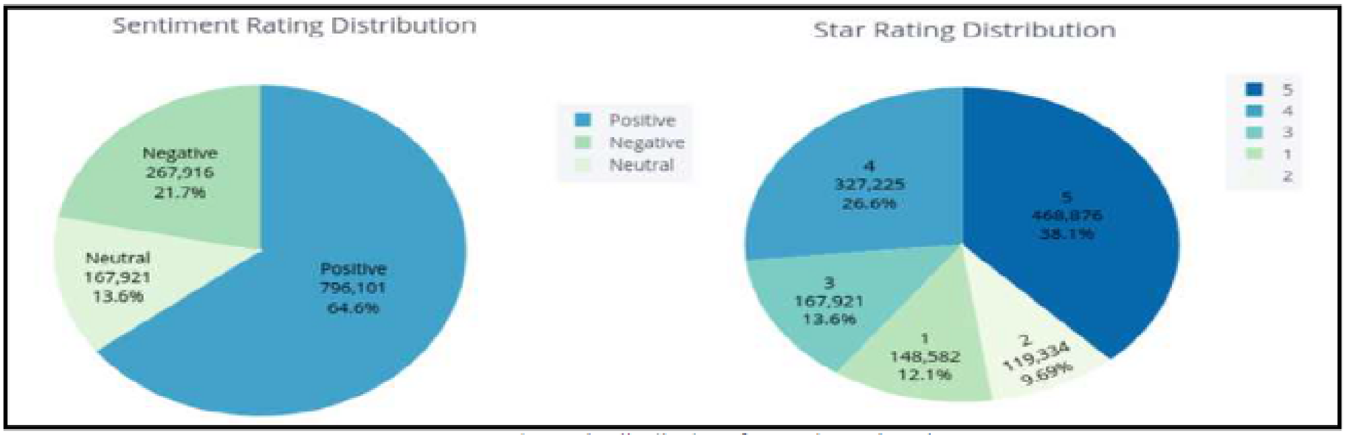

Fig-3: Sentiment and Star Rating Distribution (src : https://ieeexplore-ieee-org.libaccess.sjlibrary.org/document/9428850)

Fig-4 represents the best restaurants based on their average star rating. Thus we can see how EDA helped us to narrow down the whole dataset to get the best restaurants.

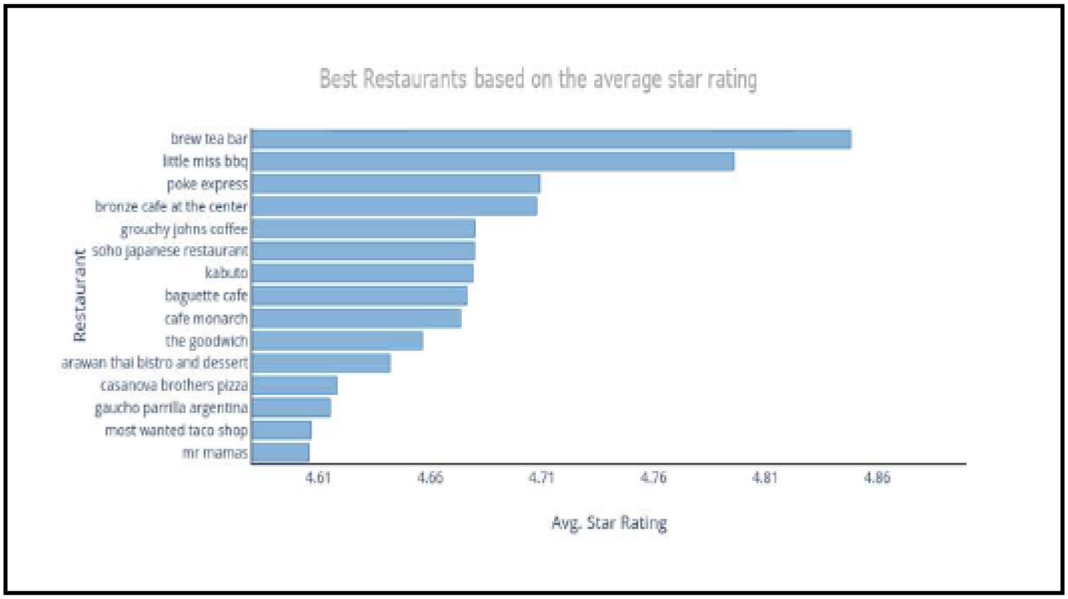

Fig -4 : **Best Restaurants based on average start rating**(src : https://ieeexplore-ieee-org.libaccess.sjlibrary.org/document/9428850)

“Brew tea bar” is the best-rated restaurant as per the average rating, which is 4.86 out of 5. On the other hand, the “KFC” restaurant represents the worst in terms of its average rating, which is 1.80 out of 5. To deep dive into branch wise, the authors have analyzed the data based on location wise in USA for KFC. Fig-5 represents the average star rating for each state in USA where KFC is located, Pennsylvania has the highest average value of the ratings. This represents the **spatial analysis** of data. This helps in understanding the branches/states which are poorly performing.

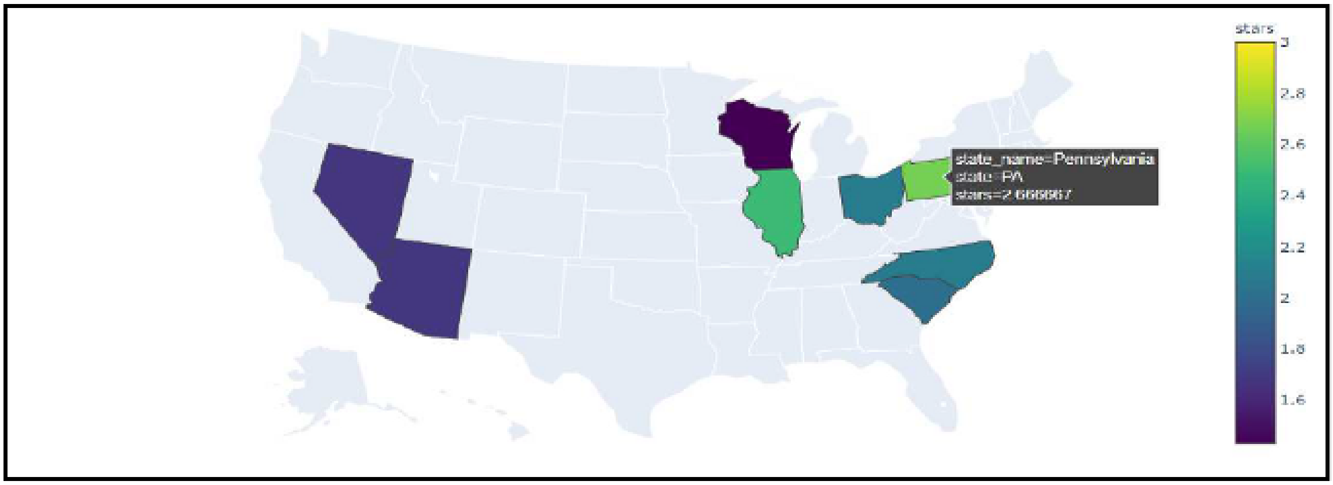

Fig-5: The average star rating for each USA state with KFC branches(src : https://ieeexplore-ieee-org.libaccess.sjlibrary.org/document/9428850)

Temporal Analysis also adds value in understanding how the customers were satisfies over the timeline. This can be done based on the review published date. The line graph in Fig-6 indicates the annual timeframe for the reviews of “KFC” from December 2005 to December 2017. Since 2014 it can be seen that the number of one-star reviews has risen sharply.

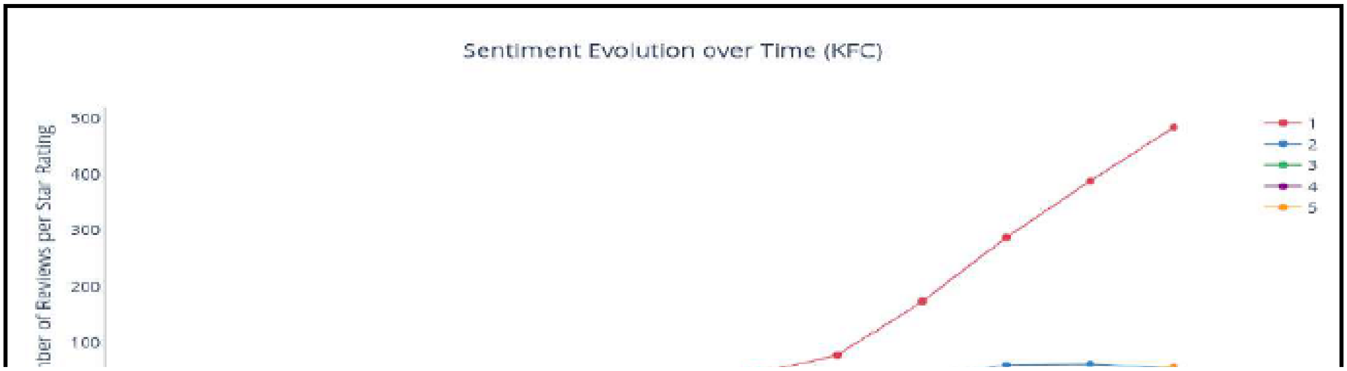

Fig-6: KFC star ratings from Dec 2005 to Dec 2017(src : https://ieeexplore-ieee-org.libaccess.sjlibrary.org/document/9428850)

Feature Extraction : To find about the most used words in positive, negative and neutral reviews, the authors has used Bag-of-words(BOW) and Count Vectorizer. First the data is preprocessed by tokenization and lemmatization. Further punctuations, personal pronouns, non-alphabetical characters are removed using the library in Python-‘spaCy’. The data was converted to numerical vectors representing the frequency of each word in the dataset. Using BOW unigram features, ‘good’ is the highest repeated word in most positive reviews. Using the bigram feature ‘great food’ is the most repeated feature.The most frequent phrase based on Trigram is “food greatservice” in positive reviews, “good not great” in neutral reviews, and “bad customer service” in negative reviews .Fig-7 shows the most frequent trigrams in positive reviews. So this kind of feature extraction will help in understanding what is good and what is not regarding customer satisfaction.

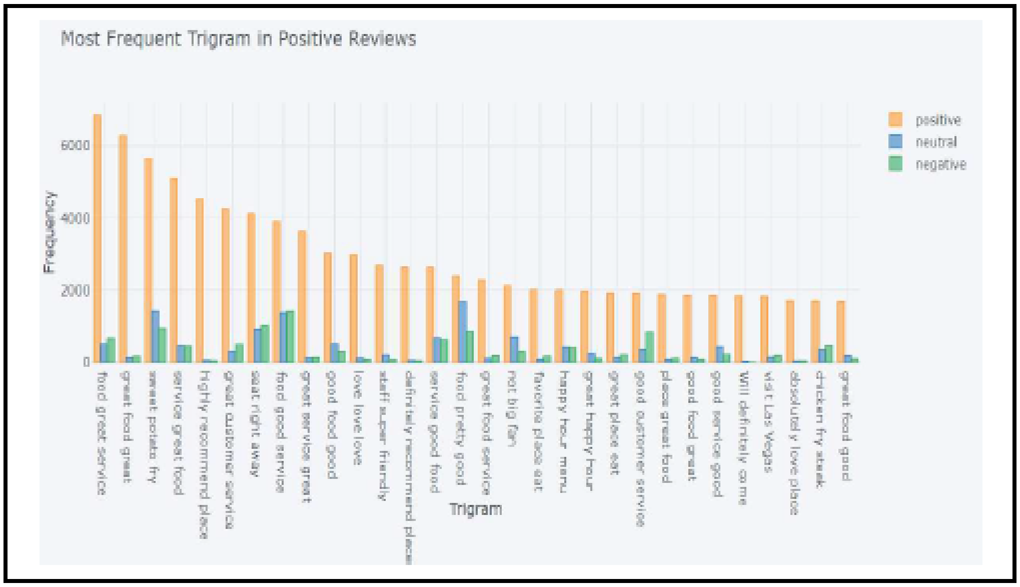

Fig-7: Most Frequent Trigram in Positive Reviews(src : https://ieeexplore-ieee-org.libaccess.sjlibrary.org/document/9428850)

Conclusion :

In order to make informed decisions, data understanding is important and EDA helps a lot in this context. EDA on the yelp dataset helped to get so many insights like, most important category, most liked restaurant type, high and low rated restaurants etc. Spatial and temporal analysis of KFC gave insights which branches are poorly performing and when did rating started to deteriorate. Extracting the phrases helped to know what are the most important things customers are looking for i.e. 'food and service'. Hence to increase customer satisfaction may be restaurants should focus on that aspect.

In my opinion..

Through this research paper, I got to learn how a huge dataset was narrowed down using EDA. Firstly, it reduced down to the top category in yelp dataset. Further analysis helped to segregate the best and worst restaurants. Spatial and temporal analysis gave very useful insight, which companies could use to their benefit and have a targeted problem resolution. I found feature extraction to be a very useful technique because it is important to understand what people are mostly saying about, for eg, in this case it was 'food and service'. So overall EDA with feature extraction has great potentiality to find out useful insights from a dataset.

Happy Reading!

References:

1. https://ieeexplore-ieee-org.libaccess.sjlibrary.org/document/9428850

2. S. Kaleru and S. Rao Dhanikonda, "Exploratory Data Analysis and Latent Dirichlet Allocation on Yelp Database," Int. J. Appl. Eng. Res., vol. 13, no.21, pp. 15035–15039, 2018.

3. https://spacy.io/

4. R. Khan and S. Urolagin, "Airline sentiment visualization, consumer loyalty measurement and prediction using twitter data," Int. J. Adv. Comput.Sci. Appl., vol. 9, no. 6, pp. 380–388, 2018.

Eda Data Mining Yelp Review