Exploratory Data Analysis and Data Mining on Yelp Restaurant Review

- By Pooja Prasannan

Objective

- How EDA help in understanding yelp data set- find top categories and narrow down the data set to get useful insights.
- Temporal and Spatial Analysis of a case study- KFC in USA.
- Identifying frequent words and phrases used in positive, negative and neutral reviews using Bag-of-Words and Count Vectorizer.

Exploratory Data Analysis

- 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

Bag-of-Words

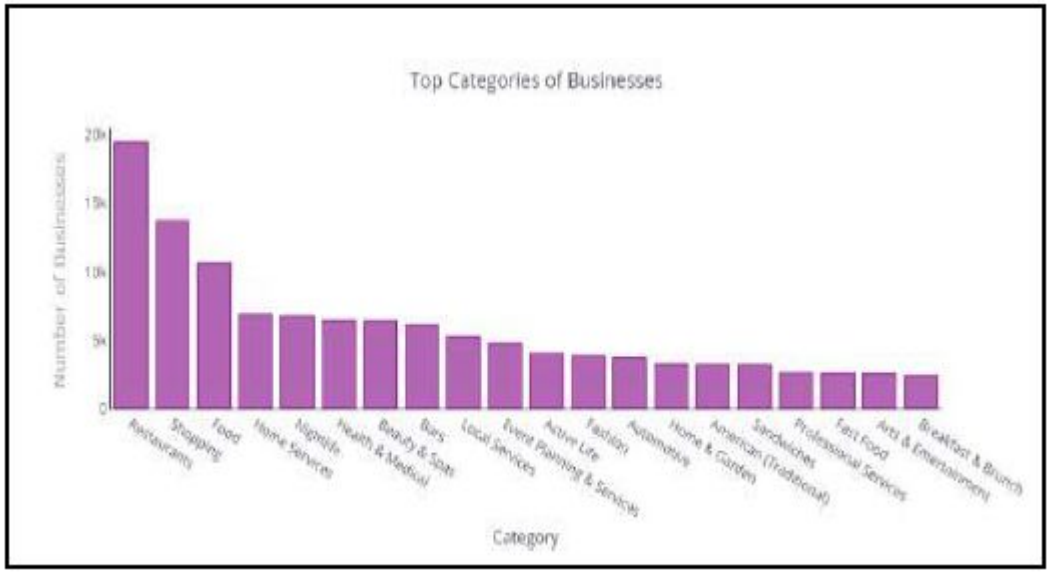
- A technique to extract textual features from texts.
- It describes the occurrence of words in a text corpus.
- It is called a bag because the order or meaning of the words doesn't matter.
- It is concerned with how many recognized terms are there in the document and their frequency
- The model's output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

Yelp Dataset:

- YELP provides crowd-based local business feedback.
- 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.

EDA on Yelp Dataset

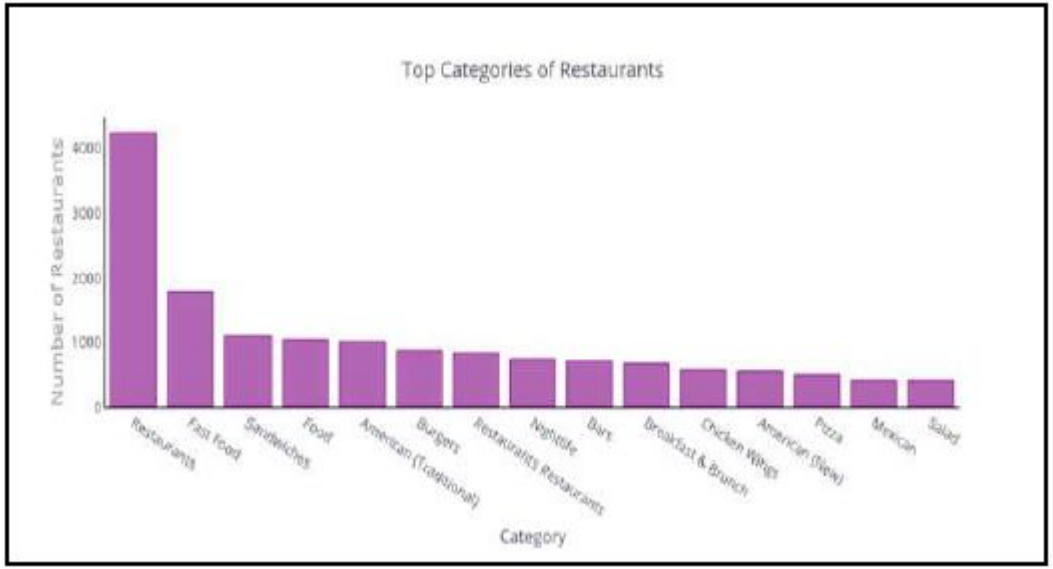
The authors have plotted the graphs using Plotly and Cufflinks library of Python.



Observation

Figure 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.

Top category among restaurants

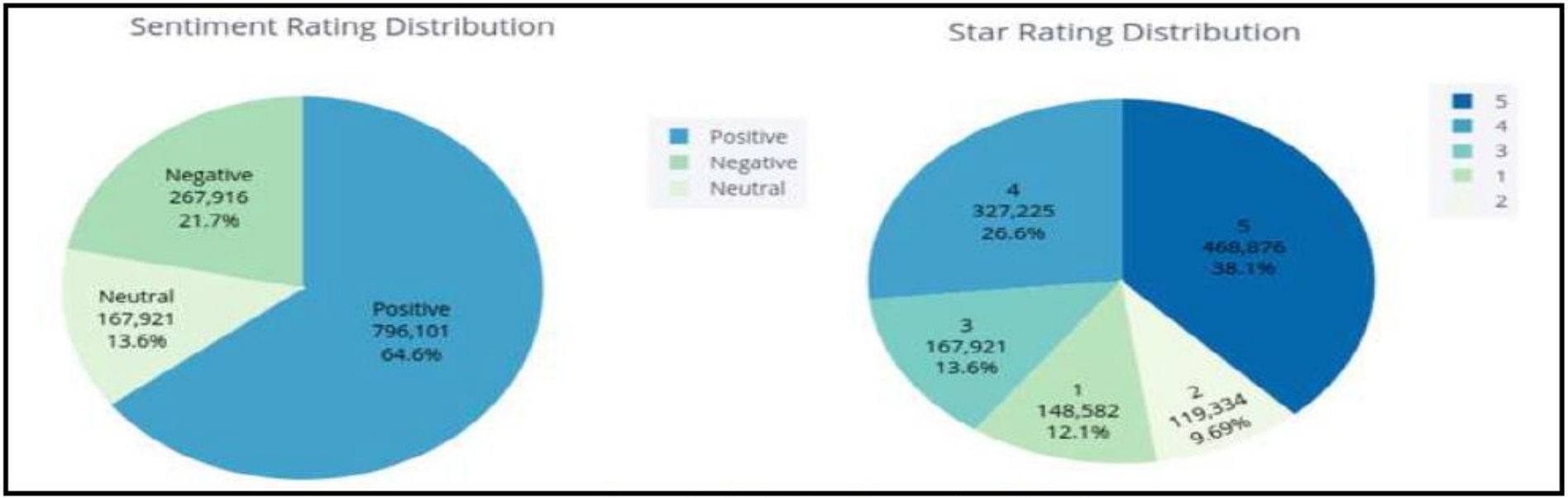


Observation

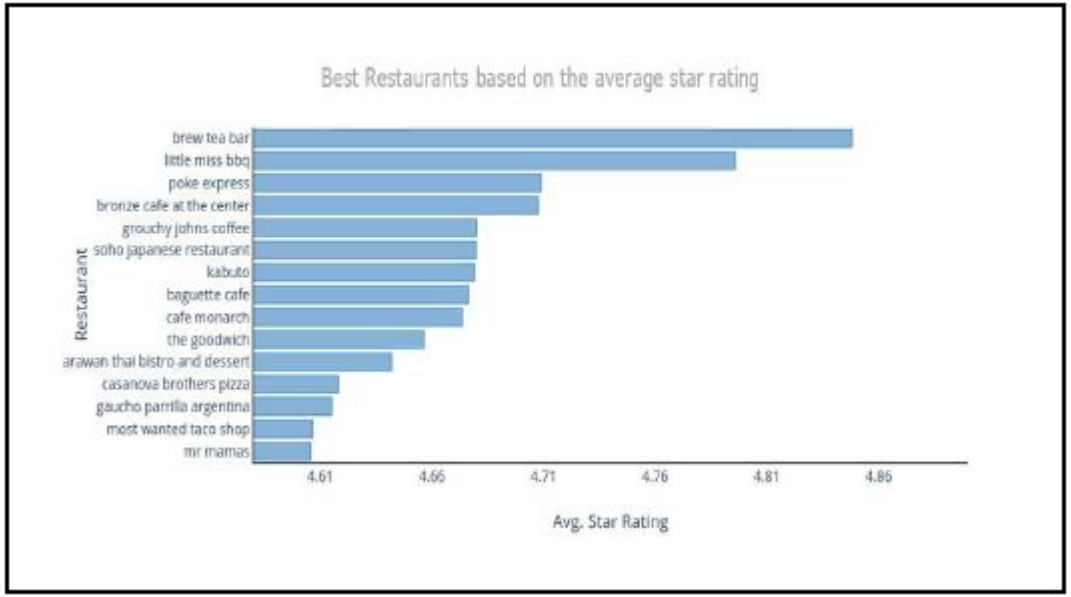
We can see the most famous categories. Fast food restaurants, followed by sandwich restaurants, then American restaurants, are the most prevalent restaurants.

Distribution of ratings

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 2–4 stars represents neutral.



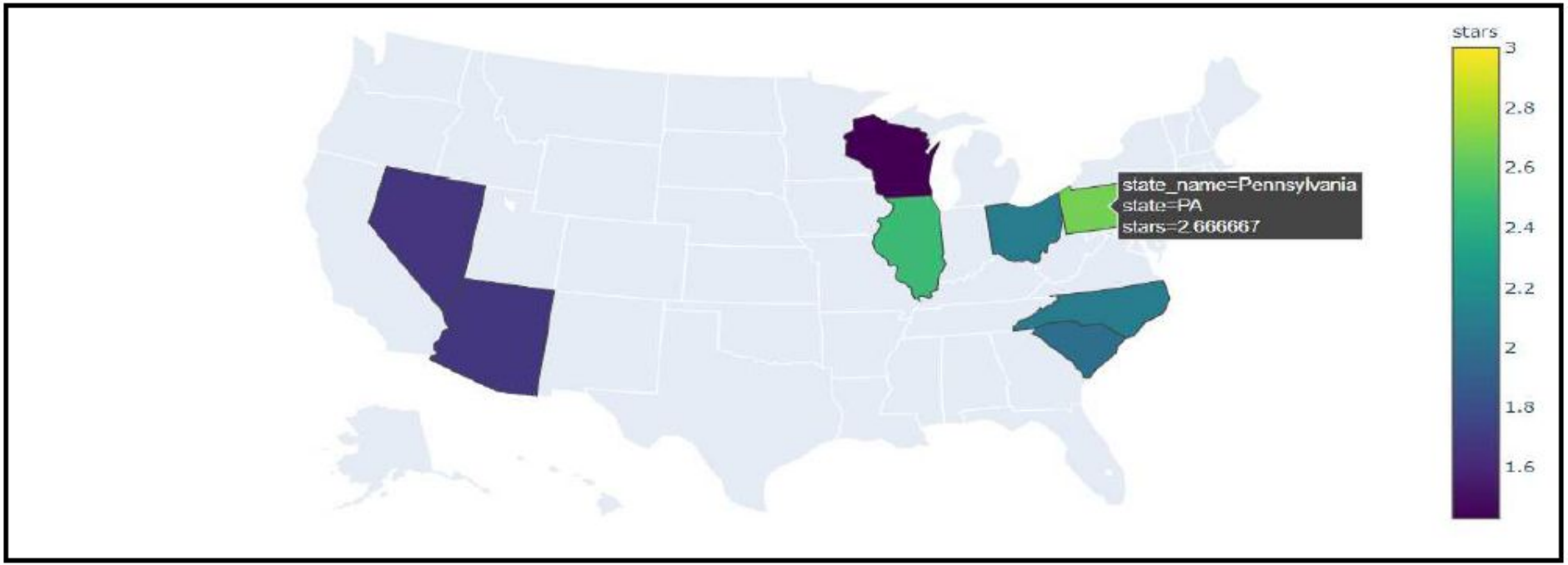
Below figure represents the best restaurants based on their average star rating. Thus we can see how EDA helped us to narrow down the whole dataset to get the best restaurants.



Observation

“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.

Case Study- KFC

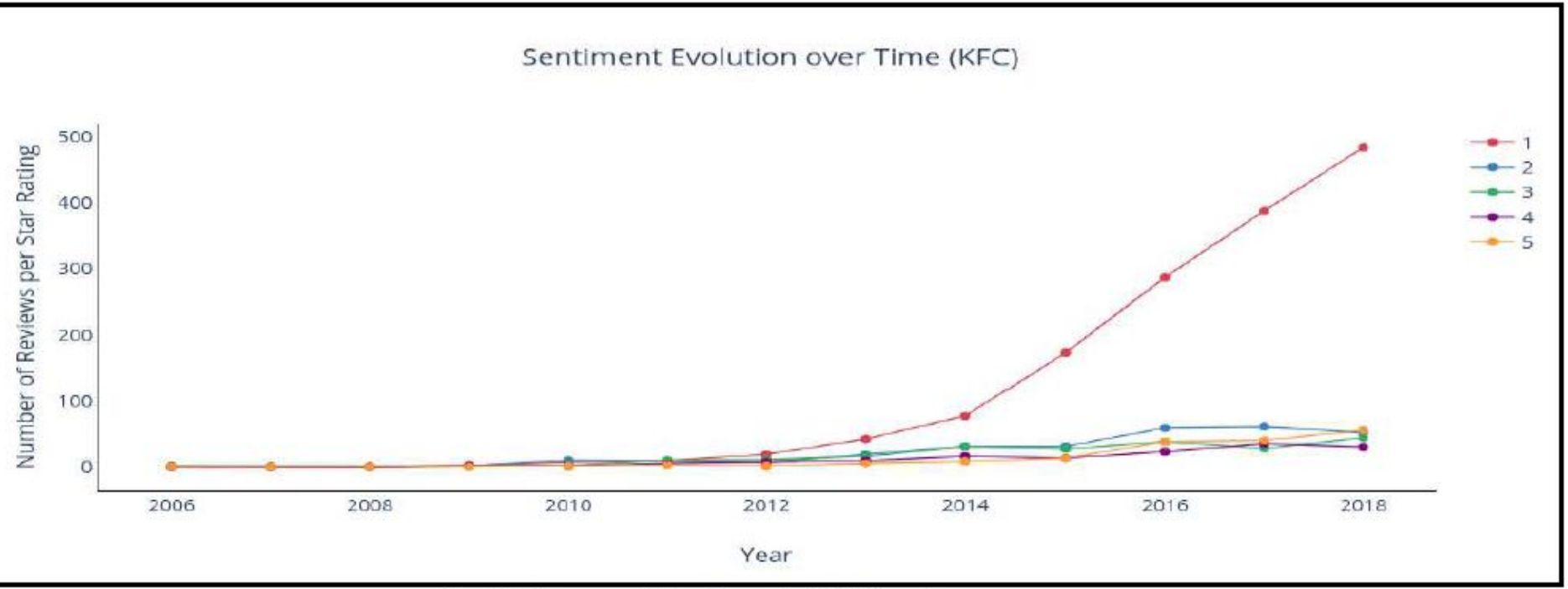


Observation

To deep dive into branch wise, the authors have analyzed the data based on location wise in USA for KFC. Figure 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.

Temporal Analysis

Temporal Analysis also adds value in understanding how the customers were satisfied 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.



Feature Extraction

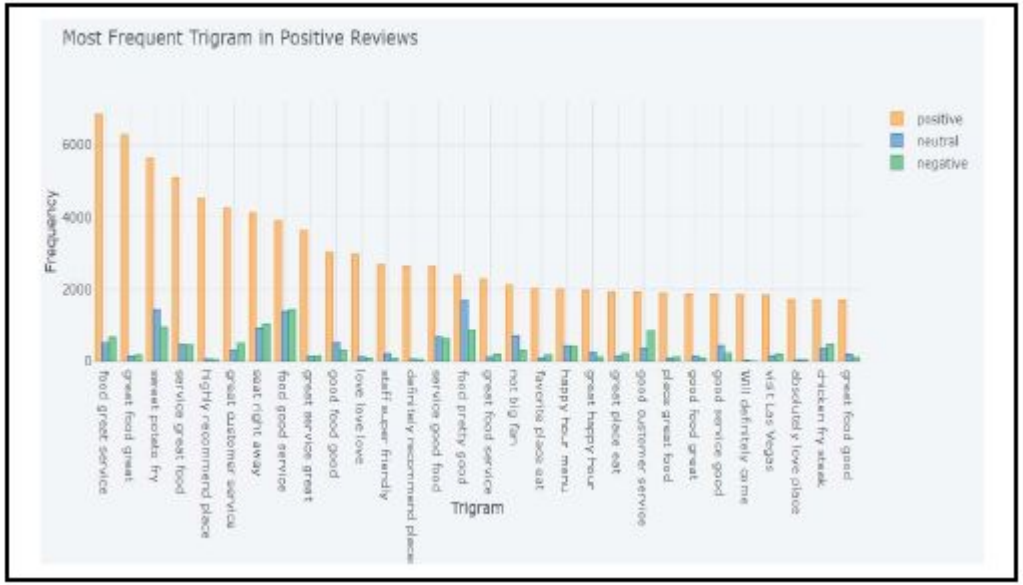
To find about the most used words in positive, negative and neutral reviews, the authors has used Bag-of-words(BOW) and Count Vectorizer. First the data is preprocessed by tokenization and lemmatization.

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.

Results

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 great service” in positive reviews, “good not great” in neutral reviews, and “bad customer service” in negative reviews .

Figure shows the most frequent trigrams in positive reviews



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.
- 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.

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.