Richa Meherwal

Netid: meherwa2

CS 410 Fall 2020

Technology Review - Latest Applications of Text Mining in evaluating customer satisfaction in various industry sectors

Introduction

In todays world, extracting value out of online reviews to analyse various aspects of customer satisfaction is deemed necessary. With the advent of online shopping and rise fo e-commerce sectors, it is now a customer right to know and understand the service/product before investing in it. Although overall rating is a quick way to help a customer figure out whether a service or product is something the customer would like to hire or buy, it is through the reviews and the detailed description in the online comments that the customer is influenced to buy or not buy a product. For example, when shopping on an e-commerce website, a customer clicks on the product displayed on the initial page by looking at the overall rating. On the next page, where the product description, reviews are provided in details, the customer quickly scrolls through the comments to see if the rating is actually a reflection of the reviews that are seen. It is like a second and final screening before buying a service/product. This is why text mining has diversified its applications to various sectors in the industry including airlines, hotels, healthcare etc. In this technology review, I will introduce a few of the application of text mining on analysing reviews in airlines and hotel industry. These two industries are a major portion of the travel industry, so we will actually be reviewing the latest research in harvesting value from the online reviews on travel industry.

Application of Text Mining in Hotels

One of the latest research has been identifying the importance of determinants of customer satisfaction in hotels using text mining.

Application of Text Mining in Airline Industry

A recent article[1] on airline customer satisfaction text mining emphasises that relevant features can be accurately extracted from OCRs as the key to identify the competitive edge that is needed for airline companies to excel. The analysis is done using LDA, a popular topic extraction technique in Text Mining on a UGC dataset. The dimensions discovered not only give an overall insight into what areas customer satisfaction is seen negatively and positive but it gives a layered analysis on what topics carry a sentiment in certain types of passenger based on location, time etc.

DATASET

This dataset comprises of 55,000 OCRs, covering over 400 airlines and passengers from 170 countries. It is collected from Air Travel Review (ATR), a widely used website for review collection for airlines industry. It has over 681 airlines and 725 airports worldwide. The final dataset had a publication date, review and answer in Yes/No to a question - "Would you recommend this airline". Some reviews consisted of contextual information such as nationality , cabin flown.

PRE-PROCESSING

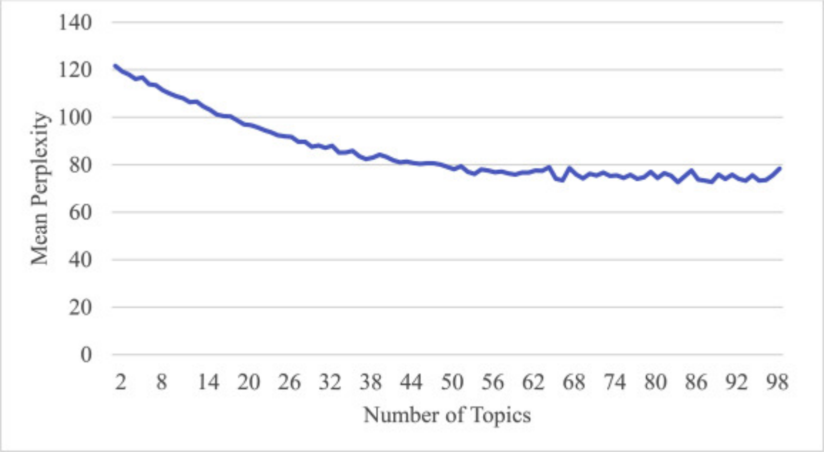
A web crawler was developed to crawl the ATR website and extract all the reviews in the html code. Challenges that were faced included 1) dealing with casual language and grammar 2) pulling out only those dimensions that give valuable information 3) to map text to number using encoding techniques 4) to deal with different languages. Using entity recognition airport names, prices, dates, flight codes, airline names were identified and replaced with a standard term starting with '_' , so as to separate satisfaction keywords from context.

TOOLS USED

Python 3.6, NLTK, sklearn

TOPIC EXTRACTION PROCESS

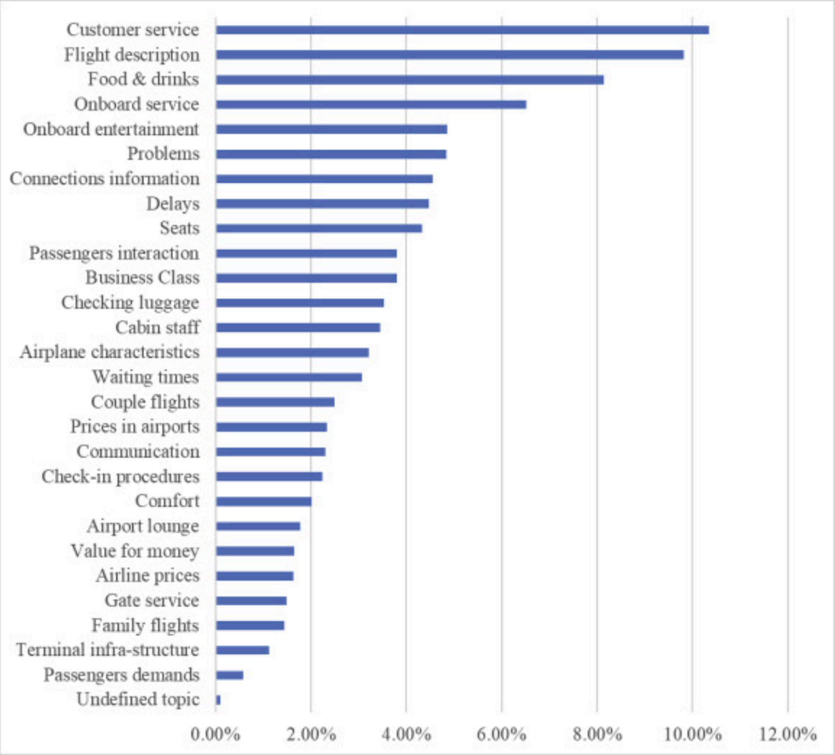
POS tagging was applied and only nouns and adjectives were kept. Nouns were used to help with identifying the topics for customer satisfaction whereas adjectives were used for sentiment analysis. A matrix was created where rows were docs/reviews and columns were TF-IDF for each term. A Latent Dirichlet Allocation model was used. In LDA, each topic can be shown through a word distribution. The challenge here was that number of topics was not known, setting less number of topics could miss out on information and more number of topics could make the model more complex. To combat this, a perplexity measure was used to identify the appropriate number of topics where lower the perplexity measure, better is the model.



Once LDA was fitted and number of topics was determined, then the word to topic likelihood was used to determine the top ranked words that are relevant to the topic. A consensus process is followed till researchers establish which words correctly describe a given topic.

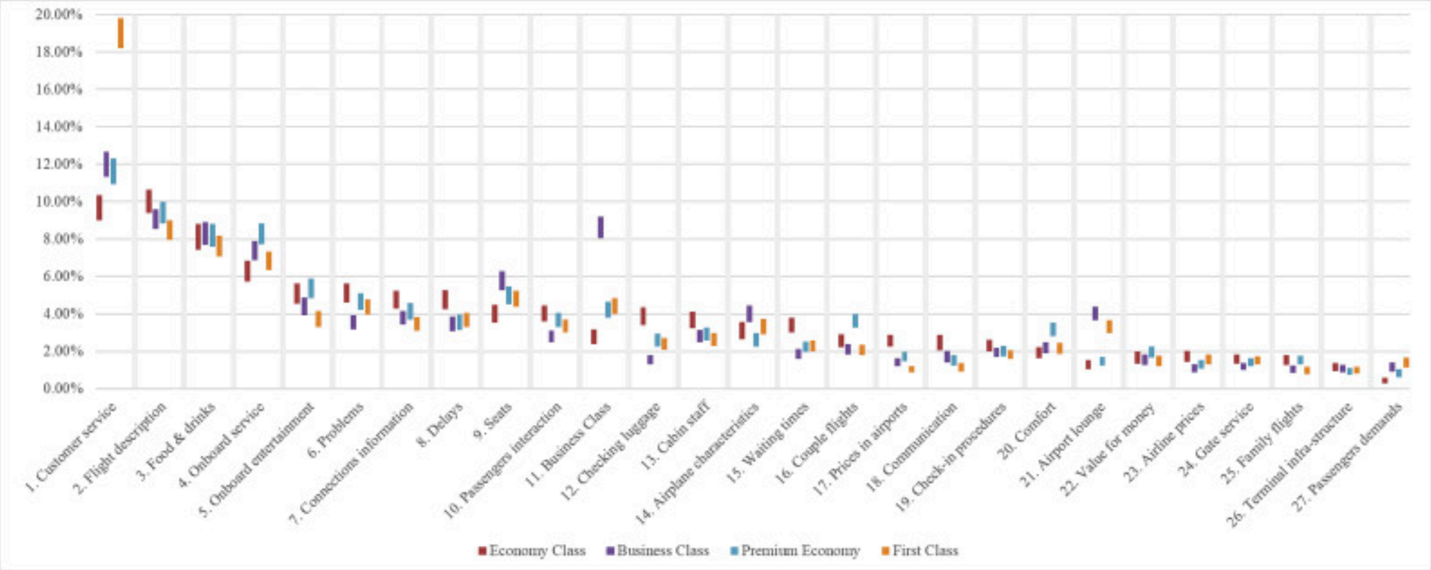
METHOD OF ANALYSIS AND ACCURACY

Distribution of topics across all reviews is noted. Then, reviews are grouped by criteria such as airline, time, passenger nationality, type of passenger, and type of cabin flown and then distribution of topics is noted for each group. For sentiment analysis, Naive Bayes Classifier is used. Labeled dataset of negative reviews, those that were rated 0 and positive reviews , those that were rated 10 are fed not the classifier. The dataset is then fitted and sentiment strength coefficients was assigned to each adjective. This provides the list of important adjectives and the sentiment scores associated with them. Similar meaning words (882 adjectives) were then grouped together and the final topic word list was now with 27 dimensions of satisfaction.



Another way the data was analysed was using confidence intervals. Dimensions are laid down like shown against the percentage CI.

Each of these CI graphs were built for each group and dimensions. It was easy to infer what group of people cared about which dimensions and also compare them with other groups. For example First class care a lot more about customer service then other cabin flown types. Similarly, business class care a lot more about airport lounge than economy and premium economy.



Dimensions and adjectives were then modelled with logistic regression to predict the recommendation with an accuracy of 79.95%.

References

https://www.sciencedirect.com/science/article/pii/S0278431916300202

References

https://www.sciencedirect.com/science/article/pii/S0278431916300202