National Research University Higher School of Economics, St. Petersburg

“Text Mining And Firms Behaviour”

by

Nadezhda Denisova, Daria Kovaleva, Alexandra Lyapina

A COURSE PAPER

SUBMITTED TO THE FACULTY OF MANAGEMENT

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE

DEGREE OF BACHELORS OF SCIENCE

Academic Advisor: Dr. Jeff Downing

St. Petersburg, Russia

June 2019

© NADEZHDA DENISOVA, DARIA KOVALEVA, ALEXANDRA LYAPINA 2019

Abstract

Over the recent years the field of Natural Language Processing has been rapidly developing as well as widely used in business. Analysing User Generated Data in the form of online reviews has allowed researchers to uncover managerial insights on topics ranging from identifying issues of a product to predicting films’ box office success. This paper is focused on analysing customer reviews on airlines in economy and business class sector. The reviews analysed in this research were posted on TripAdvisor over the course of 2018-2019. The research questions it sets are what features are discussed in reviews on economy class cabins as opposed to business class and if some of those features tend to be evaluated by customers as mostly positive or negative. The total of about 5,000 reviews on economy and same number of reviews on business class is collected and analysed via Latent Dirichlet Allocation topic modelling algorithm. Next, Pearson coefficient correlation is computed for the identified topics and review ratings. The first hypothesis stating that main features discussed in reviews on the two different cabin types tend to differ is confirmed with the Jaccard overlap coefficient for the topics being about 41%. The second hypothesis stating that there are topic and rating correlations for both cabin types is confirmed partially with economy class reviews topics having some moderate positive and negative correlations with the ratings and business class reviews having very little to no correlation with the ratings. Since the research is focused on one particular user feedback platform, only analyses the reviews on trips taken over 2018-2019 and processes all reviews with no regard to companies’ geographic location, some ideas for future research include cross-platform analysis, analysis of the shifts in consumer attitude over the course of several years and comparative research of customer feedback in different geographical regions.

*Keywords:* Natural Language Processing, Text Mining, Airline Reviews, Latent Dirichlet Allocation, Topic Extraction

CHAPTER 1: INTRODUCTION

Over the recent years big data have been widely used in business giving birth to the term business intelligence. Previous studies prove the effectiveness, relatively low cost and the reduced bias brought by big data that is used in business to gain competitive advantage, raise brand awareness and retain target audience (Yang et al., 2015).

The subject of this study - Natural Language Processing (NLP) - as a subfield of artificial intelligence has been rapidly developing over the last few years and has taken leading positions in data analytics. The concept includes text mining - finding useful insights from big data in textual form - and opinions and sentiment analysis through which attitudes and overall mood of the data is derived. All that is achieved via methods such as tokenisation, lemmatisation, parts of speech (POS) tagging and Named Entity Recognition (NER) and calculating term frequency - inverse document frequency (tf-idf) (Fernández-Gavilanes et al., 2019).

As the Internet has become one of the leading platforms for promotion it is crucial to understand needs of the target audience (Cambria et al.,2013). Product reviews are extremely important as they offer valuable information that helps product designers to understand the needs and preferences of consumers much better, and moreover they can also influence potential customers’ choice (Liu et al., 2013). Customers attitudes towards a product play vital role in its design and promotion, that is why analysing opinions is one of the key priorities of the companies. As literature claims, the more popular the brand is the more information about it can be found on the Internet (Moen et al., 2017). For that reasons this topic of research is considered to be actual and interesting for further studying. This research is focused on applying opinion mining and sentiment analysis to online user reviews in tourism sphere. Some companies such as TripAdvisor and AirBnb have built their business models on these reviews and became successful in the field. However, this course paper aims namely to study applications of reviews analysis and its effect on companies marketing strategy in Airlines business process as we found that there are much less research on this topic and not enough findings by previous researches. This gap of study should be filled as it has a big significance for full understanding of NLP applications to tourism sphere. Moreover, as the prior research is focused particularly on the factors affecting satisfaction and dissatisfaction levels of customers, this study is going also to embrace the concept of overall sentiment of comments and take into account such characteristics as people age, location and gender.

The **research questions** are as follows:

1. What are the main features of airline service that draws reviewers’ attention in economy as opposed to business class cabins?
2. Do some of those features tend to receive mostly positive customer feedback, while others - mostly negative?

Based on the research questions, the **hypotheses** are the following:

1. The main features discussed in the reviews on economy and business class cabins are different.
2. The features outlined in H1 correlate with the ratings.

The research goal is to analyse user-generated content by customers using NLP techniques and define the influence of customer opinions on companies operations in tourism sphere (namely airlines industry). The goal achievement requires the following tasks:

1. To study prior research on this topic for finding gaps and identifying trends in business intelligence applications focusing on tourism industry.
2. To develop a plan for studying including research questions defining and propose pertinent methodology for answering questions set.
3. To study NLP applications to tourism business, its trends and challenges for airlines industry.
4. To conduct the analysis using big data techniques and find out appropriate results.

This research has both descriptive and exploratory components - it focuses on description and exploration of the trends in customer attitudes towards airlines found in online reviews. The target audience of research is airlines - the aim of research is to analyse their customers reviews and opinions and thus identify its effect on airlines marketing strategies. The study is conducted on an industry level.

The structure of the research is the following: firstly, in literature review the basic overview of the applications of big data in business and particularly in travel industry as well as the concept of natural language processing (NLP) and text mining are given, based on this overview existed gaps in previous studies are identified. Next, there is the description of the research methodology and process of the collected sample of user-generated textual data in the form of reviews in order to answer to our research questions. The data sample is created of customer reviews on airlines from TripAdvisor.com with the use of a web crawler and finally we summarise the results and make conclusions. In the last section there are the results and findings obtained during our analysis and studying, the practical value of our research for companies and applicability for further research and possible limitations are determined.

CHAPTER 2: LITERATURE REVIEW

*2.1 Big Data and Business Intelligence*

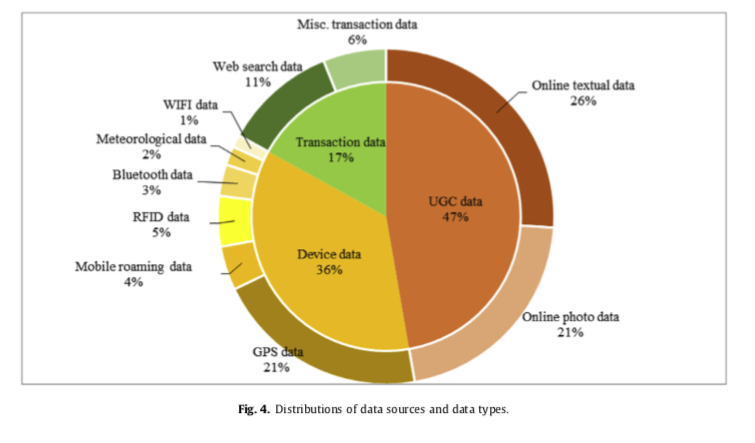
In 2001 current VP analyst of Gartner Inc. Douglas Laney in his research note characterised big data as having three dimensions - volume, variety, and velocity. Volume described the need to manage the constantly growing size of information collected and stored as it was increasingly perceived by companies as a ‘tangible asset’; variety emphasised the abundance of existing types and formats of data and the need to work on their compatibility; lastly, velocity was used to characterise the increasing speed at which data are generated and the need to be processed quickly to gain competitive advantage. Since first coined the definition has become ubiquitous and is constantly being extended by data analysts to include more V’s that better fit their work. (Douglas Laney, 2017)

With the spreading of big data such phenomenon as business intelligence and analytics has also significantly grown in popularity over the recent years. Plethora of researches have been using big data to prove that they can be effectively applied to various spheres to improve managerial decisions. What was before analysed through traditional means such as surveys and focus groups now is researched on a larger scale with the use of big data since it saves time and allows for lower costs as well as a significantly bigger sample set (Yang et al., 2015). Some of the current ways of using big data include identifying issues of a product (Rangu et al., 2017), predicting volumes of tourists (Yang et al., 2015) and box office success in film industry through online sentiment (Kim et al., 2018), and utilise big data in healthcare for decision support and prediction (Wang et al., 2018). Moreover, big data are currently highly used in social media management to conduct sentiment analysis, text mining and Natural Language Processing, predictive and content analysis (Misirlis and Vlachopoulou, 2018).

Big data is also applied to analyse external factors influencing businesses such as global events, users’ sense of belonging to them, their behaviour and activity on social media accounts, engagement of participants in the events and combining reviews. Big data analysis allows to reveal people’s incentives to actively consume social networks during global events, such as entertaining, responses to public actions and experiencing sense of belonging to a big community. As for the latter, sentiment analysis is used with geolocation techniques (as Twitter provides feature to mark location and some tweets get geotagged automatically) that define the feeling of group belonging of users and help to achieve more accurate results. However, researchers remark that such approach is novel and quite challenging (Fernández-Gavilanes et al., 2019).

*2.2 Big Data in Tourism*

“Data lies at the core of all smart tourism activities.” (Gretzel et al., 2015a, 2015b). According to an extensive literature review conducted by Li et al. (2018) there exist two major types of big data used for research in travel industry - user generated content (UGC) in the form of reviews, comments, tweets, pictures, videos, etc. posted online, device data (the most prominent example being GPS data) and operational data that include website visits, bookings, payment transactions, etc (see Fig.1).



*Figure 1*

*source: Li et al., 2018*

Li et al. (2018) explain such distribution by different levels of availability of different big data types. Transaction and device data are often private property of travel organisations and government while UGC data are mostly public and analysing them has gained popularity due to its relatively low cost (Kim et al., 2016, Li et al., 2018). Cambria et al. (2013) remarked that “the purpose of visiting and using websites has changed from read-only to read-write”, which provides for the abundance of openly available public opinion. This research will focus on UGC in the form of text reviews to explore and apply natural language processing techniques.

For example, Punel and Ermagun (2018) used UGC analysis to identify market segments in airline industry through clustering and text mining. They collected public data of 1934 Twitter users following Air New Zealand account that included profile descriptions and locations as well as the comments they post on the airline’s page. Discovered clusters provide an opportunity for the company to tailor its services for particular segments of its customers as well as see the areas where it is over- or under-represented (Punel and Ermagun, 2018).

*2.3 Text Mining*

Text mining is applied in business for reasons aimed at quality product improvement. As the number of customer reviews is growing rapidly, applying text mining allows companies to identify any issues of a constant product to understand customer needs and make decisions on enhancing products quality and raise brand awareness (Rangu et al., 2017).

In hospitality and tourism text mining has been around only for a short while. According to Li et al. (2018), the first publications on the topic in academic databases such as Web of Science and Scopus appeared around 2007 indicating the start of NLP usage in tourism. For example Pan et al. (2007) applied positive and negative word count in user reviews to infer service level provided.

Over time researchers started applying more sophisticated text mining techniques to derive useful managerial insights. Kim et al. dug deeper beyond calculating the general sentiment scores to discover the premises of customers’ attitudes towards hospitality services through co-occurrence analysis. With that tool they found out that the major reasons of travellers discontent with transportation in Paris stem from high metro and taxi pricing as well as density and dirt on the buses.

The process of deriving useful insights from user-generated textual data starts with developing a web-crawler in a programming language such as Python (Guo et al., 2017; Kim et al., 2016) that automatically scrapes data off a number of selected websites. At the next step, described as text mining, the data are cleaned and pre-processed via tokenisation (breaking text into single word-tokens), eliminating misspellings and stop-words - for example, articles and prepositions that are of no actual meaning, - then words are lemmatised (their lemma is derived to group them by their meaning), parts of speech (POS) and named entities are tagged and then various machine learning algorithms are applied to achieve particular research goals (Li et al., 2018).

According to Li et al. (2018), algorithms currently most used by researchers in user-generated textual data include Latent Dirichlet allocation (LDA) (Guo et al., 2017), sentiment analysis (Hu et al., 2017), statistical analysis (Racherla and Friske, 2012), clustering (Punel and Ermagun, 2018), text summarisation (Hu et al., 2017) and dependency modelling (Xie et al., 2014). For example, Cuicui Chen et al. (2016) combined Named Entity Recognition (NER) and clustering to identify the most and least prominent brands and analyse brand leadership, brand similarity, and their marketing performance on Instagram (Cuicui Chen et al., 2016). Hu et al. (2017) used text summarisation to fill the gaps in the previous research, that had not covered conflicting online opinions and credibility of those who post them, by focusing on the most informative sentences that summarise a review via assigning importance score based on their length, position and word content (Hu et al., 2017).

Liu and Law (2015) identified the types of insights researchers are most focused on uncovering from textual data. Opinion mining is currently one of the key topics along with the role and effect of online reviews, customer satisfaction and management and motivation. While online reviews and satisfaction are the most researched fields, opinion mining shares the second position with the remaining spheres.

Hotels capture the most attention of researchers, then come travel/tour and restaurant with the same number of research articles published on each. There also are several articles included in ‘other’ group.

*2.4 Relevance of This Research*

This coursework aims to fill some of the gaps identified by the previous researchers of user-generated textual data in travel industry.

First, past research is mostly focused on the features that influence satisfaction and dissatisfaction levels, which are expressed in the form of nouns and adjectives and adverbs associated with them, while the overall sentiment of a review was ignored. Besides focusing on n-grams that serve the aforementioned purpose we would like to also take into consideration the overall sentiment of a comment and based on that try to distinguish between sarcastic and sincere attitude with which the same words may be used. Only counting term frequencies may result in biased findings due to omitting the sentiment. We also aim to include user information such as age, location, and gender published on their profiles as proposed by Yang et al., (2015).

While there is plethora of research conducted in the hotel industry, there are much less researched niches located under the tourism umbrella, such as airlines. That is why we would like to focus on user’s reviews of various airlines to derive the main factors influencing their satisfaction as well as identify the main target segments among the customers.

CHAPTER 3: METHODOLOGY

*3.1 Population, sample size, units and sampling method*

The population of the research is the total number of airline reviews in English on TripAdvisor. The choice of the platform is based on the fact that while airline reviews on other platforms such as Skytrax, Facebook and Twitter have been widely researched in the past (Hong and Park, 2019; Lacic et al., 2016; Yao et al., 2015; Kaur and Balarkishnan, 2018; Liau and Tan, 2014) there is currently little to no research conducted on TripAdvisor data (posted on Web of Science). Therefore, this research aims to discover whether there are any platform specific features that have not been identified on the aforementioned three.

A sample unit is a single online review left by a consumer. Using sentiment analysis of the review texts it is possible to explain and forecast traveler satisfaction (Lacic et al., 2016). Therefore, these units will be used to extract and evaluate key features influencing customer satisfaction.

The total number of airline reviews in English on TripAdvisor at the time of the research is about 885,000. The number of times an airline has been reviewed ranges from one to 45,000 with 257 out of 575 companies having less than 100 reviews. To derive the most up-to-date insights in the industry, only the reviews posted within the last year (365 days) will be extracted which in total is about 100,000 units. However, due to the limitations of computational power available for the research, only a portion of the total number of reviews in English will be selected with random sampling.

Taking 100,000 as the overall population, to provide for results accurate up to ±5% with 95% probability the required sample size is about 383. However, to get a broader view on the issue and have more data to enhance the accuracy of the model the proposed sampling method is randomly selecting about 5,000 reviews on economy class cabins and the same number on business class to make it possible to generalise the findings for the whole airline industry at the aforementioned confidence level.

*3.2 Review extraction and analysis*

First, to extract the identified number of reviews, a simple web scraper written in Python (using *requests* and *BeautifulSoup* libraries) will be used. The data will have a table format with 4 columns, one for each feature specific to TripAdvisor: review title, review text, date of travel and rating.

Next, each review will undergo tokenisation, lowercasing and non-english words and characters elimination, then the words appearing in less than 10 reviews will be removed.

Sample comment before and after the cleaning:

*"The two of us had a total of 35kg in two bags, one large and one small. The small bag, full of the heavy stuff, was 10kg. But the other bag at 25kg, cost us $120. Extremely ordinary when compared to any other airline we have flown on and very money hungry. Service was otherwise no better than any other airline in the region we have used. Best advice is, if you don't wish to get skimmed of your money and you have a choice between Air New Zealand and any other airline, take the other airline."*

*“total bag larg small small bag heavi stuff bag cost extrem ordinari compar airlin fli money hungri servic better airlin region best advic wish skim money choic air new zealand airlin airlin” .*

To check H1 about the difference in main topics discussed in reviews on economy and business class cabins, unsupervised machine learning algorithm Latent Dirichlet Allocation will be used as it was proven to reveal meaningful dimensions not found by other traditional machine learning algorithms (Guo et al., 2017). The model was introduced by Blei, Ng and Jordan in 2003 and since then has been widely used in topic extraction in various research domains (Blei et al.,2003)

LDA is based on the assumption that each document can be represented as a probability distribution of topics and each topic - as a probability distribution of words associated with it. The latter makes the topics human-interpretable while the fact that both topic distributions across each document and word distributions across each topic can be summed up to unity means that they can be viewed as percentage weights which also aids the interpretation process (Blei et al., 2003). The algorithm is the following:

1. Each word in each document is randomly assigned to one of the *k* topics
2. For each document *d* it is assumed that its topics are assigned incorrectly while in all the other documents the assignment is correct
3. The proportion of words from each topic in the document *d* is calculated
4. The proportion of each topic *t* assignments throughout the whole corpus that come from each word *w* in the document is calculated
5. The two proportions are multiplied and are treated as probability based on which a new topic is assigned to the word
6. The process is repeated iteratively until a steady state is reached.

The input to the model is a bag-of-word matrix in which rows represent the documents (reviews) and columns represent the words that appear throughout the whole corpus (all reviews combined). At the intersection of each row and column is an integer value representing the number of time a word (i.e. the root of a word in any form since the documents are lemmatised and stemmed) occurs in the particular document.

The main hyperparameter required to build the model is the number of topics. In order to devise the best fitting number the ‘elbow’ method will be applied. The bag-of-words vectors representing words occurrences in the reviews will be clustered by k-means algorithm, the number of clusters ranging from 2 to 30. Then the sum of squared deviation of each vector from the center of its cluster will be computed and plotted on a graph. The optimum number of clusters is the point where the steep reduction of the squared error changes into moderate creating an ‘elbow’ angle. (Blei et al., 2003).

Next, one of the authors will assign names to the discovered topics and the other two will confirm or alter them when necessary. The main topics will be extracted from economy and business class reviews separately and then to answer the RQ1 the Jaccard coefficient will be calculated to obtain the magnitude of topic overlap:

,

where E is the topics discussed in economy class cabin reviews and B is business class cabin reviews topics.

After the main features customers tend to express their opinions on are discovered, this paper aims to test the H2 about the correlation between the topics and the ratings in both sectors. To achieve this goal, Pearson correlation coefficient will be used. Each topic is represented as a percentage value of its coverage in each of the reviews, which is a numeric variable. Each rating is an integer value ranging from 1 to 5. To test the correlation, the covariance of the two variables is divided by the product of their standard deviations:

Next, to find out more information on satisfaction and dissatisfaction factors in both cabin classes, word co-occurrence analysis will be conducted. First, the reviews will be grouped according to their rating (4-5 stars representing a generally positive experience, 1-3 stars - poor experience) which works as a naïve representation of the overall review sentiment. Next, after eliminating the stop-words and the ones occurring less than 10 times in the corpus, the reviews will be split into bigrams, i.e. pairs of two that are used together. The output we aim to produce is a list of top-15 most frequently used word pairs in negative and positive comments in both industry sectors.

The python libraries to be used for the analysis are *sklearn*, *gensim* and Stanford’s *nltk*.

*3.3 Validity and reliability of the work*

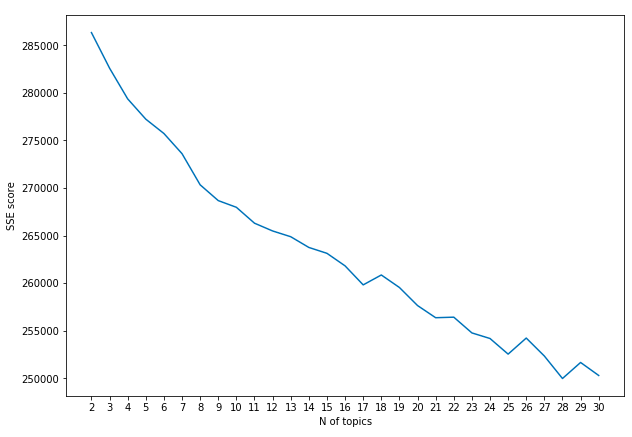
There are several threats for the validity of the research. Firstly, there is a possibility that some airlines companies stimulate people to leave negative reviews about their competitors in order to maintain their reputation and image as well as purchase positive reviews for their company. For this reason, some reviews may not reflect real opinions of customers and valid information about airlines service and work, hence the reliability of reviews is questioned. In addition, there is a possible tendency among people to mostly express negative opinions and share their experience if they faced negative situations related to airlines service. Therefore, the number of negative comments may exceed the number of positive ones that can affect the validity of this study.

CHAPTER 4: ANALYSIS

*4.1 Economy class*

After the reviews are scraped off Tripadvisor, the total number of economy and business class reviews are 64,811 and 15,760 accordingly. The number of reviews on economy class used to train the LDA model that were randomly picked from the total is 5,676.

First, the main topics discussed in reviews on economy class cabins are discovered. The sum of squared errors scores for different number of topics are shown in Figure 2.

**

*Figure 2*

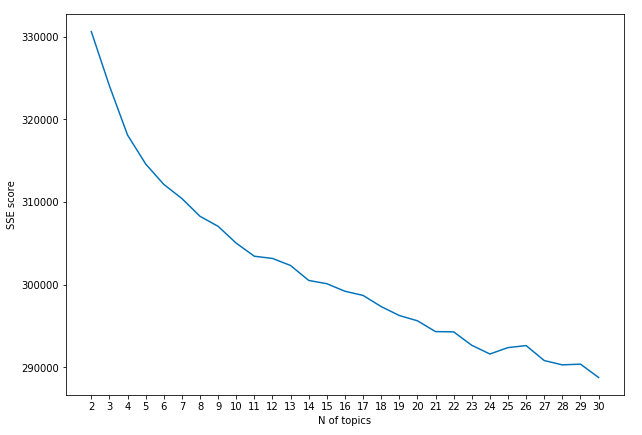
The pace of SSE reduction slows around nine topics, therefore *k* = 9 is chosen for the interpretation.

Table 1 shows the stemmed keywords for each topic arranged according to the descent of their significance to a topic as well as the researchers’ interpretations of the topics (in italics).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| TOPIC 1  *food and service* | TOPIC 2  *flight conditions* | TOPIC 3  *baggage and extra fees* | TOPIC 4  *assistance during the trip* | TOPIC 5  *delays* |
| good  airlin  food  meal  fli  servic  staff  friend  entertain  help | time  fli  seat  airlin  servic  good  great  plane  food  comfort | bag  luggag  pay  check  carri  baggag  charg  airlin  extra  fli | air  airplan  go  take  tell  host  late  good  minut  hot | hour  delay  time  airport  arriv  day  leav  fli  minut  wait |
| TOPIC 6  *crew service* | TOPIC 7  *seats* | TOPIC 8  *check-in and boarding* | TOPIC 9  *booking and cancellations* |  |
| airlin  fli  like  servic  passeng  plane  time  attend  experi  crew | seat  leg  room  sit  row  plane  hour  food  drink  board | checkin  board  pay  book  airport  time  ticket  staff  check  airlin | airlin  custom  servic  book  tell  cancel  day  say  ticket  refund |  |

*Table 1*

*4.2 Business class*

**

*Figure 3*

To train the model on business class reviews the total of 5,676 reviews is randomly selected. The sums of squared errors are higher than those of economy class clustering, *k = 8* being a possible ‘elbow’ on the graph (Fig. 3). The keywords for the topics are shown in Table 2.

|  |  |  |  |
| --- | --- | --- | --- |
| TOPIC 1  *lounge food* | TOPIC 2  *class conditions* | TOPIC 3  *booking and delays* | TOPIC 4  *positive flight experience* |
| meal  loung  serv  drink  busi  food  offer  choic  breakfast  class | class  busi  good  food  seat  comfort  loung  great  fli  economi | hour  tell  time  ticket  book  day  arriv  pay  delay  busi | fli  airlin  busi  class  air  best  new  great  travel  time |
| TOPIC 5  *cabin service and food* | TOPIC 6  *check-in and boarding* | TOPIC 7  *general experience* | TOPIC 8  *seats* |
| good  food  crew  time  great  comfort  staff  excel  friend  cabin | board  checkin  staff  airport  passeng  time  airlin  check  gate  busi | airlin  fli  busi  class  year  food  time  good  experi  thai | seat  class  busi  food  flat  sleep  like  passeng  bed  plane |

*Table 2*

*4.3 Topic overlap*

|  |  |  |
| --- | --- | --- |
| Topic | Economy | Business class |
| **food and service** | +  *0.304* | +  *0.048* |
| **flight conditions** | +  *0.399* | +  *0.023* |
| baggage and extra fees | +  *-0.151* |  |
| assistance during the trip | +  *-0.14* |  |
| delays | +  *-0.155* |  |
| crew service | +  *-0.009* |  |
| **seats** | +  *-0.02* | +  *-0.158* |
| **check-in and boarding** | +  *-0.041* | +  *-0.012* |
| **booking and cancellations** | +  *-0.502* | +  *0.04* |
| lounge food |  | +  *-0.014* |
| positive experience |  | +  *0.074* |
| general experience |  | *+*  *-0.036* |

*Table 3*

Two similar topics for business class were combined to form “general experience”. Jaccard coefficient = 41.7% which means that topics discussed in the reviews on economy and business class cabins tend to differ in more than half of the cases. In the Table 3 it is shown that topics like extra fees, baggage issues and delays only concern economy class customers. Therefore, the hypothesis about the difference in key topics between the two cabin types is confirmed on the established significance level.

*4.4 Topics and ratings correlation*

The results of the correlation coefficient calculations for economy and business class reviews are shown in tables 5 and 6 respectively. Overall, while in business class sector the discovered topics have very weak correlations with ratings (the strongest correlation being -15.8% for topic 8 - seats), they are noticeably stronger in economy sector. Highlighted in the table are the results representing moderate positive or negative correlations. Thus, topic 4 (booking and cancellations) tends to appear in negative reviews, while topics 1 (food) and 2 (flight conditions) is associated with positive customer feedback. Weak negative correlations are shown by the topics 3, 4 and 5 (baggage and extra fees, help, delays).

|  |  |
| --- | --- |
| *Topic*  *(economy)* |  |
| *1* | *0.304* |
| *2* | *0.399* |
| *3* | *-0.151* |
| *4* | *-0.14* |
| *5* | *-0.155* |
| *6* | *-0.009* |
| *7* | *-0.02* |
| *8* | *-0.041* |
| *9* | *-0.502* |

*Table 4*

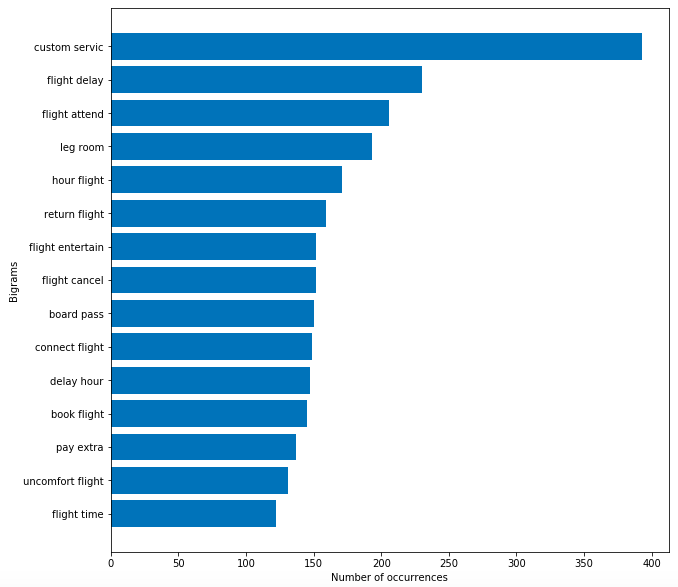
|  |  |
| --- | --- |
| *Topic*  *(business)* |  |
| *1* | *-0.014* |
| *2* | *0.023* |
| *3* | *0.04* |
| *4* | *0.074* |
| *5* | *0.048* |
| *6* | *-0.012* |
| *7* | *-0.036* |
| *8* | *-0.158* |

*Table 5*

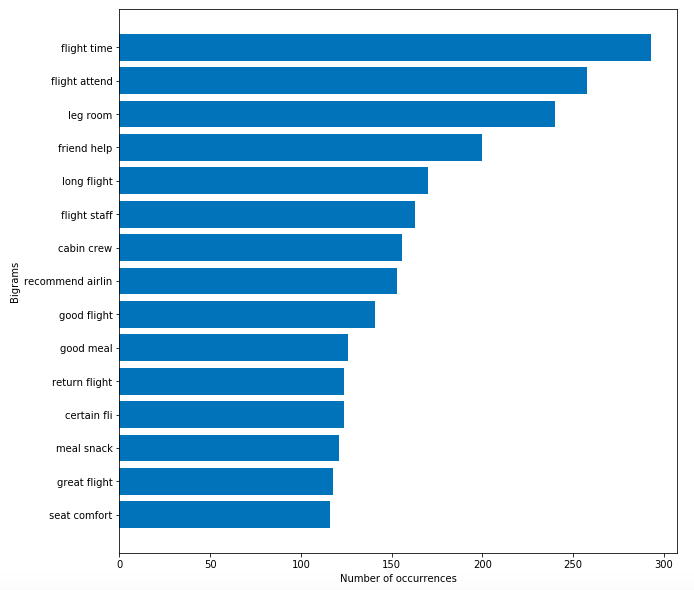
*4.5 Co-occurrence analysis*

Figures 4 and 5 show the results of bigram analysis performed on negative and positive economy class cabin reviews respectively. Customer service is the aspect causing the most significant amount of dissatisfaction among the customers. A further co-occurrence analysis revealed that customer service in negative comments is described as ‘worst’, ‘poor’ and ‘terrible’. The next most frequent bigram is ‘flight delay’, as was found out in the correlation analysis. Cancellations, booking and extra fees are also some the most frequently occurring phrases in negative reviews.

Flight time is the most frequently used bigram in positive reviews. Other possible factors of customer satisfaction include ‘good meal’, cabin crew and comfortable seats. Bigrams like ‘leg room’ and ‘flight attend’ are equally used in positive and negative comments which indicates them as important aspects of a flight without, however, giving them a particular sentiment.

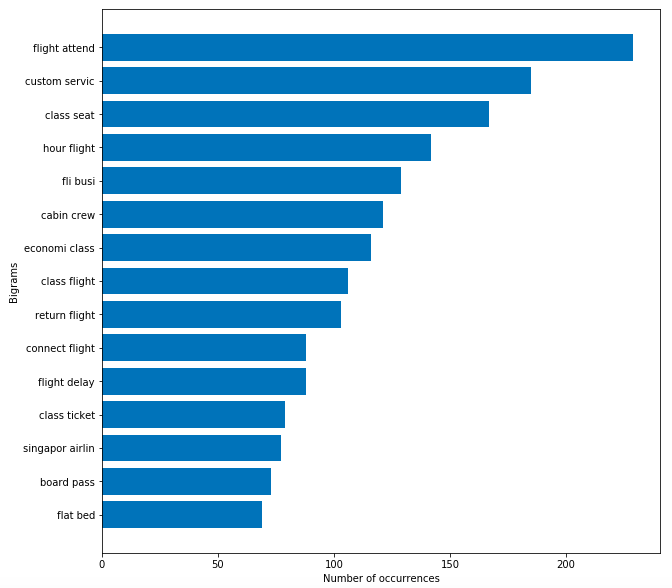


*Figure 4*

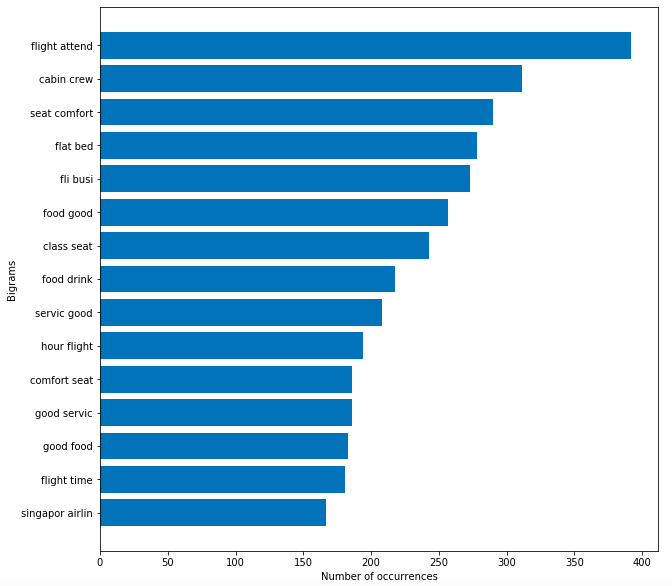


*Figure 5*

The most frequently used bigram in both negative and positive reviews on business class cabines was ‘business class’ (more than 750 mentions in both cases), therefore it was eliminated due to it being uninformative and skewing the graphs. ‘Flight attend’ being the next most frequently used bigram in both positive and negative comments (Fig. 6 and 7 accordingly) does not let us classify it into either satisfaction or dissatisfaction factors thus supporting the weak correlations between the topics and ratings (similar for ‘cabin crew’ and ‘flight bed’). However, customer service is frequently mentioned only in the negative comments, while food - only in the positive ones.



*Figure 6*



*Figure 7*

CHAPTER 5: DISCUSSION

Based on the conducted analysis, the main topics in reviews on economy and business class cabins were discovered. As for economy class cabins, 9 main topics reflecting customer opinions were revealed – food and service, flight conditions, baggage and extra fees, assistance during the trip, delays, crew service, seats, check-in and boarding, booking and cancellations, Business class reviews were divided into the following topics – lounge food, class conditions, booking and delays, positive flight experience, cabin service and food, check-in and boarding, general experience, seats. While most of the topic words revealed in economy class reviews were easy to summarise, topic words in the business class reviews were often ambiguous and repetitive (‘positive flight experience’ and ‘general experience’) which leads to a conclusion that the actual number of distinguishable issues discussed is smaller that the one defined by the ‘elbow’ method.

Jaccard coefficient that is used to identify similarity and difference of two sets is about 41% which means that topics of the reviews on economy and business class cabins tend to differ in more than half of the cases. Therefore, the Hypothesis 1 was confirmed and there is a significant difference between the features discussed in the reviews of economy and business class cabins passengers. Moreover, comparing the topic names defined for the two classes we see that economy class customers tend to focus their attention on concrete aspects of service (that they are mostly dissatisfied with, as was revealed by the correlation analysis) while business class customers leave more general reviews about the overall flight experience.

Another identified difference between the two sets of topics is that customers of business class cabins tend to express their opinion not only on regular meals provided on the board but also on lounge food issues. It might be assumed that besides the condition on board, a wide choice of menu, high quality dishes and comfortable eating conditions in the lounge are ones of the most important criteria for choosing a particular airline. Therefore, lounge food should be one of the main focuses of business class service providers. In economy class cabins there is no “lounge food” words identified and passengers pay attention only to the food service provided on the board. Such key topics as baggage and extra fees, assistance during the trip, delays, and crew service are found in economy class cabins and the correlation coefficients revealed that these topics tend to draw mostly negative attention. This might be related to the fact that economy class passengers of low-cost airlines tend to pay extra fees for their baggage or extra options that sometimes cause negative reactions, as well as that economy class customers attach importance to crew service and its help and timely informing in cases like delays or other issues.

Overall, topic and rating correlations are stronger in economy class sector compared to business class sector. Thus, among reviews by economy class passengers, “food and service” and “flight conditions” topics are marked by positive customer attitude. Moreover, the co-occurrence analysis revealed that ‘good meal’ is one of the most frequently used phrases in positive economy class reviews. Therefore, customers are mostly content with the service, meal and general flight conditions in economy class cabins. As for the negative reviews, such topics as “baggage and extra fees”, “help”, “delays” and “booking and cancellations” are associated with negative customers experience as these topics have stronger negative correlations. It is remarkable that “booking and cancellations” has the strongest negative correlation (-0.5) that might be caused by frequent complaints about customer service concerning booking issues and often flight cancellations. Thus, economy class brands should focus their attention on timely informing customers about the possible extra fees and assisting them in case of delays or cancellations. Other topics (“seats”, “crew service” and “check-in and boarding”) have little to no correlation with ratings.

Business class sector has weak correlations between the topics and ratings. The strongest negative correlation was for “seats” topic. Thus, while in economy class sector particular features tend to receive mostly positive or mostly negative feedback, in business class sector there are very few peculiarities correlated with strong satisfaction or dissatisfaction factors.

CHAPTER 6: CONCLUSION

According to the work done, it can be said that using big data can reveal meaningful insights into a company’s service. In order to meet customers’ needs companies can and should analyse reviews that their customers leave on the Internet, since customer attitude plays a great role in making work effective and efficient as it influences their demand that makes the whole company work.

Our hypotheses were the following:

1. The main features discussed in the reviews on economy and business class cabins are different.
2. The features outlined in H1 correlate with the ratings.

While doing the work the following analysis has been done. About 5,000 economy class and 5,000 business class reviews were collected via a simple web scraper written in Python. The reviews were tokenised, lowercased, lemmatised and stemmed while also eliminating non-english words and characters. The words appearing in less than 10 reviews were removed.

The first hypothesis was confirmed with Jaccard overlap coefficient being about 41%. The second hypothesis was only partially confirmed with some moderate correlations in economy sector and little to no correlations for business class.

There are several limitations to this research. Since only one platform was used the results might not be generalisable for the whole population of airline reviews on the Internet. Therefore, cross platform analysis of the reviews is one of the possible fields for the future research. Secondly, only the reviews on the trips taken within the last calendar year were analysed, while a similar analysis of reviews on trips taken over the several past years can uncover possible shifts in consumer attitude and help predict the key points that may draw consumer attention in the future. Lastly, since the reviews on airlines operating around the whole world were analysed in combination, an analysis of reviews based on the geographic factor may provide more insights for companies in particular regions.

References

Aymeric Punel, Alireza Ermagun (2018). Using Twitter network to detect market segments in the airline industry. *Journal of Air Transport Management, 73,* 67-76.

Chandrasekhar Rangu, Shuvojit Chatterjee, Srinivasa Rao Valluru (2017). Text Mining Approach for Product Quality. *Enhancement IEEE 7th International Advance Computing Conference.*

Christopher Moody (2016). Mixing Dirichlet Topic Models and Word Embeddings to Make lda2vec.

David M. Blei, Andrew Y. Ng, Michael I. Jordan (2003). *Latent Dirichlet Allocation*

David M. Blei, Andrew Y. Ng, Michael I. Jordan (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3,* 993-1022.

Douglas Laney (2017). Infonomics: *How to Monetize, Manage, and Measure Information as an Asset for Competitive Advantage.*

Erik Cambria, Björn Schuller, Yunqing Xia, Catherine Havasi (2013). New Avenues in Opinion Mining and Sentiment Analysis. *IEEE Intelligent Systems, 28,* 15-21.

Jingjing Li, Lizhi Xu, Ling Tang, Shouyang Wang, Ling Li (2018). Big data in tourism research: A literature review. *Tourism Management, 68,* 301-323.

Jorge Ale Chilet, Cuicui Chen, Yusan Lin (2016). Analyzing Social Media Marketing in the High-End Fashion Industry Using Named Entity Recognition. *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining.*

Kun Kim, Oun-joung Park, Seunghyun Yun, Haejung Yun (2016). What makes tourists feel negatively about tourism destinations? Application of hybrid text mining methodology to smart destination management. *Technological Forecasting and Social Change, 123,* 362-369.

Markus Schuckert, Xianwei Liu, Rob Law (2015). Hospitality And Tourism Online Reviews: Recent Trends And Future Directions. *Journal of Travel & Tourism Marketing, 32,* 608–621.

Michael Röder, Andreas Both, Alexander Hinneburg (2015). Exploring the Space of Topic Coherence Measures. *WSDM’15 Proceedings of the Eighth ACM International Conference on Web Search and Data Mining,* 399-408.

Milagros Fernández-Gavilanes, Jonathan Juncal-Martínez, Silvia García-Méndez, Enrique Costa-Montenegro, Francisco Javier González-Castaño (2019). Differentiating users by language and location estimation in sentiment analysis of informal text during major public events. *Expert Systems With Applications, 117,* 15-29

Nikolaos Misirlis, Maro Vlachopoulou (2018). Social media metrics and analytics in marketing – S3M: A mapping literature. *International Journal of Information Management, 38,* 270-276.

Øystein Moen, Lars Jaako Havro, Einar Bjering (2017). Online consumers reviews: Examining the moderating effects of product type and product popularity on the review impact on sales. *Cogent Business & Management, 4,* 1-20.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean (2013). Distributed Representations of Words and Phrases and their Compositionality.

Xin Yang, Bing Pan, James A. Evans, Benfu Lv (2015). Forecasting Chinese tourist volume with search engine data. *Tourism Management, 46,* 386-397.

Ya-Han Hu, Yen-Liang Chenb, Hui-Ling Choub (2017). Opinion mining from online hotel reviews – A text summarization approach. *Information Processing and Management, 53,* 436-449.

Yichuan Wang, LeeAnn Kung, Terry Anthony Byrd (2018). Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. *Technological Forecasting & Social Change, 126,* 3-13.

Ying Liu, Jian Jin, Ping Ji, Jenny A. Harding, Richard Y. K. Fung (2013). Identifying helpful online reviews: A product designer’s perspective. *Computer-Aided Design, 45,* 180-194.

Yoosin Kim, Mingon Kang, Seung Ryul Jeong (2018). Text Mining and Sentiment Analysis for Predicting Box Office Success. *KSII Transactions On Internet And Information Systems, 12, 8.*

Yue Guo, Stuart J. Barnes, Qiong Jia (2017). Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation. *Tourism Management, 59,* 467-483.

Appendix

Distribution of duties:

Nadezhda Denisova – 33,3 %

Daria Kovaleva – 33,3 %

Alexandra Lyapina – 33,3 %