PROJECT REPORT

Business IT project

Adhikari Nischal
Karttunen Marko
Fortes da Silva Joäo

Group:

Initially we started the project with 4 members, but we ended the project with 3 members. The group members are Nischal Adhikari, Marko Karttunen and João Fortes da Silva. During the project, Marko Karttunen was Data Analyst Specialist while João Fortes da Silva and Nischal Adhikari were Data Analysts. All team members come from different background and have different expertise, and each member have contributed their expertise to the project.

Marko Karttunen is a student of at Haaga-Helia University of Applied Sciences. He is interested in speech processing, speech recognition, natural language processing, text analytics, time series analysis and other machine learning techniques to uncover trends in large datasets.

João Fortes da Silva is a student of Industrial and Management Engineering at Polytechnic Institute of Porto. Despite his background, he is interested in Artificial Intelligence and has knowledge of python programming for machine learning techniques and data analysis.

Nischal Adhikari is a student of BITe at Haaga-Helia UAS with some knowledge of python programming. He is interested in Business Analytics, Artificial Intelligence, and hopes to learn more on using of python in business analytics.

Tools like Trello, Teams, GitHub, and email were used during the project. Trello was used as project management team, where team members were given specific roles based on their expertise and contribution. Teams was used as a primary tool for scheduling meetings and organizing meetings with team members. GitHub was used for the purpose of sharing code and keeping track of the progresses made on different stages of the project. Email was used for sharing links, contacting team-members, and sharing general information about the project progress.

An agile and iterative method was chosen for the project. During the project we had 3 sprint sessions. During sprint sessions we had goals that we wanted to achieve, we had team meetings and we set goals. After each sprint session we had discussion with the teacher and received feedback in order to improve our work. The third sprint session was the final one, after which we delivered the final output. Our project has been an iterative process, we improved after each meeting and we were able to produce a better output each time.

Topic and datasets:

The aim of our team project was to detect and to understand of public opinion on an airport service quality from the airport twitter data. The data covers a period of five years. We were interested in this topic for many reasons. First, people mostly communicate in digital channels such as social media. Another alternative would be to organize costly and time-consuming face-to-face interviews or to do research work at airports.

Secondly, we were wondering how do you measure efficiently the customer experience at the airport? Measuring customer experience has recently taken a center role in many work communities.

Why is it important to measure it? An excellent customer experience increases revenues. The core of customer experience is the interaction between the community and the customer. The customer experience includes emotions that arises from the customer when he or she encounters with community representatives, channels and services. The good customer experience means that the service meets expectations. The positive customer experience maintains customer loyalty and attracts new customers.

Thirdly, we were especially interested in the sentiment analysis which is one of the most popular natural language (NLP) processing applications when measuring social media. Companies and major brands want to listen very carefully what consumers say about their services and products online. The crucial factor is whether the feeling is positive or negative and how the trend develops over time. Previously, surveys were used as data collection methods. Now, as a channel, the social media provides a very good way to understand differences on the tone of consumers' emotions. For instance, feedback from customers provides a good way for product and service development.

Fourthly, the data was in an unstructured format which has traditionally been very challenging to explore. The data is ungrammatical and noisy so our task was to explore this raw data and produce valuable insights for airport service quality developers.

To put it briefly, the aim of this team project is to develop a sentiment analysis and a topic modeling methodology that can be used to discover and analyze public opinion concerning travelling in airports. We will also pay attention to changes over time. We used the VADER sentiment analysis model as our primary sentiment analysis tool.

A Vader sentiment tool is based on open source code, which is designed to analyze the tone and overall intensity of social media materials. Its strength is that teaching and test data may not be needed. It can easily understand the sentiment and its code is easy to absorb and to use. Its strength is heuristic rules when, for example, exclamation marks in tweets increase the sentiment score. The challenge, of course, is that sarcasm may not be noticed, and the same sentence may contain both positive and negative emotions.

In addition to that, the goal is to do first an exploratory data analysis and then to clean the tweets when entering the next phase, the sentiment analysis and a topic modeling in order to produce valuable insights. Moreover, our team was interested in comparing the results of two sentiment analysis tools, VADER and TextBlob. Lastly, the goal is to count which words most often occur together in tweets.

The sentiment analysis was based on tweets airport customers have sent regarding airports where they were traveling. Five airports were chosen: Dublin Airport, Heathrow Airport, Gatwick Airport, London Stansted Airport and Manchester Airport. We chose these airports as they seemed very interesting because the high amount of tweets and they are from English speaking region.

The dataset used for analysis was found from moodle. The datasets contained three excel files, one PowerPoint file and a word file. The excel file ‘Data After pre-processing’ was used for the analysis purpose. The dataset we used had 868 352 rows and 18 columns. There were some null data, however, they did not have impact on the analysis because they were removed. In total the percentage of null data were quite insignificant to our analysis. There were total of 20 501 duplicate values, which were removed.

We also changed some of column names which was done to make the data reading purpose efficient and more understandable. Some columns which were insignificant for the analysis purpose were dropped. Columns like screen name, full name, Tweet ID, App, Verified, User Since, Bio, Profile Image were dropped. As mentioned earlier, only 5 airports were chosen for the analysis purpose.

For that reason, rows that did not contain tweets related to those airports were dropped. Rows were selected using the relevant twitter tags of the airports chosen for analysis. Tags @heathrowairport, @manairport, @dubinairport, @gatwick_airport and @stn_aiport were the significant

twitter tags, thus they were used to select as the relevant airports. After dropping irrelevant columns and rows we had total of 273 188 rows and 10 columns left for the analysis.

The sentiment analysis was done using the cleaned dataset with the purpose of finding positive or negative emotion from twitter users regarding the airport they have twitted at, and different aspects of the airport. This helps understand the customer experience in depth regarding the airport service quality.

With the help of this analysis customer service managers can improve the service quality they provide to customers based on the emotion or sentiment. The VADER (Valence Aware Dictionary and Sentiment Reasoner) was used for sentiment analysis. The VADER model is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media.

During the data cleaning phase, removing stopwords was an important step for the analysis. Stopwords are the words that have very little useful information and does not add value to a sentence. In English words like ‘a’, ‘the’, ‘is’, ‘are’ are so called stopwords. They do not add any value to a sentence. Removing punctuations and tokenizing were also another steps during the cleaning phase. Tokenizing means the process of splitting a large sample of text into words.

Additionally, on top of the sentiment analysis, we also wanted to categorize and retrieve more information from such huge data. For that, topic modelling is applied, which was a powerful way to mine data and to relate, in this case, similar tweets together by topic. By doing this, we get a better view of the main keywords that relate a tweet from other.

Findings:

Sentiment analysis project A:

Project A used the tweets that were related to the five airports that was chosen by the team. The analysis provided us insights on different sentiments customers have related to these airport and its service quality. After calculating the compound score of tweets, any tweets that scored above or equal to 0.05 was positive sentiment, tweets that score below -0.05 were negative sentiments and tweets between -0.05 and 0.05 were neutral tweets.

The total number of positive tweets was 107 164 compared to the 40 384 negative tweets. It seems that customers have more positive sentiment than negative. Also, 93 256 tweets were neutral, they were neither positive nor negative.

The word cloud visualization was used to illustrate most tweeted words in positive, neutral, and negative tweets. This helps us understand more about the customer’s sentiment and why they feel the certain way based on the words they have used. It can be concluded quickly that baggage handling, car parking, delays, passport, and security checks were among the most common words.

In bar-chart we can see that London Heathrow has a slightly low trend of negative sentiment in tweets. It can said that London Stansted airport has one of the highest trends in negative sentiment. To conclude, a more detailed frequency analysis of words and topics on negative sentiment provides a good basis for airport service and product development.

Sentiment analysis project B:

We also tried to gain a better understanding of the customer sentiment from different angles and tried to understand if we could find something different or significant information compared to project A. This project was done to analyze sentiment of airport customers based on the location they have used while tweeting. Two locations that deemed important for the analysis were chosen. Locations are Manchester and London. All the tweets related to the location were selected and used as the primary source for the analysis purpose.

As in project A, Vader was chosen as the tool to analyze tweets and any tweets with the compound score of 0 were classified as neutral tweets, tweets with compound score above 0 were classified as positive tweets and tweets with compound score below 0 were classified as negative tweets.

Based on locations, the amount of positive tweets were much higher than negative tweets. The word cloud was used to gain deeper analysis of positive and negative tweets, and to understand more about the sentiments in depth. The bar chart shows us that in 2018 the number of positive tweets were quite high compared to any other years.

Comparing Vader and TextBlob:

This analysis was done to better understand two sentiment analysis tools and compare their results on the same data. Text Blob is a python library which supports complex analysis and operations on textual data. It can also be used for the sentiment analysis. The same dataset was used for both analyses.

The VADER was used in similar way as in project A. Tex tBlob calculated tweets in way that tweets with the polarity score above 0 was positive, tweets with polarity score of 0 were neutral and tweets with polarity score of below 0 were negative.

Results were slightly different. Total of 16.50% of tweets were negative in VADER compared to 13.20% negative tweets in Text Blob analysis. Similarly, 43.30% tweets were neutral in VADER compared to 52.00% using Text Blob analysis. Furthermore, 40.30% of tweets were positive using VADER compared to 34.80% using Text Blob analysis.

Counter:

This method was used to count values in tweet texts, to calculate tweets length, word count and to create ngrams in order to find out which words most often occur together. There were no big differences in tweet's length across negative and positive sentiments. When measured which words most often occur together in tweets, in bigrams, it is interesting to see that car park, hand luggage, flight delayed, security staff, baggage reclaim often occur together.

In trigrams, car park, fast track security, hand luggage, queue passport control, flight delayed hours, long stay parking, security hand luggage often occur together. So, in other words, these were airport service quality areas which are mostly tweeted and where airport customer service managers should pay attention.

Topic Modelling:

For the process of learning and extracting topics from the huge number of tweets that we had, a topic modelling was proceeded. There were two approaches used for this. First, a machine learning based method was used in which we didn't give the computer system a set of rules, rather we let the computer generate its own rules to identify topics in a corpus (large collection of tweets). An unsupervised manner was the way to go where we didn't knew the topics of our tweets and, instead, we let the system identify those topics and cluster the ones of a higher degree of similarity together and then examine the words that occur the most frequently in each cluster to get a sense of the topics at hand. For this, a LDA (Latent Dirichlet Allocation) was used. In result of this model, using as a reference a topic coherence score which is measures the average similarity between top words , the ones that have highest weights in a topic, we concluded that our dataset had 4 main topics with a coherence score of 40,54%. But maybe there was a way to solidify better our model. How? Using the a rule-based approach where it uses a set of rules to extract topics from a text. It does this by identifying keywords in each tweet in a corpus. Documents that are shorter, such as tweets, tend to fare better from rules-based approaches. One of the most common ways to perform this task is via TF-IDF (term frequency-inverse document frequency). TF-IDF looks for a word's frequency in a single text, respective to that word's use across the corpus. Together with TF-IDF, we also builded bigrams and trigrams, since that are words that when used together with other words, they get a distant meaning. After a LDA model was applied in which the best coherence score was again with 4 topics with a score of 40,76%. The score didn't change that much, but a more consolidated and truth to itself model was builded, a model that could be able to understand and process words in the way we humans use the language. In the table below you could see the topic, the 8 most weighted keywords and the number of tweets of that topic.

Topic Number	Keywords	Number of tweets
1	flight, hour, plane, morning, today, bag, minute, tomorrow	138891
2	year, passenger, security, airport, week, travel, drop, queue	45903
3	thank, time, home, love, family, holiday, number, way	54136
4	airport, people, day, parking, staff, monopoly, luggage, service	47547

Conclusions:

Collecting data from social media channels and analyzing is relatively cost-effective compared to if surveys or interviews are conducted face-to-face at airports. For instance, feedback from passengers provides a good way for product and service development.

With the help of this natural language processing analysis, airport customer service managers can improve the service quality they provide to customers based on the emotion or sentiment of the customers. A good customer experience means that the customer experience meets their expectations. The positive customer experience maintains customer loyalty and attracts new customers.

By doing this project, we have come to understand how we can use different python libraries, modules, and other tools to understand customers' sentiment and the topics around this subject better.

We have also come to understand how big of a role the natural language processing can play in the future and how it can be utilized to improve the service level and make passengers' experience better. We sincerely believe that the natural language processing is one of the focus areas in the future and being able to utilize it, for instance, customer experience managers can achieve a lot.

Sprint runs and visualizations

Sprint runs, visualizations and short descriptions are found: https://markok20.github.io/TeamProject.github.io/