# Capstone Public Project Summary

| | |
|---|---|
| Student Name | Sama Tanveer |
| Student ID | 196237 |
| Company Name | Talk A Bot |
| Project Category | Sentiment Analysis |
| Faculty Supervisor | Bence Arato |

The project I chose was a natural language processing project for a chatbot company on its incoming messages. The planned scope of the project was to classify the messages into positive and negative sentiments but during the project execution I broadened the scope to include emotions to understand if treating them separately than grouping them into (positive and negative) sentiments bring in more insight.

Chatbots have been involved in natural language processing since long but I have 2 factors that make my project unique.

a. Use of emotions in addition to sentiments

Group of emotions sum up to categorize for positive and negative sentiments. I have a code that provides both sentiment and emotions classification for the decision maker to consider while deciding. Emotions also help understand the situation better example if the emotion is a result of anger or trust.

b. Using Emojis along with text tokens

Since Emojis have gained popularity as a new language, I believed they are as important for my analysis as the English word tokens. Additionally, my rationale for using emojis instead of dropping them was that they are very accurate in their meanings, if it is a sad face it must mean sad and no other way.

Data Extraction

I extracted my data using DBeaver keeping in mind the schemas done in PgModeler. I used PostGreSQl to extract data after a thorough exploratory analysis. The details of my PostGreSql findings are present in the Appendix A of my Technical Report. While doing so I also tried how the current data infrastructure is set up and have come up with feedback for the client. I believe that they should move to a cloud service as the data they have is increasing and the natural language processing tools already available should be leveraged as they can invest time in creating what hasn't been done already. This would also make them a Digital Champions as they according to the PwC study are open to using available solutions to invest time in being innovative in new areas.

Infrastructure Recommendation

Studying the current infrastructure, the report suggests possible improvements to the current structure. Moving to a cloud based online infrastructure has been analyzed for its advantages and disadvantages and suggestions have been given on how to proceed with or without it.

Concepts

Once I had data, it was important for me to understand what Natural Language Processing concepts were. I used the following material for guidance; StackoverFlow, Book: Text Mining with R, Data Science 4: Unstructured Text Analysis course material, Eduardo Github Repository and Emoji Analysis – Hadley Emo Github Repository. Other courses like IoT, Use Case Seminars, Entrepreneurship, Digital Transformation and Developing Digital Organization also helped me use this as a business problem and suggest project applications. The internet was also used to understand Psychological theories related to sentiments and emotions.

CEU eTD Collection

Textual Analysis

The first stage of my learning involved the Data Infrastructure. After trying various iterations and joins I selected the fields useful for my analysis and extracted them in a suitable format. The next step of data cleaning was done in R where characters were removed, 'TEXT' data coming 'IN' from the chatbot was filtered and a package 'EMO' by Hadley was used to filter out languages in which Emojis were typed. The package was found after struggling with different means of incorporating Emojis in my analysis once I had realized they are a good source of Sentiment for me. My initial approach was finding a English Emoji meaning excel file and using that to replace the Emoji code with the English meaning. But this meant I lost the Emojis what were typed from non-English keyboards. The EMO package helped me read Emojis from helped me detect that most of my Emojis in my data came from languages "en", "bn", "ceb", "de", "ja" and "su".

The packages I have used for my analysis include tidytext, dplyr, stringr, stringi, tidyr, ggplot2, tidyverse, lubridate, scales, emojifont, utf8, wordcloud, widyr, igraph, ggraph, reschape2, textcat, textclean, data.table, cld2, emojifont, cld3 and RColorBrewer.

The process of this analysis is first the data cleaning discussed above followed by textual analysis. Further cleaning was done to remove system generated msgs and other messages not necessarily stop words. The textual analysis involves analysis of most frequently used words, most frequently used pairs of words and most frequent other combinations of words. This was done involving emoji data and results with and without emoji data were compared. Since we did not have many Emojis, rest of the analysis involved Emoji data included to the general dataset.

Tokenization

The next stage was tokenization in we had unnested tokens from the messages and stop words were removed. Similar textual analysis was done, and we also used Bi-grams to study combinations frequent and rare. We apply Zipfs law to get the Term Frequency for the important unique words rather than the frequently occurring ones as they bring us more insights. Interesting combinations were revealed that helped us understand the chatbot user segmentation and base our use-case suggestions based on that. Network charts and Bigram plots were made to represent the combinations and the direction of word usage.

Sentiment Analysis

After Tokenization we proceed to our Sentiment Analysis for which we used Sentiment Lexicons. Used the dataset with AFINN, BING and NRC dictionaries to see what works better for my dataset. Since BING offered more variation it was used to carry out the sentiment analysis. The BING dictionary helped accurately separate Positive and Negative Sentiments, and this helped understand the two categories for the chatbot messages coming IN. Since the classifications are based on scores, TalkABot can have ceiling and floor to have a trigger when the score reaches an undesirable range or requires any attention.

Emotions Analysis

I further enhanced my analysis beyond the scope of this project to investigate if looking at Emotions from Positive and Negative sentiments bring us more insights. This was done using NRC library in combination to our BING library to separate the emotions and plot them individually. This did bring us good insight but due to the limited data the insights are not as beneficial as the sentiments.

Conclusion, Recommendations and Use-Cases

The better insight from separating Emotions is when I try to plot it against time. This brings us more insights as the emotions change trend when there is a change in ChatBot features such as the introduction of jokes and GIFS to the messages sent OUT during the conversation. It shows that the change has been well received and increased trust and joy for the user and decreased anger contributing to an overall sentiment. This chart can be very useful for the chatbot agency as they can track progress of different initiatives through emotions in real time.

Since I did not know Hungarian and most of the client data was in Hungarian, the shortcoming of my project is that it has not been tested on actual data from clients. It is based on the English data that the company's own chatbot receives. However, the code is structured in a way that it is reusable, and the EMO package recognizes Hungarian. There are also Hungarian Lexicon dictionaries available that can be used. Recommendations are discussed in the technical report that discuss how the vocabulary expansion can be done and how the current project can be extended and improved. Since I have learned a lot from the iterative process to find the best approach, I think the project can be used as a guide for being up with the trends and working in Sprints. The project also places stress on the role of Emoji and that could be further investigated as more and more research is done for that domain. The report also suggests some use-cases for the chatbot agency on how to make the project useful for the customer, for the client and for the chatbot agency itself as well.

CEU eTD Collection

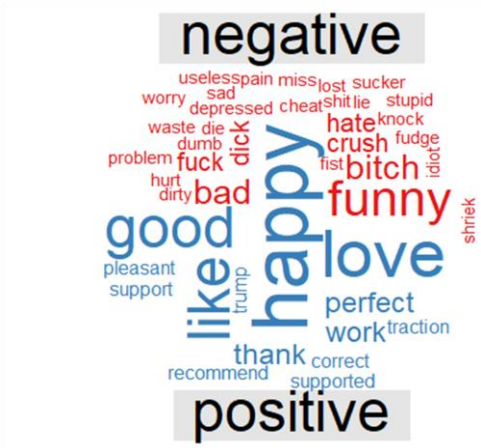
**Sentiment Response to Chatbot Messages**

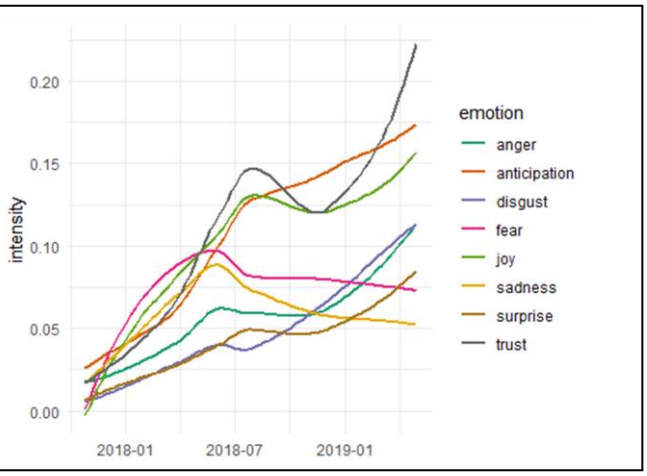
Talk-A-Bot is a smart chatbot company that has invested in making the chatting experience an enjoyable one but have not performed analytics to measure the response to its campaigns of introducing jokes and GIFS.

The objective is to use Natural Language Processing techniques to understand the msgs coming in from chatbot users and perform Sentiment Analysis to understand the response

**Conclusions**

1. Frequent negative words: Bitch, Bad, Hate and Fuck.
2. Frequent positive words: Happy, Love, Like and Good.
3. There has been a change in emotions around July 2018 where we see an increase in TRUST and a decline in FEAR.
4. The data can help us identify traits of our users that can help us in segmentation




**Recommendations**

1. Develop for other languages
2. Updates to R and Emo Hadley package
3. Vocabulary Expansion
4. Regular Campaign performance sentiment testing
5. Sentence Analysis
6. Give importance to Emojis

**Use Case: Demographic Segmentation**

Our TF_IDF analysis revealed combinations that help us identify religion or location of our chatbot users. This can be demographic segmentation tool for our clients.

**Use Case: Integrated CRM Solution**

Talk-A-Bot can spread its platform to offer a complete solution that can prioritize customers have negative sentiments

**Use Case: Immediate service and NPS Tool**

Talk-A-Bot can identify dissatisfaction to prioritize customers who require immediate service.
The satisfaction measure potential can be extended to a Net Promoter Score calculator.

**Use Case: Contribution of Emotions and Emojis to Linguistic studies**

The project identifies Emojis as a medium of emotion and the quality of data collected can be used for development in research related sciences

CEU eTD Collection

CEU eTD Collection

CEU eTD Collection

CEU eTD Collection

CEU eTD Collection

CEU eTD Collection

CEU eTD Collection