**Capstone – Team B3 – Mid-Project Report**

Members: Hanyu Chen, Yipeng Guo, Luke Hong, Ting-Ann Lu, Yuyan Ma

A) - Company Background

OneQuesh is a social media app that aims to let users express their authentic selves online. It is a recently established start-up company that believes in asking probing questions and fostering real human interactions with its users. It aims to disrupt the traditional social media space; not chasing likes and filtering hateful comments. One question a day, express your true self!

B) - Proposal & Hypothesis

We propose to achieve the following goals:

* Explore and analyze what kind of questions gain **higher popularity** ratings
* Build an algorithmic model to **predict** well performing and ‘will-be-popular’ questions

We also conduct our analysis based on our hypotheses listed as follows:

* Hypothesis 1 - Popular (interesting) Quesh may have **more responses, more characters** (initiate people to talk more)**, and more positive responses** (higher sentiment score), so-called ‘good’ questions.
* Hypothesis 2 - We suppose these good questions can generate more attention for OneQuesh, and make it a more successful company.
* Hypothesis 3 - We assume by filtering good questions out, we can find out the common sense among them and train a model to predict. Especially when we have the question bank, we can pick out 200 for OneQuesh further use.

C) - Data

With the kind support of OneQuesh, we have a couple of detailed datasets containing:

* Every day’s questions, responses, and ratings
* Users’ information: status, join-in date, date of birth, whether the profile is complete
* Specific comments for the questions (if people like the daily Quesh)

D) - Methodology

In general, we plan to use the methods and background knowledge of what we learned in our classes in the following ways:

* Sentiment score (text):
  + By conducting sentiment analysis, we expect to be able to check whether users’ responses are positive or negative (if larger than 0, then positive; Vice versa). Basically, we calculate the average sentiment score of each question. Also, we check the nature of questions by the total sentiment of the responses of questions as well.
* Character count (text):
  + By counting the amount of characters, we expect to calculate the average character count towards each question to shed lights on popularity.
* Analysis towards comments:
  + By analyzing the specific comments, we can deeply understand what a ‘good’ Quesh looks like.
* H-cluster and k-means cluster:
  + By clustering the questions, we aim to categorize them into 3 groups: great performing ones, moderate performing ones, and under-performing ones. Then, we can find the common features of what separates a great question from an underperforming question. By clustering the questions, we can get a profile of the types of questions OneQuesh deploys in their app and their numerical features to assess the performance of ratings and engagement.
* Spacy & NN prediction models:
  + To predict and find will-be popular questions (our proposed next step).

E) - Early Results

By implementing the methods and tools listed above, we achieved the following results:

We started with exploratory data analysis by creating visualizations to tell a story of the data. Also, we applied text analysis tools to get insights from our dataset. By doing so, we were able to do some feature engineering and augment our existing datasets.

* Visualization - EDA
  + Users’ information (joined-on date & age)

Chart, line chart

Description automatically generatedA picture containing chart

Description automatically generated

* + Distributions of some features towards responses (avg-rating & # of response)

Chart, bar chart

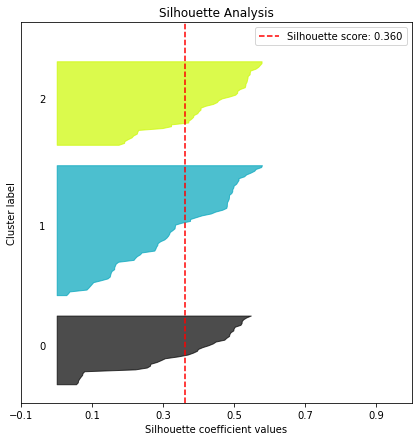
Description automatically generatedChart

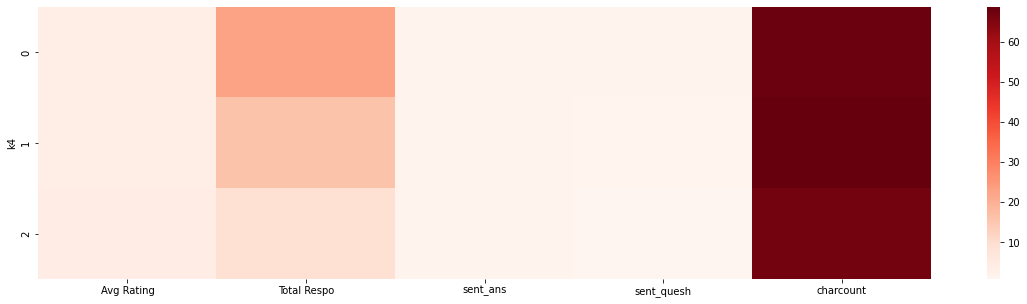
Description automatically generated

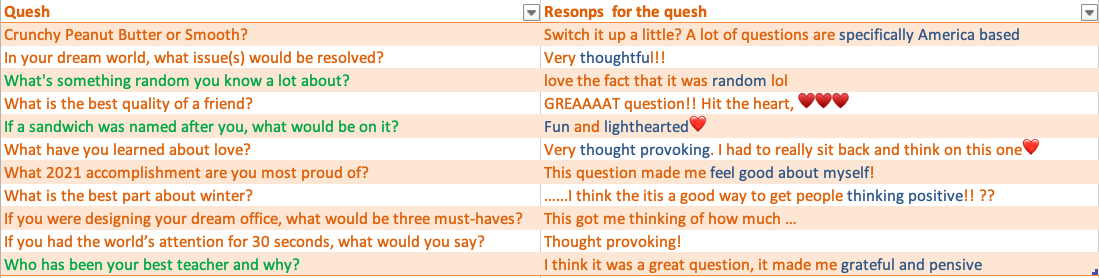
* H-Clustering using augmented dataset (including sentiment score & charcount)

Chart

Description automatically generated

* K-Means clustering



* A few insightful commons and findings (based on comments towards Quesh)
  + Green questions - Very good question (Common questions)
  + Blue keyword - insightful keywords 
* Base on those commons, we find they users loves questions are:

1. More diverse not just american based

2. Thoughtful

3. Related to Random

4. Fun and lighthearted

5. Thought provoking and pensive

6. Feel good about self

7. Thinking positive

F) - Next Steps

Since we already have many insights towards the Quesh and the responses, we should be able to create an algorithmic model to predict those will-be popular questions out of the question bank.

For now, we have filtered out those popular questions with higher number of responses, higher sentiment score, higher character count,and digging deeper into them - what contents they are sharing, what type of questions they are, what aspects that increased engagement, etc. Once we have those characteristics quantitatively, we should be able to build a predictive model.

To be more specific, we are considering converting our Quesh and related features into numerical form, label them as ‘popular’ or not, then apply deep learning skills to train a correlated model, test it against our question bank dataset and find out those ‘will-be popular questions’ to recommend to our client.

G) - Summary

By Feb 21th, we total have 2359 respons and 73 comments. We look though all of them and find out some interesting comments which might be insightful for the question creation in the future.

By the sentiment analysis and the result of H-clustering model, we could see that our hypothesis 1 is proved that the questions have more positive sentiment score can also have more responses and characters, which we call good questions.

With the results of our H-clustering model, we can use the input as a categorical value for our supervised machine learning tasks. With our supervised machine learning, we can try to predict the performance of potential questions and profile the segments to tell a story and take action.

For next step, we look into users dataset and there is what we know in toal we have 281 users and active rate is 98%, verify rate is 83% and profilo complet rate is 96%. Also we look into the top 20 most involved users most of them join the App at september of 2021 and all of them have been verified and completed the profilo.