BRAINLY

Devising a mechanism to delete the quality of poor quality content on Q&A site, Brainly

BRAINLY

Role

As a mixed method researcher, I consulted with a peer-learning platform to automate the process of eliminating poor quality content by building ML algorithms

Team

Collaborated with UX Researchers, Software Engineers, Machine learning engineers,

Duration

June 2016 - July 2017

Skills and Tools

Tools: SQL, Python, R, Github

Skills: Exploratory data analysis, Machine-learning algorithms, Building future Product Roadmap, Contextual Inquiry

Objective/Problem

Brainly, a peer-learning question-answering (Q&A) platform for K-12 community was researching on automating the deletion of inferior quality content present on their site.

To solve this problem, they collaborated with our research group where I worked as the lead of directing the project **"Automating the identification of the quality of questions on the Q&A site"**

Process

Working on this project from ground up, we divided this project into two parts:

Part 1:

Qualitative Research: We considered Brainly questions deleted from the site and deductively came up with specific categories which led to their deletion from the site by the moderators. After categorizing content based on different types, we created a code book which could essentially cover all kinds of questions across diverse categories such as **offensive, vulgar, plagiarized.**

Apart from considering the deleted content we also created a codebook for non-deleted content which include: **advise seeking, opinion seeking, factual content.**

We discussed with the team and the site moderators of Brainly regarding the different categories and made relevant modifications and later furthered our research in coming up with ways to **automatically detect inferior quality questions**.

Process

Part 2:

Quantitative Research: After establishing the different categories, we conducted literature review to find out how do different textual and contextual features contributes in determining the quality of content. The textual features consisted of the **total number of words, presence of interrogative words, present of vulgar words etc**. The non-textual features consisted of the **total number of upvotes/downvotes**, and the social capital earned by the asker/answerer on the site. scripts were written to extract textual features from the questions to perform further analysis. To dig deeper exploratory data analysis was performed understand the difference between questions deleted vs non-deleted questions by the moderators of the site.

We considered questions **deleted by the moderators as the gold-standard as that of inferior quality**. We later built **machine learning models to automatically predict the quality of questions** to reduce the workload of the moderators.

Challenges

- Identifying the quality of content is an extremely complex topic, and requires deep understand of not just the text but also users' intention. Although we devised a comprehensive codebook to categorize the quality of questions, it was difficult for the machine learning algorithms to completely identify the quality of content.
- The textual and contextual features chosen by further worked on to obtain better performance by the machine learning algorithms.
- It was difficult to match with the pace of research along with collaborating with the engineering team who were responsible for building the end to end architecture for the machine learning pipeline.

Outcomes

Brainly with the help of the content quality research came up with **premium account (Brainly Plus) model** which essentially lets its members to view unlimited top-quality answers.

The research was employed in the machine learning pipeline of the startup which tremendously helped in reducing the workload of the company

Summary

Summary:

The research experience gained by working with the startup was invaluable and was extremely fulfilling. Understanding and coming up with a way to identify inferior quality content is essential in a site which is heavily used for educational purposes by K-12 students.

The process carried out in my research was time-consuming yet was methodical. It accomplishes the set goals to a great extent. The coming of with a codebook was essential to justify what consists of inferior quality content. Moreover, applying machine learning algorithms in identifying the quality of content gave us to enough success to better understand how to devise better predictive models in future.