Eliminating poor quality content on Q&A platform

Objective

Brainly, a peer-learning question-answering (Q&A) platform for K-12 community was experiencing **low user engagement & user satisfaction** due to the inferior quality of user generated content on their site. This required frequent intervention from site moderators to identify and remove these content.

This process was time consuming and not scalable. Brainly was exploring ways to automate the process and collaborated with our research group.

I worked as the Project Lead for **“Automating the process of identifying & deleting inferior quality of questions on the Q&A site”**

Role

As a mixed method researcher, consulted peer-learning site, Brainly; helped automate the process of eliminating poor quality content by integrating ML into their platform.

Team

Cross-functional team consisting of the CEO/Cofounder, UX Researcher, Software Developers, Machine Learning Engineers

Duration

~1 year
(June ‘16 - July ‘17)

Tools & Skills

Tools: SQL, Python, R, Github

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

Method

Collaborated on the project from ground up. The research consisted of two parts:

<u>Part 1: Qualitative Research</u>

Analyzed 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, created a code book which essentially covered varieties of questions across diverse categories labeled as **offensive, vulgar, plagiarized.**

Along with considering the deleted content, also created a codebook for non-deleted content which included: **advice seeking, opinion seeking, factual content.**

Brainstormed with the site moderators regarding the different categories identified in the code book and made relevant modifications. Later furthered our research in coming up with ways to **automatically detect inferior quality questions**.

Method

Part 2: Quantitative Research

After establishing different categories of questions, conducted literature review to find out how do different textual and contextual features contribute 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.

Wrote ML scripts to extract textual features from the questions to perform further analysis. For further examination, performed exploratory data analysis to understand the difference between deleted questions vs non-deleted questions by the site moderators.

Considered questions **deleted by the moderators as the benchmark for inferior quality**. Subsequently, build **machine learning models to automatically predict the quality of new questions.**

Outcome

- Brainly transitioned into a “freemium” pricing strategy by introducing **premium account (Brainly Plus)** option, which allows users to view unlimited top-quality answers aided by our research.
- The research was deployed in the machine learning pipeline, which significantly reduced manual intervention from site moderators by automatically eliminating inferior quality content
- Reported increase in Customer Satisfaction & engagement, post implementation of our research.
- Published 4 research papers; Presented findings in Information Science conferences

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.

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.

Further information

The research in this work can be viewed in the following papers:
https://dl.acm.org/doi/abs/10.1145/3020165.3022145
https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/pra2.2017.14505401036
https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/pra2.2017.14505401081