Proposal: Ranking of Yelp customer reviews

Group Members:

Xiaoyang Sheng

Yulin Gao

Zicong Xiao

Research Problem:

The proper assessment and arrangement of customer reviews have consistently posed significant challenges in the realm of e-commerce. This is due to the fact that displaying relevant reviews can significantly impact consumer psychology, leading to increased sales and a favorable reputation. Conversely, the prominence of deceptive or misleading feedback can potentially result in financial losses and the erosion of customer trust.

Hence, the development of a method to effectively rank and present reviews can be a game-changer. Such an approach would empower sellers and e-commerce platforms to boost their sales while ensuring that customers enjoy a seamless and trustworthy shopping experience.

Data Source and description:

We will use the open Yelp dataset, accessible at (https://www.yelp.com/dataset). Yelp is a big company renowned for curating crowd-sourced reviews on various businesses. The dataset aligns seamlessly with our research objectives, offering comprehensive data that encompasses the core aspects we require.

Yelp's dataset encompasses a wealth of information, including business details such as location and attributes, full review text data with corresponding user_id and business_id, as well as user-related data that includes friend mappings and various user-associated metadata. This multi-faceted dataset equips us with the resources to approach our research problem from diverse angles.

Analysis:

Our research involves two primary analyses. Firstly, we intend to leverage the

non-textual data pertaining to businesses, reviews, and users to conduct regression analysis focused on understanding 'useful votes.' These votes are crucial as they serve as a significant indicator of a review's value and impact.Additionally, we plan to undertake Natural Language Processing (NLP) tasks to delve into the textual content of reviews, assessing potential sentiment as well as 'useful votes' mentioned above. After the preliminary analysis we intend to use transformer to accomplish this task. By bagging these two methods, we construct a comprehensive ranking algorithm, facilitating the ranking of reviews based on their vote scores or sentiment.

Our approach involves partitioning the data into training, validation, and test sets using a combination of random and temporal criteria. By doing so, we ensure that our training data incorporates early records, which tend to have more credible and established useful votes, along with other stable information. Following the training and fine-tuning of our methods, we will employ the test set to evaluate the model's performance. Subsequently, we will make necessary improvements and corrections to refine the model further.

Delieverables:

We will finally come up with an algorithm to rank the customer review appropriately and efficiently, together with our training and assessment methodology.