Yelp Fusion

Motivation for Dataset Creation

The Yelp Fusion was created to allow access to business information and the associated reviews/ratings from businesses across 32 countries. Users can search for category, location, address, phone number, price, rating, hours of operation, photos, etc. Yelp content has been used by thousands of startups and some of the largest companies in the world, including Apple, Twitter, Microsoft, BMW, and Mercedes-Benz for a broad array of products including augmented reality devices, in-car navigation, artificial intelligence, virtual assistants, transit, navigation, messaging, data visualization, local search and robotics.

Dataset Composition

Each instance is a business with all the information associated with it that is available on Yelp. For example, a business is searchable by name and relevant data such as address, phone number, rating, price, hours of operation, etc can be identified along with the search. This is an open-source tool that allows users to look up specific information about a business of interest; hence there was not a previous attempt to use this for some type of research.

Data Collection Process

Data was collected directly from the businesses themselves (they list updates every month) so as to deliver the most current and accurate local data available. New reviews, ratings, and photos are also added by active Yelp users. Data is available from 2007 to current. Data is directly observable (i.e., business specific information, users writing reviews and reporting ratings, etc). The dataset contains all possible instances. There is missing data if the business did not disclose the specific information to Yelp. Hence, it is more likely that the data is missing because it was unavailable rather than being intentionally dropped.

Data Preprocessing

The available data has not been preprocessed or cleaned. Such tasks are the responsibility of the Yelp Fusion users to make use of the dataset for their study purposes. Yelp Fusion includes access to broad content and sophisticated search tools for users and provide support for new initiatives to stem from the information available in the dataset.

Dataset Distribution

The dataset can be accessed via https://www.yelp.com/developers/documentation/v3/get_started. The user would need to obtain an API Key to access the data. There are no fees associated with access but one can only make up to 5,000 API requests per day.

Dataset Maintenance

One can reach out to api@yelp.com for further support. The Yelp team is in charge of updating the dataset directly from the information they obtain from businesses and the dataset is reviewed daily. The changes are documented in https://www.yelp.com/developers/v3/changelog. Issues with the dataset can be posted on their Github page (https://github.com/Yelp/yelp-fusion/issues) for troubleshooting.

Legal & Ethical Considerations

The Yelp dataset does not directly relate to people and their attributes but was generated by people (i.e., specific information about businesses, ratings, reviews, photos, etc). Users are aware that all the information they are inputting would be recorded. The dataset does contain photos which should mostly be of the business and/or their products but may contain unintended pictures of individuals at the business. This might be considered sensitive but since the intent is not the data regarding the individual this does not breach any confidentiality regulations. The dataset contains reviews from customers for businesses which may be considered offensive if the customers were harsh with their comments.