Talk of the Town

Group 6

Jonathan Jude Regalado, William Robinson, Justin Tapia, Emily Torres

Description:

Our team project will be conducting sentiment analysis of reviews from tourists from 11 U.S. cities with the goal of helping city municipalities optimize their tourism strategies. With access to nearly 7 million Yelp reviews, our team will be building and training an NLP model that can identify elements in tourist reviews that either praise or criticize popular tourist destinations, diving deeper than basic star ratings to really understand customer feedback. Ultimately, the goal is to give cities data-driven guidance on where to invest more resources to enhance tourism and the tourist experience, which could boost the economic impact these tourist destinations bring to the city.

Problem Statement:

Cities put a lot of time, effort, and money into attracting visitors. Tourism can be an economic engine if harnessed correctly. Consumer feedback via reviews on Google, Yelp, TripAdvisor, and the like is a gold mine of information these cities could mine to identify the things most in need of improvement as well as the things that are already attracting tourists, but this data is not easily parsed or understood.

By parsing reviews and analyzing the sentiments of the feedback, our analysis will go far beyond how many stars a site has; it will fully unpack what elements about each site people like or dislike, enabling cities to invest more in what works and discontinue or change what does not.

This is a pure NLP problem because the core mission of our project is to build and train a model that can read human-written reviews, assess the sentiment of the components of each message, and identify which specific things produced positive reactions and which produced negative reactions.

Dataset Selection:

The Yelp Open Dataset provides us with a database of nearly 7 million reviews of more than 150,000 businesses across 11 cities, all in a series of documented JSON files. This dataset is ideal for its geographic diversity, volume of language data, and relevance to our problem statement. By including the star rating that each user gave, it also gives us a point of reference when evaluating our sentiment analysis model.

By focusing on places that are popular with tourists, we are concentrating our analysis on the locations likely to have the most reviews, giving our model more to train on and making our analysis more robust.

There are some limitations to our dataset. As we only intend to focus on using the Yelp open dataset (at the moment), there are aspects of the Yelp dataset that could skew or bias the reviews to users that are actively using Yelp, which potentially could be older users (as younger generations could be using other platforms) and potentially only written in English, which would limit the data to only English-speaking tourists. Additionally, the time span of the available data in the dataset could also introduce some nuances caused by reviews that could've been written during unusual circumstances (i.e. the pandemic). We will attempt to address these limitations during the development of the model.

Expected Outcomes:

We expect to produce sentiment analyses of each review (as long as it has a written component and is not only a star rating) left for major sites at 11 cities. The analyses should identify elements of each review as positive, neutral, or negative.

For example, imagine a reviewer left a comment saying, "It was great when we finally got there, but it was a nightmare to find. Really needs more signs!!" In this case, we would want to identify both a positive sentiment towards the site as well as a negative sentiment towards signage.

Additionally, Yelp reviews afford us the ability to compare our sentiment analysis to the number of stars the reviewer gave the site. This is a ground truth of sorts that allows us to plot the positivity of our sentiment analysis versus the number of stars the reviewer gave. If we see that many of the reviews our model identified as positive are actually 1-star reviews, we know our model needs major improvement. Conversely, if we see that our model pairs 1-star reviews with negative sentiments, 2–4-star reviews with a mixture of sentiments dominated by neutral sentiments, and 5-star reviews with positive sentiments, then we know our model fits the data quite well.