DSCI 510: Final Project

I. Name and Team Member(s)

Analyzing User Sentiment Across Popular Dating Apps

Member(s): William Khuu

II. Description of the Project

This project performs a data-driven analysis on user reviews left on the big 3 dating apps (Hinge, Tinder, and Bumble). Using Vader Sentiment Analysis and Latent Dirichlet Allocation (LDA) Topic modelling, I analyzed overall user satisfaction and identified the 5 main themes driving the users' satisfaction and complaints. The results provide clear, actionable data to help app teams fix user pain points like fake profiles, safety concerns, and poor subscription value.

III. Data Collection

The data for this project was scraped directly from the Google Play Store in real-time, using the Google Play Scraper Python Library. For each app, 1000 user reviews were scraped and analyzed.

IV. Data Cleaning, Analysis, and Visualization

1. Data Cleaning and Preprocessing

The initial phase focused on transforming the raw dataset into a structured and analyzable dataset. Missing values were removed, text was normalized (lowercase, standard punctuation, and special characters were removed), stopwords (i.e., 'the', 'a', 'is') were removed, and feature flag engineering was incorporated to segment high-priority complaints. One flag revolved around safety features, and another was related to subscription issues. The cleaned and processed data were saved into a CSV file, ready for analysis.

2. Analysis and Modeling.

This phase applied two text analysis models to quantify user sentiment and discover thematic concerns for the dating apps overall. The VADER model was used to assign a sentiment category (positive, negative, neutral) to every review based on its compound score. This provided a quantitative measure of overall user emotion. The Latent Dirichlet Allocation Topic Modelling was applied to automatically discover the five major underlying themes among the reviews. The text was first tokenized and vectorized using the CountVectorizer library, which converted the text into a matrix of token counts, allowing the LDA model to run. For each topic, there were 10 defining words that were then manually interpreted to create the thematic labels.

3. Visualization and Reporting

Four visualizations were generated to communicate the findings. First, the Sentiment Distribution plot was a bar chart comparing the percentage of positive, negative, and neutral reviews across the three applications. Next, the Average Rating Plot was a bar chart comparing the mean rating for each app. The Topic Distribution Plot displayed the breakdown of the 5 LDA topics across each app, illustrating the specific reasons for user satisfaction or dissatisfaction. Finally, the Feature Flag Comparison compares the percentage of reviews containing high-priority safety and subscription issues across the apps.

Initial premise

The project aimed to understand what the main pain points are when using dating apps. Specifically, testing whether the frustration is primarily from what I believe would be the largest concerns: safety risks and aggressive monetization practices.

Key Conclusions

Sentiment vs Rating Disconnect

The VADER analysis showed that a large portion of reviews fell into the Neutral category, yet all average star ratings were low (around 2). This suggests that users who take the time to review often do so to express extreme dissatisfaction or criticism that the star rating alone fails to capture.

Two Dominant Pain Points

The feature flag analysis showed that the two most common and severe complaints are Subscription services and Safety/Scam concerns. These overshadow usability complaints and represent the biggest barriers to overall success.

Competitive Thematic Strengths

The LDA Topic Analysis revealed 5 thematic topics: Monetization and High-Cost, Positive Outcomes and Match Quality, General Usability and Comparative Discussion, Time Waste and Fake Accounts, and Profile Quality and Technical Issues.

The key takeaways from these themes are that financial frustration is severe and the prevalence of fake accounts is high. With a high volume of reviews flagged for subscription complaints, it confirms that users feel they are being manipulated into paying for basic features without any benefit. When looking at the fake account issue, users think the platform is actively facilitating deception, making the time they spend on the app feel unproductive.

The analysis paints a picture for the dating app industry by proving that low user satisfaction is driven by two systemic issues: a crisis of authenticity and crisis of value. The core impact is a requirement to shaft product strategy: resources must be aggressively allocated toward security and anti-scam measure to restore trust in uders and premium subscription models must be redesigned to offer genuine, success-focused utility instead of relying on paywalls that create financial friction.

V. Changes from the Original Proposal

My original project idea was actually using the Spotify API and analyzing the WEB API song. However, recently Spotify made changes to its developers page, and this information was no longer publicly available, making the analysis no longer possible.

This new idea was generated quite late, making another project proposal redundant, as you would not be able to review it in time (Friday, November 12).

One change I made to my original idea for this project was using the Google Play Store reviews instead of Apple App Store reviews. The iPhone is the more popular choice of device, and is what would likely represent the population of users more accurately. However, the App Store Scraper library was difficult to use and would have likely taken more time than I had to troubleshoot. To combat this, I pivoted to using the Google Play Store Scraper as it was much more accessible.

VI. Mention of Future Work

Given more time and resources, I believed using the App Store scraper would have provided much more accurate and valuable information. Also, having a larger dataset would create a greater representation of the population of people who use dating apps. Pulling the most recent 1000 can lead to lots of bias and drive an incorrect analysis. Furthermore, to enhance the analytical rigor, future work using more advanced NLP techniques like N-gram analysis and potentially time-series analysis would be beneficial for the analysis.