Plan Overview

A Data Management Plan created using DMPTool

Title: Study of Mobile App Rivals: Harnessing Machine Learning to Decode User Reviews in the Mobile App Landscape

Creator: Bhavik Patel

Affiliation: San Jose State University (sjsu.edu)

Funder: Digital Curation Centre (dcc.ac.uk)

Template: Digital Curation Centre

Project abstract:

User reviews on the app are of utmost importance as they can be utilized for accurate data-driven decision-making by upcoming users and developers to make their app stand out in the cut-throat market. The existing work majorly focuses on a single domain while analyzing the user sentiment about a particular application with limited attention given to multiple domains. An in-depth analysis of the competitive mobile app landscape across different domains is the need of the hour with the drastic increase in the mobile application market and the consumers that rely on them. The current study aims to suggest an ideal app for new users who are juggling between different Travel and Lodging apps based on their requirements by adequately analyzing the sentiments of the current users about those apps. It involves using different Machine learning algorithms such as Bi-LSTM, BERT, XLNet, and SVM that are used to perform sentiment analysis on the review data collected from 20 different mobile applications from the Google Play Store. App Functionality, Customer Service, User-Friendly, and Payment Experience are the four major topics identified using the LDA technique in Topic Modeling. The models are assessed using performance metrics like Accuracy, F1-Score, Recall, and Precision. Based on the Model Assessment, the BERT model reliably recognizes the user's sentiment and offers insightful app comparisons with 92.32% accuracy. The results achieved can save a lot of downtime for the new users in identifying the app of their choice.

Start date: 09-22-2023

End date: 12-01-2023

Last modified: 12-08-2023

Study of Mobile App Rivals: Harnessing Machine Learning to Decode User Reviews in the Mobile App Landscape

The primary dataset will consist of user reviews collected from Google Play Store, which includes information about reviewId, userName, userImage, content, score, thumbsUpCount, reviewCreatedVersion, at replyContent, repliedAt, appVersion. Two new columns will be added: AppName and DomainName for better analysis after merging the dataset.

The data will be collected using a python script, which utilizes a google play scraper library to fetch latest reviews.

The documentation will cover the following topics:

1. Collection Source
2. Collection Method
3. Anonymization Techniques
4. Metadata Schema
5. Data Quality Information
6. Usage Guidelines
7. Version Control Information

The study will emphasize user consent and privacy, anonymize personal information in user reviews, maintain transparency in the use of the data, abide by ethical norms, and request regular reviews from an ethics council in order to handle ethical difficulties.

The study will manage copyright and intellectual property rights (IP/IPR) for the use of user reviews by ensuring compliance with system terms of service, protecting privacy through the anonymization of private information, and adhering to fair use legislation. Furthermore, the study will seek legal counsel as necessary to guarantee full adherence to IP/IPR laws and regulations.

During the research, all data will be securely stored and backed up on Google Drive, taking advantage of its comprehensive cloud storage solutions to ensure data integrity and accessibility. Google Drive's automated backup settings will be configured for continuous protection of newly added and updated files. In addition to these automatic backups, The study will perform regular manual backups to serve as a secondary safeguard, thus maintaining a reliable repository of our research data.

Access to the data stored on Google Drive will be strictly managed, with permissions assigned only to authorized members of the research team to preserve confidentiality. Each team member will have individual access rights appropriate to their role in the research, ensuring a high level of data security. To manage document edits and maintain a clear audit trail, we will utilize Google Drive's versioning feature, which allows for the tracking and management of revisions. For an extra layer of security, sensitive data will be encrypted before being uploaded to the drive. These measures will ensure our data storage practices meet the highest data management standards and comply with all pertinent data protection policies.

Long-term value data, which is deserving of being kept, shared, and preserved, mostly consists of anonymized user reviews, which are a valuable resource for studying consumer behavior over long time horizons. Both the methodological documentation outlining the data collecting and processing techniques and the aggregated statistics that indicate trends and patterns are equally valuable. This guarantees that the results can be confirmed and that the study can be repeated or developed further in the future. Additionally, papers and research findings that have been

synthesized capture the key takeaways from the data and are extremely valuable for future scholarly and business use.

The dataset's long-term preservation plan calls for storing the data in a reliable, secure digital repository, transforming it into formats that are both enduring and easily accessible, and appending extensive metadata to facilitate user access. In order to combat technology obsolescence, we will plan for possible data transfers, carry out routine audits of data integrity, and create explicit access policies that adhere to legal and ethical requirements. Following through on these actions and obtaining the funds required for ongoing data curation guarantees that the dataset will be a useful resource for upcoming studies and research. A key component of the preservation plan will be community interaction, which will promote active use and referencing of the dataset in later publications.

There will be a controlled and organized method for sharing the data. This will probably entail putting the dataset in a reputable academic or institutional repository to provide widespread accessible while upholding documentation and data quality standards. The information might also be accessible through a Github repository page. The data will come with extensive metadata and understandable documentation explaining its purpose, how it was collected, and possible uses to make it easier to utilize. In appropriate cases, we will also think about releasing a data article in an academic journal that will improve the dataset's discoverability and scholarly effect by giving a comprehensive synopsis of the dataset and its importance. The final report, nevertheless, will be in the repository.

It could be essential to impose restrictions on data sharing in order to preserve privacy and comply with legal and ethical requirements. Despite being anonymized, the dataset contains user reviews, therefore it's important to make sure the sharing mechanism complies with data protection laws like GDPR or HIPAA, depending on where the data is located. Researchers who consent to use the data only for academic purposes—with no commercial usage allowed—may be granted access. Before access is allowed, a data usage agreement that outlines the conditions of use and any restrictions may occasionally be necessary. In addition, any sensitive information will be further anonymized or censored to protect personal identification and uphold ethical confidentiality norms.

We, a group of four students working without a manager, have developed a cooperative strategy for managing the data in our project. Because the tasks are divided equally among the participants, everyone can concentrate on particular elements that are essential to the success of the study. The first student's job is to supervise the data collection procedure and make sure that the information is gathered accurately and quickly. In order to guarantee the data's security and accessibility, the second member is in charge of its safe storage and routine backup. Maintaining data accuracy and quality, as well as the crucial duty of anonymizing personal data for privacy adherence, fall under the purview of the third student's function. The fourth student is in responsible of managing the metadata and paperwork in the interim, making sure that all data is accurately documented and that information is up to date. We've agreed to meet on a frequent basis in order to keep our data management activities cohesive and effective. These meetings are essential for keeping each other informed on individual development and for guaranteeing a unified and coordinated strategy for handling the data for our research.

Our data management approach requires a wide range of resources to be implemented successfully. This comprises software tools like database management systems, statistical analysis software, and cloud storage services like Google Drive that are used for data collecting, analysis, storage, and backup. Hardware resources are essential for data handling and backup procedures, especially PCs with enough processing power and storage combined with dependable and secure internet connectivity. Training materials and tutorials will be made available to guarantee our team's competence with these products. There will be a use of data encryption and anonymization methods to

ensure data security and privacy. Online solutions such as Jira, Zoom, Google Colab, and Google Drive will help with effective team collaboration and project management. Our research endeavors will also be aided by our access to the academic publications, research networks, and library databases of our university. Finally, a contingency budget will be allocated to cover unforeseen costs like new software or increased storage requirements. Our data management plan won't be successful unless these resources are efficiently coordinated and managed.

Planned Research Outputs

Model representation - "Study of Mobile App Rivals: Harnessing Machine Learning to Decode User Reviews in the Mobile App Landscape"

Comparative study on different mobile application data by utilizing Latent Dirichlet Allocation (LDA) and 4 deep learning models namely: BERT, XLNet, CNN-BiLSTM, SVM.

Planned research output details

Title	Type	Anticipated release date	Initial access level	Intended repository(ies)	Anticipated file size	License	Metadata standard(s)	May contain sensitive data?	May contain PII?
Study of Mobile App Rivals: Harnessing Machine Lea ...	Model representation	2023-12-06	Open	Comparative-Study-of-Mobile-App-Rivals	400 MB	None specified	User Review Dataset Metadata Schema	Yes	No