DSO 560 NLP

Google Play Store Apps Reviews Analysis

Contents

1

Overview

Business Problem - Dataset Review - Preprocessing

Overview – Business Problem

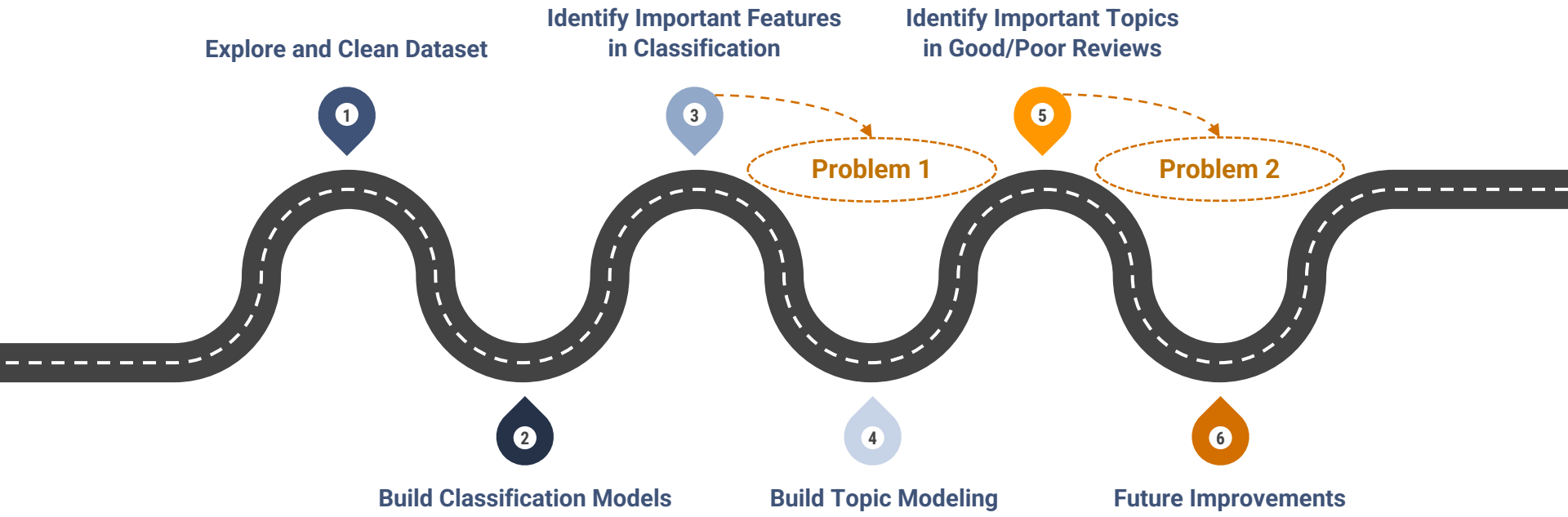
1. How to classify a review as good/poor only based on the text?
   Which features matter?

   → Classification Modeling

2. Why do users give good/poor reviews?
   How can the Apps improve?

   → Topic Modeling

Overview – Roadmap



Overview – Dataset Review

Dataset Introduction:

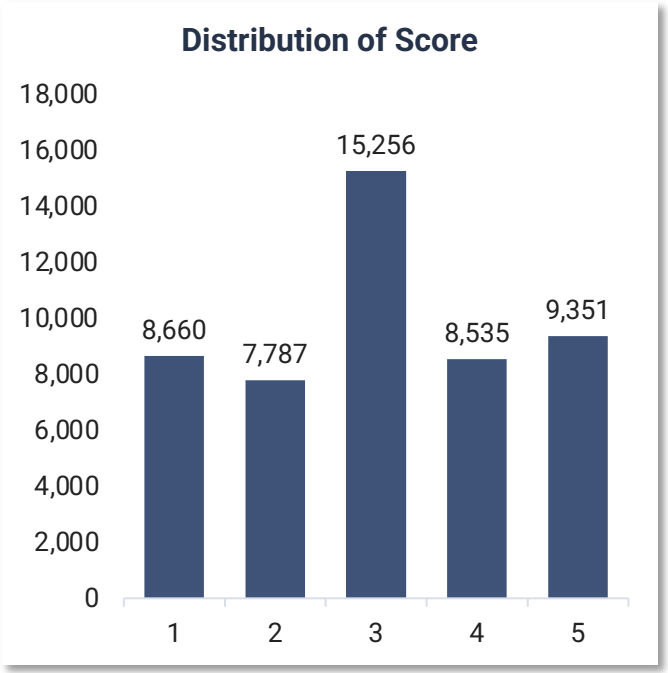
This dataset contains more than 50k reviews of Apps from Google Play Store which are categorized into Browsers, Video Players, File Managers, Mobile Payment, and Communication.

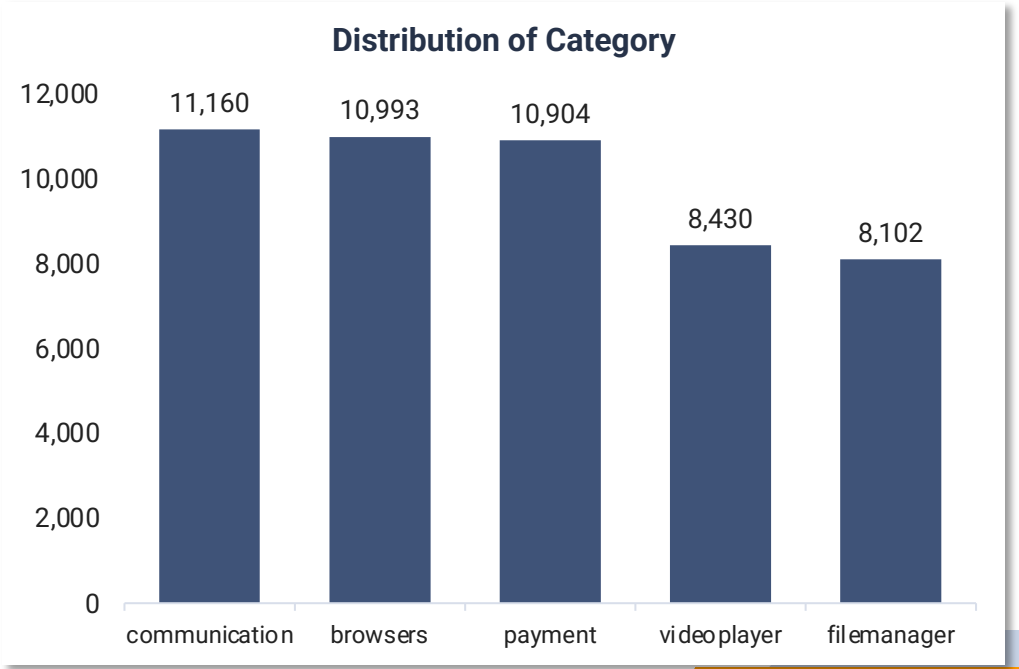
Important Features

- **Content:** the reviews of the Apps.
- **Score:** the corresponding rating of the review. (Values: 1 - 5, 1 is the lowest, 5 is the highest)
- **Category:** Apps belonging to the categories. (Values: browsers, videoplayer, filemanager, payment, and communication)

Overview – Dataset Review

The distributions of categories are similar, hence we can analyze each category respectively;



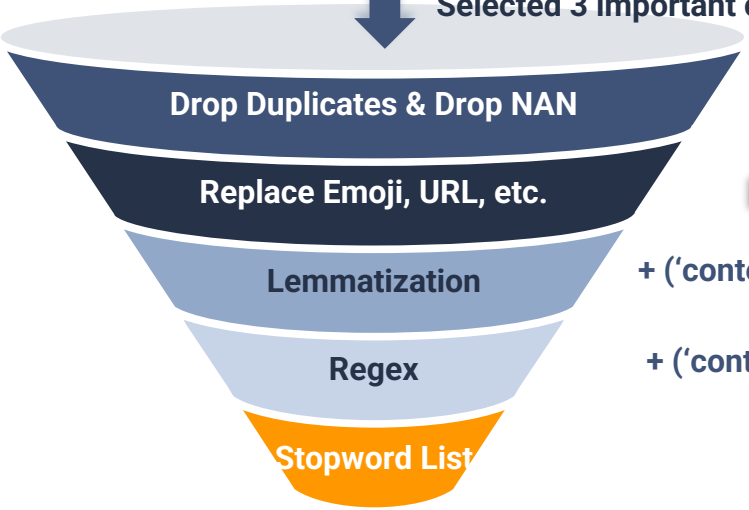


Overview – Preprocessing

Original Dataset

(55744, 13)

Selected 3 important columns



+ ('content_lemma')

+ ('content_regex')

(49589, 3)

Use Textacy

Deal with "n't" words before lemmatization - (49589, 4)

(49589, 5)

Remove "n't" words
Append some punctuations

Why do we keep "n't" words?

The "n't" words will change the sentiment of a review.

Prepared Dataset

(49589, 5)

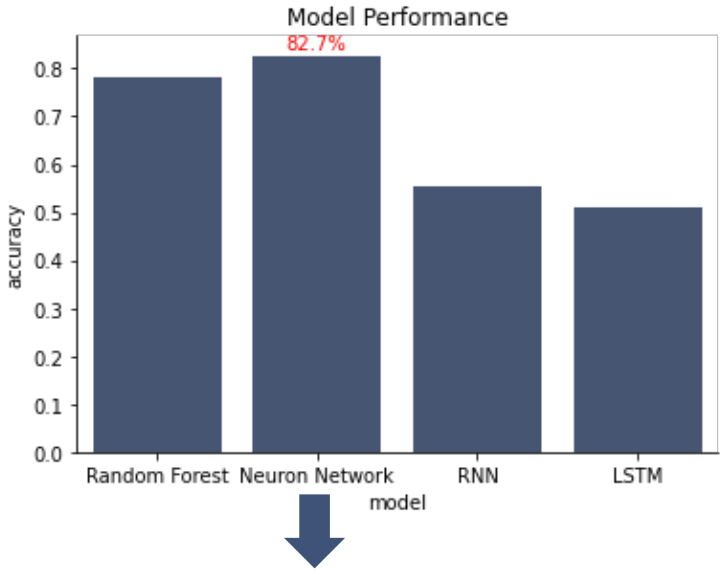
2

Classification Modeling

Deep Learning Models – Feature Importance

– Problematic Reviews

Classification Modeling – Deep Learning Models



Confusion Matrix	Predicted Positives	Predicted Negatives
Actual Positives	1,381	351
Actual Negatives	273	1,598

- **Best Result:**

A simpler deep learning architecture has the best performance with 82.7% accuracy, **outperforming** a brute random guess significantly by **32.7%**.

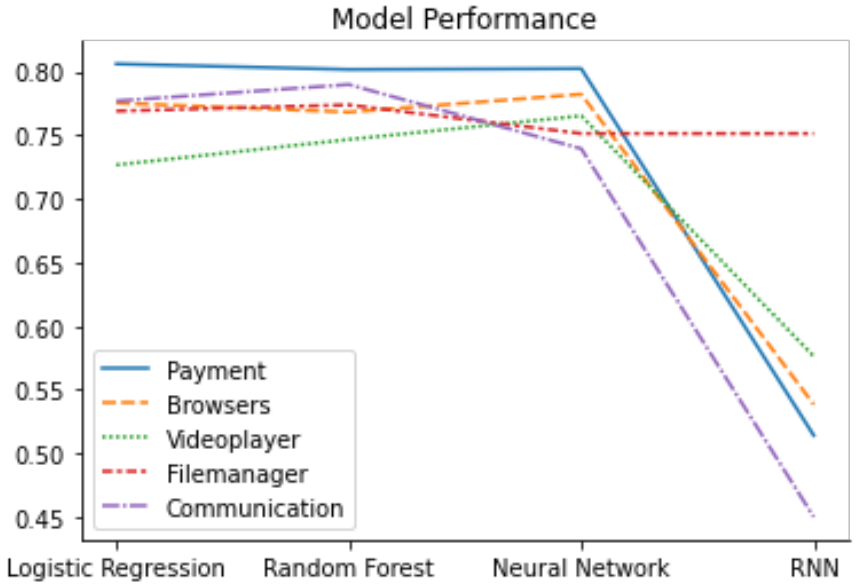
- **Models Comparison:**

Random Forest and **Neural Network** perform much better than deep learning architectures with more complexity, i.e. RNN and LSTM, which do not perform so well in our sentiment analysis of app reviews.

- **Precision-Recall Balance:**

The confusion matrix returned by this model indicates a fairly good balance between precision (79%) and recall (78%).

Classification Modeling – Neural Network



Model Architecture of Neural Network

This model was established using our own trained TF-IDF vectorization embedding with **N-gram = (1,3)**.

The model uses a sigmoid function to output the probability of a positive/negative review *(score of 5 being **positive** and score of 1 being **negative**)*.

Model Hyperparameters:

Vocab Size = 5,000, **Max Length** = 493, **Embedding Size**=50

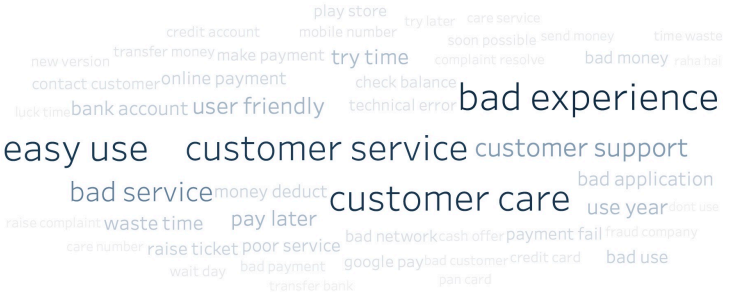
Layers: Embedding (Our Own Trained TF-IDF Vectorization), **Flatten**(), **Dense** (1 Unit Sigmoid)

Classification Modeling - Feature Importance

Overall

- Using the whole dataset, the most important features are "easy use", "please fix", "bad experience", "not work" and so on.
- In general, the success of an App is highly related to whether it is easy to use and whether there are some bugs in it.

Payment



- Using the data of payment Apps, the most important features are "customer service", "easy use" and so on.
- For payment Apps, the company should focus more on the customer service and easiness

Classification Modeling - Feature Importance

Communication

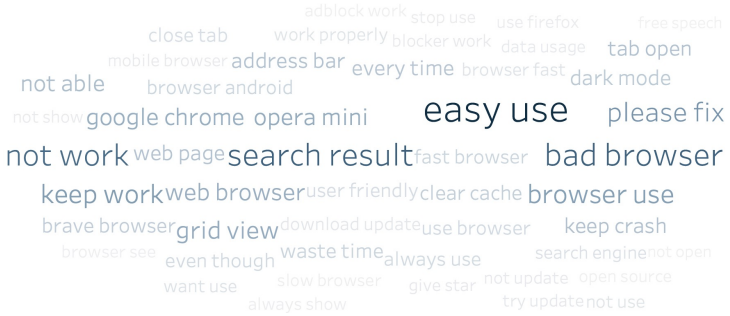
- Using the data of communication Apps, the most important features are "please fix", "good", "easy use" and so on.
- For communication Apps, the companies should focus more on fixing bugs and easiness of their Apps.

File Manager

- Using the data of file managers, the most important features are "last update", "good file manager", "nearby share" and so on.
- For file manager Apps, the company should focus more on the update of their Apps and the function of nearby sharing.

Classification Modeling - Feature Importance

Browsers



- Using the data of browsers, the most important features are "easy use", "search result", "bad browser", "grid view" and so on.
- For browser Apps, the company should focus more on easiness and improving search result of their Apps.

Video Player

- Using the data of video players, the most important features are "many ad", "not work", "not support" and so on.
- For video player Apps, the company should focus more on the support for varies format of videos and find the balance between user experience and the number of ads.

Classification Modeling - Problematic Reviews

Spurious Negative

"Why the search page isn't opening in the first go after the last update? User have to search the same thing twice unnecessarily."

Score: 5.0 (Positive) **Prediction: Negative**

Analysis - Wrong Given Score

Our model **successfully** captures the negative sentiment of this review which has a spurious 5.0 score. The true sentiment is inconsistent with the actual score given. **Thus, our model can be applied as a complementary testing tool to the existing rating system.**

Spurious Positive Prediction

*"Generaly speaking, Whatsapp is very **helpful and user friendly**... But there is a major drawback of Whatsapp..... There is no "Edit " option... So once you post your message in Whatsapp, the content of the message becomes permanent... Even if there is some spelling mistakes, you cannot correct it anymore .... But Telegram has edit option. So Telegram has an edge over Whatsapp... So if Whatsapp wants to retain its users, Edit option should be there in future upgrade of Whatsapp."*

Score: 1.0 (Negative) **Prediction: Positive**

Analysis - Wrong Prediction

There are important tokens such as 'helpful' and 'user friendly' in the review, which show frequently in our good reviews and make our model predict it as positive. But the overall signal of the review is negative since the user is much more concerned about the drawbacks of Whatsapp.

3

Topic Modeling

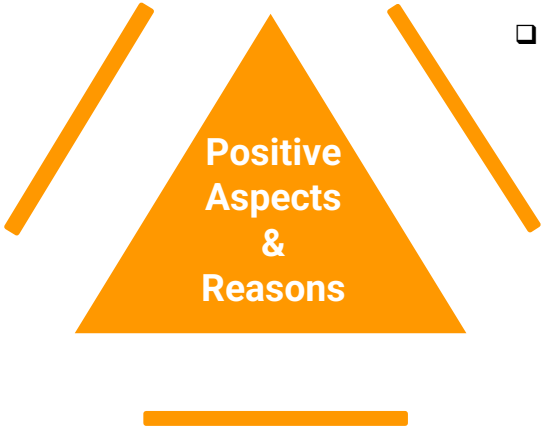
Business Insights from Topic Modeling


What is
the
Business Insights?

Topic Modeling – Payment (Positive Reviews)

Bank Money Transfer

- Improve automatic money transfer features and remind customers to set up regular frequent transactions
- Cooperate with more banks for wider coverage of the banking system
- Decrease money processing time by enhancing telecommunications with banks



Easy Features & User Friendly Interface

- Stick to intuitive and simple design styles and user-friendly interfaces

Sending Money

- Use fraud analytics algorithms to detect transaction anomaly and ensure transaction safety
- Expand identity verification methods including username, phone numbers, email, etc.

Topic Modeling – Payment (Negative Reviews)

Customer Service

- Ensure the accessibility of customer service by increasing customer hotlines and online receptionists
- Establish a formal procedures to handle customer complains and process refund requests within a specific timeframes (such as in 3 days)

Bank Account Update

- Adjust transaction fee according to the estimated time between payment applications and bank accounts
- Add email and text notification features to notify customers of the current stage of transactions

Pay Later Feature Failure

- Conduct credit scoring for profile screening and use credit metrics as reference of transaction approvals
- Increase the options of payment dates so customers can choose to repay on different dates

Technical Problems

- Monitor technical issues on a daily and weekly basis,
- Analyze the cause of these issues through daily transaction data and customer complaints and debug

Topic Modeling – Browser (Positive Reviews)

Easy To Use Features

- Streamline features that are less likely to be used and keep user interface neat and clear
- Highlight no tracking and automatic deleting browsing history options to ensure users of data privacy and security

Dark Mode & Data Saving Mode

- Design more dark mode options and improve dark mode layout for dark modes to increase visibility in different environments
- Set up an automatic reminder feature for users to transfer to data saving modes when the battery is below a certain threshold

Extension Support

- Test the compatibility with different browser extensions on a regular basis
- Improve the functionality of high-frequency browser extensions

Desktop Versions

- Adjust browser features according to different platforms to customize different platforms
- Improve the automatic syncing features of bookmarks and browsing histories

Topic Modeling – Browser (Negative Reviews)

Update Failure

- Remember user preference of automatic software update
- Avoid automatic update when users reject the options
- Report update failure cases, analyze the cause of failures and debug accordingly

AdBlock Failure

- Monitor and test the AdBlock features to check AdBlock performance on different websites
- Customize AdBlock features according to different websites and dynamically adjust the features

Forced Grid View

- Increase view options for users to choose between cascade view and grid view
- Survey and research user experience preference before conducting software updates

Topic Modeling – Video Player (Positive Reviews)

No ad

- Provide video players without intrusive ads
- Create a non-ad version for users to choose from

Many Feature

- Provide option to switch between languages for dual audio vids
- Provide features similar to PC version to increase the loyalty of existing PC users
- Add feature of pan zoom which users can freely zoom video
- Customize number of seconds to skip
- Customize audio

Support Multiple Format

- Support different formats of video such as MP4, MOV, FLV, AVI
- Support Android TV and other smart TV

Good for Movie Watching

- Ensure movie can load in quickly
- Provide automatic recognition subtitles in different languages
- Provide 4K version for video to improve the user experience in movie watching

Topic Modeling – Video Player (Negative Reviews)

Crashing in Latest update

- Ensure new version work properly before launching new version
- Fix issue like (1) when forward the video, audio doesn't stop (2) screen turn off when video is playing (3) OTG connection problem

Too many ad

- Reduce the length of ads or even provide a non-ad version to increase the user retention

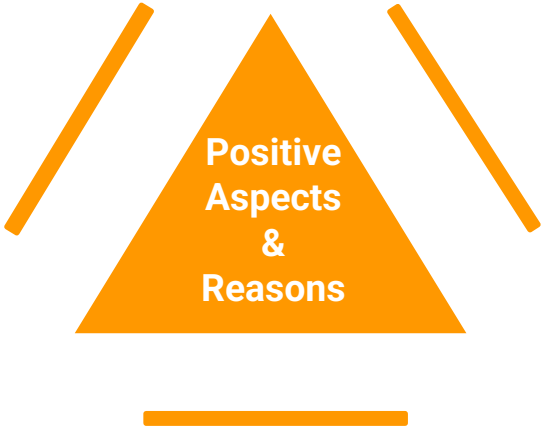
Format Not Support

- Support more video format such as MPEG4, HEV, OPUS, HEVC
- Support 4K video which is an increasing need
- Add more audio format like AAC, EAC3, AC3

Topic Modeling – Communication (Positive Reviews)

Video Call

- Support video calls in communication app, which is an increasing need
- Support group video call for a larger number of users instead of only 4
- Available across all platforms such as Android, IOS, windows, etc.
- Ensure the video call is clear, smooth, and fast



Easy to Use

- Put common features in a prominent place
- Provide user tutorials to new users
- Keep consistence when updating to new version

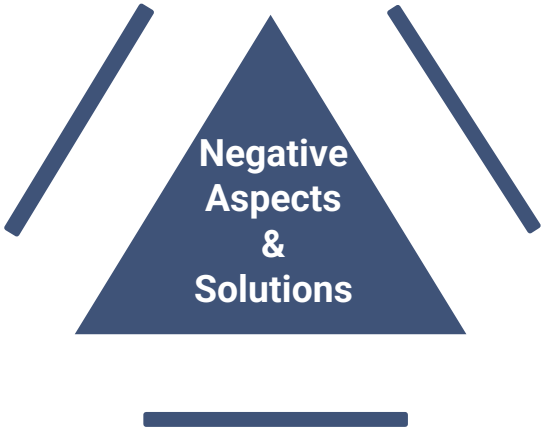
Status Option

- Provide status options for users to show their real-time status

Topic Modeling – Communication (Negative Reviews)

Video Call Problem

- Provide minimize square frame for video while in chatting page
- Fix video call auto rotate issue, provide manual rotate option
- Ensure smooth switching to cellular data when WIFI is disconnected
- Reduce battery consumption for video calls



Phone Number

- Improve security and privacy by owning users' phone number
- Allow users to sign up using email if the PC version does not require a phone number
- Provide automatic identification of phone numbers to enable users copy & paste phone numbers easily
- Provide an appropriate way to solve the "phone number already exists" problem

Message Sending Issue

- Provide option to send message to multiple users at one time
- Improve message sending speed
- Show specific reason for the issue when users couldn't send the message successfully

Topic Modeling – File Managers (Positive Reviews)

Easy to use

- Design a user-friendly and easy-to-use interface
- Highlight the most important features for users' daily use
- Improve constantly the user interface design

No ad

- Ensure app service quality without intrusive ads
- Find the balance between user experience and the number of ads to increase the profit

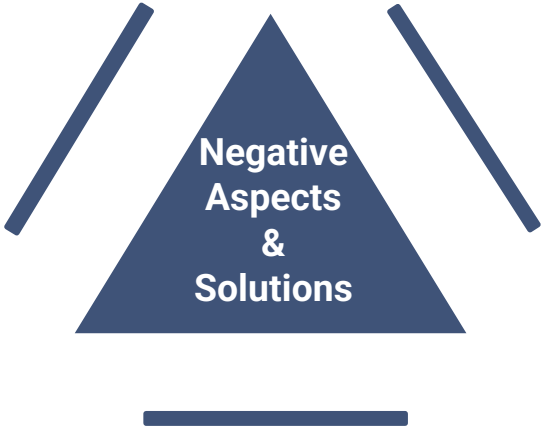
Free of Charge

- Continue to provide free file manager service to users

Topic Modeling – File Manager (Negative Reviews)

Run Slowly

- Ensure the file preprocessing performance after users use it for a long time
- Fix the problem of crashing after long-time use
- Optimize the Apps to enable user use less memory and run faster



Update Problem

- Combine the adjustment in features in several updates to reduce the number of updates

SD Card Issue

- Fix the problem of reading and writing the file from the SD card
- Ensure that the apps can detect various types of SD cards

4

Future Improvement

Future Improvement

Data Cleaning

Clean reviews more thoroughly and reasonably to achieve higher accuracy

Modeling

Try more models such as transformer and BERT to seek better classification results

Do oversampling based on the ‘thumbupcount’ to consider the impact of thumb-ups in each review

Feature Engineering

Explore how to Choose a better topic size; try to avoid the same words appearing in different topics

Topic Modelling

THANKS!

Any questions?

You can find us at

zoux@usc.edu/