ANDROID PLAY STORE DATA ANALYSIS

A project report submitted by

Vedasamhitha Challapalli,

B20CS078

for the design credits project in

B.Tech Summer Term

Indian Institute of Technology Jodhpur

Department of Computer Science and Engineering

May-July 2022

Declaration

I hereby declare that the work presented in this Project Report titled **Android Play Store Data Analysis** submitted to the Indian Institute of Technology Jodhpur in partial fulfilment of the requirements for the award of the degree of B. Tech is a bonafide record of the work carried out under the supervision of **Prof. Sumit Kalra.** The contents of this Project Report in full or in parts, have not been submitted to, and will not be submitted by me to, any other Institute or University in India or abroad for the award of any degree or diploma.

Vedasamhitha Challapalli

B20CS078

Abstract

The report contains the work done in the project titled Android Play Store Data Analysis. I have started the work by scraping the app details of 30 Android Business apps from Google Play store using the google chrome extension (web scraper). I have collected various details of the apps like rating, number of reviews, number of downloads and the minimum age for each app. I have done exploratory data analysis (EDA) on these apps. I also collected 1000 newest reviews of all ratings for each of the 30 apps. I have done topic modeling (LDA model) on these reviews and divided the reviews into 6 major topics. Based on these topics, I split them into technical and non-technical reviews. I have then done bug analysis for all the technical reviews based on the CWE documentation. Similar process has been followed for all the apps. The results, conclusions and bug analysis of the findings have been put forth in this report.

Github repository: file

Keywords:

Data Analysis, Google Play Store, EDA, Topic Modeling, LDA, Topics, Technical, Non-technical, Results, Conclusions, Analysis

CONTENTS

S.no	Title	Page No.
1	Data Scraping	5
2	Exploratory Data Analysis	5
3	Topic Modeling	6
4	Conclusion: Bug Analysis	22
5	Analysis	23
6	References	23

Github repository: file

1. DATA SCRAPING

(i) *Scraping App details:*

Thirty Android Business Apps were chosen from google play store and details like app name, rating, number of reviews, number of downloads and minimum age for each app have been scraped by an inbuilt scraper extension: https://chrome.google.com/webstore/detail/web-scraper-free-web-scra/jnhgnonknehpejjnehehllkliplmbmhn?hl=en.

The scraped csv file has been saved as: business-apps

The data was stored in a csv file and exploratory data analysis was done in a collab file and the corresponding graphs are put forth in the next section

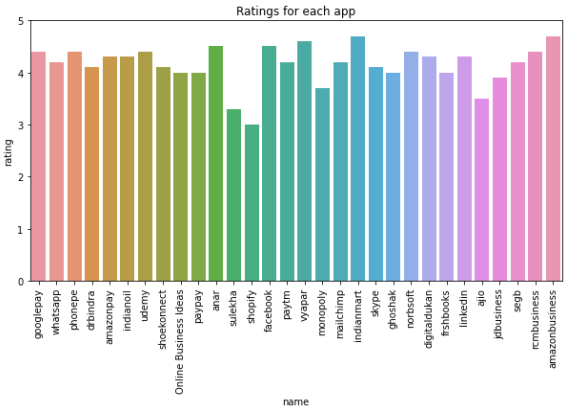
(ii) *Scraping reviews for each app:*

Thousand reviews were scraped for each of the thirty business apps. The reviews were collected based on all the sentiments (ratings) because to analyse the reviews of an app, I believed that all the ratings are necessary. I scraped this using the inbuilt python library "google-play-scraper".

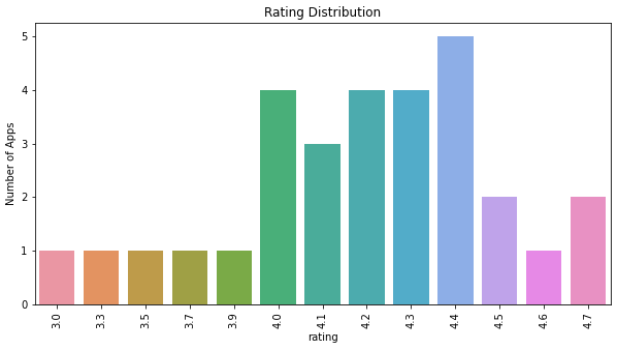
I combined all the reviews of all the apps and saved it in a csv file: reviews

I performed topic modeling on this data, specifically LDA model, split the topics into technical and non technical; and performed bug analysis on the technical issues based on the topics extracted.

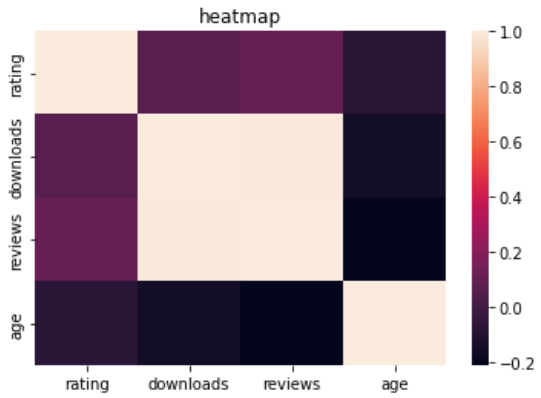
2. EXPLORATORY DATA ANALYSIS



Ratings for each app

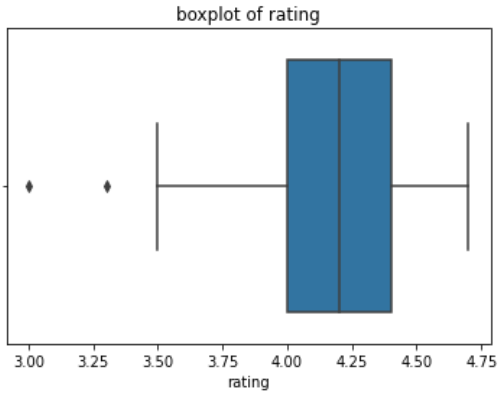


Number of apps for a particular rating (conclusion: most of the apps have rating around 4 to 4.5)



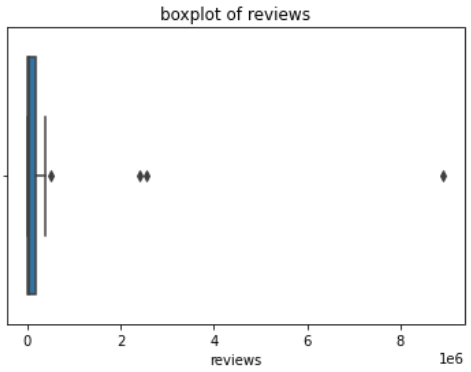
Heatmap

(conclusion: very less correlation)



Boxplot

(conclusion: most of the apps have rating between 4 to 4.5)



Boxplot

(conclusion: most of the apps have less than 1 million reviews with few outliners)

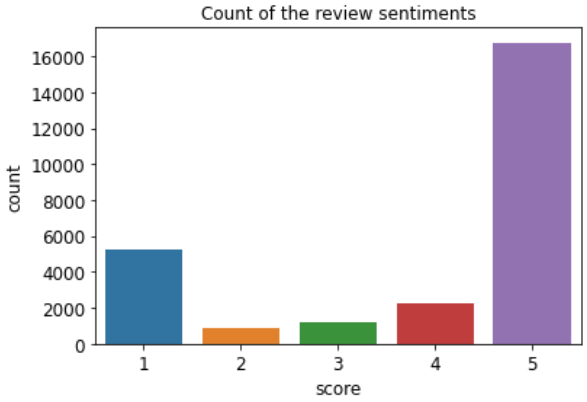
3. TOPIC MODELING

Topic modeling is recognizing the words from the topics present in the document or the corpus of data.

- LDA is a topic model that generates topics based on the word frequency from a set of documents
- WHY LDA? Because it finds accurate mixtures of topics because it is frequency based

(i) For full data (all apps together)....

SENTIMENT GRAPH:



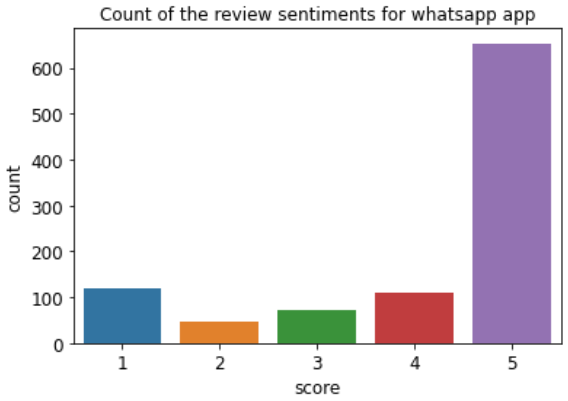
RESULTS:

* Topic 1: Not working properly, issues with updates and login **(technical)**
* Topic 2: Positive reviews (best platform, great app, thanks for the app...) **(non technical)**
* Topic 3: Like the app but issue with updates for better versions **(technical)**
* Topic 4: Price related issues **(non technical)**
* Topic 5: Recommending the app but need improvement **(non technical)**
* Topic 6: Time waste because of the fakeness of the app **(non technical)**

(ii) App wise….

(I) WHATSAPP:

SENTIMENT GRAPH:



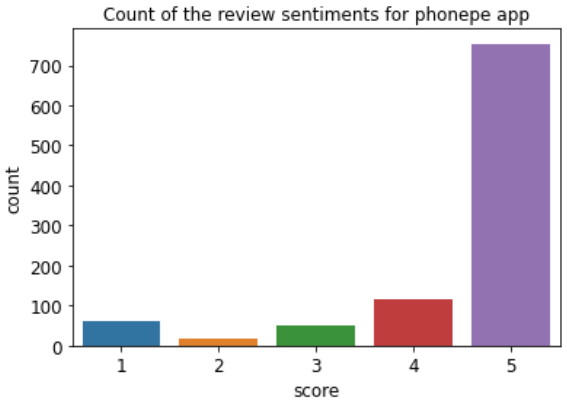
RESULTS:

* Topic 1: Love the app but installing slowly **(technical)**
* Topic 2: Easy to use but update problem **(technical)**
* Topic 3: Best app (good reviews) **(non technical)**
* Topic 4: Time taking for download **(technical, Topic 1)**
* Topic 5: Help related issues **(non technical)**

* Topic 6: Great app but problem opening a chat **(technical)**

(II) PHONEPE:

SENTIMENT GRAPH:

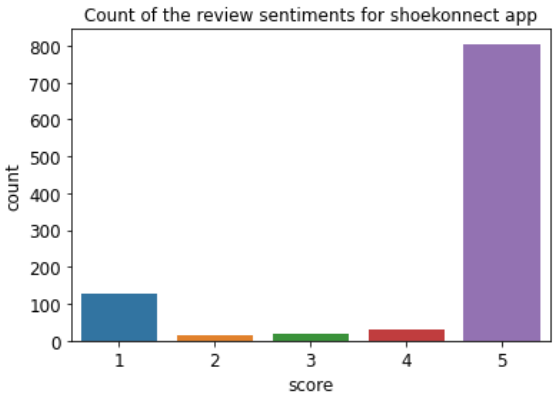


RESULTS:

* Topic 1: Positive reviews **(non technical)**
* Topic 2: Positive reviews **(non technical)**
* Topic 3: Good app but help related issues (n**on technical)**
* Topic 4: Problem in receiving payment **(technical)**
* Topic 5: Positive reviews **(non technical)**
* Topic 6: Problem with the service **(non technical)**

(III) SHOEKONNECT:

SENTIMENT GRAPH:



RESULTS:

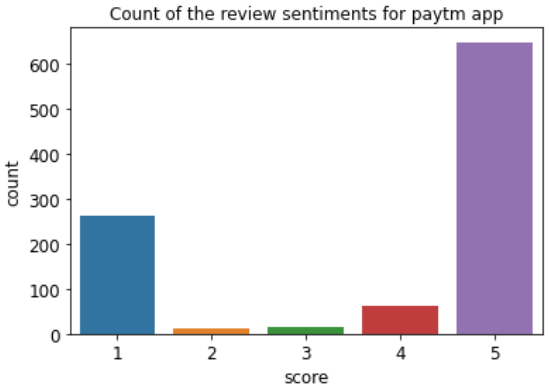
* Topic 1: Positive reviews, recommended to buy **(non technical)**
* Topic 2: Positive reviews, regarding service **(non technical)**
* Topic 3: Quality related issues **(non technical)**
* Topic 4: Positive reviews, regarding service and purchase **(non technical)**

* Topic 5: Help software related issues **(technical)**

* Topic 6: Price related issues **(non technical)**

(IV) PAYTM:

SENTIMENT GRAPH:



RESULTS:

* Topic 1: Taking time to work **(technical)**

* Topic 2: Notifications are not being received **(technical)**

* Topic 3: Easy to use **(non technical)**

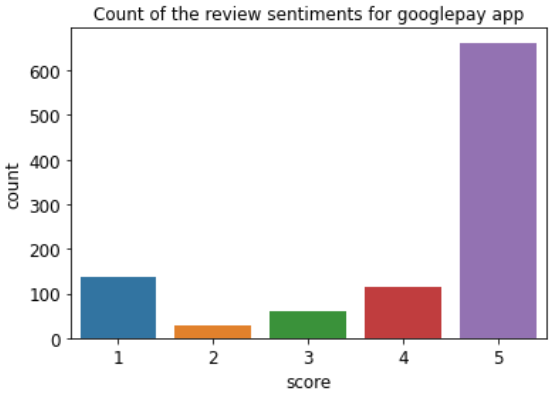
* Topic 4: Support team related issues **(non technical)**

* Topic 5: Good service but help team related issues **(non technical, topic 4)**

* Topic 6: Positive reviews **(non technical)**

(V) GOOGLEPAY:

SENTIMENT GRAPH:



RESULTS:

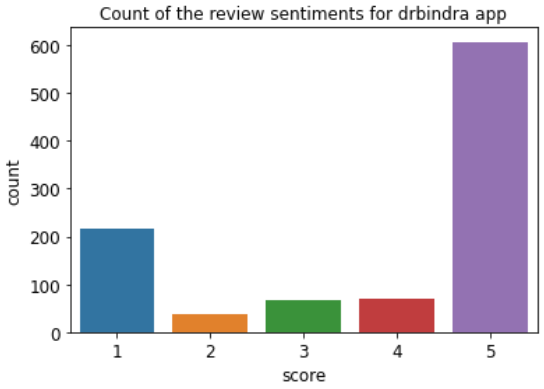
* Topic 1: Good reviews especially regarding rewards **(non technical)**

* Topic 2: Good reviews but cashback is not being received **(technical)**

* Topic 3: Positive reviews **(non technical)**
* Topic 4: Notification related issues **(technical)**
* Topic 5: Payment update related issues **(technical)**
* Topic 6: App not working properly, taking time **(technical)**

(VI) DRBINDRA:

SENTIMENT GRAPH:

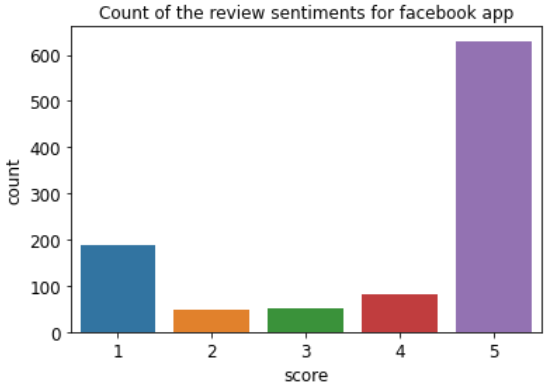


RESULTS:

* Topic 1: High data is being consumed **(technical)**
* Topic 2: Nice content **(non technical)**
* Topic 3: Video quality should be improved **(technical)**
* Topic 4: Interface related issues **(technical)**
* Topic 5: Login and update issues **(technical)**
* Topic 6: Taking time **(technical)**

(VII) FACEBOOK:

SENTIMENT GRAPH:

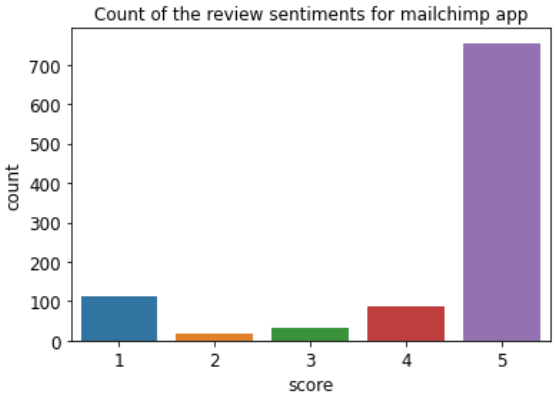


RESULTS:

* Topic 1: Reply or comment(bug) issues **(technical)**
* Topic 2: Help issues**(non technical)**
* Topic 3: Message notification being received late **(technical)**
* Topic 4: Positive reviews **(non technical)**
* Topic 5: Positive reviews regarding updated version meta **(non technical)**
* Topic 6: Issues regarding opening of updated features **(technical)**

(VIII) MAILCHIMP:

SENTIMENT GRAPH:

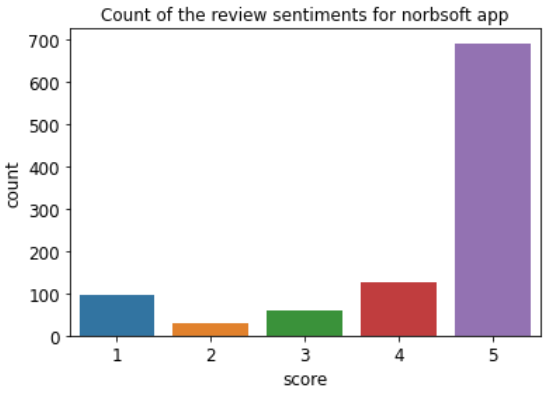


RESULTS:

* Topic 1: Positive reviews regarding experience of the app **(non technical)**
* Topic 2: Positive reviews, also requests for help **(non technical)**
* Topic 3: Positive reviews **(non technical, topic 2)**
* Topic 4: Problems using the mobile app **(technical)**
* Topic 5: Time being taken **(technical)**
* Topic 6: Positive reviews **(non technical)**

(IX) NORBSOFT:

SENTIMENT GRAPH:

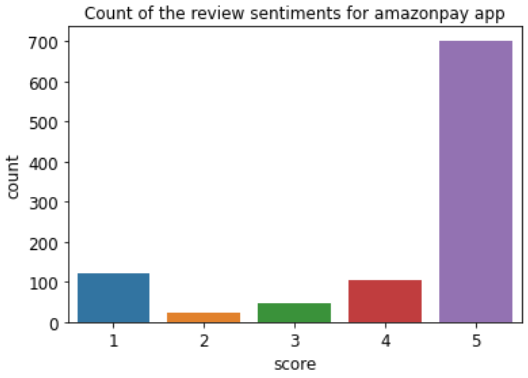


RESULTS:

* Topic 1: Error connecting properly or installing**(technical)**
* Topic 2: Slow app **(technical)**
* Topic 3: Positive reviews **(non technical)**
* Topic 4: New versions not being updated **(technical)**
* Topic 5: Download issues **(technical)**
* Topic 6: Positive reviews regarding the people's network **(non technical)**

(X) AMAZONPAY:

SENTIMENT GRAPH:

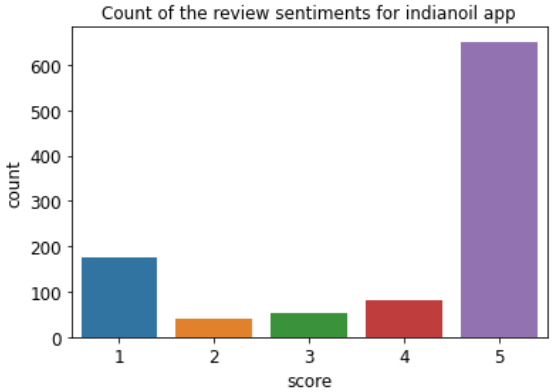


RESULTS:

* Topic 1: Good reviews especially regarding rewards **(non technical)**
* Topic 2: Voice inputs are not taken properly **(technical)**
* Topic 3: Good reviews, regarding updates **(non technical)**
* Topic 4: Good reviews especially regarding rewards **(non technical, topic 1)**
* Topic 5: Notification and help related issues **(technical and non technical)**
* Topic 6: Good reviews **(non technical)**

(XI)INDIANOIL:

SENTIMENT GRAPH:

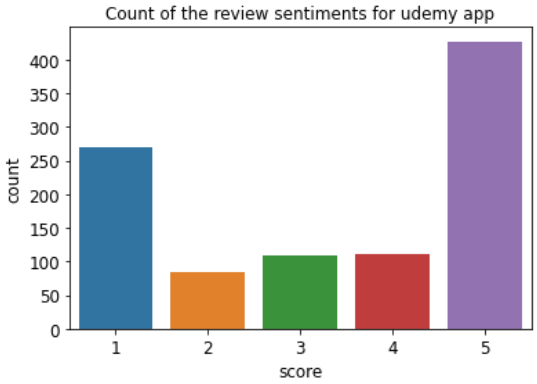


RESULTS:

* Topic 1: Super app but Download related issues **(technical)**
* Topic 2: Good app but time delayed **(technical)**
* Topic 3: Nice app but update issues **(technical)**
* Topic 4: Booking related issues **(technical)**
* Topic 5: App being hanged **(technical)**
* Topic 6: Update done slowly and also app getting hanged **(technical)**

(XII)UDEMY:

SENTIMENT GRAPH:

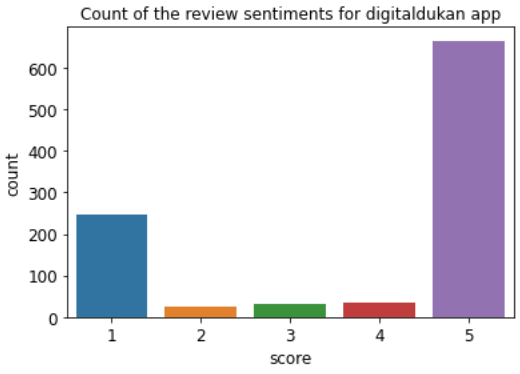


RESULTS:

* Topic 1: Good app but screenshot option needed **(technical suggestion)**
* Topic 2: Help related issues **(non technical)**
* Topic 3: Login and loading issues **(technical)**
* Topic 4: Download and update problem **(technical)**
* Topic 5: Good reviews **(non technical)**
* Topic 6: Download and loading issues **(technical, topic 4, topic 3)**

(XIII)DIGITALDUKAN:

SENTIMENT GRAPH:

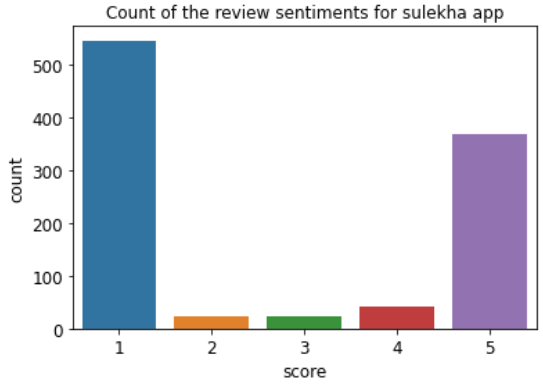


RESULTS:

* Topic 1: Good reviews **(non technical)**
* Topic 2: Time loading pages **(technical)**
* Topic 3: Downloading some feature issues **(technical)**
* Topic 4: Good reviews **(non technical)**
* Topic 5: Premium feature issues **(technical or non technical)**
* Topic 6: Fraud and fake issues **(technical)**

(XIV)SULEKHA:

SENTIMENT GRAPH:

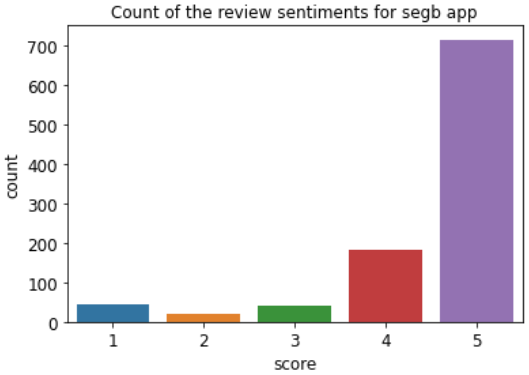


RESULTS:

* Topic 1: Good reviews but help issues **(non technical)**
* Topic 2: Refund and fraud issues **(non technical)**
* Topic 3: Time related issues **(technical)**
* Topic 4: No response to the client **(non technical)**
* Topic 5: some issue **(content not clear)**
* Topic 6: Useless investment **(non technical)**

(XV)SGB:

SENTIMENT GRAPH:

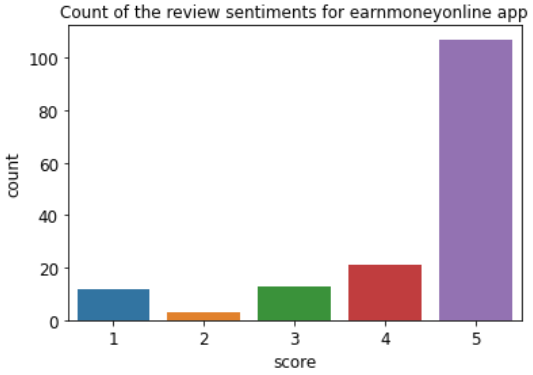


RESULTS:

mostly positive reviews

(XVI)EARNMONEYONLINE:

SENTIMENT GRAPH:

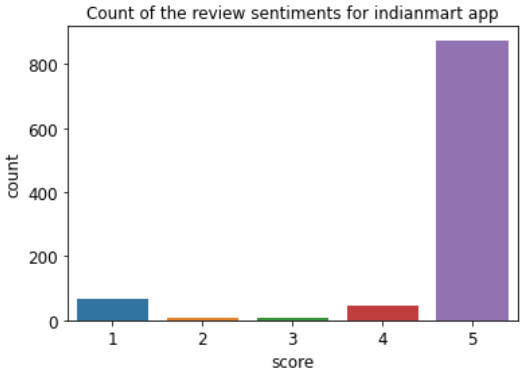


RESULTS:

mostly good reviews except help related issues (non technical)

(XVII)INDIANMART:

SENTIMENT GRAPH:

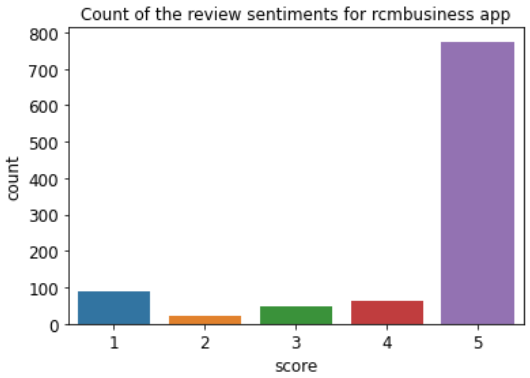


RESULTS:

* Topic 1: Easy to use app **(non technical)**
* Topic 2: Good reviews regarding service **(non technical)**
* Topic 3: Useful app **(non technical)**
* Topic 4: Good reviews but few issues regarding fakeness **(non technical)**
* Topic 5: Price related issues **(non technical)**
* Topic 6: Takes time for response **(technical)**

(XVIII)RCMBUSINESS:

SENTIMENT GRAPH:

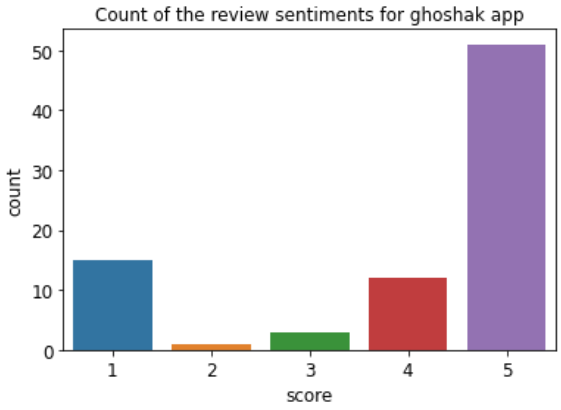


RESULTS:

* Topic 1: Good reviews regarding quality **(non technical)**
* Topic 2: Update related issues **(technical)**
* Topic 3: Time delay **(technical)**
* Topic 4: Good reviews **(non technical)**
* Topic 5: Help related issues **(non technical)**
* Topic 6: Quality improvement needed **(non technical)**

(XIX)GHOSHAK:

SENTIMENT GRAPH:

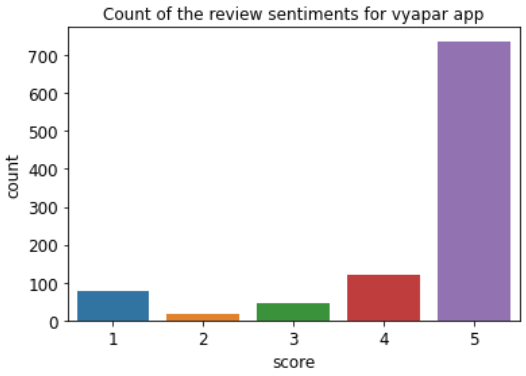


RESULTS:

* Topic 1: Good reviews with offer related issues **(non technical)**
* Topic 2: Waste of money **(non technical)**
* Topic 3: Help and support issues **(non technical)**
* Topic 4: Good reviews **(non technical)**
* Topic 5: Download issues **(technical)**
* Topic 6: Good reviews regarding service and support **(non technical)**

(XX)VYAPAR:

SENTIMENT GRAPH:

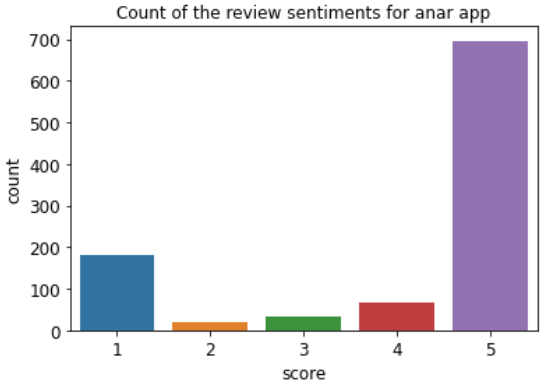


RESULTS:

* Topic 1: Good reviews regarding help team **(non technical)**
* Topic 2: Mobile version update issues **(technical)**
* Topic 3: Good reviews **(non technical)**
* Topic 4: good software **(non technical)**
* Topic 5: Invoice update issues **(technical)**
* Topic 6: Invoice and price issues **(technical)**

(XXI)ANAR:

SENTIMENT GRAPH:



RESULTS:

* Topic 1: Slowly running and download update issues **(technical)**
* Topic 2: Fake and fraud issues **(non technical)**
* Topic 3: Help feature related issues **(technical)**
* Topic 4: Good reviews **(non technical)**
* Topic 5: Good reviews **(non technical)**

* Topic 6: Price and help related issues **(technical or non technical)**

(XXII)SHOPIFY:

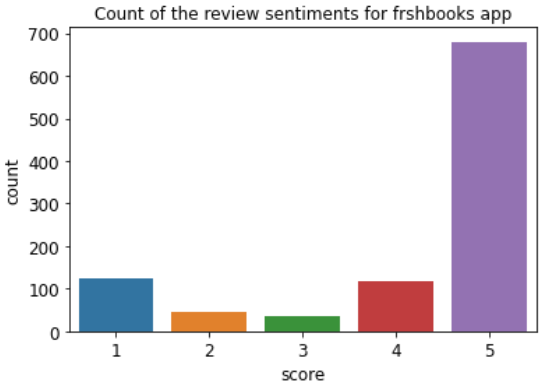
SENTIMENT GRAPH:

RESULTS:

* Topic 1: Fake product and update issues(**technical)**
* Topic 2: Fraud issues **(non technical)**
* Topic 3: Money and help issues **(non technical)**
* Topic 4: Good reviews **(non technical)**
* Topic 5: Good reviews **(non technical)**
* Topic 6: Time delay and also help team issues **(technical and non technical)**

(XXIII)FRSHBOOKS:

SENTIMENT GRAPH:

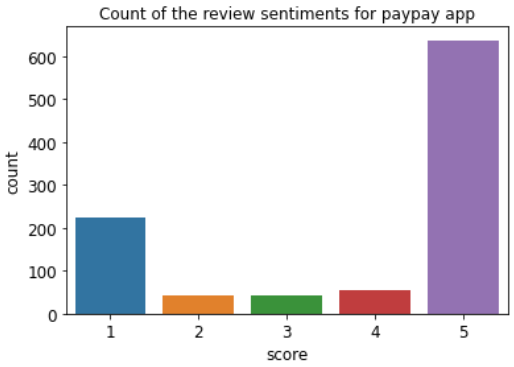


RESULTS:

Mostly positive reviews except for loading and time issues

(XXIV)PAYPAY:

SENTIMENT GRAPH:

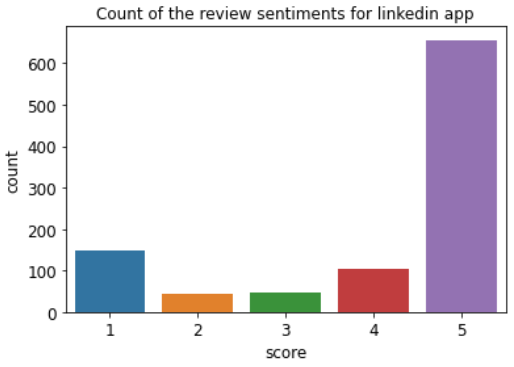


RESULTS:

* Topic 1: Good reviews **(non technical)**
* Topic 2: Mostly good reviews except for help team issues **(non technical)**
* Topic 3: Transaction or payment related issues **(non technical)**
* Topic 4: Positive reviews but Difficulty in cash transaction **(non technical)**
* Topic 5: Account opening problems and download issues **(technical)**
* Topic 6: Password and loading issues **(technical)**

(XXV)LINKEDIN:

SENTIMENT GRAPH:

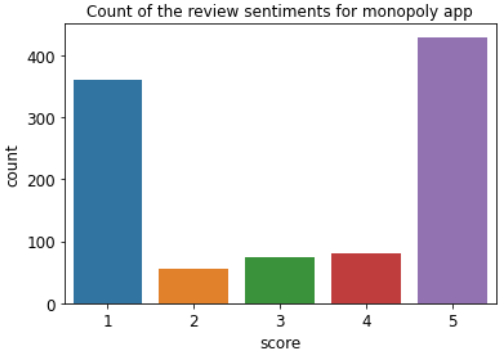


RESULTS:

* Topic 1: Best platform for job search **(non technical)**
* Topic 2: Good reviews **(non technical)**
* Topic 3: Good reviews in terms of help **(non technical)**
* Topic 4: Issues related to some restriction **(non technical)**
* Topic 5: Good reviews in terms of network and service **(non technical)**
* Topic 6: Good reviews

(XXVI)MONOPOLY:

SENTIMENT GRAPH:

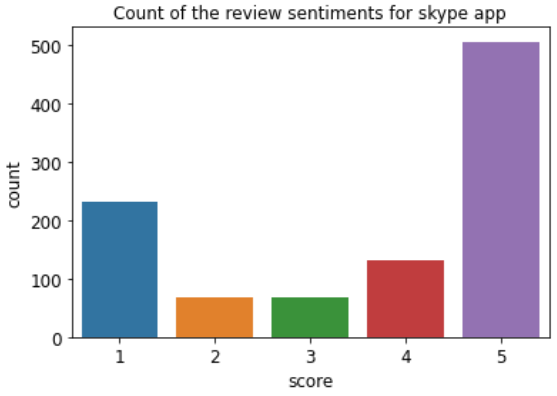


RESULTS:

* Topic 1: Network issues **(technical)**
* Topic 2: Good reviews **(non technical)**
* Topic 3: Update issues **(technical)**
* Topic 4: Development issues **(technical)**
* Topic 5: Bad app in terms of the game **(non technical)**
* Topic 6: Update and crashing issues **(technical)**

(XXVII)SKYPE:

SENTIMENT GRAPH:

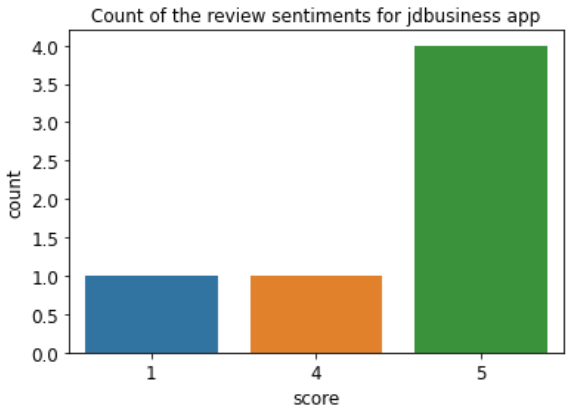


RESULTS:

* Topic 1: Update and notification issues **(technical)**
* Topic 2: Good reviews **(non techncial)**
* Topic 3: Login issues in app **(technical)**
* Topic 4: Good reviews **(non technical)**
* Topic 5: Connection issues **(technical)**
* Topic 6: Sign in and permission issues **(non technical)**

(XXVIII)JDBUSINESS:

SENTIMENT GRAPH:

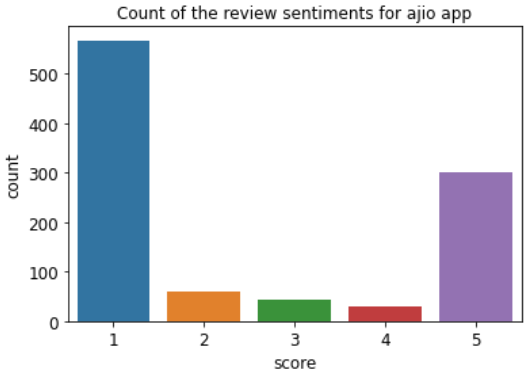


RESULTS:

Only 5 newest reviews, all of them being very vague and unclear, it is difficult to divide the reviews into topics

(XXIX)AJIO:

SENTIMENT GRAPH:

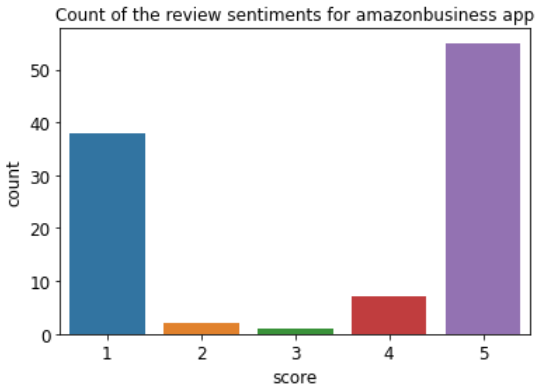


RESULTS:

* Topic 1: Delivery and return issues **(non technical)**
* Topic 2: Order and price issues **(non technical)**
* Topic 3: Registration issues **(technical)**
* Topic 4: Slow **(technical)**
* Topic 5: Quality issues **(non technical)**
* Topic 6: Login and update issues **(technical)**

(XXX)AMAZONBUSINESS:

SENTIMENT GRAPH:



RESULTS:

* Topic 1: Best experience with app but opening issues **(technical)**
* Topic 2: Opening issues **(Topic 1, technical)**
* Topic 3: Topic 2 and topic 1
* Topic 4: Not working properly **(technical)**
* Topic 5: Issues with custom status **(technical)**
* Topic 6: Custom status issues **(technical)**

4. CONCLUSION : BUG ANALYSIS

Software bugs are mostly classified based on CWE analysis.

CWE™ is a community-developed list of software and hardware weakness types. It serves as a common language, a measuring stick for security tools, and as a baseline for weakness identification, mitigation, and prevention efforts.

Most commonly available bugs are:

Login issue	Audit / Logging Errors - (1210)
Time delay	Complexity Issues - (1226)
Update issue	CWE-1277: Firmware Not Updateable
Problem opening chat (for whatsapp)	Behavioral Problems - (438)
Payment issues	Business Logic Errors - (840)
Software issues	User Interface Security Issues - (355)
Notification Issues	Behavioral Problems - (438)
Cashback issues	Resource Management Errors - (399)

Login issue	Audit / Logging Errors - (1210)
Time delay	Complexity Issues - (1226)
Update issue	CWE-1277: Firmware Not Updateable
Problem opening chat (for whatsapp)	Behavioral Problems - (438)
opening issues	Authentication Errors - (1211)
app not working properly	Behavioral Problems - (438)
Registration issues	Authorization Errors - (1212)
network issues	Signal Errors - (387)
fake and fraud issues	CWE-295: Improper Certificate Validation
crashing issues	CWE-248: Uncaught Exception
Permission issues	Permission Issues - (275)

5. ANALYSIS

Most of the reviews were good on the whole but there are few bugs in every app. Most of the apps had update issues and time delay issues, which means that the most common errors are Complexity Issues - (1226) and Complexity Issues - (1226). Most of the payment apps had Payment errors, so by default they had Business Logic Errors - (840) error. Also, quite a few apps also had crashing error: CWE-248: Uncaught Exception. Some uncommon errors are opening and login issues, also network errors and permission errors were also present in a few apps.Most of the social media apps like facebook and linkedin have update errors, so CWE-1277: Firmware Not Updateable is the most common bug in them.

Github repository: file

6. REFERENCES

[1]https://monkeylearn.com/blog/introduction-to-topic-modeling/#:~:text=Topic%20modeling%20is%20an%20unsupervised,characterize%20a%20set%20of%20documents.

[2]https://en.wikipedia.org/wiki/Topic_model

[3]https://towardsdatascience.com/topic-modeling-and-latent-dirichlet-allocation-in-python-9bf156893c24

[4]https://towardsdatascience.com/end-to-end-topic-modeling-in-python-latent-dirichlet-allocation-lda-35ce4ed6b3e0

[5]https://www.analyticsvidhya.com/blog/2016/08/beginners-guide-to-topic-modeling-in-python/

[6]https://www.toptal.com/python/topic-modeling-python

[7]https://ourcodingclub.github.io/tutorials/topic-modelling-python/

[8]https://griddb.net/en/blog/topic-modeling-with-lda-using-python-and-griddb/