Exploratory analysis of Brazilian mCommerce users to create a recommendation system based on content similarity

Mariana de Oliveira Goncalves Rodrigues[1]; Prof. Ms. Fernando Freire Vasconcelos.

[1] Fließstraße, 14, Niederschöneweide; 12439, Berlin, Berlin, Germany
[2] University of Sao Paulo. Advisor USP/Esalq /Doctoral student FEA-USP. University of Sao Paulo – Prof. Luciano Gualberto Ave., 908 – Butantã; 05508-010.

Exploratory analysis of Brazilian mCommerce users to create a recommendation system based on content similarity

Summary

The objective of this study was to analyze Brazilian mCommerce users to develop a mobile application recommendation system, with the goal of increasing the visibility of lesser-known apps among Android users. The study utilized demographic data, investment figures, reviews, and information on the total number of organic and paid downloads from the general ranking of shopping category apps in Brazil. This was to conduct a demographic survey to better understand user behavior. Furthermore, the research revealed into the correlation between high advertising investments and app popularity, proving that the 10 most downloaded apps also had over 50% of the total investments in paid traffic. Using the API from the world's largest mobile data platform, data.ai, it was possible to extract this critical data, allowing for comprehensive analysis. By utilizing content similarity techniques, such as the TF-IDF matrix, the system was able to successfully recommend similar apps. It was concluded that the developed recommendation system was a useful tool for increasing the visibility of lesser-known apps and allowed users to discover new app options.
Keywords : Recommendation system, TF-IDF matrix, content similarity, mCommerce demographic analysis, Python, app recommendation, investment correlation.

Introduction

It's hard to imagine the modern world without accessing your smartphone at least once, whether for communication, research or to stay informed about recent events. Since 2007, when Apple launched its first smartphone mobile device, the iPhone, and Google announced its operating system, Android, the market has continued to grow and open up new possibilities.

According to research carried out by the Pew Research Center (2016), the number of smartphone users soared in 2013, reaching more than 25% growth in countries emerging countries such as Brazil, Chile and Malaysia; also accelerating the use of the internet, which reduced the gap between developed and emerging countries.

Between 2013 and 2021, the smartphone market in Brazil grew to the point of placing the country in fifth place in the number of users in the world. The year 2020 closed with 234 million mobile phone accesses, which represents an increase of 7.39 million accesses compared to the same period in 2019. The COVID-19 pandemic was one of the probable factors for the increase in accesses, since there was a transfer of in-person activities to the online environment during this period. (ANATEL, 2020).

According to the research Economic and social impact of Android in Brazil (MOURA, Lívia, CAMARGO, Gustavo, 2020) carried out by Google in partnership with the global consultancy Bain & Company, Android is present in more than 90% of smartphones in the country, being responsible for the democratization of internet access in Brazil, rising from 41% to 70% between 2010 and 2018. During this period, 24 million Brazilians accessed the internet for the first time. This is mainly due to the free nature of this operating system and also to the fact that it is open and free of charge, which allows its free use by many smartphone manufacturers. Another factor for Android's sovereignty in Brazil is the high price of devices from its main competitor: Apple.

The annual “State of Mobile 2022” study by the Data.ai platform, the leading tool for application data analysts, indicates that in 2021, around 2 million new applications were launched, of which 77% were for Google Play. In 2020, a total of 230 billion downloads were made globally, generating revenue of $170 billion in transactions made within applications. Although it is a global market with many developers, only 233 apps have accumulated revenues of over $100 million, with only 13 exceeding the value of $1 billion. Brazil was highlighted in the study as it reached the 4th country with the highest volume of downloads and the 14th in revenue, in addition to being the one that spends the most time on mobile phones, exceeding 5 hours a day.

Regarding advertisers, the study also revealed some important figures. With so many options and a vast market, the competition for user attention is large, involving billion-dollar

investments in advertising, which are close to $295 billion dollars in 2021. There are expectations of exceeding $350 billion dollars globally in 2022.

Appsflyer, a well-known app attribution and analytics platform, has also released its eCommerce- focused study called " State of the Game." of eCommerce (2021 *)"* , which points out that, in relation to e-Commerce Apps alone, global spending on ads for user acquisition was $5.4 billion between the last quarter of 2020 and the first quarter of 2021. Behind only the United States, Brazil represents the second largest investing market and the largest market on Android, with 19% of the total installations of shopping apps in the world.

Data.ai platform and presented in this paper show that large technology conglomerates are among those that have invested the most in advertising, such as, for example, the Chinese companies *Shopee* , *Shein, the* Argentine *MercadoLivre* ; and the American Amazon. Among the Brazilian companies are MagazineLuiza and Lojas Americanas. By analyzing the ranking of the most downloaded applications, it is possible, in principle, to make a correlation between the largest advertisers and the applications with the largest volumes of downloads.

This research aims to analyze this correlation and the profile of Brazilian m- Commerce users , and thus generate a recommendation system using content similarity for applications that are not among the top 10 most downloaded mCommerce apps on the Android system. For this purpose, demographic and dimensional data are used with the aim of generating greater visibility.

Material and Methods

This research collected data provided by the Data.ai platform, one of the largest platforms used by analysts and marketing professionals focused on the app market, for the development of exploratory analysis of mCommerce users in Brazil and for the recommendation system. After collection, the data was pre-processed to fit the necessary analysis standards, removing null and incomplete values. Finally, the recommendation system was generated based on content similarity.

1.1 Data Collection

Data collection was done using the SQL language, through the Data.ai API (application programming interface) , which are stored in the cloud in the “Snowflake” program (cloud computing company).

The research was divided into four distinct samples, all related to apps in the shopping category and available for the Brazilian market through the Google Play store. The data collected is from July to September 2022.

The first sample, called the General Sample, contains data related to the ranking of applications, description, number of active users, total number of downloads, and information about the number of downloads generated by each paid media. The sample contains a total of 347 observations and 44 variables.

The second sample is related to demographic data, age range and gender, totaling 2101 samples across 344 different applications.

The third sample included a total of 19,092 observations with user reviews and ratings ranging from one to five, conducted between July 16 and 22, 2022. These ratings were made for the applications that occupied the top three positions in the e-commerce category in the same period, which are, respectively, Shopee, Shein and Mercado Livre.

The fourth sample includes data on advertisements made by each application between July and September 2022, totaling 389 observations, with information on the number of advertisements made by the applications and the number of platforms used for advertisements.

1.2 Preprocessing

Jupyter Notebooks" tool was used together with the Python language and libraries such as Pandas and Numpy to perform an initial screening of the data, thus excluding those that would not be necessary for any analysis.

In general, texts in their natural form are not well formatted or standardized. Pre-processing involved using several techniques to convert the text into a sequence of well-structured linguistic components so that other NLP systems and applications can interpret them correctly (SARKAR, 2016). During the process, comments with HTML links, repeated evaluations, stop words (words of low semantic relevance) were removed and, in order to concentrate the sample on adjectives, verbs were removed, using the Python package called Spac.

To analyze the advertising investment data, we also used the paid download data collected from the General Sample. When combining the two samples, some null values were found, which were discarded, resulting in 317 observations. In addition, the values of the three months were added for each observation, resulting in a total of 118 observations:

Table 1. Total estimated advertising spend of top 10 apps

Ranking Position	Application Name	Estimating Investment in User Acquisition	Number of creatives	Announced platforms	Paid downloads	Estimated number of ads
First	Shein	$ 30,598,014	129,344	992	13,659,828	44,665,746
Second	Shopee	$28,303,396	47,926	12360	12,635,445	197,838,154
Third	MercadoLivre	$8,630,543	12.216	846	3,852,921	3,257,938
Room	Magalu	$9,728,987	44,637	5129	4,343,298	87,635,892
Fifth	American	$10,255,647	35,894	624	4,578,414	7,843,272
Sixth	Amazon	$9,066,509	23,725	443	4,047,549	3,195,293
Seventh	Bahia Houses	$7,977,110	22,471	216	3,561,210	1,853,635
Eighth	Aliexpress	$7,640,223	63,842	1101	3,410,814	24,836,256
Ninth	OLX	$6,407,909	1,833	45	2,860,674	29,796
Tenth	The Apothecary	$3,173,063	3.074	20	1,416,546	14,768

Source: original research data

According to the Mobile Shopping Apps Report (2022) released by Liftoff and Singular, around 85% of advertising spend is aimed at new users who are downloading the app for the first time. The average user acquisition cost, known as the CPI (cost per install) model, is around $2.24. With the data collected, it was possible to estimate the investment value by multiplying the number of paid downloads by the average CPI value.

1.3 Defining the model for the recommendation system

One of the main techniques used is through content similarity based on TF-IDF, as it is simple to implement, efficient and widely used when there is a greater limitation in the volume of data (Muthurasu et al, 2019). Several studies have been conducted using this same technique, mainly for recommendation systems for streaming movies and series. One of them was carried out by Chiny and collaborators (2021) that used data from titles and descriptions and, unlike the streaming studied, demographic data was also considered so that a difference could be generated in the analysis, something that was also considered for the present study.

According to the "ASO Guide 2020", a survey conducted by TheTool and PickASO , two companies specialized in managing and creating tools for digital marketing professionals, one of the most used techniques to generate visibility for applications is *ASO* – "App Store Optimization *"* . *According* to the aforementioned survey, there are two factors that influence the ranking: " on metadata " *,* which are factors that can be modified by developers, such as the application title, description, genre and subgenre, and keywords; the second factor is the "off metada *",* which are the factors over which there is no control and which depend almost exclusively on investments, such as, for example, the number of active users and total downloads.

To refine the recommendation system and show relevant apps to its users, Google uses internal data from the Android device itself and from actions performed within the Google Play app, such as purchasing an app or service, search history, and installed apps. (Support Google, 2023). To develop the recommendation system, the two factors relevant to ASO were considered: the title, the app's subgenre, and the main audience, whether female or male (on-metadata); and the number of active users (off-metadata). In the latter, the values were divided into segments according to their percentile; the technique was used to convert quantitative data into qualitative data and thus use it in the recommendation algorithm.

A methodology based on content similarity was used, using the TF-IDF matrix, which stands for *Term Frequency – Inverse Document* Frequency, which consists of an effective statistical instrument for determining the relevance of terms in a text.

$$idf_t = log\ \frac{N}{df_t}$$

where t is a specific term in the document, N is the total number of documents, and df_t is the number of documents that contain the term. IDF measures the rarity of a term in a collection of documents. To calculate term frequency:

$$TF = \frac{numero\ de\ vezes\ em\ que\ o\ termo\ aparece\ no\ documento}{numero\ total\ de\ termos\ no\ documento}$$

The TF-IDF matrix aims to convert each text, also called a document, into a vector. Each component of this vector is related to a specific term and its respective value is determined by a weight, which represents the relevance of this term in the context of the document.

The weight is a combination of the frequency terms with the inverse frequency of the document to produce a different weight for each document, which can be translated into the equation below (SCHÜTZE; MANNING; RAGHAVAN, 2008):

$$tf - idf_{t,d} = tf_{t,d} \,.\, idf_t$$

where tf – idf is the tf-idf value for each term (t) and the number of times it appears in the document is equal to the frequency of the term in the document multiplied by the IDF of the specific term. Multiplying TF by IDF results in a numerical value which indicates the importance and relevance of the term in the document.

To obtain the TD-IDF matrix, the python library was used *sklearn* call TfidfVectorizer .

Results and discussion

2.1 Descriptive statistics

With a total of 240 observations after pre-processing the data, some general data were extracted from the apps in the Shopping category. These include total downloads, paid

downloads, organic downloads (which did not originate from advertising), active users, and sessions per user, which can be interpreted as a retention metric, since they indicate how many times the user opened the app.

From the standard deviation, it is possible to observe that there is a great variability between the data and the presence of discrepancies in the sample in practically all metrics, except for session per user, for which the data has a distribution closer to the average, that is, user retention is consistent among all applications, regardless of whether they are in the top 10 or not. However, the biggest difference is due to the number of downloads, especially those generated in a paid manner.

Table 2. Key metrics from overall Shopping app data

Metrics	Total downloads	Paid downloads	Organic Downloads	Active users	Session per User
Average	306,532	142.207	164,325	1,295,025	11,12
Standard Deviation	908,860	457,762	460,495	5,079,521	10.39
First Quartile	18.193	3.813	8,080	13,382	5.88
Second Quartile	46,864	18.110	26,342	40,682	8.02
Third Quartile	235,796	88,679	134,853	217,497	11.56
Maximum	8,674,918	4,553,276	4,121,642	44,589,200	100.74

Source: original research data

The total number of downloads for the apps in the top 10 is 37,462,837, which corresponds to 51% of the total, with the remaining 49% divided between 230 apps. Among paid downloads, this number rises to 53% of the total. The graph below details the number of total downloads divided by the apps in the top 10:

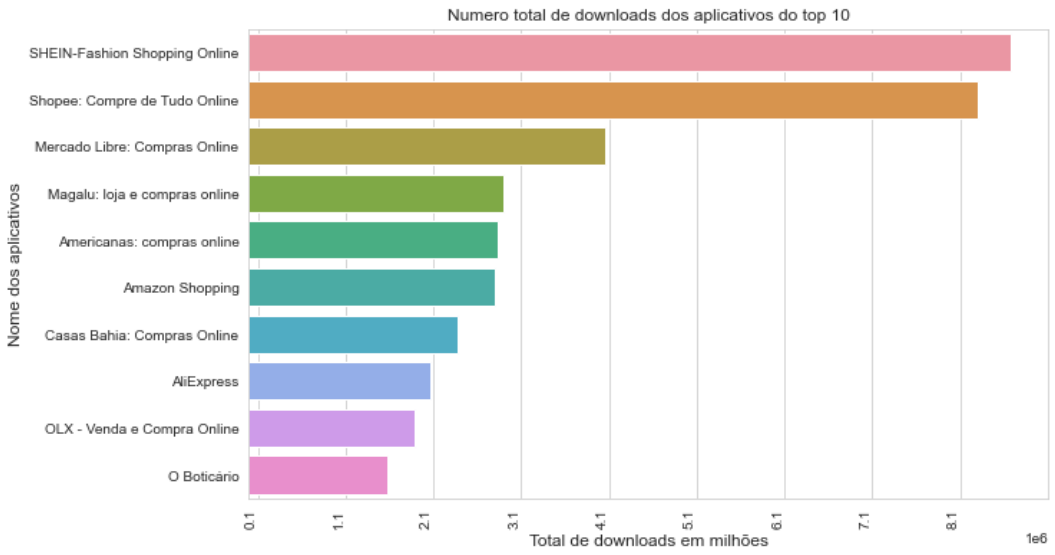

Figure 1. Total downloads of the top 10 most downloaded apps in the Google Play Shopping category
Source: Original research data

2.2 Exploratory analysis of demographic data and user reviews

An analysis was carried out to identify the main characteristics of users of mCommerce applications in Brazil in order to draw up a user profile.

The demographic data collected shows that 54.5% of users are women and 66.1% of the total are between 25 and 44 years old.

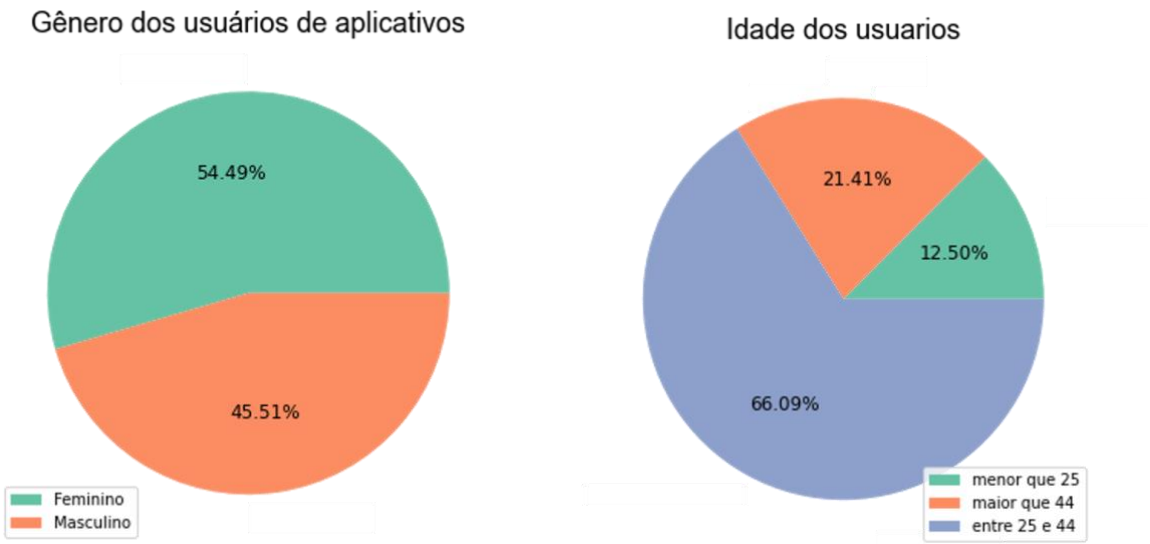

Figure 2. Gender and age range of shopping app users in Brazil
Source: Original research data

The data shows that women are more likely to use applications from the E-Commerce (Retailer) subgenre, which are generally those that do not have physical stores, only online. For men, the most used applications are also E-Commerce (Retailer), however, with a greater distribution among the other subgenres, as shown in the table below:

Table 3. Frequency of mCommerce subgenres among men and women

Subgenre	Women	M. Frequency	Men	H. Frequency
E-Commerce (Retailer)	53	50.00%	25	38.46%
E-Commerce (B2C)	24	22.64%	14	21.54%
Other Shopping	10	9.43%	5	7.69%
E-Commerce (C2C)	9	8.49%	7	10.77%
Coupons and Rewards	7	6.60%	11	16.92%
Resell	2	1.89%	1	1.54%

Pharmacy and Drugstore	1	0.94%	1	1.54%
BNPL	0	0.00%	1	1.54%

Source: original research data

The sample of user reviews and ratings provided data to better understand the level of user satisfaction and what they like or dislike most about the applications positioned among the three most popular, which are, in order: Shein (com.zzkko), Shopee (com.shopee.br) and Mercado Livre (com.mercadolibre).

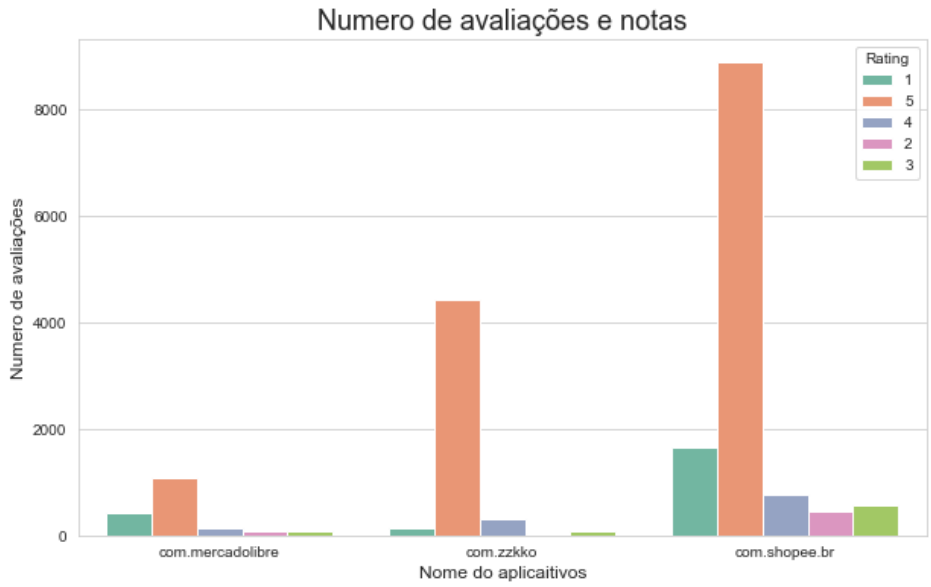

Figure 3. User ratings for the three most downloaded apps in the mCommerce category
Source: Original research data

Approximately 75% of the sample received a maximum score of 5, indicating that users are very satisfied with the services offered by the apps. The disparity in the number of reviews left by users is also notable, with Shopee , despite being in second place in the ranking, having almost twice as many positive reviews as the first-placed app, Shein. Mercado Livre also stands out as the app with the highest proportion of negative reviews compared to the others.

In order to understand what users like or dislike about apps, an analysis of comments on the Google Play page was carried out. of the three applications. The table below shows the count of the 10 most used words by users who left the highest and lowest rating.

In the negative reviews, it is possible to see that there is dissatisfaction regarding refunds, delivery and ordering . “Seller” is also a frequent word, which may justify the fact that the “Mercado Livre” application has a higher proportion of negative ratings, given that it is a *marketplace place ,* that is, there are several sellers who use the platform to offer their products. Regarding positive reviews, it is worth noting the use of words that reinforce the

feeling of satisfaction, such as “good”, “great” or “best”; the term “price” appeared several times in a positive way, indicating that it is one of the points of greatest interest to users.

Some words are repeated in both positive and negative reviews, such as “purchase”, “product” and “application”.

Table 4 and 5. Frequency of mCommerce subgenres among men and women

Positive Words	Count
Good	3154
Very	2041
Excellent	2016
Buy	1968
Product	1654
Application	1124
All	1051
Better	954
Quality	903
Price	877

Negative Words	Count
Buy	679
Product	474
No	359
Application	359
Nothing	334
Day	317
Reimbursement	292
Now	270
Delivery	241
Order	229

Source: original research data

2.3 Correlation between advertising investment and ranking

In order to quantify the correlation between high investment and positioning in a ranking, the Spearman coefficient was used , which is a measure of association between ordinal variables whose result varies between -1 and 1, with -1 being a perfect negative linear association between the variables and 1 being a perfect positive association between the variables from -1 to 1 (Fávero and Belfiore, 2017).

The correlation between ranking position and estimated investment was calculated using Spearman 's coefficient . The result was -0.93, i.e., the lower the investment, the further away from the top of the ranking the app is. The result corroborates the premise of the research, which proposes to recommend apps that have a higher investment based on the user profile.

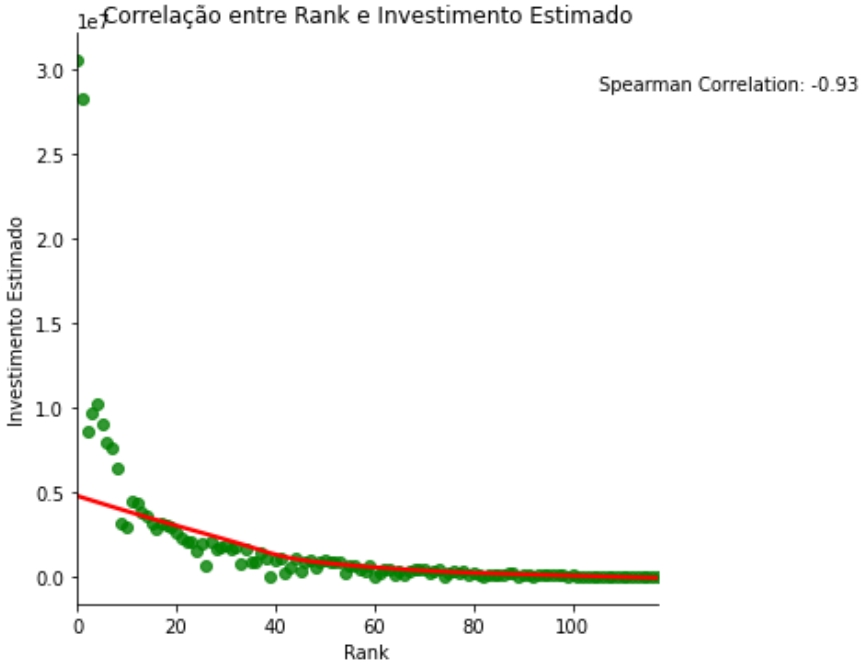

Figure 4. Graph of correlation between ranking position and investment

Source: Original research data

2.4 Recommendation model

To create the app recommendation model, we used the names of the apps available on Google Play and the information on the apps' subgenres, which are optimized to obtain a better ranking in the platform's search engine, using "ondata" ASO methods, as previously mentioned. In addition, demographic data related to the predominant gender of the users was collected to increase the accuracy of the results. Information about the number of active users was also used, using their percentile so that it could also isolate the largest apps from the smallest ones. In order to avoid duplication, an exclusion logic was added to prevent an app from recommending itself, since the highest similarity index would be between itself and another app.

A column was created in the dataframe that consists of the combination of the values mentioned above. From this column, it was possible to perform the similarity analysis between the applications and create the TF-IDF matrix. The resulting matrix has dimensions of 240x240 and presents the respective weights between the applications. For better understanding, below is a sample of the matrix as an example:

Table 6. Sample TD-IDF matrix with affinity index between applications

Application Name	Shein – Fashion Shopping Online	Shopee: Buy everything online	MercadoLibre: Online shopping	Magalu: store and online shopping	Americanas: online shopping
Shein Fashion Shopping Online	1.00	0.18	0.16	0.21	0.21
Shopee: Buy everything online	0.18	1.00	0.24	0.17	0.17
MercadoLibre: Online shopping	0.16	0.24	1.00	0.31	0.32
Magalu: store and online shopping	0.21	0.17	0.32	1.00	1.00
Americanas: online shopping	0.23	0.19	0.35	0.35	0.35

Source: original research data

A function was generated to obtain the 5 most similar applications considering the affinity index and the result; when searching for the application “Shein Fashion Shopping Online” the result was the following:

Table 7. Sample TD-IDF matrix with affinity index between applications

Ranking position	AppName	Percentile	Affinity
225	NewChic – Fashion Online	1	0.40
89	Hibobi – Kids fashion online	3	0.37
156	Light in the box – online shopping	2	0.34
88	Banggood – Online Shopping	4	0.32
222	Fashion Nova	1	0.31

Source: original research data

The result of the function gave the answers to the applications that most resemble “ Shein Fashion Shopping Online” and that are outside the top 10 most downloaded and their respective weight or, here defined as affinity, which varies from 1 to 0, with 1 being when the text is completely the same as the initial word. With this result, it is possible to offer the user new alternatives of places to make their purchases and thus further expand the range of downloaded applications.

Conclusion

Online shopping, particularly via mobile devices, has seen a significant increase in recent years. User behavior on mCommerce platforms is key for companies and developers to optimize their offerings and interfaces.

The results of the study presented highlight the user profile of shopping apps on Android devices, as well as the disparity in advertising investments for a small portion of the more than 200 apps available, which reinforces the monopoly of large companies in relation to user ranking.

ASO techniques are widely used in the *mobile marketing market* , with frequent changes to app titles and descriptions to make them easier for users to find. The TF-IDF technique allowed for the simplified evaluation of names and other variables to generate recommendations for users, allowing for a wider range of options when searching for a specific app.

It is suggested to implement new variables, such as descriptions and user reviews, increasing the amount of text and giving more context about what each application is, the applied algorithm will become more efficient considering not only the name, demographic data and number of downloads but also services provided by the applications, thus being able to bring the results even closer to the initial application.

References

National Telecommunications Agency [ANATEL]. 2020. Telecommunications sector monitoring report. Available at: <https://sei.anatel.gov.br/sei/modulos/pesquisa/md_pesq_documento_consulta_externa.php?eEP-wqk1skrd8hSlk5Z3rN4EVg9uLJqrLYJw_9INcO4NT86aq4DZSJMWh9gBoilhtRgvXnEhjT6dqYhPLeIC2xMriZOLrD6LEYnf1psEzlLJAq9-LHel_G9fbuXRs7UR >. Accessed on: October 21, 2022

Appsflyer. 2021. The state of eCommerce apps in Brazil. Available at: <https://www.appsflyer.com/pt/infograms/state-of-ecommerce-apps-brazil-2021/>. Access on : 23 Oct 2022

Chiny , M; Chihab , M, Bencharef, O; Chinab, Y. Netflix recommendation system based on TF-IDF and cosine similarity algorithms. Available at: <https://pdfs.semanticscholar.org/22fd/585a75a52264bf2f3ecaabd7a53d5a2ef465.pdf>. Accessed on 22 Mar 2023

Data.ai. 2022. State of Mobile 2022. Available at: <https://www.data.ai/en/go/state-of-mobile-2022 > . Accessed on: 28 Sep 2022

Fávero, LP; Belfiore, P. 2017. Data analysis. Statistics and multivariate modeling with Excel, SPSS and Stata. 1st ed. Elsevier Editora Ltda. Rio de Janeiro, RJ, Brazil .

Google Support . 2023. User policies for apps and digital content. Available at: <https://support.google.com/googleplay/answer/11416267?hl=en&co=GENIE.Platform%3DAndroid#zippy=%2Cdata-collection%2Cdata-types>. Accessed on April 1, 2023

Liftoff. 2022. Mobile Shopping Apps Report. Upcoming Trends in E-commerce and Delivery. Available at: <https://info.liftoff.io/pt-br/2022-mobile-shopping-apps-report?utm_source=pr&utm_medium=article&utm_campaign=DL-2022-ShoppingApps-PR&utm_content=Portuguese>. Accessed on: January 10, 2023

Moura, Livia; Camargo, Gustavo. 2020. Economic and social impact of Android in Brazil . Available at: <

https://www.bain.com/contentassets/a9200a057a0241b8963c05a9b09e33fe/impactos-do-android-no-brasil.pdf. >. Accessed on: October 4, 2022

Muthurasu, N; Rengaraj, N; Mohan, K.C. 2019. Movie recommendation system using Term Frequency-Inverse Document Frequency and cosine similarity method. Available at: <https://www.ijrte.org/wp-content/uploads/papers/v7i6s3/F1018376S19.pdf>. Access on 28 Apr 2023

Pew Research Center. 2016. Smartphone Ownership and Internet Usage Continues to Climb in Emerging Economies. Available at: < https://www.diapoimansi.gr/PDF/pew_research%201.pdf >. Accessed on: 20 Oct 2022

PickAso and TheTool.2020.ASO Guide 2020. Available at <https://thetool.io/wp-content/uploads/2020/04/aso-app-store-optimization-guide-2020-pickaso-thetool.pdf>. Access on : 11 Mar 2023

Sark ar, D. 2016. Text Analytics with Python a Practical Real-World Approach to Gaining Actionable Insights from Your Data. 1ed. Apress, Bangalore, Karnataka, India.

Schütze , H; Manning, CD; Raghavan, P. 2008. Introduction to information retrieval. Online edition . Cambridge University Press. Available at https://nlp.stanford.edu/IR-book/pdf/06vect.pdf. Accessed on: March 16, 2023