WILEY | Hindawi

Research Article
Analysis and Classification of Mobile Apps Using Topic Modeling: A Case Study on Google Play Arabic Apps

Ahlam Fuad [ID] **and Maha Al-Yahya**

Department of Information Technology, College of Computer and Information Sciences, King Saud University, P.O. Box 145111, Zip 4545, Riyadh, Saudi Arabia

Correspondence should be addressed to Ahlam Fuad; 439204463@student.ksu.edu.sa

Mobile app stores provide an extremely rich source of information on app descriptions, characteristics, and usage, and analyzing these data provides insights and a deeper understanding of the nature of apps. However, manual analysis of this vast amount of information on mobile apps is not a simple and straightforward task; it is costly in terms of human effort and time. Computational methods such as topic modeling can provide an efficient and satisfactory approach to mobile app information analysis. Topic modeling is a type of statistical modeling technique for discovering abstract topics that occur in a set of documents. This study explores the relationship between features of Arabic apps and investigates how well the current predefined Google Play app categories represent the type and genre of Arabic mobile apps. Based on the textual app description analysis, we aim to design and develop a sustainable classification system using the Latent Dirichlet Allocation (LDA) method of topic modeling in order to cover the Arabic apps classification in Google Play app store. Our study supports the hypothesis that the textual app descriptions are effective in suggesting new categories for Arabic mobile apps in Google Play app store. Also, the results indicated that the current classification on Google Play app store is not suitable for our case study "Arabic apps," as well as it is not sustainable, as it can not cover the new app types including Arabic apps. This study offers an important contribution to Arabic app analysis and design, to improve app search and exploration in several domains such as business, marketing, and technical development. Furthermore, it provides insights for the future of Arabic app research and provides guidance for the development of an Arabic app dashboard that will support users on how to select an app based on their specific needs.

1. Introduction

Mobile applications, commonly referred to as mobile apps or apps, are software programs that are designed to run on mobile devices. In recent years, the number of apps available has dramatically increased. This is due to the rapid and significant development of smart and mobile devices, as well as the availability of mobile app stores, which contain a plethora of mobile apps in different areas, including education, consumer services, and entertainment. There are various mobile app stores such as Google Play app store, Apple app store, Windows Phone store, and others that contain free and paid mobile apps [1]. Google Play app store is the official app store for mobile apps designed for mobile platforms that run on Android operating system (Android

OS). It contains information on available apps [2], where users can browse apps, read app descriptions, and download apps from the store.

The number of currently available apps in Google Play app store is approximately 2.9 million [3]. Thus, Google play app store has a big portion of 86.1% of the app store market as of 2019 [4] According to statistics published on app downloads and usage in 2018 [5], Google Play app store has a total of 76 billion downloads. The fast growth of Google Play app store is due to its dependence on Android OS, the operating system that drives app publishing [5]. Mobile app stores provide an extremely rich source of information on app characteristics and usage, which can be categorized into technical and nontechnical attributes [6]. The general description of apps contain technical information in free text

format such as app name, app size (in bytes), platform version, and app version and nontechnical information, such as category, user rating, number of installs, number of reviews, app price, and app description (free text description). This textual information can be analyzed and processed using data extraction techniques to gain insights and obtain a deeper understanding of the nature of the relationship between the technical and nontechnical attributes of apps.

Classification of apps supports the user in the exploration of the available apps. In a classification, each app is classified according to a relevant category or class. Although app developers can classify their apps, this may sometimes mislead users as it might be incorrectly classified. Additionally, each store relies on its classification scheme, depending on some aspects such as store functionality and its goals. In case users want to explore and find an app suitable for their needs, it is more useful to have a unique predefined classification scheme by which all apps are classified [7].

Several studies have been conducted to analyze the landscape of English mobile apps [4, 6, 8–12]. However, based on the literature review, no previous study has analyzed Arabic mobile apps. Hence, this study aims to focus on the case of analyzing the relationships between the most important technical and nontechnical features of Arabic apps. It also aims to investigate how well the current predefined Google Play app categories represent the type and genre of Arabic mobile apps.

To accomplish our aim, data are extracted and analyzed from a sample collection of Arabic apps in Google Play app store. The textual data of apps are analyzed using topic modeling to elicit the relationships among texts and, thus, find an appropriate topic. The topic holds a set of words that frequently occur together. In our dataset, there are two textual attributes, the app name and the app description. Thus, to serve our objective, the app description is used in the categorization task. The questions that this research addresses are as follows:

(1) Is there a relationship between the technical and nontechnical features of Arabic apps in Google Play app store?

(2) How well do the current predefined Google Play app categories represent the type and genre of Arabic apps? How can they be improved using topic modeling?

The major contributions of this study are as follows: first, it focuses on the extraction and analysis of Arabic apps from Google Play app store, for which no previous similar studies have been conducted; second, it presents a methodology for extracting and analyzing Arabic apps from Google Play app store. Third, it presents a new sustainable classification system for the Arabic apps in Google Play app store which adapts with both the existing and new Arabic apps.

The findings of this study can support app users, developers, and app store administrators, as well as offer an important contribution to app analysis and design to improve app search and exploration in several domains such as

(1) Business and marketing: better marketing for Arabic apps

(2) Social: understanding the usage of Arabic apps, finding the characteristics of popular Arabic apps, i.e., which category is most popular and why?

(3) Technical developers: finding the gap in the app market and targeting that gap with suitable apps

(4) Understanding Arabic user preferences: the analysis of Arabic apps in Google Play app store will provide insights for the future of Arabic app research and provide guidance for the development of Arabic app dashboard that will support users on how to select an app based on their specific needs

The remainder of this paper is organized as follows: Section 2 describes the previous work related to this study, Section 3 illustrates the research methodology used to answer our research questions, Section 4 presents the results and discussion of our findings, and finally, Section 5 provides the conclusion of this research and avenues for future work.

2. Literature Review

Our study relates to different studies in some important research areas. The first step in such studies is app analysis which is important to study the features that effect the apps' popularity, the number of downloads or user ratings. Using the analysis outcomes, developers or even the market can categoriese their apps based on what is suitable to them, as well as the user to create a successful journey inside market. In our study, we choose to study the Arabic apps in Google Play app store as a case study due to the lack of studies in this area. Therefore, in this section, we group the related work into three groups: app feature analysis studies, app classification studies, and Arabic app studies.

2.1. App Features Analysis. This work is closely related to app feature studies as the focus is on analysing the apps features, specifically the analysis of Google Play app store being a benefit to the user, developer, and even for the platform. Mobile app stores provide an extremely rich source of information on app characteristics and usage, which called app features, such as app name, app size, user rating, and number of installs. Many studies tackled the analysis of apps in the Google Play app store.

In a recent study [4], a descriptive analysis was performed on about 2,67,000 apps' data crawled from the Google Play app store. They found that there is no strong correlation among ratings, reviews, price, size, and install attributes, indicating a demand for gathering other attributes. Their analysis found that the most found apps in the store are the education apps while the musical games are rare. Regards to the number of reviews, action game apps are often reviewed while events are reviewed rarely. Furthermore, they found that racing game apps are frequently installed while event apps are rarely installed. According to ratings, book and reference apps are the highly rated while

map and navigation apps are the least rated. Also, role play games are high in both price and size, but event apps are less in price and tool apps are the smallest in size. This study may help app developers in making good decisions that can participate in the modification of the apps' reachability.

The work described in [8] shows an empirical analysis of both free and paid apps on the top 100 free and grossing apps chart on Apple's Korean app store, using Weibull parametric survival analysis—one of the most popular parametric regression models. The top grossing apps means the apps with the highest revenue. Weibull parametric survival model provides an estimate of baseline hazard function and coefficients for covariates [13], which is effective in determining the reliability characteristics of data. In their study, they found that user rating, ranking, and app size had the greatest impact on the apps' survival, especially free apps.

Another method for app feature analysis is data mining. For example, Finkelstein et al. [9] showed how app features (technical and nontechnical) could be extracted from a large collection of apps obtained from the BlackBerry World using data mining techniques, to study the relationships between business, technical, and social aspects of app stores. The results of their study indicate a strong correlation between the user rating and the number of app downloads. However, there is a correlation neither between price and rating nor between price and downloads. The study also indicates how the findings can be used to guide BlackBerry app developers and managers. Another study [10] explored the relationships between rating, price, and popularity of both free and paid apps, using data mining and natural language processing techniques to extract app information and features from BlackBerry World. They concluded that the degree of correlation between popularity, price, and rating differs based on category. Similar to the work presented in [9], they also highlighted a strong correlation between app rating and popularity (i.e., number of downloads), although the ratings of free apps are higher than those of paid apps. The study also showed that there is no correlation between the price of paid apps and either its popularity or rating.

In addition, the work presented in [11] discussed the problem of app sampling on a dataset comprising a total of 32108 paid apps and a total of 9999 free apps. The author used topic modeling to extract publicly available data from the app store to detect and analyze the relationships between technical attributes, such as software features, and nontechnical attributes, such as rating and popularity information, using Spearman's Rank correlation for correlation analysis. The author obtained a large amount of data from different app stores, including Google Play app store, Windows app store, and Blackberry World. Similar to other studies, the findings indicate a strong correlation between ratings and the number of downloads. However, neither the price and rating nor the price and download rank of apps correlated.

Furthermore, the study presented in [12] investigated the most significant factors for identifying high-rated apps by applying random forest classifier. They mined 1492 high-rated and low-rated free apps from the Google Play app store and found that high-rated apps were statistically significantly different based on several factors. The experiment shows that the most significant factors are app size, number of promotional images displayed on the app page, and app version.

In general, the results from the studies presented above show that there is a strong correlation between app rating and the number of downloads (installs). Although the nature of data studied in the app feature analysis literature is similar to that in this study, the application scopes are different as we focus on Arabic apps only and do not include apps in other languages.

2.2. Classification of Apps.
App classification and categorization refers to the task of placing apps in specific categories so that users can find them within a specific category. There are several methods for app classification, including assigning apps to predefined categories using contextual information [14, 15], using machine learning to classify apps [7, 16], neural language models [17], information retrieval and agglomerative hierarchical clustering [18], and data-driven topic modeling [19].

Zhu et al. [14, 15] aimed to improve the quality of app classification using additional web knowledge from web search engines. The authors extracted some contextual mobile app features from context-rich device logs of mobile users. Their solution enabled the automatic classification of an app into a predefined category in Apple app store taxonomy.

Meanwhile, the work described in [7] presents a system used to classify apps into user-defined classification schemes, using Support Vector Machines (SVMs) to classify apps into 50 classes based on the textual features generated from app descriptions and app categories as defined in Google Play app store or Apple app store. Their results indicate that performing automated app classification using SVM can provide better results than manual classification. A key limitation of this research is that it relies only on the textual description available in app stores, and may mislead the user if the textual description is not useful or meaningless.

Additionally, the study in [16] introduced a novel framework called SimApp for detecting similar apps using machine learning. SimApp examines multimodal heterogeneous data in app stores. The authors proposed an online kernel learning algorithm to learn the optimal combination of similarity functions of multiple modalities. They built many kernel functions to measure app similarity, each kernel function corresponding to a specific kind of the collected data (name, category, developer, update, permissions, description, content rating, images, size, and reviews). The evaluation of the proposed system is based on a real-world dataset crawled from Google Play app store. The results demonstrate that SimApp is effective and promising for use in many applications such as app recommendation, app search, and app categorization.

Using language models, the study presented in [17] describes a novel approach of classifying apps from app stores into predefined taxonomies (classification hierarchies). They achieved good results by using a language model

to learn the app embeddings, using a neural language model and then using the k-nearest neighbour classification approach to classify apps into appropriate categories.

Furthermore, using agglomerative hierarchical clustering, the work described in [18] proposed a novel technique to cluster apps by measuring the similarity between features extracted from their descriptions. The author mined 17877 apps from BlackBerry World and Google Play app store in 2014. The features extracted were used to cluster the apps using agglomerative hierarchical clustering. Their findings showed that this technique improved the existing categorization quality of both BlackBerry World and Google Play app store.

Using web mining and feature extraction, the authors in [20] proposed a method based on semisupervised machine learning and web mining techniques that automatically arrange apps based on their functionality. The categorization system is based on features extraction from more than 6000 app descriptions, which are enriched with the information mined from the web based on the app name. Thus, they achieved better results. Similarly, another data-driven model is the work described in [19] which presents a FRamework for App Categorization (FRAC+) based on a data-driven topic model. This proposed system automatically suggests the categories appropriate for the app store, as well as detects miscategorized apps. Based on their findings, FRAC+ outperformed some popular methods such as multiclass SVM, one-class SVM, LDA, and k-means++ clustering algorithm in most settings, especially for larger data sizes. Moreover, the experimental results of FRAC+ on Google Play app store show that game apps are more detected as miscategorized apps.

The work in [21] evaluated the available features of a specific app category in Google Play app store related to diet and nutrition. They used a data mining technique to mine the apps and then used comparative analysis to identify these apps. One of the most interesting findings is that users recommended apps that exceed \$0.99 than those less than \$0.99. This study can be generalized to other categories in Google Play app store by changing the scope, goal, and research questions.

Although the reviewed literature shows a wide variety of methods and techniques for automated app classification, none of them specifically focuses on classification of Arabic apps. Hence, there is a need for this study to explore how well the current predefined Google Play app categories represent the type and genre of Arabic mobile apps.

2.3. Arabic App Studies. In this study, we focus on Arabic apps analysis and classification, as very few existing studies are reported in the literature. In a recent study [22], the authors investigated existing Islamic apps available on Google Play app store where Islamic apps are the most popular religious apps in Google Play app store. They tackled the issue of the absence of a religion category in the classification of the store and the miscategorization of Islamic apps. Therefore, they suggested a new categorization for the Islamic apps based on their common features they studied.

These features were extracted from the dataset to easily classify the apps, such as app ratings, the number of downloads, the current predefined category in Google Play app store, and languages and particular app features based on each category. Furthermore, they studied the concerns associated with using Islamic apps such as the authenticity of using Islamic apps' contents. The findings of the thematic analysis of the data they collected suggested 5 different categories for the Islamic apps, Quran, Qibla/Prayer Time, Hadith, Zakat, and Supplications. However, their study was limited to 300 apps.

Another recent study explored the Quranic Arabic vocabulary learning apps available in Google Play app store in order to determine their common features and the teaching activities offered to the users. They collected their data from the apps' information in the store, as well as from testing apps, and they used a content analysis approach to achieve their study goals. Their findings indicated that there are few Quranic Arabic vocabulary apps compared to the English vocabulary apps. Although developing Quranic Arabic vocabulary apps are still in their infancy, they are on the right track in offering the quality of end-user apps [23], which includes lessons, exercises, assessments, and many other motivational elements in order to attract users.

The work presented in [24] investigated "weight-loss" apps in Arab countries and the need for improvement. They considered whether current Arabic weight-loss apps have features that adhere to evidence-informed practices. They collected their dataset from the six app stores: Apple iTunes app store, Google Play app store, BlackBerry World, Samsung Galaxy store, Nokia Ovi, and Microsoft store. The authors systematically searched these stores for collecting the weight and diet apps using Arabic keywords. Then, they used a statistical approach called the latent class analysis in order to find subtypes of the related latent classes from multivariate categorical data. Their results showed serious weaknesses in the available Arabic weight-loss apps because they need to provide more features based on the best available evidence in the context of Arabic culture. The most important limitation is that most of the app content in the study was in English, with relatively few apps having Arabic content.

Furthermore, the study in [25] reviewed all available Arabic diabetes apps from Google Play app store and Apple app store. The author firstly identified Arabic keywords and then obtained the relevant apps using the app web interface and the respective stores on iPhone and Android devices. Detailed information on each app (i.e., description, price, and functionality) was extracted and then reviewed. The outcomes of the study indicate a need for more Arabic diabetes apps due to the small number of Arabic diabetes apps compared to English diabetes apps. However, the scope of the study was limited because only Google Play app store and Apple app store were considered. This study is limited only to the free apps that focus on diabetes and are in the Arabic language, so it is limited to a certain category of apps.

Unfortunately, the strategy of collecting data in [22, 23–25] was based on using some keywords; thus, there is no specific predefined category in Google Play app store

which makes our study essential to build a strong and easy base to help in classification and then access the Arabic apps. In contrast, our study is based on building good classification systems for Arabic apps; thus, users can access any types of apps using our proposed categories without depending on specific keywords.

Based on the Arabic studies reviews, the scope of the previous studies was found to be limited in terms of app categories considered, as well as the low number of apps. Hence, it is clear that there is a lack of studies focusing on the analysis of Arabic apps in Google Play app store. Table 1 shows a summary for these studies discussed in the last section regarding the platform and scope compared to our study.

3. Methodology

We adopted a seven-phase approach to analyze Arabic mobile apps on Google Play app store, as shown in Figure 1. The first phase is the crawling phase, which involves the crawling and extraction of raw data from Google Play app store. The second phase is the filtering phase, which focuses on filtering the data to extract only Arabic apps. In the third phase (data extraction), the results of the previous phase are parsed to retrieve all available attributes for each app, including name, category, size, price, reviews, ratings, installs, and textual descriptions. In the fourth phase (data cleaning), duplicate rows are removed, and other data-cleaning tasks are performed. In the fifth phase (data formatting), the parsed attributes are saved into a Comma Separated Values (CSV) file. In the sixth phase, the data in the CSV file are analyzed, and the relationship between app attributes is studied using statistical analysis methods. Finally, in the last phase, topic modeling is applied to the textual descriptions of the Arabic apps and new representative categories are suggested.

3.1. Phase 1: Crawling. The first step of the overall process is the collection of mobile app metadata from Google Play app store, using a quantitative method. We implemented a web crawling system to collect the raw web page data from Google Play app store. A total of 282183 apps in different languages were crawled using Scrapy tool [26].

3.2. Phase 2: Filtering. The crawling process produced a large number of raw data representing all the apps in Google Play app store in different languages. In this phase, we filtered the apps to extract Arabic apps whose names/titles are in Arabic or have mixed names of Arabic and English. This phase resulted in the reduction of total extracted apps to 13682 Arabic apps.

3.3. Phase 3: Data Extraction. In this phase, we extracted Arabic app features by parsing their raw data according to a set of search rules. These rules are based on HTML-tags, each tag indicating a unique signature for each attribute of interest. For example, to retrieve the name of an app, we searched for the value of the {span} HTML tag, its attributes "id = name," and its class "class = AHFaub." After that, we developed search rules to capture all information about app name, app size, app version, platform version, app description, app category, app rating, number of app installs, number of app reviews, and app price. These extracted features will be analyzed later in phase 5 to study the relationships among them. The extracted features of the Arabic apps are as follows:

(i) Technical features: app name, app size, app version, and app platform version

(ii) Nontechnical features: description, category, user rating, number of installs, price, and number of reviews

3.4. Phase 4: Data Cleaning. In this phase, the extracted Arabic app features from the previous phase are cleaned. Data cleaning is an essential step for correctly analyzing the data and understanding the relationships among them to gain useful insights. Three steps were involved in our data preparation:

(i) First, we removed duplicate rows automatically using python script, resulting in a total of 13282 unique rows

(ii) Second, we converted each string and categorical attribute into numerical values using python script

(iii) Third, we resolved the problem of missing values using either the median or mean values, based on the number of missing values for an attribute

3.5. Phase 5: Data Formatting. In this phase, the extracted app features from the previous phase are stored and exported as a CSV file containing all relevant metadata related to Arabic apps.

3.6. Phase 6: Analysis. In this phase, we analyzed the Arabic apps and their features using statistical analysis methods to study the relationship among the features, i.e., the degree to which the features are associated with one another. We used frequency distribution and correlation for the analysis. During this phase, we performed app classification analysis to answer the first part of our second research question that concerns how well the current predefined Google Play app categories represent the type and genre of the Arabic apps developed. In the app classification and analysis, we reviewed each app description to see how well it matches the category assigned. We selected a sample that constitutes 5% of the most dominant categories and manually reviewed both app name and app description to discover the extent to which they match the associated category.

3.7. Phase 7: App Classification. For app classification, we use automated methods (topic modeling) to classify the Arabic apps into new categories based on the textual

TABLE 1: Summary of Arabic app studies.

Study	Platform	App categories	App numbers
[22]	Google Play app store	Islamic apps	300
[23]	Google Play app store	Quranic Arabic vocabulary learning apps	11
[24]	Apple iTunes app store, Google Play app store, BlackBerry World, Samsung Galaxy store, Nokia Ovi, and Microsoft store	"Weight-loss" apps	65
[25]	Google Play app store and Apple app store	Arabic diabetes apps	18
Our study	Google Play app store	All available categories	13282
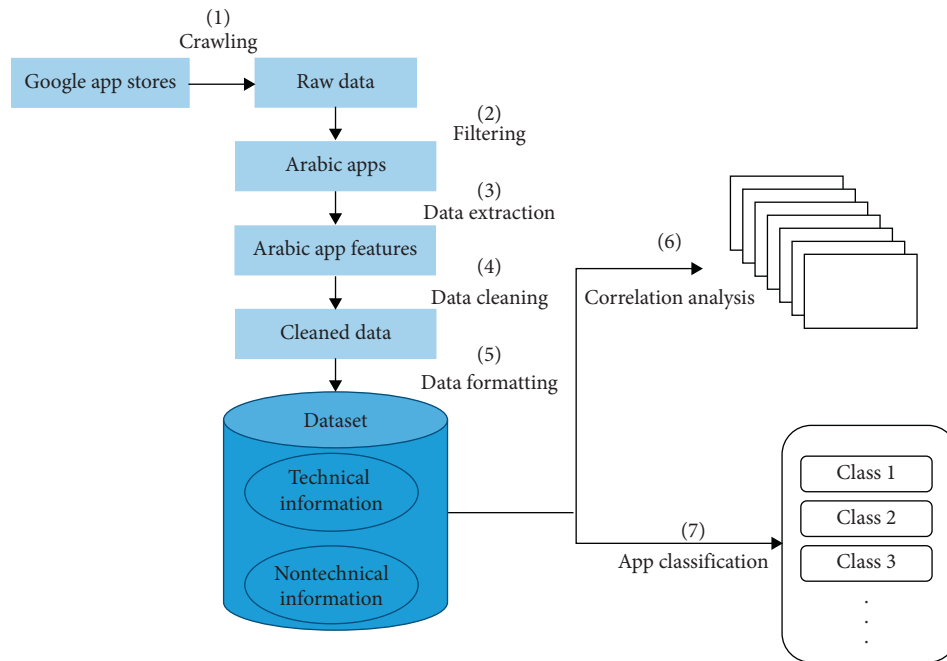
FIGURE 1: Overall app analysis of Arabic apps in Google Play app store. A seven-phase approach, including crawling, filtering, data extraction, data cleaning, data formatting, statistical analysis, and app classification.
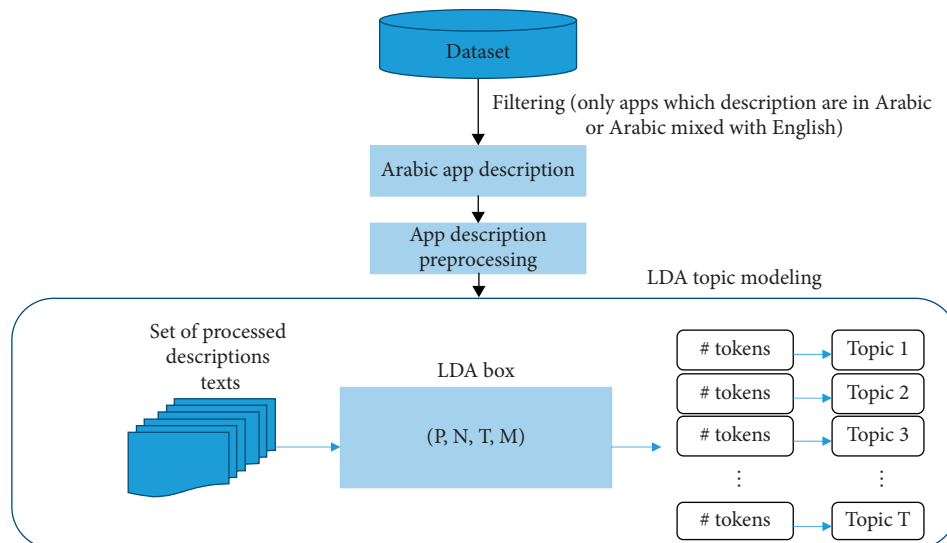
FIGURE 2: Arabic app classification using LDA algorithm.

description provided by the app developer, as shown in Figure 2.

We first filtered the apps to include only apps whose descriptions are in Arabic or have mixed descriptions of Arabic and English. This phase resulted in a filtered dataset of 7657 Arabic app descriptions. Subsequently, each description in the filtered dataset was preprocessed automatically using python script based on the following:

(i) Remove punctuations and diacritics that may appear in the description.

(ii) Remove special characters and emoji icons, as some descriptions may use special characters and emoji icons.

(iii) Remove non-Arabic characters, as some descriptions have a mix of Arabic, English, and other languages words. Only Arabic characters are kept.

(iv) Normalize text by removing Tatweel which stretch connective letters in medial positions, replace Hamzat Alif (آ, إ, أ) with a bare Alif (ا), replace Dotless yaa (ي) with Yaa (ى), and finally, replace Taa Marbutah (ة) with Haa (ه).

(v) Tokenize text and convert the description into a list of tokens.

(vi) Remove stop words, such as words commonly used in Arabic (هذا,إن,سوف).

After app description preprocessing, we applied topic modeling on the preprocessed app descriptions using latent Dirichlet allocation (LDA) algorithm, using Gensim library [27].

Topic modeling is a type of statistical modeling technique for discovering abstract topics that occur in a set of documents [28]. LDA is one of the used methods of topic modeling which proved its success in the area of discovering topic modeling [29]. LDA is an unsupervised generative probabilistic model of topics in which documents are composed of a varied distribution of latent topics. Each topic is considered a discrete probability distribution that indicates the probability that each word appears on a given topic.

To apply the LDA algorithm, we assume each document M is composed of a number of words N. We also set the parameters as follows: number_topics $T = 10$ and number_passes $P = 50$. Based on the parameters we set and the tokens output from LDA, 10 new topics were suggested according to our dataset for Arabic apps in Google Play app store.

4. Results and Discussion

To present and discuss the results obtained from our study, we used Microsoft Power BI tool [30]. This section will explore the research questions of our study.

Our first research question concerns about whether there is a relationship between the technical and nontechnical features of Arabic apps in Google Play app store. The results of our analysis included several interesting findings as follows:

(1) The most dominant app categories in Google Play app store according to number of Arabic apps are Family, Books and Reference, Lifestyle, Entertainment, and Education, with percentages of 28.8%, 14.5%, 10.6, 8.82%, and 8.71%, respectively, as shown in Figure 3. These categories are considered generic; thus, most developers tend to classify their apps in these categories. They usually avoid adding their apps to more specific categories due to the lack of apps in these categories, as well as their inability to attract users' attention.

(2) Based on the previous result, we also found that the top five categories according to the number of Arabic app installs (downloads) are Education, Family, Books and Reference, Lifestyle, and Entertainment, as illustrated in Figure 4. We found that these five categories based on number of app installs are the same as the categories based on the number of apps shown in Figure 3. This result supports the previous result and indicates that users tend to install apps from generic categories rather than specific categories.

(3) Regarding the relationship between the user rating of the app and number of app installs, the number of Arabic app installs increase with high ratings, indicating a positive relationship, as shown in Figure 5. This means that users tend to install the apps which have higher ratings more than the others. This result supports the results presented in the literature [6, 8, 9]. This is logical because people tend to have confidence in popular apps which have more install numbers and ratings.

(4) Regarding the relationship between app size and number of app installs, the results show that users tend to install Arabic apps which have smaller sizes, as illustrated in Figure 6. However, there are exceptions; for example, some large-size apps have large number of installation such as "The holy Quran–القرآن الكريم," whose size is $89 * 10^6$, with more than $5 * 10^6$ of installs. Because Quran is the holy book for Muslims, users tend to install more Quran apps and similar apps regardless of their size.

(5) There is no clear relationship between app size and number of app reviews, as shown in Figure 7. Some smaller size apps have a large number of reviews regardless their reviews (positive or negative). This result may need more investigation in the future by deep exploring the nature of reviews (positive or negative).

(6) Based on the correlation matrix between app features presented in Figure 8, the number of installs and number of reviews are correlated more than the other features. Thus, Arabic apps reviewed by more users increase their downloads from Google Play app store.

The second research question concerns how well the current predefined Google Play app categories represent the
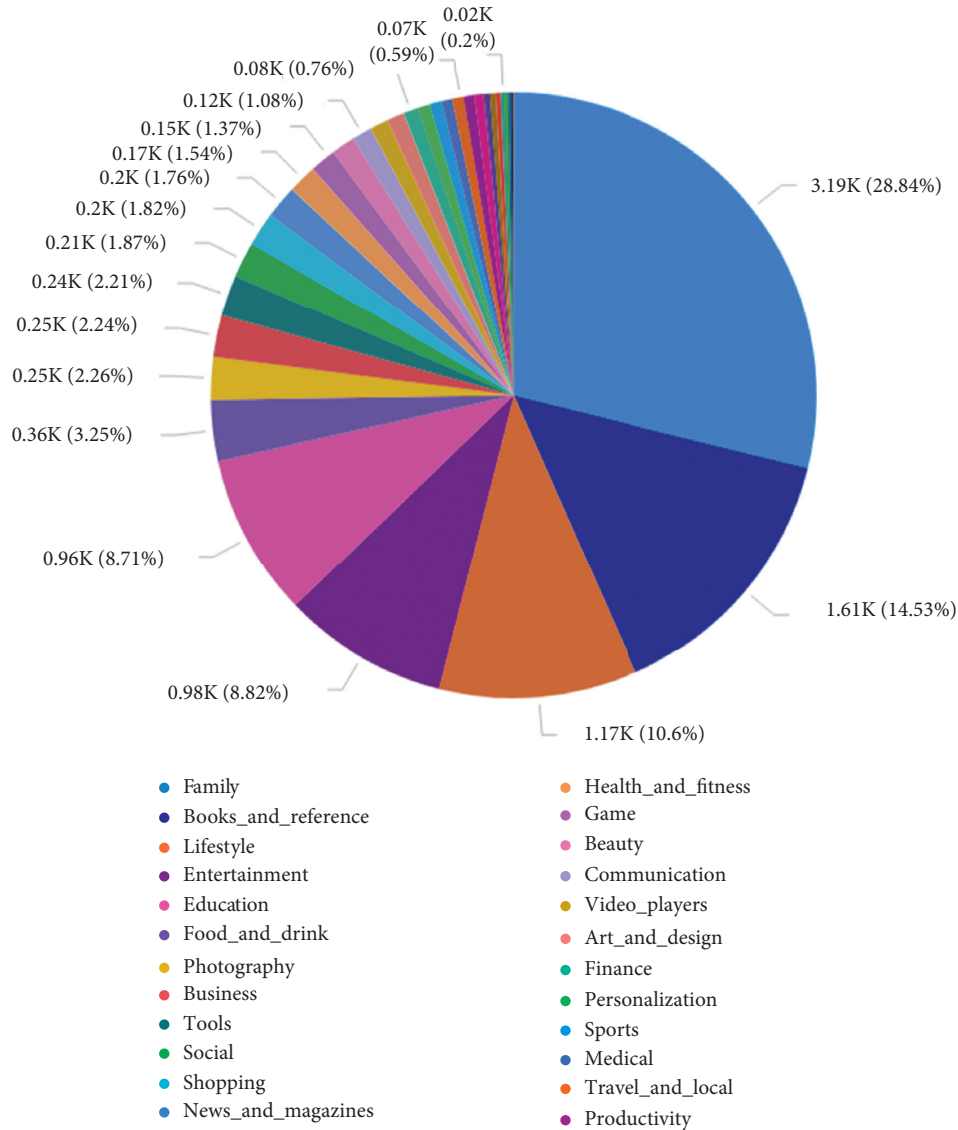
FIGURE 3: Number of Arabic apps per category in Google Play app store.

type and genre of the Arabic apps developed and whether it can be improved using topic modeling. The results of the app classification analysis show that a high percentage of the selected apps are miscategorized and could be better classified under new proposed categories. For example, some Islamic apps are miscategorized under different categories, such as Music and Audio, Lifestyle, Communication, and Entertainment. The Educating Mushaf "المصحف المعلم" app was miscategorized under communication, whereas it was classified under Quran apps using LDA topic modeling.

Another case of miscategorization is the hadith app "الأربعون النووية," which was categorized under lifestyle, whereas it was classified under Islamic apps using topic modeling. The reason for this miscategorization is due to the insufficiency of the Google Play app store categorization scheme, especially for Islamic apps. Additionally, developers may select the wrong categories for their apps because they misunderstand the predefined classification scheme of Google Play app store. Topic modeling has provided a new

and more meaningful system for the categorization of Arabic apps to match their description. Hence, the current predefined classification scheme of Google Play app store does not match the description of some Arabic apps.

Using LDA algorithm of topic modeling to suggest new classifications for Arabic apps, we identified a set of 10 main topics, which are Prayers and Supplication (Doa'a and Azkar), The Holy Quran, Education, Health, Islamic, Beauty and Cookery, Occasions, Music, Poetry and Novels, and Games. Table 2 presents these suggested topics and the Arabic tokens and English translation, which have been identified after applying the LDA algorithm. These tokens are the most frequent words under each topic in our dataset. Although these words help us to suggest the related topic, some of them are not suitable for this topic. Reviewing the descriptions of the apps, the reasons for this may be due to the fact that some descriptions are not written very well, with a single repeated sentence, or the use nonrelevant words.
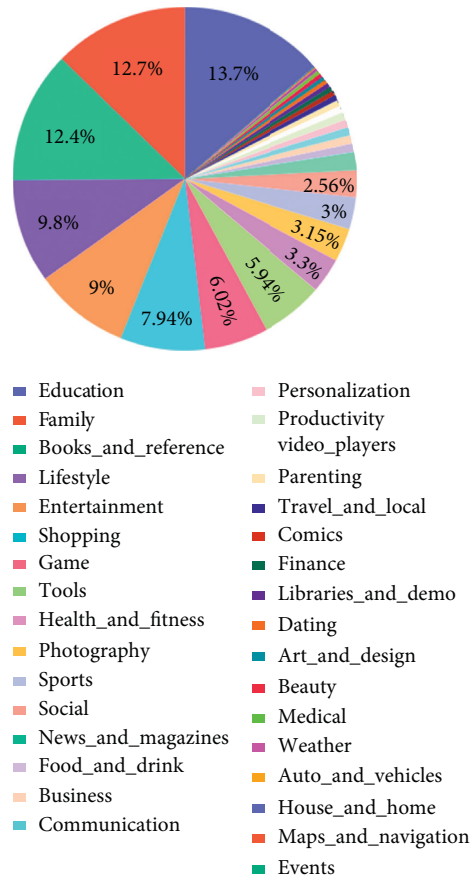
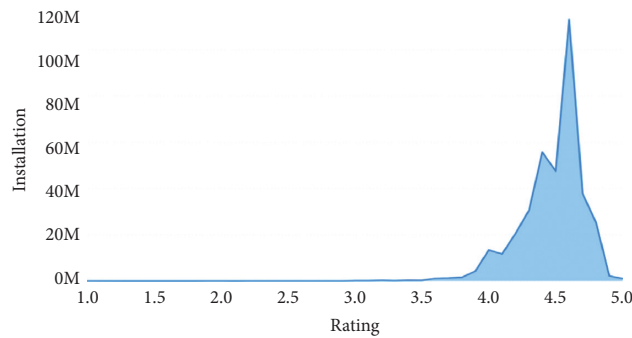FIGURE 4: Number of installed Arabic apps per category in Google Play app store.

FIGURE 5: Relationship between app rating and the number of installs.
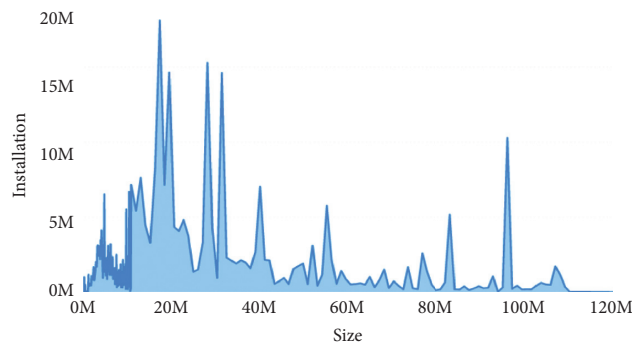

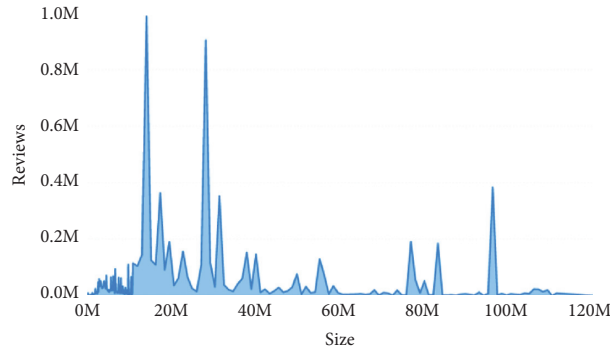FIGURE 6: Relationship between app size and the number of installs.

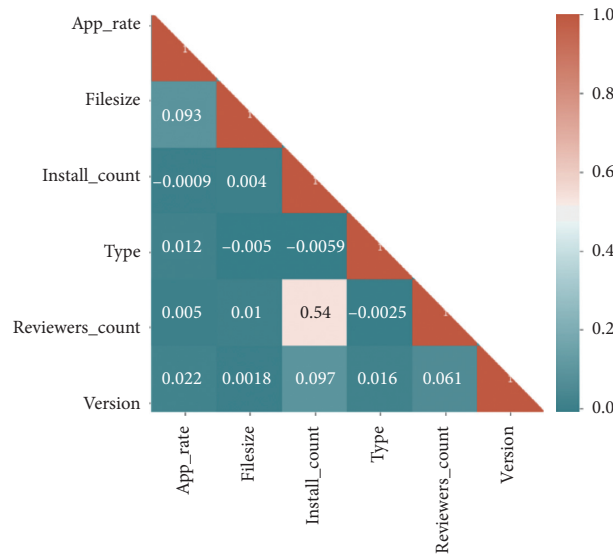FIGURE 7: Relationship between app size and the number of reviews.

FIGURE 8: Correlation matrix between app features.

TABLE 2: Suggested app classification identified after applying the LDA algorithm.

Topic	Arabic tokens	English translated tokens
Topic 1: prayers and supplication (Doa'a and Azkar)	دعاء, أدعية, اسلامية, اذكار, دينية, الجمعة,الصباح, المساء,رواية, قصص	Prayer, supplications, Islamic, religious, Friday, morning, evening, narration, and stories
Topic 2: the Holy Quran	القران, الكريم, سوره, قران, بصوت, صوت, الشيخ, كريم, ايات, عبد	The Holy Quran, surah, the Quran, with the voice of, voice, Sheikh, Karim, verses, and Abd (slave)
Topic 3: education	تعلم, اللغة, تعليم, الانجليزيية, العربية, للأطفال, حرف, الكلمات, الفرنسية, اطفال	Learning, language, education, English, Arabic, for children, letter, words, French, and kids
Topic 4: health	فوائد, الوزن, علاج, رجيم, البطن, تمارين, الحمل, بالأعشاب, الجسم, تخسيس	Benefits, weight, treatment, diet, abdomen, exercises, pregnancy, herbs, body, and slimming
Topic 5: Islamic	الله, الشرعية, الرقية, رقيه, محمد, الرسول, السحر, الكتاب, الشيخ, ايات	The God, legitimacy, ruqyah, Al-ruqyah, Muhammad, the messenger, magic, the book, the sheikh, and verses
Topic 6: beauty and cookery	طريقة, عمل, وصفات, حلويات, الشعر, تحضير, كيفية, الدجاج, سهلة, طرق	Method, work, recipes, desserts, hair, preparation, how to, chicken, easy, and ways
Topic 7: occasions	رمضان, عيد, العيد, الاضحى, تهنئة, شهر, تفسير, نكت, صور, الشعراوي	Ramadan, Eid, the Eid, Al-Adha, congratulation, month, interpretation, jokes, pictures, and Al-Shaarawi
Topic 8: music	أغاني, شيلات, حزينة, شيلة, أطفال, رنات, اغنية, موسيقى, عربيه, عراقيه	Songs, folk songs, sad, children, ringtones, music, Arabic, and Iraqi
Topic 9: poetry and novels	صور, قصص, روايات, الحب, كلام, رسائل, كلمات, رومانسية, حكم, اجمل	Pictures, stories, novels, love, words, messages, romance, proverbs, and more beautiful
Topic 10: games	الصور, امكانية, لعبة, العالم, مشاركة, سيارات, اسئلة, ألغاز, العربية, وظائف	Pictures, possibility, game, the world, participation, cars, questions, puzzles, Arabic, and jobs

Although some of the suggested categories are already part of the Google Play app store categories, our topic modeling approach revealed new categories that could better match app descriptions in the store. These new categories include (1) Doa'a and Azkar, (2) Quran, (3) Occasions, and (4) Islamic. The results of our experiment show that topic modeling can improve the categorization of Arabic apps on Google Play app store.

These results indicated that our new classification system of Arabic apps in Google Play app store is more effective, as well as sustainable due to its ability to adapt with the new apps that are not have a specific category in the store such as Covid-19 apps. These new apps also do not have a well-descriptive category in Google Play app store in English which can be under the "pandemic" category.

5. Conclusions

To conclude, this study aims to fill a gap in the research literature on app information analysis, as no previous study focuses specifically on the case of Arabic apps analysis in the Google Play app store. In addition, it provides a method of proposing new categories for classifying Arabic mobile apps.

In this paper, we analyzed 13279 Arabic apps in Google Play app store. The results of the app data analysis revealed several important findings. The most dominant categories by number of apps are Family, Books and Reference, Lifestyle, Entertainment, and Education. However, when considering the number of installs (downloads), Education has the top rank, followed by Family, Books and Reference, Lifestyle, and Entertainment. This study also showed that there is a positive relationship between app rating and the number of installs. Furthermore, users tend to install small size apps, as well as free or low-cost apps. In addition, the results show that there is a positive correlation between the number of installs and number of reviews.

Furthermore, we used the LDA algorithm to design a sustainable categorization system using the textual app descriptions in order to suggest new categories for Arabic mobile apps in Google Play app store. The results of app classification analysis showed that a high percentage of the selected apps are miscategorized and could be better classified under a new proposed category. This miscategorization can be due to two reasons. The first reason is a misunderstanding of the predefined classification scheme of Google Play app store. The second reason is the lack of a suitable category for some Arabic apps. This study shows that using automated categorization methods such as topic modeling, textual app descriptions are effective in suggesting categories for classifying Arabic mobile apps. This can help Arabic app developers to choose suitable categories when they add new apps to Google Play app store. For example, automatically generated keywords can be suggested to the app developers when adding apps to the app store. A limitation in this study is the size of the dataset; the results will be more generalized with a larger dataset which includes all the Arabic apps in the Google Play app store including the new apps which were added recently such as pandemic-related apps.

In future work, we will consider the review types (positive, negative, and neutral) while studying the relationship between app installs and reviews. Furthermore, our study has a number of important implications for future practice as we need to focus on the analysis of Arabic app data from other app stores and compare the results with the results of this study. Moreover, the results can be generalized for non-Arabic apps as a sustainable system for automated app categorization.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

References

[1] I. Malavolta, S. Ruberto, T. Soru, and V. Terragni, "Hybrid mobile apps in the google play store: an exploratory investigation," in *Proceedings of the 2015 2nd ACM International Conference on Mobile Software Engineering and Systems*, pp. 56–59, Florence, Italy, May 2015.

[2] App Download and Usage Statistics 2019-Business of Apps, 2019, https://www.businessofapps.com/data/app-statistics/.

[3] Statista, "Google play store: number of apps 2009-2017," 2018, https://www.statista.com/statistics/266210/number-of-available-applications-in-the-google-play-store/.

[4] P. B. Prakash Reddy and R. Nallabolu, "Machine learning based descriptive statistical analysis on Google play store mobile applications," in *Proceedings of the 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*, pp. 647–655, Coimbatore, India, July 2020.

[5] A. Dogtiev, "App download and usage statistics (2018)-business of apps: app download and usage statistics," 2018, http://www.businessofapps.com/data/app-statistics/.

[6] M. Harman, Y. Yue Jia, and Y. Yuanyuan Zhang, "App store mining and analysis: MSR for app stores," in *Proceedings of the 2012 9th IEEE Working Conference on Mining Software Repositories (MSR)*, pp. 108–111, Zurich, Switzerland, June 2012.

[7] G. Berardi, A. Esuli, T. Fagni, and F. Sebastiani, "Multi-store metadata-based supervised mobile app classification," in *Proceedings of the 30th Annual ACM Symposium on Applied Computing*, pp. 585–588, Salamanca, Spain, April 2015.

[8] E.-Y. Jung, C. Baek, and J.-D. Lee, "Product survival analysis for the app store," *Marketing Letters*, vol. 23, no. 4, pp. 929–941, 2012.

[9] A. Finkelstein, M. Harman, Y. Jia, F. Sarro, and Y. Zhang, "Mining app stores: extracting technical, business and customer rating information for analysis and prediction," *UCL Research Notes*, vol. 13, p. 21, 2013.

[10] A. Finkelstein, M. Harman, Y. Jia, W. Martin, F. Sarro, and Y. Zhang, "Investigating the relationship between price,

rating, and popularity in the Blackberry world app store," *Information and Software Technology*, vol. 87, pp. 119–139, 2017.

[11] W. Martin, "App store analysis for software engineering," Doctoral dissertation, University College London, London, UK, 2017.

[12] Y. Tian, M. Nagappan, D. Lo, and A. E. Hassan, "What are the characteristics of high-rated apps? A case study on free android applications," in *Proceedings of the 2015 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, pp. 301–310, Bremen, Germany, October 2015.

[13] Z. Zhang, "Parametric regression model for survival data: Weibull regression model as an example," *Annals of Translational Medicine*, vol. 4, no. 24, p. 484, 2016.

[14] H. Zhu, H. Cao, E. Chen, H. Xiong, and J. Tian, "Exploiting enriched contextual information for mobile app classification," in *Proceedings of the 21st ACM International Conference on Information and Knowledge Management—CIKM'12*, pp. 1617–1621, Maui, HI, USA, October 2012.

[15] H. Zhu, E. Chen, H. Xiong, H. Cao, and J. Tian, "Mobile app classification with enriched contextual information," *IEEE Transactions on Mobile Computing*, vol. 13, no. 7, pp. 1550–1563, 2014.

[16] N. Chen, S. C. H. Hoi, S. Li, and X. Xiao, "SimApp: a framework for detecting similar mobile applications by online kernel learning," in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pp. 305–314, Shanghai, China, February 2015.

[17] V. Radosavljevic, "Smartphone app categorization for interest targeting in advertising marketplace," in *Proceedings of the the 25th International Conference Companion*, pp. 93-94, Montréal, Canada, May 2017.

[18] A. A. Al-Subaihin, F. Sarro, S. Black et al., "Clustering mobile apps based on mined textual features," in *Proceedings of the 10th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, pp. 1–10, Bolzano, Italy, September 2016.

[19] D. Surian, S. Seneviratne, A. Seneviratne, and S. Chawla, "App miscategorization detection: a case study on google play," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 8, pp. 1591–1604, 2017.

[20] D. Lavid Ben Lulu and T. Kuflik, "Wise mobile icons organization: apps taxonomy classification using functionality mining to ease apps finding," *Mobile Information Systems*, vol. 2016, Article ID 3083450, 22 pages, 2016.

[21] H. Schumer, C. Amadi, and A. Joshi, "Evaluating the dietary and nutritional apps in the google play store," *Healthcare Informatics Research*, vol. 24, no. 1, pp. 38–45, 2018.

[22] A. Hameed, H. A. Ahmed, and N. Z. Bawany, "Survey, analysis and issues of islamic android apps," *Elkawnie*, vol. 5, no. 1, pp. 1–15, 2019.

[23] A. R. Salam, F. Y. Mustaffa, N. B. M. Sharif, and A. Sirri, "Features and teaching/learning activities used in educational android mobile applications to teach quranic Arabic vocabulary," *International Journal of Engineering and Advanced Technology*, vol. 8, no. 5, pp. 1184–1187, 2019.

[24] R. E. Amalraj, A. Sathiaseelan, A. S. Al-Khalifa, A. A. Alnasser, and D. Marais, "Do Arabic weight-loss apps adhere to evidence-informed practices?" *Translational Behavioral Medicine*, vol. 6, no. 3, pp. 396–402, 2015.

[25] D. Alhuwail, "Diabetes applications for Arabic speakers: a critical review of available apps for android and IOS operated smartphones," *Studies in Health Technology and Informatics*, vol. 225, pp. 587–591, 2016.

[26] Scrapy | A Fast and Powerful Scraping and Web Crawling Framework, 2019, https://scrapy.org/.

[27] Gensim: models.ldamodel—latent Dirichlet Allocation, 2020, https://radimrehurek.com/gensim/models/ldamodel.html..

[28] Topic Modeling and Latent Dirichlet Allocation (LDA) in Python, 2020, https://towardsdatascience.com/topic-modeling-and-latent-dirichlet-allocation-in-python-9bf156893c24..

[29] R. Alghamdi and K. Alfalqi, "A survey of topic modeling in text mining," *International Journal of Advanced Computer Science and Applications*, vol. 6, no. 1, pp. 147–153, 2015.

[30] Power BI Desktop—Interactive Reports | Microsoft Power BI, https://powerbi.microsoft.com/en-us/desktop/..