GLARE: Google Apps Arabic Reviews Dataset

Fatima AlGhamdi[1], Reem Mohammed[1], Hend Al-Khalifa[12]and Areeb Alowisheq[13]

[1]National Center for Artificial Intelligence (NCAI), Saudi Data and Artificial Intelligence Authority (SDAIA), Riyadh 12391, Saudi Arabia

[2]College of Computer and Information Sciences, King Saud University, Riyadh 11495, Saudi Arabia

[3]College of Computer and Information Sciences, Imam Mohammad Ibn Saud Islamic University, Riyadh 13318, Saudi Arabia

{falghamdi, rmohammed}@sdaia.gov.sa, hendk@ksu.edu.sa, aaalowisheq@imamu.edu.sa

Abstract

This paper introduces GLARE an Arabic Apps Reviews dataset collected from Saudi Google PlayStore. It consists of 76M reviews, 69M of which are Arabic reviews of 9,980 Android Applications. We present the data collection methodology, along with a detailed Exploratory Data Analysis (EDA) and Feature Engineering on the gathered reviews. We also highlight possible use cases and benefits of the dataset.

Keywords: Natural Language Processing, Arabic, Reviews, Data Analysis, Data Collection, Feature Engineering, Google PlayStore.

1. Introduction

Various Natural Language Processing (NLP) tasks such as: Topic Modeling, Sentiment Analysis (SA) and Aspect Based Sentiment Analysis (ABSA), require good quality datasets (Nassif et al., 2021). But there are issues that lie behind the availability of such data when it comes to low resource languages like the Arabic language (Alsarsour et al., 2018).

The Arabic language is a complex and rich language. It has different forms depending on the region (dialects) or usage. It can be classified into: Dialectal Arabic (DA), Modern Standard Arabic (MSA) and Classical Arabic (CA) (Abdul-Mageed et al., 2018). Dealing with DA can be challenging compared to MSA, as the former has a shortfall of different NLP tools and resources built to support it (Al-Ayyoub et al., 2019). Nevertheless, the language itself lacks more diverse resources since many of the recent available datasets are from social media platforms, especially Twitter (Nassif et al., 2021).

In this work, we present Google App Arabic Reviews dataset (GLARE). A dataset of Android Applications reviews collected from the Saudi Google PlayStore [1]. It differs from other available datasets in size, and to best of our knowledge, it is the largest Arabic reviews dataset to date, with 76M reviews covering 9,980 Android Applications. Although, similar datasets such as (Al-Shamani et al., 2022) exist, yet with 51K reviews its significantly smaller than GLARE. We believe that GLARE will be a good contribution to the Arabic NLP community for two reasons: its huge size and its domain. Compared to tweets, the character limit for App reviews on Google play is longer, as up to 4000 characters are permitted compared to 280 characters on Twitter, this allows for more expressive sentences. Thus, the dataset will be helpful to many researchers who are looking to utilize it for NLP tasks such as Sentiment Analysis or Aspect Based Sentiment Analysis, as a good volume of data is needed to perform any Artificial Intelligence (AI) related practice (Ahmed et al., 2022)

The rest of the paper is organized as follows: Section 2 reviews related work in the domain of app reviews, Section 3 presents our data collection methodology, Section 4 analyses GLARE dataset, Section 5 further explores GLARE dataset by applying feature engineering to extract additional features, Section 6 highlights possible use cases and benefits of the dataset. Finally, Section 7 concludes the paper with future work.

2. Related Work

Researchers are continuously contributing high-quality resources for the Arabic NLP community. Masader (Alyafeai et al., 2021) provides a public catalogue for over 200 Arabic NLP datasets. The majority of these datasets were from social media platforms, predominantly from Twitter.

As examples of Twitter datasets, (Haouari et al., 2020) published the first Arabic tweets dataset related to COVID-19. They collected approximately 2.7M tweets using Twitter search API. Similarly, (Mulki and Ghanem, 2021) released a dataset targeting the "Levantine" Arabic dialect. The dataset is intended to be used for the misogyny language detection task. Another work by (Alharbi et al., 2020) where they published a benchmark dataset that consists of 95k annotated tweets with sentiment labels of multiple Arabic dialects.

As for other types of datasets, (Einea et al., 2019) aimed to provide an Arabic dataset that can be used for text classification/categorization tasks. The dataset consisted of Arabic news articles gathered from different news portals, it consists of 200K articles distributed between 7 categories. (Elnagar et al., 2018) published an Arabic hotel reviews dataset. The dataset was collected to be used for SA tasks and it had approximately 38K reviews covering 1,858 hotels. Likewise, (Aly and Atiya, 2013) collected

[1]https://play.google.com/store/

over 63K Arabic book reviews from the website Goodreads [2] representing 2,143 books. Another related work was by (Al-Smadi et al., 2015) where they published a benchmark dataset using a subset of book reviews from the previously mentioned work but in contrast, it was manually annotated with the intention for it to be used for ABSA task, the dataset contained 1,513 reviews. A final example, (Ali et al., 2021) presented an Arabic dataset that could be used for fact-checking. They crawled 6,222 claims from 5 Arabic fact-checking websites and were shared to the public.

Given the previous work in Arabic dataset creation, our intention in this work is to release a dataset that is not only large in size but also from an under-represented source for Arabic data such as App Stores Reviews. We believe that our dataset can benefit both the NLP and Software Development communities.

3. Data Collection Methodology

3.1. Approach

GLARE dataset was harvested from Google PlayStore using google-play-scraper[3] library in Nodejs for crawling and its python[4] version was used for scraping reviews and their metadata. We scraped reviews from top 200 free apps from each main and sub category in the Saudi Google PlayStore, which resulted to a total of 59 categories and over 11K apps. We chose to scrape free apps since they are accessible to all and; hence, should have more reviews than paid apps. After dropping duplicated apps, we ended up with a total of 9,980 unique apps. The number of retrieved reviews is over 76M reviews with a total size of 17 Gigabytes (including apps metadata). After applying pre-processing steps, which include dropping duplicates and null review content, removing symbols, numbers and noise, and keeping only Arabic reviews, we ended up with over 69M Arabic app reviews. The raw and engineered 5 datasets are described in Table 1 and available for download via GitHub[5] and Hugging Face[6]. A summary of GLARE dataset is presented in Table 2.

3.2. Apps Metadata

Metadata extracted from Google PlayStore holds useful information about the application that can be used to derive insights using statistical analysis and machine learning approaches. These data include app rating (score) at the time of data scraping, application ID and URL in the PlayStore, icon image URL, and app summary. An overview of the collected and engineered metadata is presented in Listing 1.

[2] https://www.goodreads.com/

[3] https://github.com/facundoolano/google-play-scraper

[4] https://github.com/JoMingyu/google-play-scraper

[5] https://github.com/Fatima-Gh/GLARE

[6] https://huggingface.co/datasets/Fatima-Gh/GLARE

Data Type	File Name	File Size
raw	apps	4.1 MB
raw	reviews	17 GB
raw	categories	4.3 MB
engineered	apps	3.8 MB
engineered	reviews	21.9 GB
engineered	vocabulary	530.5 MB

Table 1: An Overview of GLARE Raw and Engineered Data.

Store	Apps	Reviews	Size	Period
Google PlayStore	9,980	76M	17 GB	March 21 - April 21

Table 2: Statistics of Collected Data.

3.3. Reviews Metadata

The reviews metadata also offers valuable information such as user rating associated with the review, number of users that agree with the reviewer, user display name and review app version. This data is extracted and published along with the reviews. A more descriptive overview of the raw and engineered reviews metadata can be found in Listing 2.

4. Dataset Analysis

To understand the properties of GLARE dataset and the insights that can be derived from the collected data, we conduct the following descriptive analysis:

- The distribution of reviews ratings and the number of users that agree with the reviewer.
- Developers engagement with users.

```
"raw":
{
    "title": "application name",
    "app_id": "application unique identifier",
    "url": "application url at Google PlayStore",
    "icon": "url for image object",
    "developer": "developer name",
    "developer_id": "developer unique identifier",
    "summary": "short description of the application",
    "rating": "application accumulated rating"
},

"engineered":
{
   "reviews_count": "number of reviews for the app",
   "categories": "list of app categories",
   "categories_count": "number of app categories"

}
```

Listing 1: Apps Metadata

```
"raw":
{
    "at": "review datetime",
    "content": "review text",
    "replied_at": "developer reply datetime",
    "reply_content": "developer reply content",
    "review_created_version": "user application
            version during the time of review",
    "review_id": "review unique identifier",
    "rating": "user rating",
    "thumbs_up_count": "number of users that agree
            with the reviewer",
    "user_name": "user display name",
    "app_id": "application unique identifier"
},

"engineered":
{
    "tokenized_review": "list of words in review",
    "words_count":"number of words in review"
}
```

Listing 2: Reviews Metadata

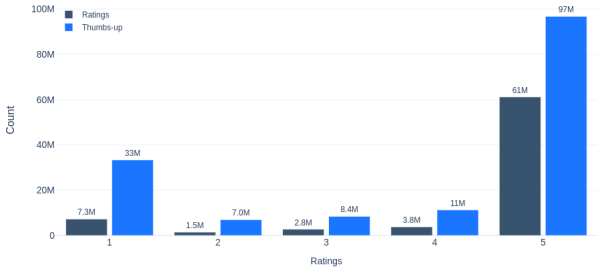

Figure 1: Statistics of Thumbs-up with respect to Ratings Distribution.

4.1. Ratings and Thumbs-up Count Distribution

4.1.1. Ratings

To write a review in Google PlayStore, it is mandatory for the user to provide a rating or score of the application that ranges from 1 to 5. Analysing ratings of the reviews and their effect on various properties such as inciting developers to reply or how user ratings change overtime can be of help in software maintenance and evolution life-cycle (Dąbrowski et al., 2022). In our dataset, we found that the ratings are skewed greatly with over 80% of the reviews having 5 stars. Additional statistics of ratings distribution is presented in Figure 1.

4.1.2. Thumbs-up

Google PlayStore provides user-to-user engagement functionality through a voting mechanism. Any user can view and up-vote in agreement with other users' reviews. This feature provides useful insights of the general sentiments of an application's customer population. Over 98% of the apps had reviews with thumbs-ups, with a total of 8.1M reviews. The highest thumbs-up for a review was 67K while the lowest was of 1.

4.1.3. Thumbs-up and Ratings Distribution

To show a descriptive analysis of the distribution of the previously mentioned features, reviews' ratings were mapped with up-votes. The results are shown in Figure 1.

4.2. Developers Reply to Users

Developers engagement with the users is defined as developers replying to users' reviews in the app store. The affect of such interaction can provide valuable analysis on customer behavioral patterns when service providers engage with them. In GLARE dataset, about 48% of apps had engaged with customer reviews with a total of 3.7M developers' reply. The highest developer engagement was by Azar App, a video chat and livestream application, with over 203K developers replies. The reviews ratings and developer engagement distribution is presented in Figure 2.

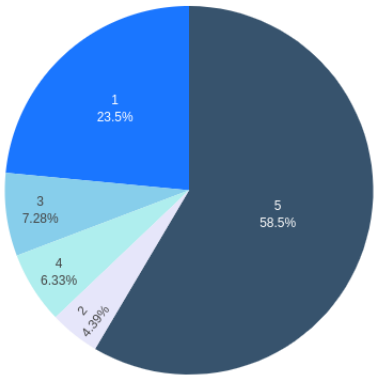

Figure 2: Percentage of Developers Engagement with respect to Reviews Ratings (1 to 5).

5. Feature Engineering

To further explore our GLARE dataset, feature engineering methods were applied to extract additional features from the raw data that can be utilized for machine learning and NLP modeling tasks. Some of these methods include engineering a vocabulary dataset, identifying words count per review and reviews count per app and engineering duplicated apps. The results of the feature engineering process is shown in Table 3.

5.1. Term Dictionary

To identify words' statistics and noise distribution in the dataset, we constructed term frequency dictionary using CountVectorizer (Pedregosa et al., 2011) with its default settings that split words on white-space and punctuation and keep words with characters count greater or equal to 2. Features (words) and their occurrences are extracted from all the reviews in GLARE dataset to construct a vocabulary dictionary. The total number of unique words are 8.7M, with 6.9M words only appearing once in the reviews dataset. Noise in the vocabulary constitutes around

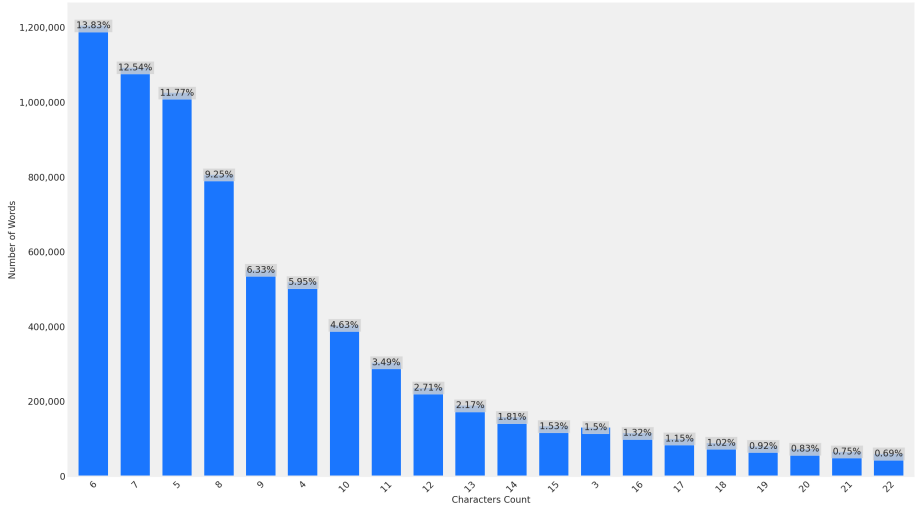

Figure 3: Top 80% Most Frequent Characters Length per Word with respect to the Total Number of Words in the Vocabulary Dataset.

Raw and Engineered Features	**Result**
Total Number of Reviews	76,512,077
Total Number of Non-empty Reviews	76,387,928
Vocabulary Count	8,707,776
Tokens with Term Frequency = 1	6,960,449
Longest Review Sequence (Tokens Count)	753
Shortest Review Sequence (Tokens Count)	1
Number of Reviews with One Word	30,953,303
Percentage of Reviews with One Word	40.5%
Reviews Replied by Developers	3,798,136
Reviews with 5-star Ratings	61,184,348
Highest Number of Reviews per App	3,207,530
Lowest Number of Reviews per App	1
Number of Apps with One Review	216
Longest Character Count for a Word	2,383
Maximum Number of Categories per App	4
Number of Main and Sub Categories	59
Number of Categories Combinations	213

Table 3: A Summary of Raw and Engineered Features Statistics.

17% of the words, while 15% of the vocabulary words had digits; hence, we define noise as any character that is non-Arabic alphanumeric. We use this information to understand how much noise exist in the vocabulary set and

Figure 4: Top 200 Most Frequent Words.

what cleaning steps could be applied on GLARE dataset. The longest word based on character count was 2,383 character long, followed by 1,557, 1,362, 1,340, and 1,317 characters long. We found that all the words with very high character count had only appeared once in the whole dataset. 13.83% of the words in the vocabulary have a length of 6 characters as shown in Figure 3. As for the top frequent terms, the word جدا *jidan* has the highest term frequency with 12M occurrences followed by ممتاز *mumtaz* with 5.6M occurrences and جميل *jameel* with 5M occurrences 4.

5.2. Tokenization and Review Length

To understand words per review distribution, the reviews were tokenized on white-space, and words count was calculated by getting the length of the tokenized review. The longest review based on words sequence was 753 words long, while the shortest review consist of 1 word. The top longest reviews mainly consisted of noise, which includes symbols, digits, and emojis, and repeated terms. More than 40% of the reviews are of length 1, which in total amounts to over 30.9M reviews. Samples from the top longest reviews are shown in Figure 5

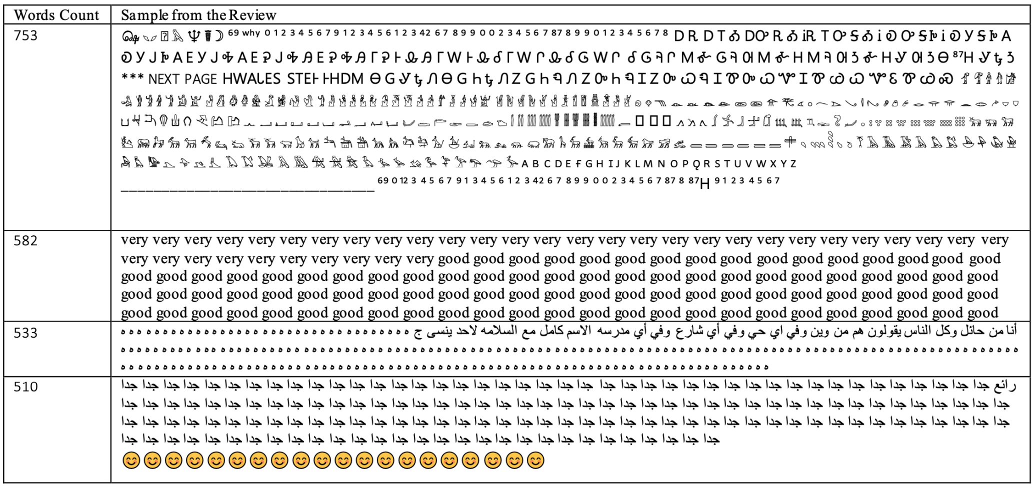

Figure 5: Samples from the Top Longest Reviews.

5.3. Duplicated Apps

While collecting apps from the Saudi Google PlayStore, each app was mapped to a set of apps' categories, i.e. apps on PlayStore can have more than one category label. The number of extracted apps was 11,736 apps from 59 categories. The application ID was used to scrape reviews, and duplicated apps were only scraped once. Duplicated apps were handled by keeping only one instance of the app. The final list of selected applications consists of 9,980 unique apps, which was engineered by removing duplicated apps that had exactly the same values in all the properties. As for apps that had different ratings, we averaged the ratings after dropping duplicates with rating of 0 to avoid skewing the average, and then we kept only one instance of the app with the updated rating.

5.4. Reviews Count per App

In GLARE dataset, the highest number of reviews for an app is 3.2M reviews for PUBG, followed by YouTube with 2.7M reviews, and Facebook with 2.5M reviews. While the lowest number of reviews for an app was 1 review. 2.2% of the dataset apps had only one review, this equals to 216 apps. Additional information of top apps based on review count is presented in Figure 6

5.5. Apps Categories

Google PlayStore allows application developers to label applications into more than one category from a set of over 50 main and sub categories. For GLARE dataset, we focused on 59 categories that we retrieved from the scraper library3 to help in crawling and scraping apps and their reviews. Apps with only one category represent the majority of the apps in GLARE dataset with 91.2% or 9,103 apps, followed by 673 apps with 2 categories, 184 apps with 3 categories and 20 apps with 4 categories. Since apps can have more than one category, a total of 213 different category combinations were formed. Samples of the formed application categories sets are shown in Table 4.

6. Dataset Benefits and Usage

Plenty of NLP tasks can be explored using the GLARE dataset. One suggested application can be applying Opinion Mining aka Sentiment Analysis (SA) on the reviews using their ratings. Another application is using the dataset for Text Classification by classifying applications based on their category and using the review as an input. Also the application and development of spam detectors for reviews will improve the review quality. Another interesting NLP tasks would be the study of demographic language variation as some apps attract certain demographics. Other Machine Learning (ML) tasks can performed as well such as Applications' Ranking Prediction in App Stores.

App Store reviews can work as a method of connection between software developers and their users. It can help software engineers from the step of specifying software requirements to the testing phase (Dąbrowski et al., 2022). The dataset can also be of help to software developers in analyzing their applications. It can aid them during the process of understanding their audience, and debugging errors. In addition, they can also benefit from the dataset when releasing, editing or deleting software features. It is believed that many software engineers perform maintenance measures based on the App Stores reviews (Al-Subaihin et al., 2019).

7. Conclusion and Future Work

In this paper, we presented GLARE: Google Apps Arabic Reviews dataset. GLARE consists of 76M reviews, of which 69M are Arabic Reviews, across 9,980 Android Applications collected from the Saudi Google PlayStore. We showcased the contents of the dataset through performing EDA. We also applied feature engineering to the dataset to extract more helpful aspects from the data. As a future contribution, we are aiming to use the dataset in building a domain specific Arabic Language Model. The Language Model can be used in modeling NLP tasks that are related to software or smart devices applications' reviews. We are also planning to tackle the Arabic Aspect Based Sentiment Analysis (AABSA) task using the Arabic collected

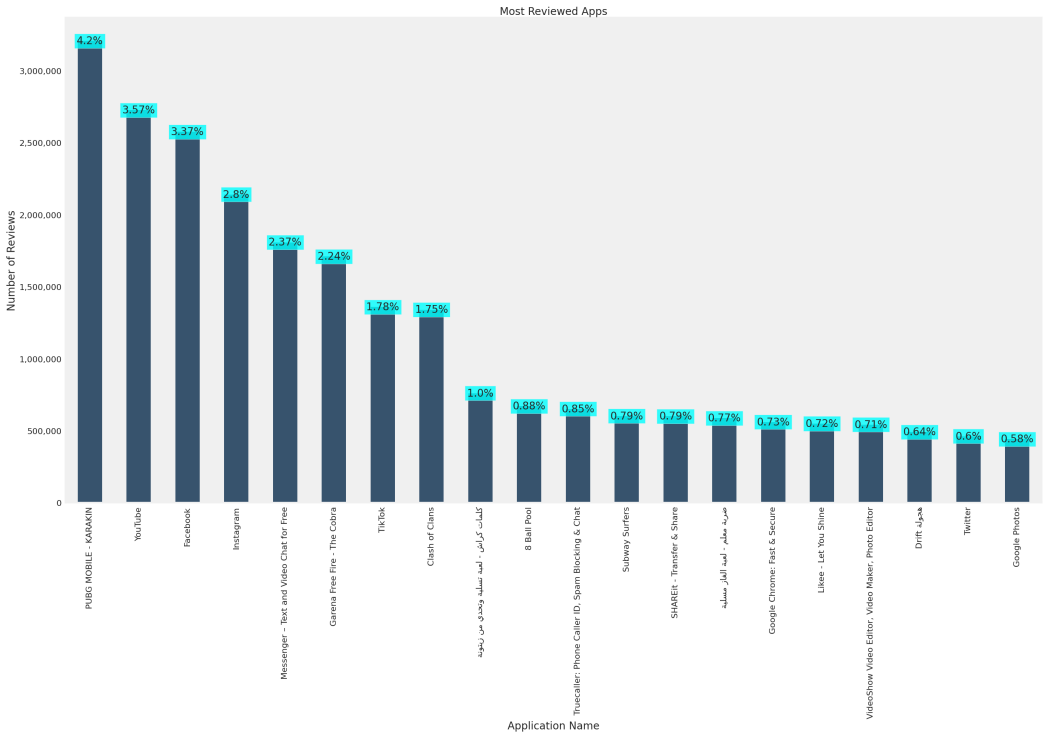

Figure 6: Percentage of Reviews in Top 20 Most Reviewed Apps.

Category Combination	No. of Apps	No. of Reviews
(maps_and_navigation)	190	274,901
(family_create, art_and_design)	11	6,046
(android_wear, health_and_fitness, application)	1	904
(game_music, family, game, family_musicvideo)	1	9,656

Table 4: Sample of Categories.

reviews, as it is one of the NLP tasks that need more exploration (Nassif et al., 2021). Another future contribution is to create a benchmark dataset using GLARE for the tasks of Aspect Term Extraction (ATE), Aspect Category Detection (ACD) and Sentiment Analysis.

8. Copyrights

9. References

Abdul-Mageed, M., Alhuzali, H., and Elaraby, M. (2018). You tweet what you speak: A city-level dataset of arabic dialects. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Ahmed, A., Ali, N., Alzubaidi, M., Zaghouani, W., Abdalrazaq, A. A., and Househ, M. (2022). Freely available arabic corpora: A scoping review. *Computer Methods and Programs in Biomedicine*, 2:100049.

Al-Ayyoub, M., Khamaiseh, A. A., Jararweh, Y., and Al-Kabi, M. N. (2019). A comprehensive survey of arabic sentiment analysis. *Information processing & management*, 56(2):320–342.

Al-Shamani, M., Al-Sarem, M., Saeed, F., and Almutairi, W. (2022). Designing an arabic google play store user review dataset for detecting app requirement issues. In *Advances on Smart and Soft Computing*, pages 133–143. Springer.

Al-Smadi, M., Qawasmeh, O., Talafha, B., and Quwaider, M. (2015). Human annotated arabic dataset of book reviews for aspect based sentiment analysis. In *2015 3rd International Conference on Future Internet of Things and Cloud*, pages 726–730. IEEE.

Al-Subaihin, A. A., Sarro, F., Black, S., Capra, L., and Harman, M. (2019). App store effects on software engineering practices. *IEEE Transactions on Software Engineering*, 47(2):300–319.

Alharbi, B., Alamro, H., Alshehri, M., Khayyat, Z., Kalkatawi, M., Jaber, I. I., and Zhang, X. (2020). Asad: A twitter-based benchmark arabic sentiment analysis dataset. *arXiv preprint arXiv:2011.00578*.

Ali, Z. S., Mansour, W., Elsayed, T., and Al-Ali, A. (2021). Arafacts: the first large arabic dataset of naturally occurring claims. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 231–236.

Alsarsour, I., Mohamed, E., Suwaileh, R., and Elsayed, T. (2018). Dart: A large dataset of dialectal arabic tweets. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Aly, M. and Atiya, A. (2013). Labr: A large scale arabic book reviews dataset. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 494–498.

Alyafeai, Z., Masoud, M., Ghaleb, M., and Al-shaibani, M. S. (2021). Masader: Metadata sourcing for arabic text and speech data resources. *arXiv preprint arXiv:2110.06744*.

Dąbrowski, J., Letier, E., Perini, A., and Susi, A. (2022). Analysing app reviews for software engineering: a systematic literature review. *Empirical Software Engineering*, 27(2):1–63.

Einea, O., Elnagar, A., and Al Debsi, R. (2019). Sanad: Single-label arabic news articles dataset for automatic text categorization. *Data in brief*, 25:104076.

Elnagar, A., Khalifa, Y. S., and Einea, A. (2018). Hotel arabic-reviews dataset construction for sentiment analysis applications. In *Intelligent Natural Language Processing: Trends and Applications*, pages 35–52. Springer.

Haouari, F., Hasanain, M., Suwaileh, R., and Elsayed, T. (2020). Arcov-19: The first arabic covid-19 twitter dataset with propagation networks. *arXiv preprint arXiv:2004.05861*.

Mulki, H. and Ghanem, B. (2021). Let-mi: An arabic levantine twitter dataset for misogynistic language. *arXiv preprint arXiv:2103.10195*.

Nassif, A. B., Elnagar, A., Shahin, I., and Henno, S. (2021). Deep learning for arabic subjective sentiment analysis: Challenges and research opportunities. *Applied Soft Computing*, 98:106836.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.