Arabic App Reviews: Analysis and Classification

Authors: Othman Aljeezani, Dorieh Alomari, and Irfan Ahmad

Publisher: ACM Transactions on Asian and Low-Resource Language Information Processing (February 2025)

Created by: Hassan Alsayhah

Motivation

- Global number of smartphones risen from 2.5 billion to 3.2 billion (2016 to 2023).

- In the first quarter of 2020:
  - Google Play Store offered 2.56 million apps.
  - Apple’s App Store offered 1.85 million apps.

- With this large number of apps, developers should continuously improve their apps

Motivation (cont.)

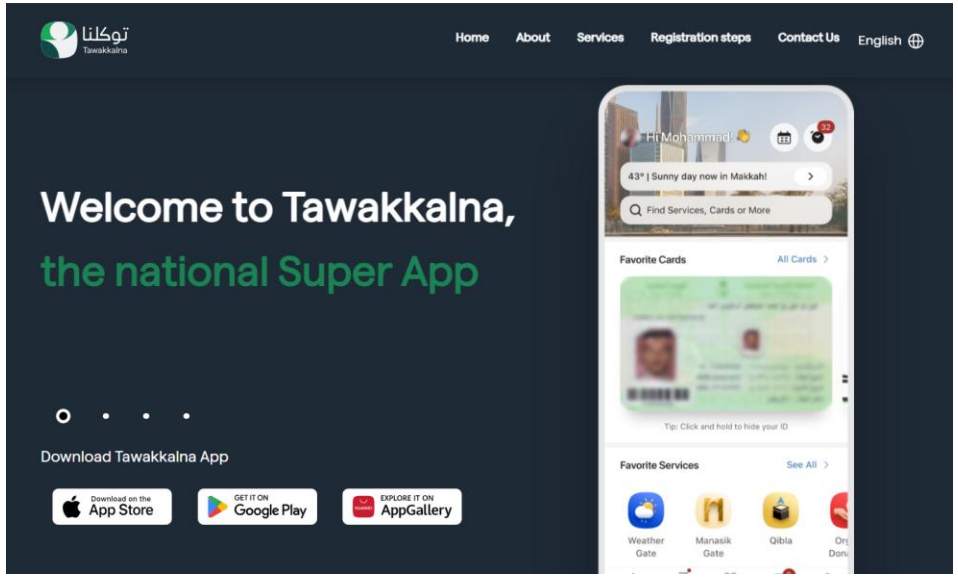
- App reviews provide valuable information including:
  - Bug reports
  - Features requests
  - Issues related to User Interface (UI) and User Experience (UX)

- Those reviews mostly paired with star rating (1-5) to reflect user satisfaction.

Motivation (cont.)

- A significant challenge in the vast volume of reviews for each app.

- Example of the challenge:
  - Instagram App on Apple's App Store: ~23 million reviews
  - Tawakkalna app: 90 thousand reviews



Research Directions of App Reviews

1. Sentiment Analysis
   - classifying the text of an app review based on the mood or emotion of the author (app user).

2. App Review Classification:
   - Categorizing the reviews based on the topics they address.

   - Categories examples:
     - Bug reports, feature requests, user experience, and rating.

     - Functional bug, functional demand, and non-functional requests.

Related Work

Related Work Datasets

Table 1. Summary of the Datasets Used for Arabic App Reviews Analysis and Classification

Reference	Year	Dataset	Task	Dialect	Apps Category	Number of Apps	Source	Number of Reviews
Chader et al. [17]	2021	Collected by Authors	Sentiment Analysis	Algerian	–	–	Google Store	50,000
Saudy et al. [54]	2022	MASR	Sentiment Analysis	Egyptian	9 Categories	12	Google Store	2,469
Voskergian and Saheb [63]	2022	AMAR_ABSA	Multi-label Aspect-Based Sentiment Analysis	Multi-Dialects	Musical Apps	3	Google Store	100,000
Al-Hagree and Al-Gaphari [5]	2022	Collected by Authors	Sentiment Analysis	Yemeni	Bank Apps	8	Google Store	3,192
Hadwan et al. [32]	2022	Arb-AppsReview-V1	Sentiment Analysis	Multi-Dialects	COVID-19 Government Apps	6	Google Store	7,759
Hadwan et al. [33]	2022	Arb-AppsReview-V2	Sentiment Analysis	Multi-Dialects	COVID-19 Government Apps	6	Google Store	51,767
Ramzy and Ibrahim [52]	2024	Collected by Authors	Sentiment Analysis	Multi-Dialects	COVID-19 Government Apps	18	Google Store	114,599

Related Work Summary

Table 2. Summary of the Works on Arabic App Review Analysis and Classification

Reference	Dataset	Task	Best Model	Results
Chader et al. [17]	Collected by Authors	Sentiment Analysis	SVM	Acc= 72%
Saudy et al. [54]	MASR	Sentiment Analysis	Hybrid RF-LR-MLP	Acc= 74.8% F1-score= 73.5%
Voskergian and Saheb [63]	AMAR_ABSA	Multi-label Aspect-Based Sentiment Analysis	MNB	Acc= 64.42% Hamming Loss= 0.1
Al-Hagree and Al-Gaphari [5]	Collected by Authors	Sentiment Analysis	NB	Acc=89.65% F1-score= 88.25%
Hadwan et al. [32]	Arb-AppsReview-V1	Sentiment Analysis	KNN	Acc= 78.46% F1-score= 78.96%
Hadwan et al. [33]	Arb-AppsReview-V2	Sentiment Analysis	SVM	Acc= 94.38% F1-score= 94.30%
Al-Hagree and Al-Gaphari [6]	Arb-AppsReview-V2	Sentiment Analysis	NB-LD	Acc=96.4%
Al-Shalabi et al. [8]	Arb-AppsReview-V2	Sentiment Analysis	KNN-LD	Acc=88.11% F1-score= 73.53%
Ramzy and Ibrahim [52]	Collected by Authors	Sentiment Analysis	ANN	Acc= 89% F1-score= 89%

Research Gaps

- **Underexplored Arabic App Review Classification:**
  - Existing studies focus on sentiment analysis of Arabic app reviews.
  - There is minimal attention paid to classification tasks.

- **Underutilization of RNNs and Transformers:**
  - RNNs and transformers have established themselves as paramount models in NLP.
  - Their application in the analysis and classification of Arabic app reviews remains underexplored.

Main Contributions

1. Presenting the App User Review in Arabic (AURA) dataset
   - a comprehensive public dataset for Arabic app review analysis and classification.
2. Evaluating the performance of RNNs and pretrained transformer models for **sentiment analysis** of Arabic app reviews.
3. Evaluating the performance of RNNs and pretrained transformer models for **classification** of Arabic app reviews.
4. Investigating the role and efficacy of preprocessing, focal loss, and data augmentation techniques in enhancing the performance of DL models, particularly in the context of Arabic app review classification.

AURA Dataset

- Focusing on widely used apps among Arab users.
- Selection Criteria for Mobile Apps:
  1. Select popular apps from both platforms (Android and IOS).
  2. Apps were selected from different categories.
  3. Select government apps that are developed by the government of Saudi Arabia.

- 306 selected apps.
  - 191 Android apps
  - 115 IOS apps
  - Included governmental apps are 42 from both platforms

AURA Dataset (cont.)

- Data Collection Techniques:

- Manual web scrapping
  - to obtain official top apps list from both platforms
  - on official Saudi national portal for governmental services.
- Google Play Scraper (Python tool)
- App Store Scraper (Python tool)

AURA-Sentiment Dataset

- Two trained people labeled 400 random reviews. 100 per star rating except for 3 star.
- The 400 reviews were automatically labeled using two approaches.
- First Approach
  - Considering all stars
- Second Approach
  - Considering only 1 and 5 stars.

Table 4. Comparison of Two Approaches for Automatically Labeling the Sentiments of App Reviews

Approach	# of Reviews	Correctly Classified	Wrongly Classified	Accuracy
First Approach	100 / star	321	79	0.80
Second Approach	100 / star	179	21	0.90

AURA-Sentiment Dataset (cont.)

Table 5. AURA-Sentiment Dataset Sizes During Preparation Steps

Platform	Original Size	Short Reviews Excluded	After Balancing
Android	418,804	192,171	14,850
iOS	34,435	23,746	14,850
Total	453,239	215,917	29,700

Table 7. Important Statistics from the AURA-Sentiment Dataset

Field	Value
Number of words	337,816
Number of unique words	51,094
Maximum length	576
Minimum length	3
Average length	11

AURA-Classification Dataset

Table 8. Popular Categories of App Reviews from the Literature

Category Name	Usage Count	References
Improvement request	8	[28, 31, 62]
Bug report	6	[28, 42, 62]
Rating	5	[28, 31, 42]
Others	3	[28, 47, 62]

Table 9. Selected App Review Categories with the Description of Each Category

Category Name	Category Description
Improvement Request	Requesting new features, recommending enhancements in future versions of the app, asking for content (e.g., books and movies), and suggesting modifications for existing features.
Bug Report	Reporting problems in the app (e.g., crashes and errors).
Rating	Expressing opinion about the app by praising or dispraising.
Others	Reviews that does not fit any of the categories above (e.g., spam and noise reviews).

AURA-Classification Dataset (cont.)

- Using a platform to label 2900 reviews. Then considered majority vote.

Table 10. Samples of the Labeling Output Using the Appen Platform

No.	App Review	Translation	Judgments				
			1	2	3	4	5
1	أتمنى إضافة خرائط التغطية في البرنامج لم تعد موجودة	I wish to add coverage maps in the app; they are no longer available.	IR	IR	IR	IR	IR
2	التطبيق لايعمل ويطلب التحديث	The app does not work and asks for an update.	BR	BR	BR	BR	BR
3	شي خرافي يستحق اكثر من خمسة نجوم	Amazing, deserves more than five stars.	R	R	R	R	R
4	سبحان الله و الحمد لله و لا اله الا الله والله أكبر	Glory be to Allah, praise be to Allah, there is no god but Allah, and Allah is the Greatest.	O	O	O	O	O
5	ممتاز لمنع التشتيت	Excellent for preventing distraction.	R	R	R	R	O
6	مع الاسف التطبيق لا يعمل ما الحل	Unfortunately, the app does not work; what is the solution?	BR	BR	BR	BR	O
7	شكرا تويو الى الامام	Thank you, Toyota; keep going forward.	O	R	R	R	O
8	التطبيق مميز واتمنى يدعم تسجيل الدخول عن طريق البصمه	The app is great, and I wish it supported login through fingerprint.	R	R	R	IR	IR

AURA-Classification Dataset (cont.)

- The analysis revealed a Consensus Agreement of 67%.

- Consensus Agreement is a metric that evaluates how often the majority of annotators agree on the same label for a given sample

$$C_t = \frac{\sqrt{\sum_{i=1}^{K} \left(R_{i,t} - \frac{100}{N}\right)^2}}{\sqrt{\frac{K-1}{K}}},$$

AURA-Classification Dataset (cont.)

Table 12. Important Statistics from the AURA-Classification Dataset

Field	Value
Number of words	39,300
Number of unique words	12,337
Maximum length	576
Minimum length	3
Average length	13

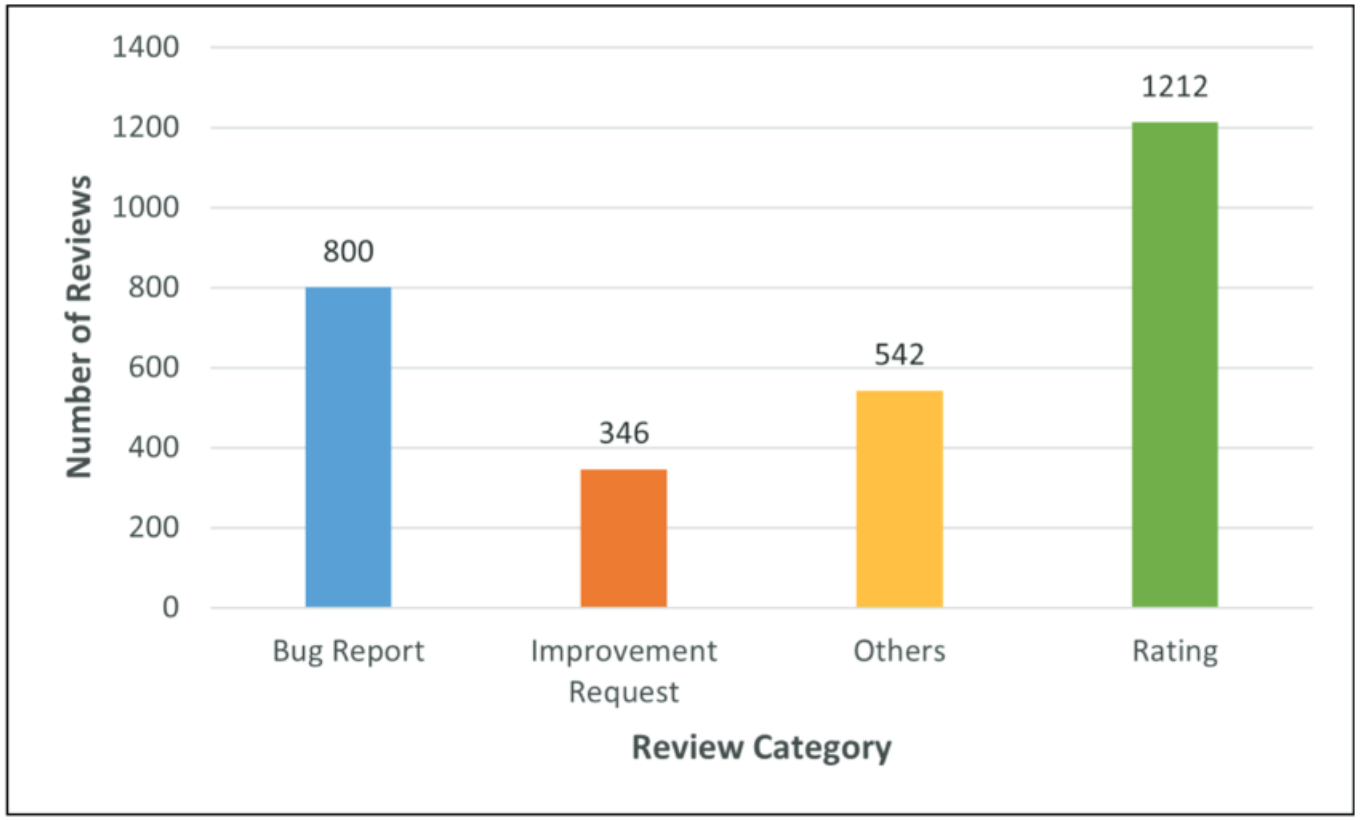


Fig. 3. Distribution of categories in the AURA-Classification dataset.

Methodology-Data Preprocessing

- Text Cleaning
  - Remove all uniform resource locators (URLs).
  - Remove numbers.
  - Remove punctuation characters.
  - Strip diacritical marks (also called Tashkeel), and strip Kashida which is a type of justification (also called Tatweel).
  - Normalize Arabic Letters: convert letters to original form (أ, إ, آ to ا).
  - Remove non-Arabic words.
- Stopword Removal
  - Using a number of Arabic stopwrods lists.
- Lemmatization
  - Converting different forms of a word to its root word

Methodology-Data Preprocessing (cont.)

Table 13. Example of Applying Preprocessing Steps on Our Dataset

Original App Review	أقولُ شكراً جِزيلاً عَلىَ هَذا التَطبيِق الجَميـــــــل والرائع!!
After Text Cleaning	أقول شكرا جزيلا على هذا التطبيق الجميل والرائع
After Stopwords Removal	أقول شكرا جزيلا التطبيق الجميل والرائع
After Normalization	اقول شكرا جزيلا التطبيق الجميل والراءع
After Lemmatization	اقول شكر جزل تطبيق جميل راءع

The number of unique words was reduced in

- the sentiment analysis dataset by 63%
- the review classification dataset by 68%

Methodology-Experiment

- Sentiment Analysis Experiments:
  - Training: 11,700
  - Validation: 8,000
  - Testing: 10,000
- F1-score is the macro

Table 20. Models Performance on the AURA-Sentiment Dataset

Model	Accuracy	F1-score
BiGRU- Trained Embedding	0.86	0.86
BiGRU-Pre-trained Embedding	0.86	0.86
BiGRU-Pre-trained Embedding + Fine-tuning	0.87	0.87
AraBERT	0.88	0.88
MarBERT	**0.89**	**0.89**
CamelBERT	0.88	0.88

Methodology-Experiment (cont.)

- Review Classification Experiments:
  - Training: 1,000
  - Validation: 900
  - Testing: 1,000

- F1-score is the macro

Table 22. Models Performance on the AURA-Classification Dataset

Model	Accuracy	F1-score
BiGRU + Loss	0.58	0.32
BiGRU + Embed	0.65	0.50
BiGRU + Embed + Loss	0.64	0.51
BiGRU + Embed + Loss + Undersampling	0.56	0.51
BiGRU + Embed + Loss + FA ×10	0.66	0.60
BiGRU + Embed + Loss + BA ×10	0.66	0.60
AraBERT	0.66	0.61
MarBERT	**0.67**	**0.62**
CamelBERT	0.65	0.60

Error Analysis

Table 23. Sample Misclassified Reviews in the AURA-Sentiment Dataset

ID	Review	Translation	Actual Label	Predicted Label
1	تطبيق رائع جدا	A very wonderful application	Negative	Positive
2	تحديث سيئ جدا	A very bad update	Positive	Negative
3	انها رائعه لكن مملـه	It's wonderful but boring	Negative	Positive
4	حلو بس الاعلانات كثير	Nice, but there are too many ads	Positive	Negative
5	اللعبه لم تعد جميله	The game is no longer beautiful	Negative	Positive
6	لعبه محتاجه زكاء	A game that requires intelligence	Positive	Negative

Error Analysis (cont.)

Table 24. Sample Misclassified Reviews from the AURA-Classification Dataset

ID	Review	Translation	Actual Label	Predicted Label
1	جميل وسهل الاستخدام وفعال	Beautiful, easy to use, and effective	Others	Rating
2	نطلب اعاده الاغاني الخاصه بشركه روتانا	We request the return of the songs owned by Rotana company	Others	Improvement Request
3	ارجو تطويرها وهي ممتعه	I hope it gets developed further, it's enjoyable	Rating	Improvement Request
4	ارجو توقيف اشعار هذي اللعبه زفت	Please stop the notifications of this crappy game	Rating	Improvement Request
5	يعمل في الخلفيه وبدون اعلانات	It works in the background and without ads	Rating	Bug Report
6	ما بيرضي يحط صور ممكن تلاقولي حل	It doesn't allow me to post pictures, can you find me a solution	Bug Report	Others

Summary

- The AURA dataset is a comprehensive resource of Arabic app user reviews, available in two versions, AURA-Sentiment, and AURA-Classification.
- Various RNN-based models were trained from scratch for app review sentiment analysis, with the BiGRU model using pretrained embeddings delivering the best performance. Furthermore, different transformer-based pretrained models were tested for the task, with MarBert achieving the highest performance, marked by an F1-score of 0.89.
- For Arabic app review text classification, the BiGRU model combined with pretrained embeddings, focal loss, and data augmentation provided the best results among RNN models trained from scratch. The overall best result, an F1-score of 0.62, was obtained using the pretrained MarBert model fine-tuned for the task.