

Kingdom of Saudi Arabia
Ministry of Higher Education
University of Tabuk
College of Computers & Information Technology



ال المملكة العربية السعودية
وزارة التعليم العالي
جامعة تبوك
كلية الحاسوب وتقنية المعلومات

Data Analysis Report

**Project:
User Experience Insights
from the Sehhaty App
Reviews**



**Supervisor by
Dr. Omar Asiri**



Dataset Overview After Cleaning

Following data cleaning, a comprehensive descriptive analysis was independently run on the Arabic and English datasets in order to identify their structure, completeness, and adequacy for further analysis.

Such an early audit helps to establish data quality and identify the remaining issues before performing more intense analytical procedures such as classification and sentiment analysis.



● Arabic Dataset

Total Rows: 86,120

Total Columns: 17 (*after removing duplicate column names*)

Missing Cells (%): 22.99%

Rows Containing Emojis: 3,776

Total Emojis Detected: 17,009

File Name: Database_Arabic_Cleaning.xlsx

The Arabic corpus is also larger than the English one with greater activity from Arabic-speaking users.

Though about **23%** of the dataset has missing values, the percentage is acceptable considering the size of the dataset and the verbosity of online reviews.

The high density of emojis reflects a strong degree of emotional expression in the text that enhances the following sentiment and thematic analysis.

Dataset Overview



Total Rows

34,441

Sheet: Sheet1 | File: After Database_English_Cleaning.xlsx



Total Columns

17

After de-duplicating names



Missing Cells (%)

24.91%



Emoji Rows

2,020

Total emojis: 7,544

Total Rows: 34,441

Total Columns: 17 (*after removing duplicate column names*)

Missing Cells (%): 24.91%

Rows Containing Emojis: 2,020

Total Emojis Detected: 7,544

File Name: Database_English_Cleaning.xlsx

 The English dataset shows a slightly higher percentage of missing values (~25%), which is typical in user-generated text data.

The presence of over two thousand emoji-containing rows highlights the expressive and emotional nature of user feedback valuable for sentiment analysis.

Overall the dataset is clean consistent and ready for subsequent stages of analysis.

DataFrame Information

Dataset Composition and Structure

To further assess the composition and structure of each dataset post-cleaning, the **DataFrame.info()** function was used.

This step provides information regarding the number of columns, data types, and completeness of every column (i.e., how much non-null values are remaining).

DataFrame Info

```
RangeIndex: 86120 entries, 0 to 86119
Data columns (total 17 columns):
 #   Column           Non-Null Count  Dtype  
 --- 
 0   UserName        86111 non-null    object  
 1   Rating          86120 non-null    int64  
 2   Content         86120 non-null    object  
 3   DeveloperReply  30192 non-null    object  
 4   Review_Year     86120 non-null    int64  
 5   Review_Month   86120 non-null    int64  
 6   Review_Day      86120 non-null    int64  
 7   Review_Hour_12  86120 non-null    int64  
 8   Review_AM_PM   86120 non-null    object  
 9   Review_Period  86120 non-null    object  
 10  Comment_Length 86120 non-null    int64  
 11  Reply_Year     30192 non-null    float64 
 12  Reply_Month   30192 non-null    float64 
 13  Reply_Day      30192 non-null    float64 
 14  Reply_Hour_12  30192 non-null    float64 
 15  Reply_AM_PM   30192 non-null    object  
 16  Content_Arabic_Clean 85088 non-null    object  
dtypes: float64(4), int64(6), object(7)
memory usage: 11.2+ MB
```

Arabic Dataset

Arabic Dataset Summary

Total Columns: 17

Total Entries: 86,120

Memory Usage: 11.2 MB

Data Types:

- Integer (int64):** 6 columns (*mainly date and rating fields*)
- Float (float64):** 4 columns (*reply-related numeric fields*)
- Object (text/string):** 7 columns (*usernames, content, periods, etc.*)

Most Complete Columns:

Rating, Content, Review_Year, Review_Month, Review_Day, and Comment_Length

Columns with Missing Values:

DeveloperReply, all Reply_ columns, and Content_Arabic_Clean

 This indicates that while most review-related data are complete missing values are concentrated in the reply-related columns (as not all reviews received developer responses). The text column `Content_Arabic_Clean` also shows slight data loss after preprocessing ($\approx 1.5\%$).

DataFrame Info

```
RangeIndex: 34441 entries, 0 to 34440
Data columns (total 17 columns):
 #   Column           Non-Null Count   Dtype  
 --- 
 0   UserName        34441 non-null    object  
 1   Rating          34441 non-null    int64  
 2   Content         34438 non-null    object  
 3   DeveloperReply  10218 non-null    object  
 4   Review_Year     34441 non-null    int64  
 5   Review_Month   34441 non-null    int64  
 6   Review_Day      34441 non-null    int64  
 7   Review_Hour_12  34441 non-null    int64  
 8   Review_AM_PM   34441 non-null    object  
 9   Review_Period  34441 non-null    object  
 10  Comment_Length 34441 non-null    int64  
 11  Reply_Year     10218 non-null    float64 
 12  Reply_Month   10218 non-null    float64 
 13  Reply_Day      10218 non-null    float64 
 14  Reply_Hour_12  10218 non-null    float64 
 15  Reply_AM_PM   10218 non-null    object  
 16  Content_English_Clean 33938 non-null    object  
dtypes: float64(4), int64(6), object(7)
memory usage: 4.5+ MB
```

English Dataset

- **Total Columns:** 17
- **Total Entries:** 34,441
- **Memory Usage:** 4.5 MB
- **Data Types:**
 - *Integer (int64):* 6 columns
 - *Float (float64):* 4 columns
 - *Object:* 7 columns
- **Most Complete Columns:**
Rating, Content, Review_Year, Review_Month, Review_Day, Comment_Length
- **Columns with Missing Values:**
DeveloperReply, all Reply_ fields, and Content_English_Clean

 Similar to the Arabic dataset, missing values mainly appear in developer reply-related fields which is expected since only a small subset of reviews received replies. The cleaned text column retains over **98%** of the total entries showing effective text cleaning with minimal data loss.

Summary

The info() test ensures that both of these datasets are of the same structure same number of columns the same data types and with similar missing value patterns.

This consistent design facilitates the following steps such as data concatenation feature extraction and model training for both languages.

Handling of Missing Values

Once it was established that missing values were present a thorough examination was done to see which columns they were in and the level of missing data for each.

Dtypes & Nulls					
Column	Dtype	Non-Null	Nulls	Null %	
⌚ Reply_Hour_12	float64	30192	55928	64.940000	
🟡 Reply_AM_PM	object	30192	55928	64.940000	
💬 DeveloperReply	object	30192	55928	64.940000	
📅 Reply_Day	float64	30192	55928	64.940000	
📅 Reply_Month	float64	30192	55928	64.940000	
📅 Reply_Year	float64	30192	55928	64.940000	
✍️ Content_Arabic_Clean	object	85088	1032	1.200000	
👤 UserName	object	86111	9	0.010000	
⌚ Review_Hour_12	int64	86120	0	0.000000	
⭐ Rating	int64	86120	0	0.000000	
🟡 Review_AM_PM	object	86120	0	0.000000	
📅 Review_Day	int64	86120	0	0.000000	
📅 Review_Month	int64	86120	0	0.000000	
📅 Review_Year	int64	86120	0	0.000000	
📄 Content	object	86120	0	0.000000	
🔢 Comment_Length	int64	86120	0	0.000000	
⌚ Review_Period	object	86120	0	0.000000	

Arabic Dataset

DeveloperReply, Reply_Year, Reply_Month, Reply_Day, Reply_Hour_12, and Reply_AM_PM.

This is as anticipated since the majority of user reviews were not replied to by app developers.

Less percentage of missing values (approximately 1.2%) were experienced in the Content_Arabic_Clean column due to text preprocessing operations that had removed blank or incorrect inputs.

The remaining columns are almost complete with almost no missing values such as 0.01% for the UserName column.

Column	Dtype	Non-Null	Nulls	Null %
⌚ Reply_Hour_12	float64	10218	24223	70.330000
⌚ Reply_AM_PM	object	10218	24223	70.330000
💬 DeveloperReply	object	10218	24223	70.330000
📅 Reply_Day	float64	10218	24223	70.330000
📅 Reply_Month	float64	10218	24223	70.330000
📅 Reply_Year	float64	10218	24223	70.330000
✍️ Content_English_Clean	object	33938	503	1.460000
📝 Content	object	34438	3	0.010000
⌚ Review_Hour_12	int64	34441	0	0.000000
⭐ Rating	int64	34441	0	0.000000
⌚ Review_AM_PM	object	34441	0	0.000000
👤 UserName	object	34441	0	0.000000
📅 Review_Day	int64	34441	0	0.000000
📅 Review_Month	int64	34441	0	0.000000
📅 Review_Year	int64	34441	0	0.000000
🔢 Comment_Length	int64	34441	0	0.000000
⌚ Review_Period	object	34441	0	0.000000

● English Dataset

The same pattern was observed in the English dataset too.

There are around 70.3% missing values for columns that deal with developer responses while only 1.46% are present for the cleaned text column (Content_English_Clean).

All other columns such as Content Rating and date columns are complete and consistent.

✿ Comparison of Missing and Duplicate Data

Both the Arabic and English datasets were checked to find out the presence of missing entries and duplicate records



Top Missing Columns

Column	Missing	Missing %
DeveloperReply	55,928	64.94%
Reply_Day	55,928	64.94%
Reply_Hour_12	55,928	64.94%
Reply_Year	55,928	64.94%
Reply_AM_PM	55,928	64.94%
Reply_Month	55,928	64.94%
Content_Arabic_Clean	1,032	1.20%
UserName	9	0.01%



Duplicates

Metric	Value
Duplicate rows	0

- In the Arabic data, the majority of missing values, appeared in approximately 64.9%, were in **reply columns** like DeveloperReply, Reply_Year, etc. The **clean text field (Content_Arabic_Clean)** had missing values in 1.2%, and **duplicate records** were absent in the data.



Top Missing Columns

Column	Missing	Missing %
DeveloperReply	24,223	70.33%
Reply_Day	24,223	70.33%
Reply_Hour_12	24,223	70.33%
Reply_Year	24,223	70.33%
Reply_AM_PM	24,223	70.33%
Reply_Month	24,223	70.33%
Content_English_Clean	503	1.46%
Content	3	0.01%



Duplicates

Metric	Value
Duplicate rows	0

- In the English data, a similar trend was observed, with approximately **70.3%** of the missing values found in **reply-related columns** and only **1.46%** missing within the **Content_English_Clean** field. The **main Content column** was nearly complete and **no duplicate entries** were detected. Overall both datasets demonstrated a **high level of data quality** with missing information confined mainly to **non-critical reply columns** while all reviews remained **unique and non-repetitive**.

Columns by Type

Columns by Type

Type	Count
Numeric	10
Object	7
Boolean	0

- The Arabic dataset has **10 numerical columns** and **7 text (object) columns** with no Boolean variables.
This harmonic blend of **quantitative and qualitative data** is a sign of a well-ordered structure comprised of numerical indicators such as ratings and dates, and **descriptive review text**.

Columns by Type

Type	Count
Numeric	10
Object	7
Boolean	0

- Just like the English data set contains **10 numerical and 7 text (object) columns** with no Boolean variables.
Such structural equivalence between these two datasets allows for **comparable analysis** and **training in models** and ensures comparability and consistency in both languages

Column Quality

Column Quality Report

Column	Filled %	Unique	Top sample	Suggestion
⭐ Rating	100.000000	5	5	Keep
📝 Content	100.000000	33439	ممتاز	Keep
📅 Review_Year	100.000000	7	2021	Keep
📅 Review_Month	100.000000	12	7	Keep
📅 Review_Day	100.000000	31	23	Keep
⌚ Review_Hour_12	100.000000	12	1	Keep
🕒 Review_AM_PM	100.000000	2	PM	Keep
🕒 Review_Period	100.000000	4	Afternoon	Keep
💬 Comment_Length	100.000000	438	5	Keep
👤 UserName	99.990000	75952	مستخدم Google	Keep
✍️ Content_Arabic_Clean	98.800000	30983	ممتاز	Keep
💬 DeveloperReply	35.060000	128	شكراً لمشاركة رأيك بخصوص لغة المراجع الصناعية	Review
📅 Reply_Year	35.060000	6	2023.0	Review
📅 Reply_Month	35.060000	12	12.0	Review
📅 Reply_Day	35.060000	31	17.0	Review
⌚ Reply_Hour_12	35.060000	12	8.0	Review
🕒 Reply_AM_PM	35.060000	2	PM	Review

- In the Arabic dataset, **most columns are entirely complete (100%)**, particularly the key variables such as **Rating**, **Content**, and the **date-related fields** (*Review_Year*, *Review_Month*, *Review_Day*)..

The **preprocessed text column (Content_Arabic_Clean)** indicates **98.8% completeness**, or that fewer than a negligible amount of entries had been removed in preprocessing.

Developer reply columns like **DeveloperReply** and **Reply_Year** are **approximately 35%** occupied which makes sense given that the majority of users never received a reply from the developers.



Column Quality Report

Column	Filled %	Unique	Top sample	Suggestion
User Name	100.000000	32509	A Google user	Keep
Rating	100.000000	5	5	Keep
Review_Year	100.000000	7	2021	Keep
Review_Month	100.000000	12	7	Keep
Review_Day	100.000000	31	23	Keep
Review_Hour_12	100.000000	12	8	Keep
Review_AM_PM	100.000000	2	PM	Keep
Review_Period	100.000000	4	Morning	Keep
Comment_Length	100.000000	417	4	Keep
Content	99.990000	14458	Good	Keep
Content_English_Clean	98.540000	12251	good	Keep
DeveloperReply	29.670000	69	we would like to thank for your feedback	Review
Reply_Year	29.670000	6	2022.0	Review
Reply_Month	29.670000	12	12.0	Review
Reply_Day	29.670000	31	7.0	Review
Reply_Hour_12	29.670000	12	8.0	Review
Reply_AM_PM	29.670000	2	PM	Review

- The English dataset was also in a **similar trend**, with the majority of **review columns** — such as UserName, Rating, Content, Review_Year, Review_Month, and Review_Day — being complete or almost complete.
The column of text after **cleaning (Content_English_Clean)** was **98.5% complete**, representing good preprocessing with minimal data loss.

Monthly KPIs

Monthly KPIs (single table)

Period	Reviews	Avg_Rating	Reply_Rows	Reply_Rate_%
2019-09	30	1.90	0	0.00%
2019-10	73	2.48	0	0.00%
2019-11	69	2.33	0	0.00%
2019-12	17	2.71	0	0.00%
2020-01	50	2.40	0	0.00%
2020-02	61	2.61	0	0.00%
2020-03	37	3.46	0	0.00%
2020-04	20	3.00	0	0.00%
2020-05	39	3.05	0	0.00%
2020-06	194	1.88	0	0.00%
2020-07	236	3.13	28	11.86%
2020-08	303	3.09	30	9.90%
2020-09	188	3.88	13	6.91%
2020-10	176	3.84	1	0.57%
2020-11	238	4.11	0	0.00%
2020-12	868	4.31	3	0.35%
2021-01	1,473	4.56	22	1.49%
2021-02	1,966	4.55	160	8.14%
2021-03	7,016	4.49	5	0.07%
2021-04	1,866	3.91	4	0.21%
2021-05	5,826	4.37	8	0.14%
2021-06	8,640	4.37	179	2.07%
2021-07	8,576	3.93	226	2.64%
2021-08	7,082	4.31	670	9.46%
2021-09	1,273	3.94	201	15.79%
2021-10	1,592	4.13	414	26.01%
2021-11	896	4.40	678	75.67%
2021-12	3,213	3.97	559	17.40%

Arabic Dataset Trend Overview

In the Arabic dataset, review volume showed a steady increase over time with a clear spike beginning at some point in **mid-2020**.

Average user rating also raised substantially — from approximately **2.0–3.0** for the period **2019–2020** to **4.3–4.5** for the year **2021**, which is a sign of greater user satisfaction with the app.

The developer response rate was zero at first (around **0%**), but it started to exhibit more consistently from **late 2020** and later to as much as **75.6%** in certain months. This means a better developer interaction and response to user input.

Monthly KPIs (single table)

Period	Reviews	Avg_Rating	Reply_Rows	Reply_Rate_%
2019-11	1	1.00	0	0.00%
2019-12	1	5.00	0	0.00%
2020-01	4	2.00	0	0.00%
2020-03	4	1.00	0	0.00%
2020-05	2	3.50	0	0.00%
2020-06	46	1.35	0	0.00%
2020-07	48	1.73	4	8.33%
2020-08	58	3.03	7	12.07%
2020-09	38	3.18	1	2.63%
2020-10	38	2.66	0	0.00%
2020-11	66	3.83	0	0.00%
2020-12	279	4.45	1	0.36%
2021-01	410	4.45	7	1.71%
2021-02	358	4.40	58	16.20%
2021-03	3,265	4.06	6	0.18%
2021-04	1,230	3.22	0	0.00%
2021-05	2,373	4.19	3	0.13%
2021-06	4,667	4.16	129	2.76%
2021-07	4,975	3.72	127	2.55%
2021-08	2,379	4.26	230	9.67%
2021-09	324	3.90	62	19.14%
2021-10	423	4.00	87	20.57%
2021-11	239	4.43	200	83.68%
2021-12	3,082	3.79	414	13.43%
2022-01	592	3.83	230	38.85%

English Dataset Trend Overview

The English dataset also reflected the same trend of increase.

Both **review counts** and **average ratings** increased significantly in **2021**, with ratings primarily in the range of **3.8 to 4.5**.

The **developer's response rate** also increased step by step — from nearly **0% in 2020** to over **80%** towards the end of **2021** and the early months of **2022**, reflecting improved communication with **English-speaking clients**.

Both sets of data indicate **general increasing user behavior, extremely high levels of satisfaction, and tighter user-developer relationships** as time passes.

⭐ Rating Distribution

⭐ Rating Distribution		
Rating	Count	Percent
1	12,085	14.03%
2	1,822	2.12%
3	2,837	3.29%
4	4,139	4.81%
5	65,237	75.75%

📦 Rating Buckets		
Segment	Count	Percent
Negative (1–2)	13,907	16.15%
Neutral (3)	2,837	3.29%
Positive (4–5)	69,376	80.56%

● Arabic Dataset Ratings Overview

In the Arabic dataset, most of the users provided excellent ratings with **75.7%** of all reviews being **5-star**.

16.1% of these comments were **negative (1–2 stars)** and **3.3%** were **neutral (3 stars)**.

These figures confirm that the Arabic-speaking users generally reported **satisfaction with the app**, as seen through an **extremely high level of approval and good user experience**.

⭐ Rating Distribution		
Rating	Count	Percent
1	5,558	16.14%
2	1,023	2.97%
3	1,561	4.53%
4	2,858	8.30%
5	23,441	68.06%

📦 Rating Buckets		
Segment	Count	Percent
Negative (1–2)	6,581	19.11%
Neutral (3)	1,561	4.53%
Positive (4–5)	26,299	76.36%

● The English results followed the same trend, but with the slightly more balanced feedback.

Approximately **76.3%** of the reviews were **4–5 stars and positive**, and **19.1%** were **1–2 stars and negative**, with **4.5%** being **neutral**.

While criticism was somewhat extended among the **English-speaking readers**, overall sentiment was **good and robust** in both sets.

Overall, the **rating distribution** across every dataset is consistent with the **largest**

levels of user satisfaction, where most comments support positive sentiment for ease of use and usability of the app.

💬 Most Frequent Words



● Arabic Dataset

The most frequently used word in the Arabic sample was "ممتاز" (excellent) and was utilized **30,062 times**, and subsequently by "جداً" (very), "التطبيق" (the app), and "جيد" (good).

All these common words were conveying **positive feelings** such as "ممتاز" (excellent), " رائع" (great), and "جميل" (beautiful), which conveyed the **general satisfaction of the users with the app**.

This language pattern is concurrent with the **high proportion of positive sentiment** from the above analysis.

abc

Unigrams — Top 12



Most frequent word

good

English Dataset

In the English dataset, the most frequently used word was “**good**” which appeared **14,595 times** followed by “**app**” “**very**” “**nice**” and “**excellent**.”

As with the Arabic dataset **English-speaking users** primarily employed **positive expressions** reflecting **overall satisfaction with the app’s quality and performance.**

Taken together both datasets indicate that the **most common words convey positivity** supporting the conclusion that **users generally experienced the app as beneficial and satisfying.**

💬 Most Frequent Bigrams



● Arabic Dataset

The most frequent **bigram** in the Arabic dataset was "ممتاز جدا" (**excellent very**) with a count of **5,290**, followed by "لا يعمل" (**very good**) and "not working"). Other frequent bigrams such as "جميل جدا" (**very beautiful**) and "تطبيق ممتاز" (**excellent app**) reflect users' positive experience and overall satisfaction with the performance of the app.



Bigrams — Top 12



Most frequent bigram

very good



● English Dataset

The most frequent bigram was **very good** appearing **2,439 times**, and this was followed by **good app** and **not working**

As in the Arabic set, most English bigrams report positive feedback, such as **very good**, **very nice**, and **good service**, though some mention issues such as not working

In general, both sets indicate user feedback is generally positive with only few mentions of issues

Most Frequent Trigrams



Arabic Dataset

The most common trigram in the Arabic corpus was "التطبيق لا يعمل" (**the app does not work**) and appeared **804 times**, followed by "أكثر من رائع" (**more than wonderful**) and "تطبيق ممتاز جداً" (**excellent app**).

"تطبيق ممتاز جداً" and "أكثر من رائع", "تطبيق رائع جداً" convey positive feedback others such as "التطبيق لا يعمل" reflect user discontent or technical problems.

This combination illustrates a balance **between satisfaction and frustration**, showing that both **positive experience and app issues were reported** by customers in their Arabic reviews.

Trigrams — Top 12



English Dataset

The most frequent **trigram** in the English dataset was “**very good app**” which appeared **379 times** followed by “**app not working**” and “**not working after**”. Similar to the Arabic dataset the English reviews also presented a **mix of positive and negative feedback** — expressions like “**very good app**” “**very nice app**” and “**very good service**” reflected **user satisfaction** while phrases such as “**app not working**” and “**not working properly**” pointed to **common technical complaints**.

Overall both datasets indicate that **most users were satisfied** although a portion reported issues related to app functionality and performance.

💬 RAW vs CLEAN Comparison

RAW vs CLEAN — Comparison RAW: sehhaty_reviews_google_arabic.xlsx (Sheet1) • CLEAN: Database_Arabic_Cleaning.xlsx (Sheet1)

★ Δ Avg Rating
-0.47%

✉ Δ Reply Rate
0.44 pp

📅 Δ Avg Monthly Reviews
-4.79%

📊 Comparison Table (Δ % vs RAW)

Metric	RAW	CLEAN	Δ %
Rows	90,450	86,120	-4.79%
Columns	16	17	6.25%
Missing cells (%)	24.52%	22.99%	-1.53 pp
Duplicate rows	1	0	-100.00%
Numeric columns	10	10	0.00%
Object columns	6	7	16.67%
Boolean columns	0	0	—
Non-null text rows (comments/clean)	90,449	85,088	-5.93%
Average rating	4.28	4.26	-0.47%
Reply rate % (overall)	34.62%	35.06%	0.44 pp
Avg monthly reviews	1,256	1,196	-4.79%
Avg monthly reply %	52.96%	53.39%	0.43 pp

EDA complete. Results (if saved) are in: /content/eda_outputs.

- For the Arabic dataset, **there were 4.79% fewer rows after cleaning procedures** primarily due to the removal of duplicate and irrelevant data.
Percentage of missing cells reduced by a **marginal amount of 1.53 percentage points** and all the duplicate rows were eliminated successfully.

RAW vs CLEAN — Comparison RAW: Before_sehhaty_reviews_google_english_aligned.xlsx (Sheet1) • CLEAN: After_Database_English_Cleaning.xlsx (Sheet1)

★ Δ Avg Rating
0.99%

✉ Δ Reply Rate
-0.80 pp

📅 Δ Avg Monthly Reviews
-5.32%

📊 Comparison Table (Δ % vs RAW)

Metric	RAW	CLEAN	Δ %
Rows	37,956	34,441	-9.26%
Columns	16	17	6.25%
Missing cells (%)	26.07%	24.91%	-1.16 pp
Duplicate rows	0	0	—
Numeric columns	10	10	0.00%
Object columns	6	7	16.67%
Boolean columns	0	0	—
Non-null text rows (comments/clean)	37,953	33,938	-10.58%
Average rating	4.05	4.09	0.99%
Reply rate % (overall)	30.47%	29.67%	-0.80 pp
Avg monthly reviews	527	499	-5.32%
Avg monthly reply %	59.07%	61.56%	2.49 pp

EDA complete. Results (if saved) are in: /content/eda_outputs.

- In the dataset of English, **total number of rows decreased by 9.26%** after cleaning the **data primarily because empty** and invalid values were removed

The rate of missing **cells was decreased by 1.16** percentage points and no repeated rows were found in the initial or cleaned dataset

Notably, the mean rating **increased by 0.99%**, meaning that the cleaning procedure acted to cleanse the dataset by removing unrelated or noisy reviews

❖ Introduction to Data Analysis

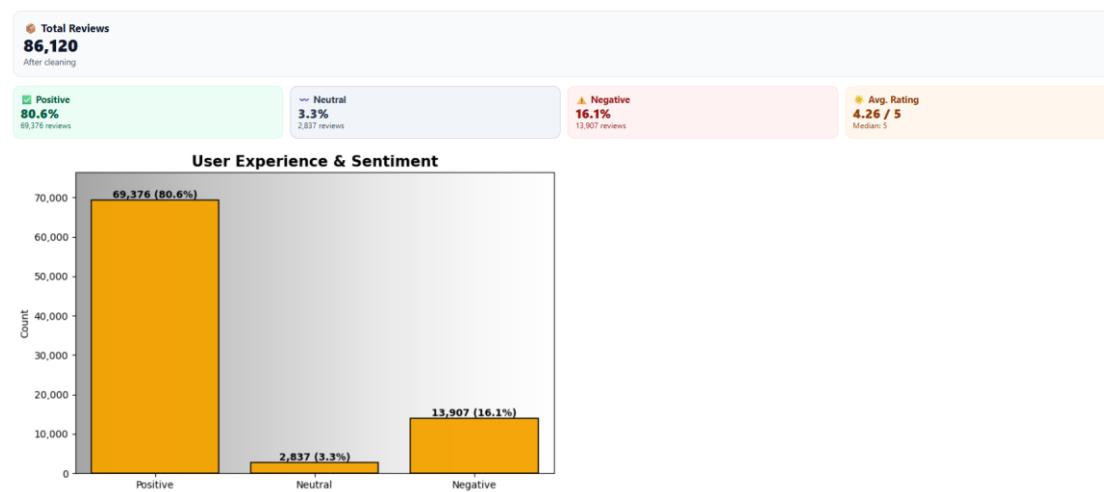
Following the completion of **data cleaning** and **initial inspection** the dataset was prepared for a **comprehensive analytical phase**.

This stage aims to uncover **patterns behaviors and sentiment trends** expressed in user reviews providing a **deeper understanding of user experiences** with the application.

The analysis begins with an examination of the **overall sentiment distribution** — identifying **positive, negative and neutral reviews** — to assess **general user satisfaction** through ratings.

This serves as the foundation for later stages, which delve into **linguistic trends emerging themes and actionable insights** that can guide **app improvement** and **enhance user experience**.

💬 User Experience and Sentiment Analysis

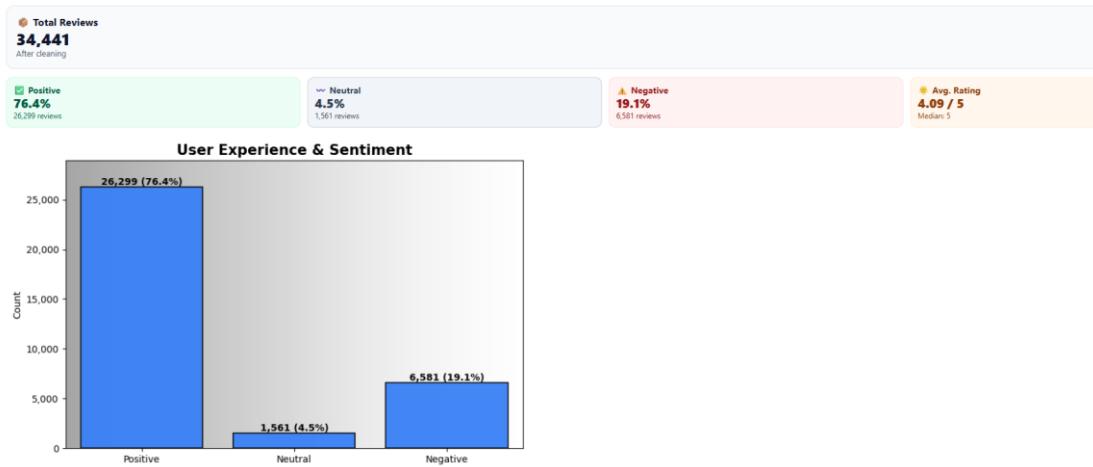


- Following data cleaning, **86,120 Arabic reviews** were in view for analysis to be included.

The vast majority by far — **80.6% (69,376 reviews)** — gave **positive opinions**, while **16.1% (13,907 reviews)** gave **negative opinions** and **3.3% (2,837 reviews)** gave **neutral opinions**.

The **4.26/5 overall rating** represents an **overall positive user perception** of the app's **function and quality**.

Such findings indicate that **most of the users provided positive comments** about the **performance and functionality** of the app, while others have **complained or reported** against some issues.



- In the English dataset, a total of **34,441 reviews** were analyzed after the data cleaning stage.
Of these **76.4% (26,299 reviews)** were **positive** **19.1% (6,581 reviews)** were **negative** and **4.5% (1,561 reviews)** were **neutral**.
The **average rating of 4.09 out of 5** reflects a **high level of satisfaction** among **English-speaking users**.
As with the Arabic dataset **most comments were positive** though a notable portion mentioned **issues such as app crashes or login difficulties**.
Overall both datasets demonstrate **strong user satisfaction** with the **Arabic dataset exhibiting slightly higher positivity** compared to the English dataset.

✖ Technical Performance Analysis

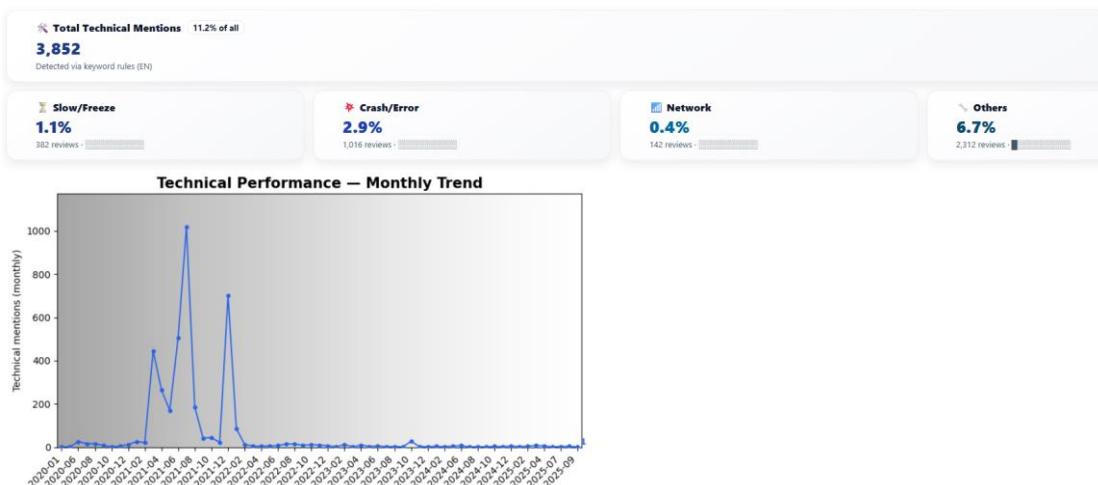


● Arabic Dataset

A total of **7,035 reviews** were identified as **technical-related** representing **8.2%** of all Arabic reviews.

Among these **slow performance or freezing issues** accounted for **1.5% (1,267 reviews)** while **crashes or system errors** were reported in **0.9% (791 reviews)** and **network-related problems** appeared in **0.4% (309 reviews)**.

Other **unspecified technical issues** made up the largest share **5.4% (4,668 reviews)**. The **monthly trend** revealed several noticeable peaks particularly between **mid-2021 and late 2021** suggesting a period of **technical instability** experienced by users. After this phase **technical complaints gradually decreased** and remained **relatively stable** thereafter.



● English Dataset

In the English dataset, a total of **3,852 reviews** — approximately **11.2%** of all entries — included **technical-related feedback**.

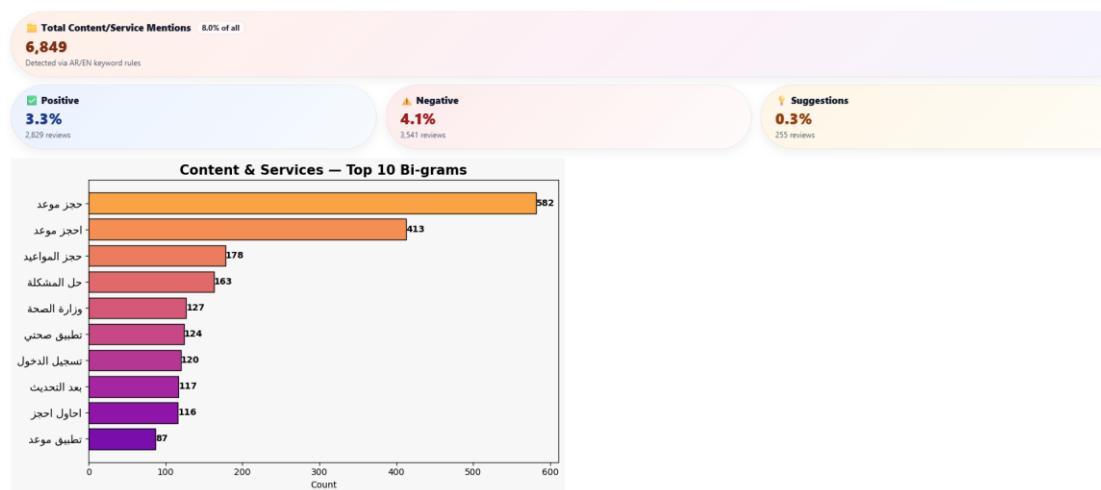
The most frequently reported issue was **app crashes or system errors**, accounting for **2.9% (1,016 reviews)** followed by **slow or freezing performance** at **1.1% (382 reviews)** and **network connectivity problems** at **0.4% (142 reviews)**.

Other **unspecified technical issues** made up **6.7% (2,312 reviews)**.

The **trend analysis** revealed a pattern similar to the Arabic dataset with a **notable surge in mid to late 2021** suggesting a period of **app performance challenges** that later stabilized.

Overall, **technical-related concerns were relatively limited** compared to the total number of reviews across both datasets indicating that **most users did not encounter major technical difficulties**.

Content and Services Analysis



Arabic Dataset

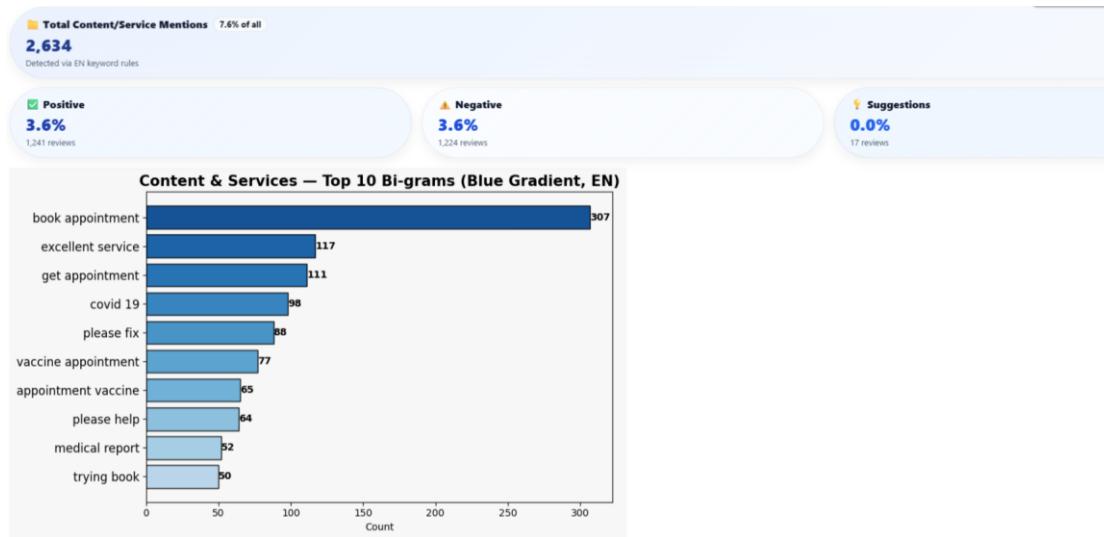
A total of **6,849 reviews** — accounting for **8.0%** of all Arabic entries — discussed **content or service-related aspects** of the app.

Among these **3.3% (2,829 reviews)** expressed **positive feedback**, **4.1% (3,541 reviews)** contained **negative comments** and **0.3% (255 reviews)** included **user suggestions** for improvement.

The most frequent bigrams such as "**حجز موعد**" and "**احجز موعد**" emphasized the **importance of appointment booking features** within the application.

Other commonly mentioned phrases — including "**حل المشكلة**" (**solve the problem**), "**وزارة الصحة**" (**Ministry of Health**), and "**تطبيقات صحتي**" (**Sehhaty App**) — reflected recurring discussions about issue resolution, service quality, and institutional support.

Overall feedback related to **content and services** was **mixed** showing both **user appreciation** for the app's functionality in managing appointments and **concerns** about **technical or procedural difficulties** that users continued to face.



● English Dataset

A total of **2,634 reviews** — representing **7.6%** of all English entries — discussed **content or service-related aspects** of the application.

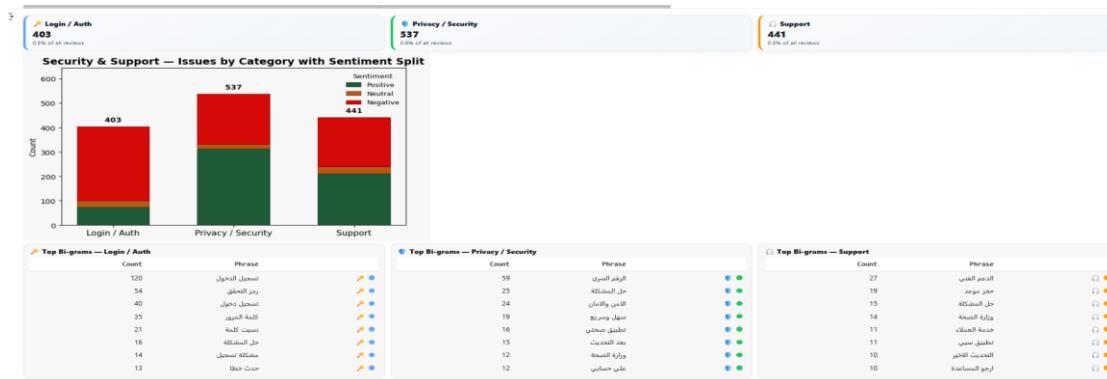
Within this group **3.6%** (**1,241 reviews**) expressed **positive opinions** **3.6%** (**1,224 reviews**) provided **negative feedback** and a small fraction (**0.0%** – **17 reviews**) contained **direct user suggestions**.

The most frequent bigrams, including "**book appointment**", "**excellent service**", and "**get appointment**" highlight the **central importance of the appointment booking function** for English-speaking users.

Additional frequently mentioned terms such as "**please fix**" "**covid 19**" "**vaccine appointment**" and "**medical report**" demonstrate active **user engagement with health-related services and requests for technical or procedural improvements**.

Overall, feedback in the **English reviews** appeared balanced reflecting both **appreciation for the app's convenience and health utility** and **concerns about appointment management and system reliability**.

❖ Security and Support Analysis



● Arabic Dataset

Security and support-related issues accounted for a relatively small segment of the dataset, totaling **1,381 reviews** — approximately **1.6%** of all Arabic entries.

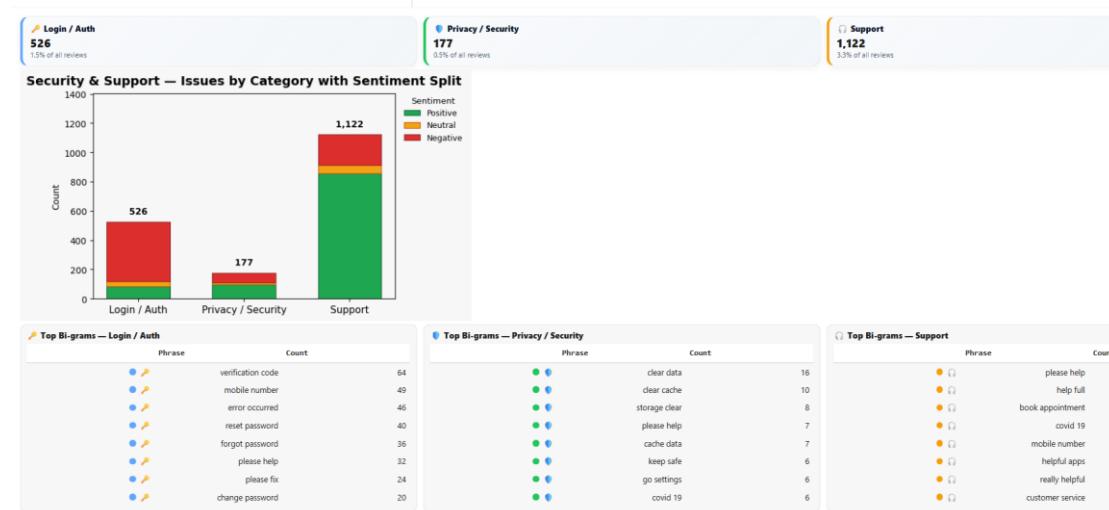
This included **403** reviews related to **login or authentication**, **537** concerning **privacy and security**, and **441** addressing **technical support**.

Frequent login-related bigrams such as "تسجيل الدخول" (login), "رمز التحقق" (verification code), and "كلمة المرور" (password) highlighted user difficulties in accessing accounts or verifying identity.

Mentions of **privacy-related phrases** like "الامن والأمان" (PIN), "الرقم السري" (security and safety), and "بعد التحديث" (after the update) reflected concerns about data protection and changes following system updates.

Support-related expressions such as "الدعم الفني" (technical support) and "حل المشكلة" (fix the issue) indicated that many users sought customer assistance to resolve technical challenges.

Overall the **sentiment distribution** was largely **negative**, suggesting **user frustration** with login barriers, password recovery, and system-related errors.



English Dataset

A total of **1,825** reviews — approximately **2.3%** of all English entries — addressed **security or support-related topics**, including **526 login/authentication**, **177 privacy/security**, and **1,122 support-related comments**.

Frequent login bigrams such as "verification code", "error occurred", and "reset password" revealed user difficulties with authentication and account access.

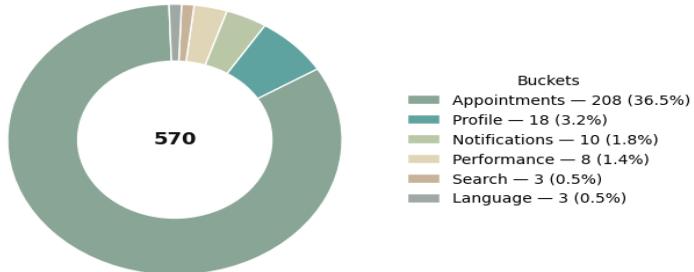
Privacy-related terms like "clear data" and "keep safe" reflected concerns regarding data handling and personal security.

Support-related mentions — including "please help", "helpful apps", and "customer service" — often carried a **positive sentiment**, showing user appreciation for responsive assistance, even when login or privacy issues were encountered.

Compared with the Arabic dataset, **English users provided more balanced feedback** recognizing both **technical challenges** and the **effectiveness of the support system**.

💡 User Suggestions Analysis

Suggestions Distribution (FULL DATA)



⌚ Summary table (FULL DATA + compact)

Bucket	Items (FULL)	Share of ALL suggestions %	Top 8 bi-grams
Appointments	208	36.49	جز موعد (32)، حل المشكلة (18)، ترجو حل (13)، احجز موعد (12)، اكتب من (11)، لا يدخل (11)، سبني جدا (10)
Profile	18	3.16	فري وضعيه (2)، على حسابي (2)، تسجيل دخول (2)، حدث خطأ (2)، راتب ممك (2)، مدخل في (2)، ارجو حل (2)، طرحو اصلاح (2)
Notifications	10	1.75	ان التطبيق (2)، يعطي تنبية (2)، تنبية ان (2)، يلحق عليه (2)، الان صنكر (2)، ازلة عن (2)، من امسة (2)، اذلة جد (2)
Performance	8	1.40	تكت التطبيق (2)، سرعة حل (2)، معذرة جدا (1)، اقدر (1)، قاتقه بطيء (1)، جدا وعديله (1)، المعلمات زلا (1)
Search	3	0.53	تطبيق داخل (1)، داخل تطبيق (1)، تطبيق مستقر (1)، مستقر (1)، عدد (1)، عدد اليمت (1)، اليمت عن (1)، عن براء (1)
Language	3	0.53	في خطأ (2)، من الفعل (1)، التطبيقات المائية (1)، المائية والتي (1)، والتي ته (1)، تم توفيرها (1)، توفيرها للمراسن (1)، المراسن والاجنبى (1)

◉ Arabic Dataset

The Arabic dataset included **570 user suggestions** representing approximately **0.7%** of all Arabic reviews.

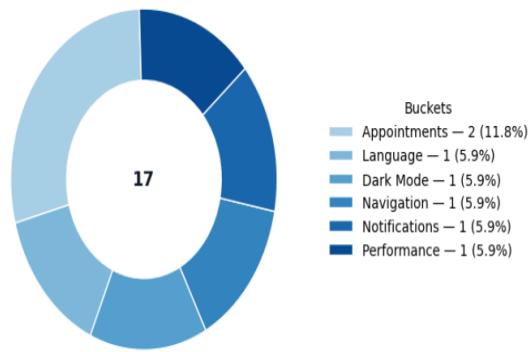
More than **36%** of these suggestions were related to **appointment services** primarily focusing on **improving the booking process, reducing technical glitches and enhancing system responsiveness**.

Frequent phrases such as "**جز موعد**" (**book appointment**), "**حل المشكلة**" (**fix the issue**), and "**ارجو اصلاح**" (**please fix**) reflected users' interest in **smoother functionality** and **fewer errors** during the appointment process.

Additional suggestion areas covered **user profiles (3.2%)**, **notifications (1.8%)**, **performance (1.4%)**, **search (0.5%)** and **language preferences (0.5%)**.

Overall users provided **constructive and improvement-oriented feedback**, emphasizing the importance of **system efficiency**, **bug resolution**, and a **more user-friendly appointment interface**.

Suggestions Distribution (FULL DATA)



Summary table (FULL DATA • compact)

Bucket	Items (FULL)	Share of ALL suggestions %	Top 8 bi-grams
Appointments	2	11.76	absolutely trash (1x), trash trying (1x), trying pas (1x), pas weeks (1x), weeks get (1x), get appointment (1x), appointment vaccine (1x), vaccine getting (1x)
Language	1	5.88	arabic know (2x), want last (1x), first second (1x), last first (1x), covid 19 (1x), 19 vaccine (1x), vaccine update (1x), update helpless (1x)
Dark Mode	1	5.88	please add (1x), add dark (1x), dark mode (1x), mode pages (1x), pages white (1x), white background (1x), background harm (1x), harm eyes (1x)
Navigation	1	5.88	doesn't deserve (1x), deserve one (1x), one rate (1x), rate please (1x), please add (1x), add rate (1x), rate doesn't open (1x)
Notifications	1	5.88	suggestion since (1x), since update (1x), update apps (1x), apps sign (1x), sign everytime (1x), everytime sign (1x), sign appears (1x), appears update (1x)
Performance	1	5.88	suggestion sehaty (1x), sehaty developers (1x), developers kindly (1x), kindly change (1x), change method (1x), method update (1x), update mobile (1x), mobile number (1x)

English Dataset

The English dataset contained **17 user suggestions** accounting for only **0.05%** of all English reviews — a significantly smaller proportion compared to the Arabic dataset. Most of these suggestions focused on **appointment-related improvements (11.8%)**, while others addressed **language settings, dark mode navigation notifications, and performance** each representing around **5.9%**.

Commonly used phrases such as "**add dark mode**" "**please update**" and "**kindly change method**" highlighted users' interest in **enhancing visual comfort, technical optimization and overall usability**.

Although **few in number** these suggestions offer **valuable insights** into potential areas for **improvement** particularly in **interface personalization** and **system functionality updates**.

⚙️ App Not Working Mentions

📊 Summary — "App not working"

✳️ **Phrases:** التطبيق لا يعمل

📊 **Total Mentions:** 806

🏆 **Peak Year:** 2021

📝 **Count in Peak:** 597

📅🔥 Mentions per Year

Year Mentions

2019 2

2020 32

2021 597

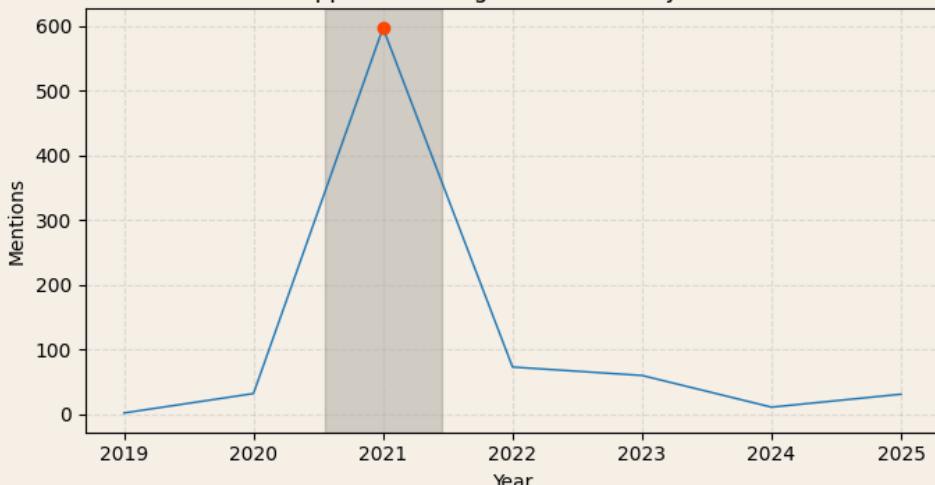
2022 73

2023 60

2024 11

2025 31

"App not working" — Mentions by Year



🌐 Arabic Dataset

"التطبيق لا يعمل" The Arabic dataset recorded a total of **806 mentions** of the phrase (app not working).

The issue reached its **peak in 2021** with **597 mentions** representing the highest number across all years.

Mentions dropped significantly afterward — from **73 in 2022** and **60 in 2023**, to only **11 in 2024** and **31 in 2025**.

This sharp decline suggests that most technical issues were addressed after

2021 likely following major system updates or maintenance improvements. The concentration of complaints in 2021 aligns with the period of rapid app expansion and increased user traffic during the pandemic.

Summary — EN: app not working

Total Mentions: 307
Peak Year: 2021
Count in Peak: 290

Mentions per Year — EN: app not working

Year	Mentions
2020	1
2021	290
2022	9
2023	5
2025	2



English Dataset

The English dataset included **307 mentions** of "app not working" also showing a clear **peak in 2021** with **290 mentions**.

After 2021 the frequency sharply decreased — to **9 in 2022**, **5 in 2023** and **2 in 2025**.

This consistent downward trend mirrors the Arabic dataset indicating a cross-language improvement in app stability and fewer functional errors over time.

The early spike reflects early-stage performance issues while the later stability suggests effective system optimization and better user experience.

Developer Replies

Summary — Developer Replies

 Total reviews: 86,120

 Total replies: 30,192

 Reply rate: 35.06%

 Columns: ID = — • Comment =  Content • Reply =  DeveloperReply

Sample of Developer Replies (n=30)

1. ID: 1

 User Comment: شكرًا جزيلاً لكم على خدمة المجتمع وخدمة المقيمين كان الله في عونكم وسدّد خطاكُم ووفقكم على فعل الخبر وشكراً لكم
 Developer Reply: شكرًا لمشاركتنا رأيك، نتمنى لكم دوام الصحة والعافية

2. ID: 2

 User Comment: جيد
 Developer Reply: شكرًا لمشاركتنا رأيك، نتمنى لكم دوام الصحة والعافية

3. ID: 3

 User Comment: جده جداً
 Developer Reply: شكرًا لمشاركتنا رأيك، نتمنى لكم دوام الصحة والعافية

4. ID: 4

 User Comment: ها
 Developer Reply: شكرًا لمشاركتنا رأيك، نتمنى لكم دوام الصحة والعافية

5. ID: 5

 User Comment: في مشكله في قراءه العلامات الحبويه تيجي رساله تم حظر التطبيق
 Developer Reply: أو عبر بريدينا : شاكرين لك وقتكم وتقيمكم، كما نعتذر عن الصعوبات التي تواجهك لطفاً تواصل معنا عبر حسابنا الرسمي في توينر "Support@sehhaty.sa"

6. ID: 6

 User Comment: اشكر القائمين على تطبيق صحتي كل يوم يتجدد وينتطور
 Developer Reply: شكرًا لمشاركتنا رأيك، نتمنى لكم دوام الصحة والعافية

7. ID: 7

 User Comment: كان البرنامج في بدايته كامل الخدمات أما الان فأصبح لا ينفذ جميع الخدمات مثال التحاليل لا تنزل بالموقع الزيارات لا تسجل المواقع عن بعد بخراجل ... البرنامج قبل التحدث مع الدكторون وأهم شي لا توج طريقة للتواصل مع القائمين على
 Developer Reply: أو عبر X : شاكرين لك وقتكم وتقيمكم، كما نعتذر عن الصعوبات التي تواجهك لطفاً تواصل معنا عبر حسابنا الرسمي في منصة "Support@sehhaty.sa" أو عبر بريدينا الإلكتروني: Support@sehhaty.sa"

Arabic Dataset

The Arabic dataset included **30,192 developer replies** out of **86,120 total reviews** with an overall **reply rate of 35.06%**.

Most replies were short polite and standardized commonly expressing appreciation such as:

"شكراً لمشاركتك، ونتمنى لك دوام الصحة والعافية"

This pattern indicates a **formal and consistent communication style**, focusing on gratitude and well-being.

Some replies also included **technical guidance**, especially for problem-related comments — for instance, suggesting users contact support via official email or Twitter handle (@Sehhatyapp).

This shows a balance between **automated courtesy responses** and **manual support engagement** reflecting an effort to maintain professionalism and responsiveness.

Summary — Developer Replies

 **Total reviews:** 34,441

 **Total replies:** 10,218

 **Reply rate:** 29.67%

 **Columns:** ID = — • Comment =  Content • Reply =  DeveloperReply

Sample of Developer Replies (n=30)

1. ID: 1

 **User Comment:** good

 **Developer Reply:** we would like to thank for your feedback on the app and sharing your experience with us.

2. ID: 2

 **User Comment:** nice app

 **Developer Reply:** we would like to thank for your feedback on the app and sharing your experience with us

3. ID: 3

 **User Comment:** Best

 **Developer Reply:** we would like to thank for your feedback on the app and sharing your experience with us

4. ID: 4

 **User Comment:** good

 **Developer Reply:** we would like to thank for your feedback on the app and sharing your experience with us.

5. ID: 5

 **User Comment:** excellent , like always . ❤️❤️❤️

 **Developer Reply:** we would like to thank for your feedback on the app and sharing your experience with us.

6. ID: 6

 **User Comment:** Good Apps

 **Developer Reply:** Thanks, for the feedback on your experience we sincerely appreciate your insight.

7. ID: 7

 **User Comment:** good

 **Developer Reply:** we would like to thank for your feedback on the app and sharing your experience with us.

English Dataset

The English dataset contained **10,218 developer replies** from **34,441 total reviews**, with a **reply rate of 29.67%** — slightly lower than the Arabic dataset.

Most replies followed a uniform format, such as:

"We would like to thank you for your feedback on the app and sharing your experience with us."

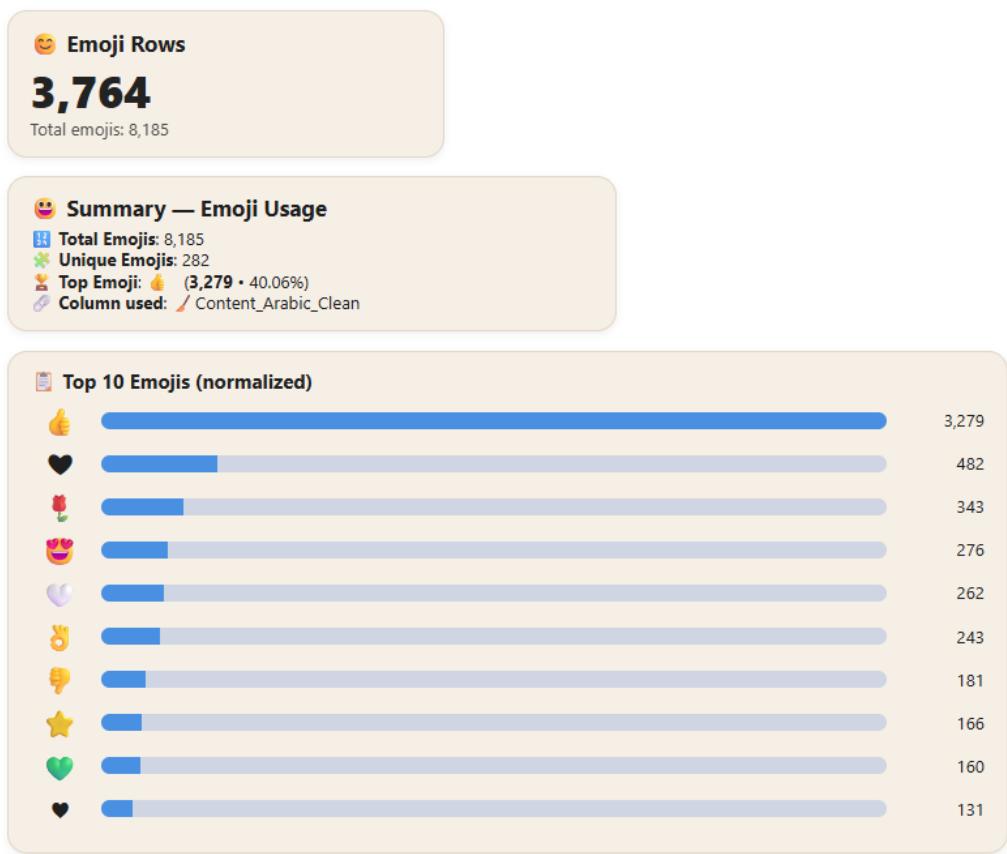
The tone was **formal and appreciative**, emphasizing acknowledgment of feedback rather than detailed problem-solving.

However, some variations appeared, such as:

"Thanks for the feedback on your experience. We sincerely appreciate your insight. This suggests that while English replies are also largely standardized, there is occasional personalization for user engagement.

Overall both datasets demonstrate a **strong customer communication strategy** though Arabic replies show slightly more **contextual responsiveness** to user issues.

😊 Emoji Usage Analysis



● Arabic Dataset

A total of **8,185 emojis** were detected across **3,764 reviews** representing rich emotional engagement from users.

There were **282 unique emojis** indicating a wide range of emotional expression.

The most used emoji was **👍 (thumbs up)** appearing **3,279 times (40.06%)**, followed by **❤️ (black heart)** and **🌹 (rose)**.

These symbols reflect **positive appreciation, respect, and gratitude** commonly used to express satisfaction and emotional connection with the app or its services.

Other frequent emojis such as **😍 (heart eyes)**, **💕 (white heart)** and **👏 (clapping hands)** further emphasize **positive emotional tone and encouragement**.

Overall emoji use in Arabic reviews suggests **high engagement and emotional positivity** particularly among satisfied users.

😊 Reviews containing emojis

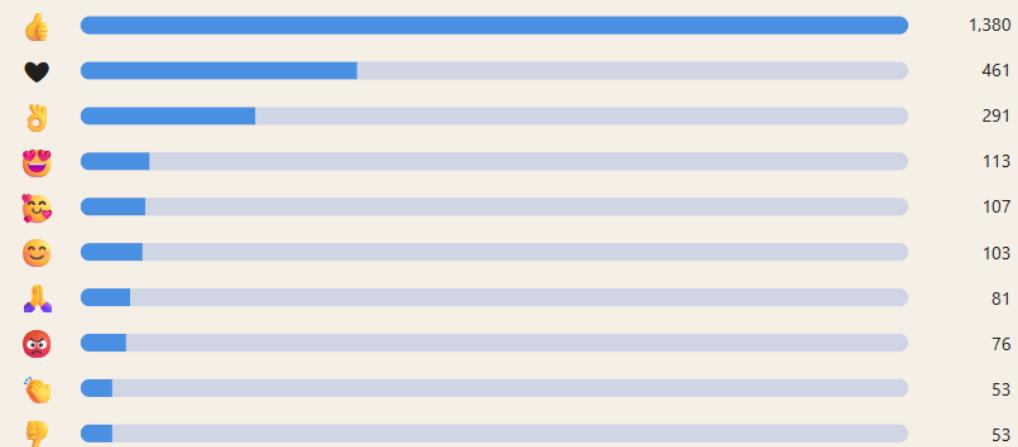
2,023

Total emojis found: 4,102

emoji Summary — Emoji Usage

>Total Emojis: 4,102
Unique Emojis: 228
Top Emoji: 🤝 (1,380 • 33.64%)
Column used: Content

Top 10 Emojis (normalized)



English Dataset

The English dataset contained **4,102 emojis** across **2,023 reviews**, with **228 unique emojis** identified.

Similarly, the **👉 emoji** was the most dominant (1,380 times, 33.64%) followed by **❤️**, **👌** and **😍**.

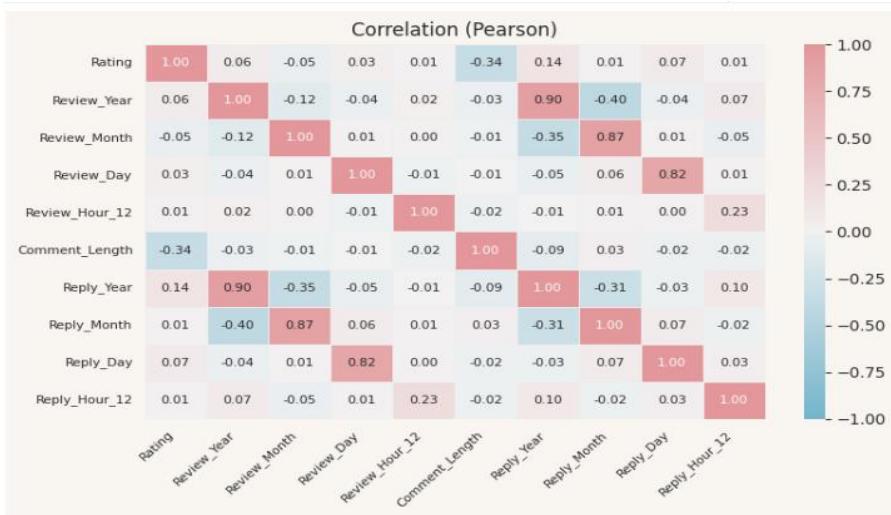
These results indicate that English-speaking users also used emojis to convey **positivity, approval, and friendliness**.

While the volume of emojis was lower than in Arabic reviews, the emotional intent remained largely **positive and appreciative**.

Emojis such as **😊**, **🥰**, and **👏** highlight the friendly and encouraging tone users maintain when interacting with the app.

Together both datasets reveal that emoji use serves as a **universal indicator of user satisfaction**, reinforcing textual sentiment and offering valuable cues for emotional analysis.

Correlation Analysis



Correlation Matrix (table)

	Rating	Review_Year	Review_Month	Review_Day	Review_Hour_12	Comment_Length	Reply_Year	Reply_Month	Reply_Day	Reply_Hour_12
Rating	1.00	0.06	-0.05	0.03	0.01	-0.34	0.14	0.01	0.07	0.01
Review_Year	0.06	1.00	-0.12	-0.04	0.02	-0.03	0.90	-0.40	-0.04	0.07
Review_Month	-0.05	-0.12	1.00	0.01	0.00	-0.01	-0.35	0.87	0.01	-0.05
Review_Day	0.03	-0.04	0.01	1.00	-0.01	-0.01	-0.05	0.06	0.82	0.01
Review_Hour_12	0.01	0.02	0.00	-0.01	1.00	-0.02	-0.01	-0.01	0.01	0.23
Comment_Length	-0.34	-0.03	-0.01	-0.01	-0.02	1.00	-0.09	0.03	-0.02	-0.02
Reply_Year	0.14	0.90	-0.35	-0.05	-0.01	-0.09	1.00	-0.31	-0.03	0.10
Reply_Month	0.01	-0.40	0.87	0.06	0.01	0.03	-0.31	1.00	0.07	-0.02
Reply_Day	0.07	-0.04	0.01	0.82	0.00	-0.02	-0.03	0.07	1.00	0.03
Reply_Hour_12	0.01	0.07	-0.05	0.01	0.23	-0.02	0.10	-0.02	0.03	1.00

Summary — Correlations

Numeric columns used: 10

Top positive correlations

Review_Year × Reply_Year — $r = 0.90$ (Very strong positive)
Review_Month × Reply_Month — $r = 0.87$ (Very strong positive)
Review_Day × Reply_Day — $r = 0.82$ (Very strong positive)

Top negative correlations

Review_Year × Reply_Month — $r = -0.40$ (Strong negative)
Review_Month × Reply_Year — $r = -0.35$ (Moderate negative)
Rating × Comment_Length — $r = -0.34$ (Moderate negative)

Top correlations with Rating

Rating × Comment_Length — $r = -0.34$ (Moderate negative)
Rating × Reply_Year — $r = 0.14$ (Weak/Negligible positive)
Rating × Reply_Day — $r = 0.07$ (Weak/Negligible positive)
Rating × Review_Year — $r = 0.06$ (Weak/Negligible positive)
Rating × Review_Month — $r = -0.05$ (Weak/Negligible negative)

Quick guide: $|r| \geq 0.70$ = Very strong • $0.40-0.69$ = Strong • $0.20-0.39$ = Moderate • < 0.20 = Weak/Negligible

Arabic Dataset

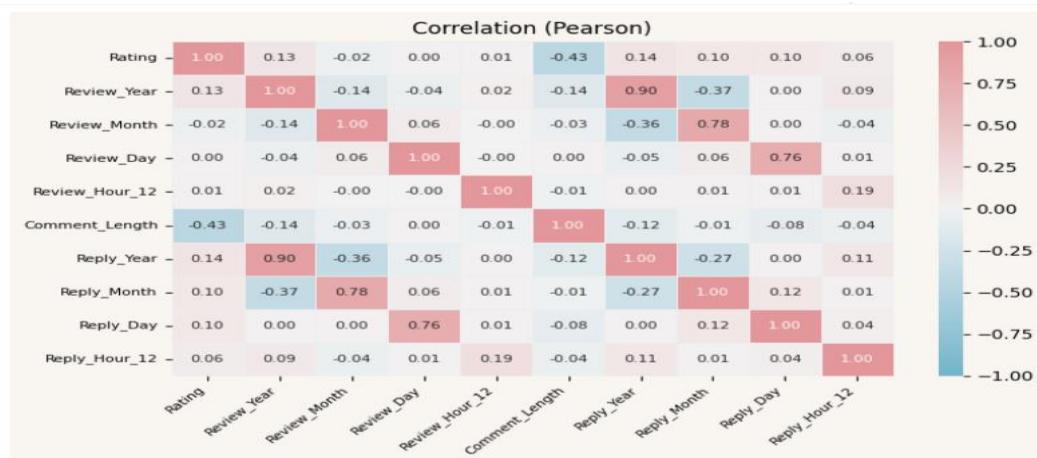
The correlation matrix revealed several **very strong positive relationships** among time-related variables particularly:

- Review Year × Reply Year ($r = 0.90$)** — A very strong positive correlation indicating that most replies were provided within the same year as the reviews.
- Review Month × Reply Month ($r = 0.87$)** — A very strong positive correlation reflecting consistent reply activity aligned with review posting periods.

- **Review Day × Reply Day ($r = 0.82$)** — A very strong positive correlation, showing that replies often occurred on or near the same day the review was posted.

In contrast, notable **negative correlations** were observed:

- **Review Year × Reply Month ($r = -0.40$)** — A strong negative correlation suggesting minor temporal variations in some replies.
- **Review Month × Reply Year ($r = -0.35$)** — A moderate negative correlation.
- **Rating × Comment Length ($r = -0.34$)** — A moderate negative relationship, indicating that **shorter reviews tended to have higher ratings** while **longer reviews were often more critical** and detailed.



Correlation Matrix (table)

	Rating	Review_Year	Review_Month	Review_Day	Review_Hour_12	Comment_Length	Reply_Year	Reply_Month	Reply_Day	Reply_Hour_12
Rating	1.00	0.13	-0.02	0.00	0.01	-0.43	0.14	0.10	0.10	0.06
Review_Year	0.13	1.00	-0.14	-0.04	0.02	-0.14	0.90	-0.37	0.00	0.09
Review_Month	-0.02	-0.14	1.00	0.06	-0.00	-0.03	-0.36	0.78	0.00	-0.04
Review_Day	0.00	-0.04	0.06	1.00	-0.00	0.00	-0.05	0.06	0.76	0.01
Review_Hour_12	0.01	0.02	-0.00	-0.00	1.00	-0.01	0.00	0.01	0.01	0.19
Comment_Length	-0.43	-0.14	-0.03	0.00	-0.01	1.00	-0.12	-0.01	-0.08	-0.04
Reply_Year	0.14	0.90	-0.36	-0.05	0.00	-0.12	1.00	-0.27	0.00	0.11
Reply_Month	0.10	-0.37	0.78	0.06	0.01	-0.01	-0.27	1.00	0.12	0.01
Reply_Day	0.10	0.00	0.00	0.76	0.01	-0.08	0.00	0.12	1.00	0.04
Reply_Hour_12	0.06	0.09	-0.04	0.01	0.19	-0.04	0.11	0.01	0.04	1.00

Summary — Correlations

Numeric columns used: 10

Top positive correlations

Review_Year × Reply_Year — $r = 0.90$ (Very strong positive)
 Review_Month × Reply_Month — $r = 0.78$ (Very strong positive)
 Review_Day × Reply_Day — $r = 0.76$ (Very strong positive)

Top negative correlations

Rating × Comment_Length — $r = -0.43$ (Strong negative)
 Review_Year × Reply_Month — $r = -0.37$ (Moderate negative)
 Review_Month × Reply_Year — $r = -0.36$ (Moderate negative)

Top correlations with Rating

Rating × Comment_Length — $r = -0.43$ (Strong negative)
 Rating × Reply_Year — $r = 0.14$ (Weak/Negligible positive)
 Rating × Review_Year — $r = 0.13$ (Weak/Negligible positive)
 Rating × Reply_Day — $r = 0.10$ (Weak/Negligible positive)
 Rating × Reply_Month — $r = 0.10$ (Weak/Negligible positive)

Quick guide: $|r| \geq 0.70$ = Very strong • $0.40-0.69$ = Strong • $0.20-0.39$ = Moderate • < 0.20 = Weak/Negligible

English Dataset

A similar pattern was observed, with strong temporal consistency:

- **Review Year × Reply Year ($r = 0.90$)** — Very strong positive correlation.
- **Review Month × Reply Month ($r = 0.78$)** — Very strong positive correlation.
- **Review Day × Reply Day ($r = 0.76$)** — Very strong positive correlation.

As for negative relationships:

- **Comment Length × Rating ($r = -0.43$)** — A moderate negative correlation, reinforcing that **positive reviews were shorter and more concise**, while **negative ones tended to be longer and more descriptive**.
- **Review Year × Reply Month ($r = -0.37$)** — A moderate negative correlation, suggesting some delays or inconsistencies in developer replies during specific periods.

Overall Interpretation

The analysis indicates that **time-based variables (year, month, day)** are highly correlated showing a **consistent and timely reply pattern** from developers. Meanwhile the **negative correlation between comment length and rating** highlights a behavioral trend: satisfied users express themselves briefly whereas dissatisfied users tend to write longer more detailed feedback. Overall the correlation results demonstrate **stable engagement between users and developers over time** with natural variations reflecting real-world interaction patterns.

Conclusion

The preliminary analysis revealed meaningful patterns and clear trends across both datasets, highlighting users' overall satisfaction, key technical challenges, and interaction dynamics.

These findings not only demonstrate the quality and readiness of the cleaned data but also set the stage for deeper analytical exploration in the next phase.