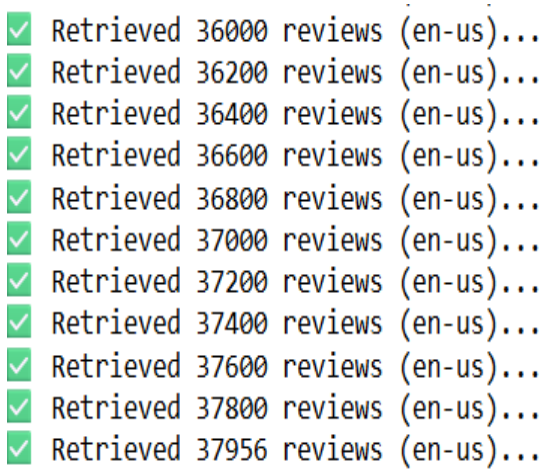
✨ Draft Report: Data Cleaning Sehhaty App Reviews

1. Overview

There were **128,406** user reviews collected using the Sehhaty mobile application. Data cleaning was aimed to preprocess for providing quality by removing invalid values, eliminating short or absent reviews, deleting duplicates, and handling missing values, thereby preparing the dataset for future analysis. Reviews were collected from the Google Play Store between **2019 and 2025**, providing a longitudinal dataset in its entirety. Arabic and English reviews were cleaned and split apart individually with language-specific stopword lists and error correction rules to meet linguistic consistency and prevent cross-language contamination. This laid a strong foundation for the following analyses such as sentiment analysis and topic modeling



2. Cleaning Steps

1️⃣ **Invalid Values Removal**

✅ Ratings were restricted to the valid range **[1–5]**.

2️⃣ **Text Cleaning**

✅ Special symbols, repeated characters, diacritics, and URLs were removed.

✅ Language-specific stopword lists were applied to eliminate common phrases such as greetings and filler words, reducing linguistic noise.

✅ Emojis were intentionally retained, as they are important for expressing sentiment and emotional context and thus improve the accuracy of sentiment analysis.

```
# ---------- Stopwords ----------
STOPWORDS_AR_SENTIMENT = {
    'وعليكم','عليكم','السلام عليكم','السلام',
    'مساء الخير','صباح الخير',
    'مرحبا','اهلاً','اهلا','اهلين','هلا',
    'اللهم','امين','آمين','يارب',
    'كذا','ترى','اصلا','اصلاً','اصلًا',
    'بس','يعني','اوكي','اوك','كمان',
    'موه','امم','ممم','خخخ','خخ','هههههه','هههه','هه','
    'ال','...','…',
    'apk','play','http','https','www','com','org',
    '500','11','666',
    'وبركاتة','ورحمه','ورحمة','ذا'
}
```

```
# ================ English Stopwords & Fixes =======
STOPWORDS_EN_SENTIMENT = {
    'the','a','an','and','or','but','to','of','in','on
    'it','its','this','that','these','those','there','
    'also','just','still',
    'i','me','my','we','our','you','your','they','them
    'be','am','is','are','was','were','been','being',
    'do','did','does','doing',
    'have','has','had','having',
    'will','would','should','could','can','may','might
    'if','then','than','because','while','when','where
    'etc','etc.','etc..',
    'hi','hello','or','is','igigi'
}
```

3. Text Normalization and Spelling Corrections

📝 Instead of discarding short reviews, a **dictionary of common spelling corrections (COMMON_FIX)** was applied for both Arabic and English text.

- ✅ This step converted frequently misspelled words into their correct forms
  Arabic: ممتازة ⟶ ممتازه ، رائع ⟶ رايع
- **English:** dont ⟶ don't

✨ By using correction rather than deletion, even short but meaningful reviews were preserved, ensuring that no potentially important words or information for the analysis were lost.

```
# ---------- Common Fix ----------
COMMON_FIX = {
    'ممتاز':'ممتاز','ممتاز':'ممتاز','متاز':'ممتاز','ممتاز':'كمتاز','ممتازة':'
    'رايع ممتاز':'رائع ممتاز',
    'ممتاز جدا':'ممتاز جدا','ممتازجدا':'ممتاز جدا','ممتازجداً':'ممتازجداً',
    'رائع':'رائع','رايعه':'رائعة','رااائع':'رائع','رااائع':'رائع',
    'روووعة':'روعة','روعة':'روعة','رائع جدا':'رائع جدا',
    'جميل':'جميل','جميلل':'جميل','جميله':'جميلة','جميلة':'جميلة',
    'جدا':'جداً','جدن':'جداً','جيد جدن':'جيد جداً',
    'حلو':'حلو','حلووو':'حلو','حلوووو':'حلو',
    'سيئ':'سيئ','سيء':'سيئ','سيئ':'سيئة','سيئه':'سيئة',
    'سيي':'سيئ',
    'سي':'سيئ',
    'سي جدا':'سيئ جدا',
    'بطئ':'بطيء','بطيئ':'بطيء',
    'بطي':'بطيء',
    'التمام':'تمام','ال تمام':'تمام','ال كثير':'الكثير',
    'مشاء':'ما شاء','شكرا':'شكراً','شكرن':'شكراً',
    'ممتاز جدا':'مميه مميه',
    'واش كركم':'اشكركم','م يفتح':'اشكركم',
    'مشكله':'المشكلة','المشكله':'المشكلة',
    'التطبيق لايعمل':'التطبيق لا يعمل',
    '👍':'👍','👍':'👍','👍':'👍','👍':'👍','👍':'👍',
}
```

```
COMMON_FIX_EN = {
    "dont":"don't","cant":"can't","wont":"won't","didn
    "isnt":"isn't","wasnt":"wasn't","shouldnt":"should
    "im":"i'm","ive":"i've","id":"i'd","ill":"i'll","u
    "plz":"please","thx":"thanks","thanx":"thanks",
    "definately":"definitely","definetly":"definitely"
    "adress":"address","addres":"address","occured":"o
    "enviroment":"environment","acheive":"achieve","wi
    "apointment":"appointment","appointement":"appoint
    "dose":"does","gud":"good","bcoz":"because","bcz":
    "aap":"app","goid":"good"
}
```

📸 Snapshot of the Preprocessing Code

🖼️ The images above show a snapshot of the preprocessing code, demonstrating the use of **language-specific stopword lists** and **spelling correction dictionaries (COMMON_FIX)** for both Arabic and English.

✅ These steps ensured **linguistic consistency** by removing irrelevant words and normalizing frequent misspellings before analysis.

Arabic — Stopwords & Corrections

Original	After	Status	Action
آمين	(removed)	❌	Delete
ابل	(removed)	❌	Delete
اصلا	(removed)	❌	Delete
اصلاً	(removed)	❌	Delete
اصلًا	(removed)	❌	Delete
الخ	(removed)	❌	Delete
الخ.	(removed)	❌	Delete
السلام	(removed)	❌	Delete
السلام عليكم	(removed)	❌	Delete
الصراحة	(removed)	❌	Delete
الله	(removed)	❌	Delete
اللهم	(removed)	❌	Delete
بطئ	بطيء	✔	Correct
بطي	بطيء	✔	Correct
بطيئ	بطيء	✔	Correct
جميلل	جميل	✔	Correct
جميله	جميلة	✔	Correct
جمييل	جميل	✔	Correct
جمييله	جميلة	✔	Correct
جمييييل	جميل	✔	Correct
حلووو	حلو	✔	Correct
حلووووو	حلو	✔	Correct
راائع	رائع	✔	Correct
راااائع	رائع	✔	Correct

English — Stopwords & Corrections

Original	After	Status	Action
a	(removed)	❌	Delete
also	(removed)	❌	Delete
am	(removed)	❌	Delete
an	(removed)	❌	Delete
and	(removed)	❌	Delete
are	(removed)	❌	Delete
as	(removed)	❌	Delete
at	(removed)	❌	Delete
be	(removed)	❌	Delete
aap	app	✔	Correct
accomodate	accommodate	✔	Correct
acheive	achieve	✔	Correct
addres	address	✔	Correct
adress	address	✔	Correct
apointment	appointment	✔	Correct
appointement	appointment	✔	Correct
apponitment	appointment	✔	Correct
bc	because	✔	Correct
bcoz	because	✔	Correct
bcz	because	✔	Correct
cant	can't	✔	Correct

🎨 **Illustration of Preprocessing Steps**

📋 The lists above illustrate examples of text preprocessing steps:

- ❌ **Delete:** Certain words were removed as stopwords.
- ✅ **Correct:** Other words were corrected for spelling and wording.

✨ These actions helped keep the dataset **clean, normalized, and ready** for subsequent analysis.

4. Duplicates Removal

♻️ **Process:**

- 🗂️ Duplicates were removed based on **(👤 UserName + 📝 Content)**.
- 🚫 Generic placeholders such as *("A Google user" / "مستخدم Google")* were ignored during duplicate detection.
- ✅ Only the first occurrence for each pair was kept.
- 🧻 Only duplicate rows were removed; no other columns were modified.

📊 **Results:**

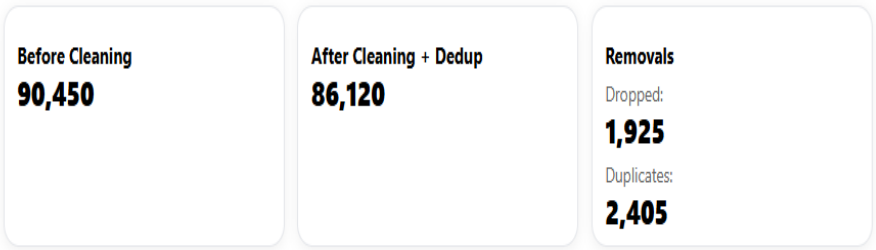
- **Arabic:** pair count **30 → 1** after deduplication.
- **English:** pair count **3 → 1** after deduplication.

📊 **Cleaned & Deduplicated Results (Arabic & English)**

🔹 Shows record counts **before and after cleaning** for both languages.

🔹 Displays how duplicate pairs were removed (**keeping only the first occurrence** for each pair).

🔹 Highlights **samples of cleaned Arabic and English reviews**.



Arabic — Before (User + CLEAN) | pair count: 30

UserName	Content_Arabic_Clean
ابو محمد	ممتاز
ابو محمد	ممتاز
ابو محمد	ممتاز
ابو محمد	ممتاز
ابو محمد	ممتاز
ابو محمد	ممتاز
ابو محمد	ممتاز
ابو محمد	ممتاز
ابو محمد	ممتاز

Arabic — After (dedup → 1)

UserName	Content_Arabic_Clean
ابو محمد	ممتاز

📊 Sample: Output (Before vs Cleaned)

📌 Before_Content	🧹 Cleaned_Content
ممتاز جدا	ممتاز جدا
ممتاز جداً	ممتاز جدا
ممتاز	ممتاز
ممتاز جدا جداً	ممتاز جدا
ممتاز ومفيد	ممتاز ومفيد
رايع جدا ويختصر الوقت	رائع جدا ويختصر الوقت
ممتاذ	ممتاز
جيد جداً	جيد جدا
ممتاز جداً	ممتاز جدا
حلووووو	حلو
ممتاز	ممتاز
ممتاز للغاية	ممتاز للغاية

Before Cleaning
37,956

After Cleaning + Dedup
34,441

Removals
Dropped:
2,985
Duplicates:
310

English — Before (User + CLEAN) | pair count: 3

UserName	Clean_Content
Shahid Ali	good
Shahid Ali	good
Shahid Ali	good

English — After (dedup → 1)

UserName	Clean_Content
Shahid Ali	good

📊 **Sample: Output (Before vs Cleaned)**

📌 Before_Content	🧹 Cleaned_Content
good	good
thank you so much for helping us inshallah 👍	thank so much helping us inshallah 👍
good	good
So good	so good
Excellent	excellent
good	good
google play store very good app internet work very fast	google play store very good app internet work very fast
most beautiful aap	most beautiful app
nice	nice
good	good
Very good	very good
good luck	good luck

5. Handling Missing Values

✅ The blank text columns were replaced with NA.

✅ Date columns containing blank cells were converted to text format and filled with NA.

✅ The dataset is now completely free of empty or missing values.

6. Manual Checking and Cleaning

🧹 In addition to the automated treatment, a **manual verification** was performed to detect and delete or replace any remaining non-Arabic or non-English text.

🧹 This step ensured that the dataset was **standardized, consistent, and fully ready** for further analyses.

ম ন্না ঙঙ ত শ؛ یہ ایپ اچھی ہے ا ل ح م د ل ل ہ ا ل ح لے سے بھی کھراب

3. Results

- 📥 **Before cleaning:** 128,406 rows
- 📤 **After cleaning:** 120,561 rows
- 🗑️ **Removed:** 7,845 rows (6.1%)

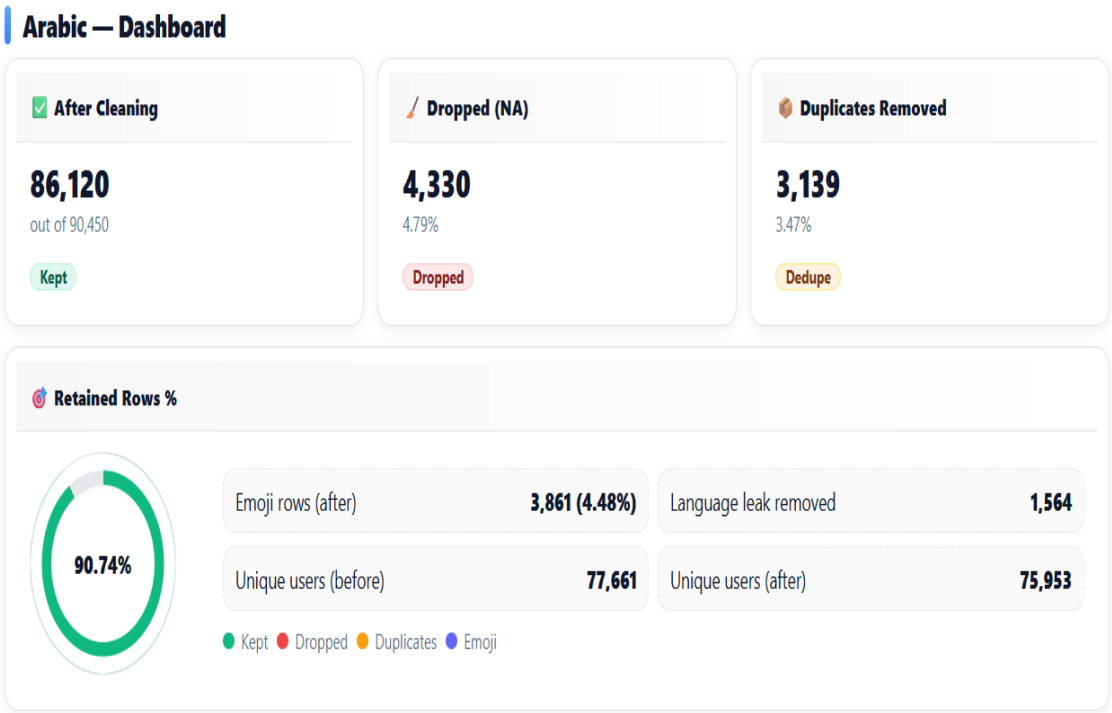
📊 **Language Distribution (After):**

- SA **Arabic:** 90,450 → 86,120 (removed 4,330 | 4.8%)
- GB **English:** 37,956 → 34,441 (removed 3,515 | 9.3%)

📊 Data Cleaning Dashboard

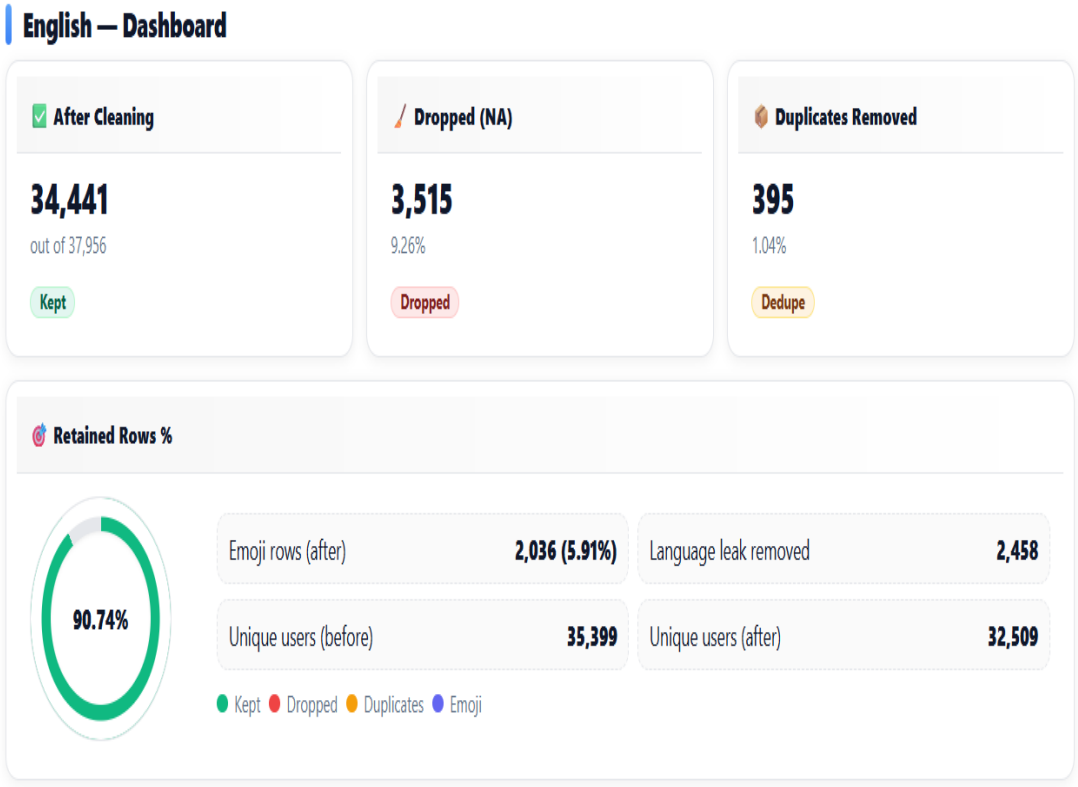
An interactive dashboard was created to summarize the results of the data cleaning process for both Arabic and English reviews:

- 📊 **Key metrics displayed:** number of rows before and after cleaning, dropped rows (NA), duplicates removed, and percentage of retained rows.
- 💡 **Additional insights:** unique users before and after cleaning, percentage of rows containing emojis, and language-leak removal counts.
- 🎯 **Outcome:** These dashboards provide a clear visual representation of how the dataset was standardized and prepared for analysis.



Summary

Metric	Value	Extra
Total Rows (Before)	**90,450**	
Valid Rows (After)	**86,120**	95.21%
Dropped (NA)	**4,330**	
Duplicates Removed	**3,139**	
Language leak removed	**1,564**	
Unique Users (Before)	**77,661**	
Unique Users (After)	**75,953**	
Emoji Rows (After)	**3,861**	4.48%
Avg Length (Before)	**21**	
Avg Length (After)	**21**	
Median Length (After)	**7**	
URLs (Before)	**0**	
URLs (After)	**0**	
URLs Removed	**0**	



Summary

Metric	Value	Extra
Total Rows (Before)	**37,956**	
Valid Rows (After)	**34,441**	90.74%
Dropped (NA)	**3,515**	
Duplicates Removed	**395**	
Language leak removed	**2,458**	
Unique Users (Before)	**35,399**	
Unique Users (After)	**32,509**	
Emoji Rows (After)	**2,036**	5.91%
Avg Length (Before)	**26**	
Avg Length (After)	**20**	
Median Length (After)	**9**	
URLs (Before)	**0**	
URLs (After)	**0**	
URLs Removed	**0**	

4. Quality Assurance

✅ **No missing values:** empty/blank entries were standardized to NA.

✅ **Ratings check:** all ratings are restricted to the valid range [1–5].

✅ **No duplicates remain:** *(UserName + Clean_Content)* pairs were deduplicated — first occurrence kept; generic placeholders such as *"A Google user / مستخدم Google"* were ignored during duplicate detection.

🔎 **Language Leakage Check:**

• SA **Arabic:** 0 English-only rows (NA-like blanks excluded).

• GB **English:** 0 rows containing Arabic letters.

🔎 **Arabizi Removal Check:**

The dataset was additionally scanned for **Arabizi words** (transliterated Arabic written in Latin letters, e.g., *alhamdulillah*, *mashallah*, *tmam*).

📊 **Sample Before/After Cleaning:**

(Table below shows the first 15 affected rows before and after cleaning — demonstrating that Arabizi words were removed while keeping meaningful English words intact).

🔎 Number of rows containing Arabizi: 223

	Row_Number	Before	After
0	3	thank so much helping us inshallah 👍	thank so much helping us 👍
1	213	alhamdulillah	NA
2	291	tmam	NA
3	324	acha	NA
4	475	mashallah good excellent 👍	good excellent 👍
5	558	alhamdulillah	NA
6	585	masha allah very good app	very good app
7	678	alhamdulillah 's good	s good
8	832	mashallah very very good	very very good
9	1025	very good mashallah	very good
10	1240	alhamdulillah	NA
11	1345	alhamdulillah	NA
12	1370	alhamdulillah allahuakbar	NA
13	1408	masha allah very nice	very nice
14	1453	mashallah	NA

✅ **Verification Result:**
The scan confirmed that **0 rows containing Arabizi words remain** — indicating that the dataset is fully clean and safe for downstream analysis.

✅ The file is clean — no Arabizi words detected.

✅ **Status:** CLEAN 100% for both Arabic and English.

🔎 **Final Quality Check**
The following tables present the final quality assurance results after cleaning and deduplication:

- SA **Arabic Data:** 86,120 rows retained, 0 duplicates, 0 English-only rows, all ratings within [1–5].
- GB **English Data:** 34,441 rows retained, 0 duplicates, 0 rows with Arabic letters, all ratings within [1–5].

✅ **Final Status:** Both Arabic and English datasets are confirmed **CLEAN 100%** and fully ready for analysis.

Arabic — After Cleaning

Arabic — After Cleaning

	Rows (After)	Dup (User+Content) — After	Dup (Content) — After	Dup (Clean Col) — After	Leak Rows (after)	Missing Rating %	Rating Dist (top)	User Col	Content Col	Clean Col	Rating Col
0	86,120	0	54,910	56,449	0	0.00%	{"1": 12085, "2": 1822, "3": 2837, "4": 4139, "5": 65237}	UserName	Content	Content_Arabic_Clean	Rating

English — After Cleaning

English — After Cleaning

	Rows (After)	Dup (User+Content) — After	Dup (Content) — After	Dup (Clean Col) — After	Leak Rows (after)	Missing Rating %	Rating Dist (top)	User Col	Content Col	Clean Col	Rating Col
1	34,441	0	21,030	22,923	0	0.00%	{"1": 5558, "2": 1023, "3": 1561, "4": 2858, "5": 23441}	UserName	Content	Content_English_Clean	Rating

Quality Check (READ ONLY — No NA Reclassification)

Language	Rows_Total	NonEmpty_Content	English-only (in Arabic)	Arabic-letters (in English)	User+Content Dups (rows)	UC Check	Status
Arabic	86,120	86,120	0	-	0	OK	CLEAN 100%
English	34,441	34,441	-	0	0	OK	CLEAN 100%

5. Attachments

Sample of Cleaned Data: ⬇️ *(screenshot)*

Final Cleaned File: *Sehhaty_Database_Clean* — ready for analysis ✅

Available in two formats: Excel (.xlsx) and CSV (.csv)

Arabic Sample — Important Columns

UserName	Rating	Content	DeveloperReply	Review_Year	Review_Month	Review_Day	Review_Hour_12	Content_Arabic_Clean
حميدالله عبد الغفار بديع الزمن	1	ممتاز جدا	شكراً لمشاركتنا رأيك ،تمنى لكم دوام الصحة والعافية	2025	8	7	7	ممتاز جدا
Emad ahmed	5	ممتاز جداً	شكراً لمشاركتنا رأيك ،تمنى لكم دوام الصحة والعافية	2025	8	7	7	ممتاز جدا
samira posaily	5	ممتاز	شكراً لمشاركتنا رأيك ،تمنى لكم دوام الصحة والعافية	2025	8	7	5	ممتاز
للجروح أرواااح	5	ممتاز جدا جداً	شكراً لمشاركتنا رأيك ،تمنى لكم دوام الصحة والعافية	2025	8	7	3	ممتاز جدا جدا
زياد الكندي	5	ممتاز ومفيد	شكراً لمشاركتنا رأيك ،تمنى لكم دوام الصحة والعافية	2025	8	7	1	ممتاز ومفيد
Mohamed Ahmad Elbealy	5	رائع جدا ويختصر الوقت	شكراً لمشاركتنا رأيك ،تمنى لكم دوام الصحة والعافية	2025	8	7	12	رائع جدا ويختصر الوقت
الحاج حبابلاوي	5	ممتاز	شكراً لمشاركتنا رأيك ،تمنى لكم دوام الصحة والعافية	2025	8	7	11	ممتاز
Mahmoud Saeed	4	جيد جداً	شكراً لمشاركتنا رأيك ،تمنى لكم دوام الصحة والعافية	2025	8	7	9	جيد جدا
علي	5	ممتاز جداً	شكراً لمشاركتنا رأيك ،تمنى لكم دوام الصحة والعافية	2025	8	7	6	ممتاز جدا
Zz Bb	5	حلووو	شكراً لمشاركتنا رأيك ،تمنى لكم دوام الصحة والعافية	2025	8	7	5	حلو
ابو ياسر	5	ممتاز	شكراً لمشاركتنا رأيك ،تمنى لكم دوام الصحة والعافية	2025	8	7	4	ممتاز
علاء هاني	5	ممتاز للغاية	شكراً لمشاركتنا رأيك ،تمنى لكم دوام الصحة والعافية	2025	8	7	2	ممتاز للغاية

.

English Sample — Important Columns

UserName	Rating	Content	DeveloperReply	Review_Year	Review_Month	Review_Day	Review_Hour_12	Content_English_Clean
Khan Sohrab	3	good	we would like to thank for your feedback on the app and sharing your experience with us.	2025	9	3	12	good
Mine 14	5	thank you so much for helping us inshallah 👍	we would like to thank for your feedback on the app and sharing your experience with us.	2025	9	2	9	thank so much helping us inshallah 👍
Saleem Sulemani	5	good	we would like to thank for your feedback on the app and sharing your experience with us.	2025	9	2	9	good
Aftab Choudhary	5	So good	we would like to thank for your feedback on the app and sharing your experience with us.	2025	9	2	9	so good
Fahad Alswailim	5	Excellent	we would like to thank for your feedback on the app and sharing your experience with us.	2025	9	2	8	excellent
Ansari Aslam	5	good	we would like to thank for your feedback on the app and sharing your experience with us.	2025	9	2	4	good
Aaftab Rizvi	5	google play store very good app internet work very fast	we would like to thank for your feedback on the app and sharing your experience with us.	2025	9	2	3	google play store very good app internet work very fast

In [ ]:
```
📊 Total number of columns: 17
👤 UserName, ⭐ Rating, 📝 Content, 💬 DeveloperReply,
🗓️ Review_Year, 🗓️ Review_Month, 🗓️ Review_Day,
⏰ Review_Hour_12, 🌓 Review_AM_PM, 🕰️ Review_Period,
🔢 Comment_Length, 🗓️ Reply_Year, 🗓️ Reply_Month, 🗓️ Reply_Day,
⏰ Reply_Hour_12, 🌓 Reply_AM_PM,
🧹 Content_Arabic_Clean / 🧹 Content_English_Clean
```

📌 *The tables shown here represent **only a sample subset** of the columns, focusing on displaying the cleaned text column (Content_Arabic_Clean or Content_English_Clean). In the full dataset, there are **17 columns** to provide comprehensive coverage and enable broader analysis.*

Figure 1. Summary of data cleaning results — showing **before and after cleaning**, **language distribution**, and **kept vs. removed reviews**.

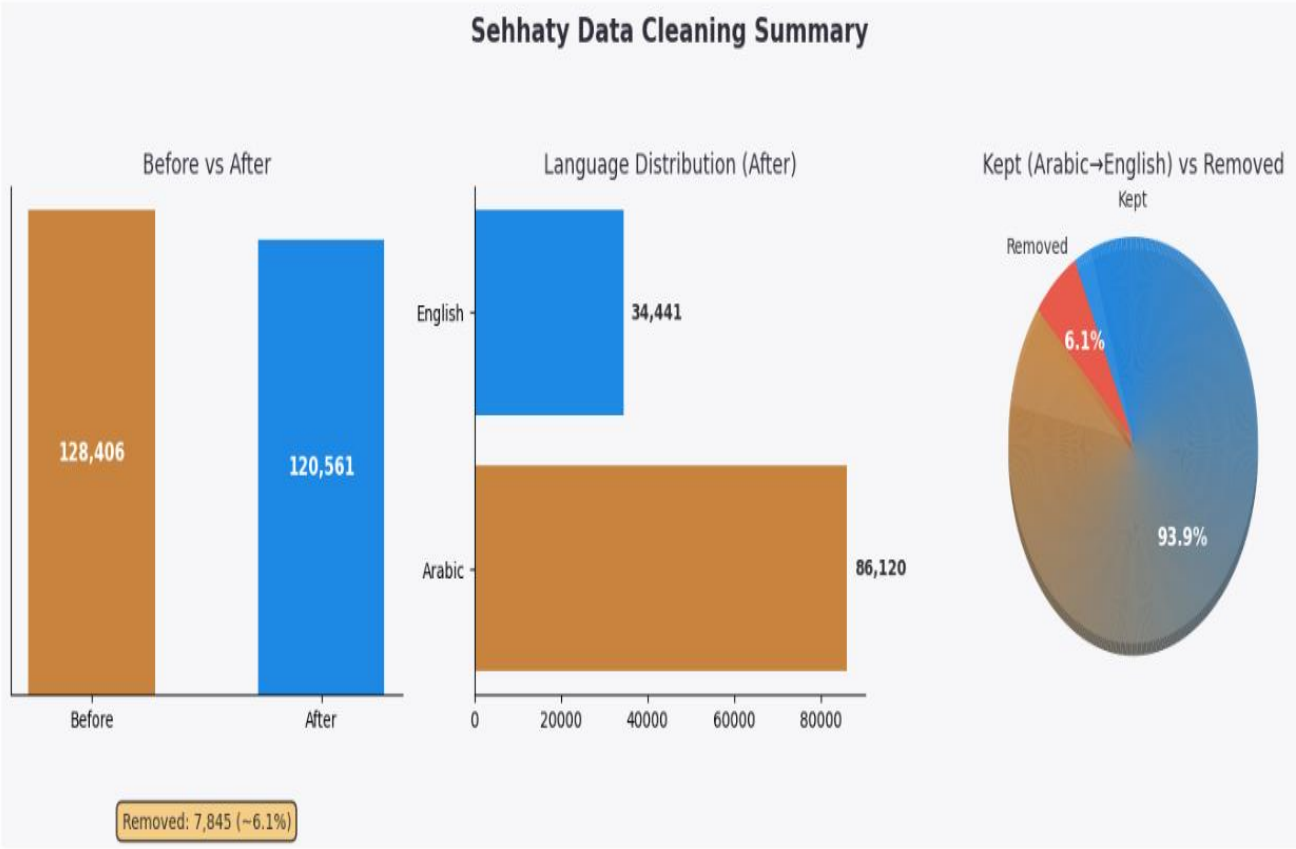
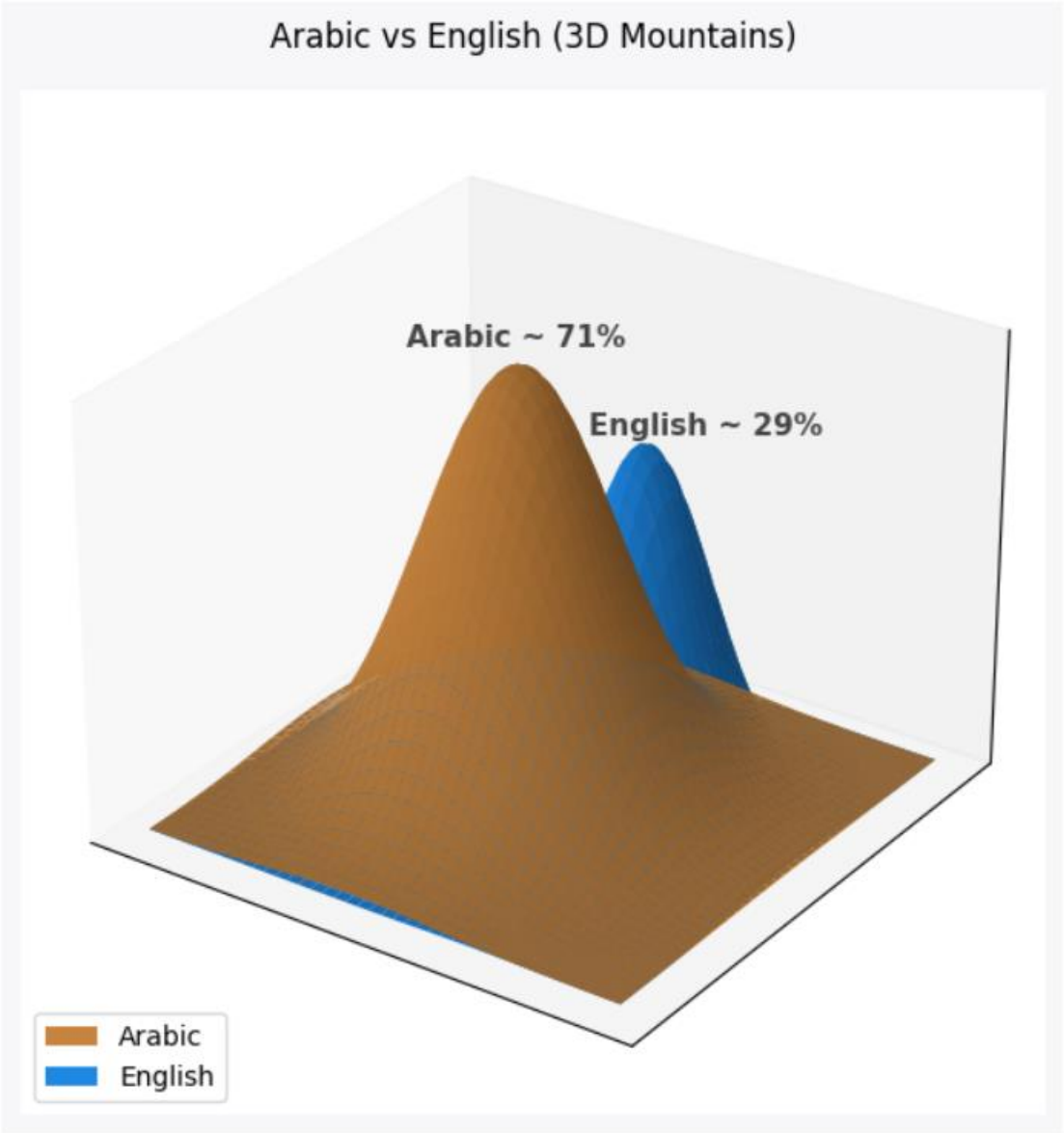


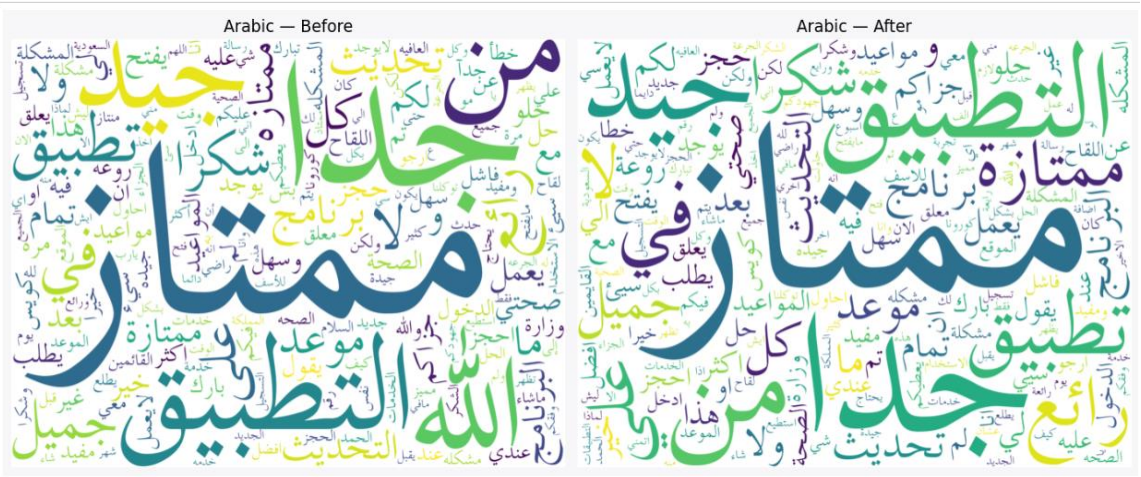
Figure 2. Distribution of the dataset by language using a 3D mountain plot. The figure shows that **Arabic reviews represent the majority (~71%)**, while **English reviews account for a smaller portion (~29%)**.



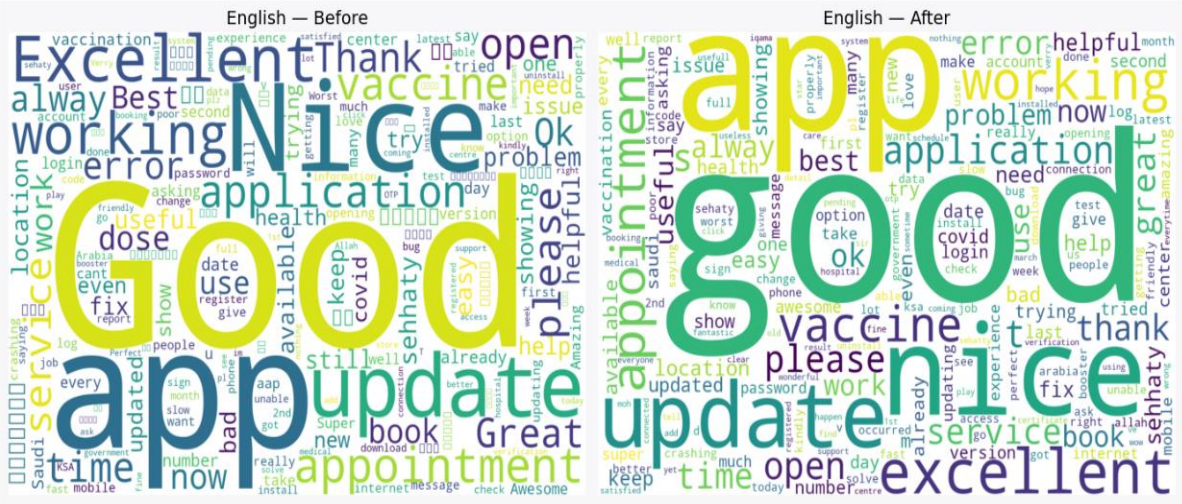

```
=== Data Cleaning Summary ===

Before  |  Arabic: 90,450   |  English: 37,956   |  Total: 128,406
After   |  Arabic: 86,120   |  English: 34,441   |  Total: 120,561
Removed |  7,845  (6.1%)
Kept    |  120,561  (93.9%)
```

Figure 3. Word Clouds (Before vs. After)
The word clouds highlight the most frequent terms in the reviews **before and after cleaning**. After cleaning, noise (e.g., greetings, links, repetitions) was removed, making key terms such as *ممتاز / التطبيق* and *app* / *good* / *update* more prominent and easier to analyze.



.



👉 **This visualization clearly demonstrates the effectiveness of the cleaning process** in enhancing the quality and focus of the data.

📊 **The cleaning process refined the *Sehhaty* datasets**, ensuring **completeness, consistency, and readiness** for future analysis.