**Datasets Drive Link**

[**https://bit.ly/EgyptianDatasets**](https://bit.ly/EgyptianDatasets)

**All Datasets have 3 columns (Text, Year, Source)**

**This file has the following format**

* **Source Name**
* **Year**
* **Information**
* **File Name**
* **Rows Number**
* **Link**

**1. Arabic Online Commentary (AOC)**

**2010**

Online version of Arabic newspaper: Al-Youm Al-Sabe’.

**688,550** **rows** AOC\_youm7\_articles (ar)

**1,048,576** **rows** AOC\_youm7\_comments

**333,225** **rows** RestOf\_AOC\_youm7\_comments

**564,854 rows** AOC\_subtitles\_authors

<http://cs.jhu.edu/~ozaidan/AOC/>

**2. Arabic Political Tweets**

**2019**

Twitter hashtag (That’s enough Sisi)

**431,452 rows**

Political Tweets

<https://www.kaggle.com/datasets/hazemshokry/egypt-arabic-politics-tweets>

**3. ar\_arz\_wiki\_corpus**

**2017**

Arabic (Modern Standard) and Egyptian Arabic dialect comparable documents from Wikipedia

**9126 rows**

Arabic\_Egyptian\_Wikipedia (ar)

<https://www.kaggle.com/datasets/mksaad/arb-egy-cmp-corpus>

<https://github.com/motazsaad/egy-arb-dialect-id>

<https://ieeexplore.ieee.org/abstract/document/8038320>

**4. Arabic Egyptian Tweets**

**2019**

Egyptian tweets cover a blend of different general topics (sentiment analysis)

**40,000 rows**

Egyptian Tweets

<https://datasetsearch.research.google.com/search?src=0&query=egyptian%20dialect&docid=L2cvMTFqOTl4bHQ1MQ%3D%3D>

<https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/LBXV9O>

<https://search.datacite.org/works/10.7910/dvn/lbxv9o>

**5. QCRI**

**2018**

Manually segmented and POS tagged tweets

**350 rows**

QCRI

<https://alt.qcri.org/resources/da_resources/>

<https://github.com/qcri/dialectal_arabic_resources>

**6. Topic Extraction Data**

**2019**

Egyptian dialect tweets used for topic extraction and topic modelling research

**2,256 rows** TE\_News

**2,052 rows** TE\_Sports

**2,358 rows** TE\_Telecom

**2,065 rows** TE\_Tweets

<https://github.com/nadaaym/Topic-extraction-data>

**7. TaghreedT**

**2021**

Egyptian Dialect Corpus (EDC) from Facebook

**13,740 rows**

TaghreedT

<https://github.com/TaghreedT/EDC>

<https://core.ac.uk/download/pdf/286366884.pdf>

**8. ArabicReddit**

**2021**

Arabic dataset of reddit titles and comments from Arabs and Egypt subreddits

**10,129 rows**

Reddit

<https://www.kaggle.com/datasets/abadrawy/arabicreddit>