This data was obtained from:

Mubarak, Hamdy. (2018). ***Build Fast and Accurate Lemmatization for Arabic.*** Obtained from: *ACL ANTHOLOGY.* [*https://aclanthology.org/L18-1181/*](https://aclanthology.org/L18-1181/)*.*

The data is produced from 70 news articles in different genres (politics, economics, health, science and technology, sports, arts, and culture.) written between 2013 and 2015.

Afterwards, spelling corrections were marked, and lemmas were provided with full

Diacritization, by an expert arabic linguist.

We used this data in fine-tuning a Bert model, to perform lemmatization.

Read Metadata for description.