# Arabic Poetry Classification based on Era using Machine Learning with Word and Document Embedding

Manal Anetallah Alsahafi

## Dataset Type

My dataset is a primary data type.First, I scraped the data from the site (adab: https://adab.com/). Then , I chose from the categories of the data on the site the ages which is Islamic and Modern, the type of writing is poetry, and the language of writing is formal (fusahaa). Finally, I fetched the title of the poem, the poet and the content of the poem from 300 web pages for each era, and put them in Pandas DataFrame to save them.

## Methodology

The main steps of the NLP for our proposed system are shown in Table 1.

| The Task | Yes/No for Apply Task | The Reason |
|---|---|---|
| **Clean text**<br>1- Remove non-Arabic content and numbers<br><br>2-Remove all diacritics, elongation punctuation marks<br><br>3-Remove extra spaces | Yes | Because the data collected from the web |
| **Normalize words**<br><br>such as “أ،إ،آ” to “ا“<br>and “ة “to “ه | No | Because it can affect the contextual meaning for some words such as<br>كره =! كرة<br>فأر =! فار |
| **Text segmentation** | No | It divides the document into sentences by predicting the |

| | | |
|---|---|---|
| | | potential end of sentence of the punctuation marks.<br>I will use tokenization to divide the text based on space |
| **Tokenization** | Yes | For simplifying the document exploration |
| **Remove Stop Words** | Yes | Stop words considere as a noise in the poems |
| **Stemming** | Yes | To facilitate further processing and increase the acurencey |
| **POS** | Yes | To capture the syntactic structure of  a word in a document |
| **Text Representations (a.k.a., Feature Extraction)** | Yes | To turning text to features numerical vectors for a Machine Learning model |