Arabic Text Simplification

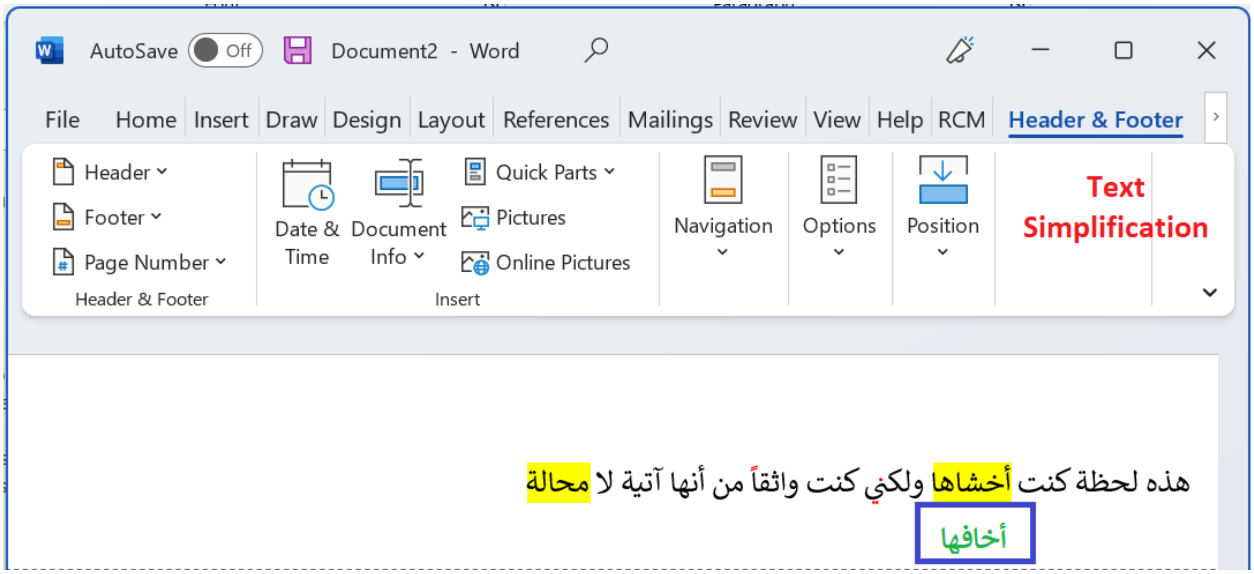
Project Description: Text simplification (**TS**) aims to reduce the complexity of a given text while maintaining the intended meaning. TS is useful for students, children, adults with learning disabilities, and non-native speakers.

Usually, the simplified sentences are shorter with easier words and grammar than original sentences. And if possible, the language of the output text is plain, without idioms, metaphors, complex terminologies, etc.

Arabic is considered as a low-resource language in many domains (including the area of TS). Thus, working on Arabic TS is highly needed for educational purposes and also for NLP applications like understanding, parsing, etc.

We will explore different resources for English, Arabic and other languages to extract relevant data, including **machine translation**, **BERT embeddings**, and **ChatGPT**.

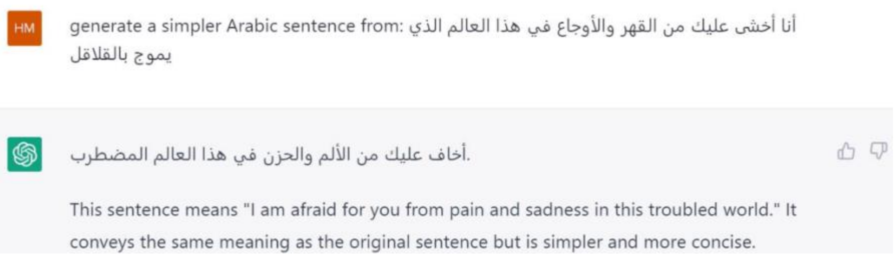
We will prepare data for linguistic annotation of complex sentences and their simplifications. We will build an **Add-on** to common editors (e.g. MS-Word and Google Docs) to identify complex words and suggest simpler alternatives. Predicting sentence and word complexities will be done using ML/DL.



هذه لحظة كنت أخشاها ولكني كنت واثقاً من أنها آتية لا محالة

أخافها

ChatGPT Output:



generate a simpler Arabic sentence from: أنا أخشى عليك من القهر والأوجاع في هذا العالم الذي يموج بالقلاقل

أخاف عليك من الألم والحزن في هذا العالم المضطرب.

This sentence means "I am afraid for you from pain and sadness in this troubled world." It conveys the same meaning as the original sentence but is simpler and more concise.

Project Type: Research and Engineering

Duties/Activities:

- Conduct literature review.
- Explore existing data, ex: English Wikipedia, and Simple English Wikipedia, WordNet, etc.
- Build and train ML/DL models to predict complex words in a given text
- Build an interface (web, add-on) to highlight complex words and give simpler alternatives

Required Skills:

- Programming experience in Python
- Experience in data management, building classifiers, experimentation and evaluation, Web interface.

Preferred Intern Academic Level:

- Junior/Senior CS and IT majors.

Learning Opportunities: You will learn about sentence and word complexity, harvest the internet for relevant resources, build an ML/DL model to predict sentence and word complexity, and build a user interface to show results.

Expected Team Size: 2-4 people.

Mentors: Hamdy Mubarak (hmubarak@hbku.edu.qa)

Samir Abdaljalil (sabdaljalil@hbku.edu.qa)

Dates: Open for Batch 1 (from May 7 to June 29) and Batch 2 (from Jun 4 to Aug 3)

Resources:

Simple English Wikipedia: A New Text Simplification Task

https://aclanthology.org/P11-2117.pdf

Arabic Word-level Readability Visualization for Assisted Text Simplification

https://arxiv.org/pdf/2210.10672.pdf

Automatic Difficulty Classification of Arabic Sentences

https://arxiv.org/pdf/2103.04386.pdf