

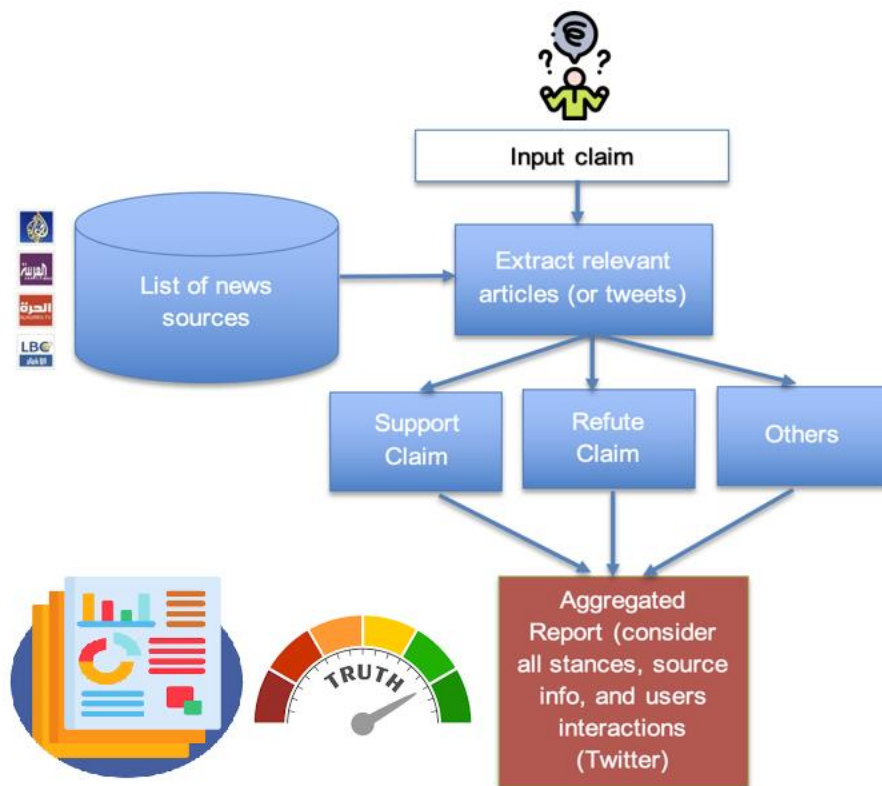
## Arabic News Clustering for Factuality Estimation

**Project Description:** Given the recent increase of false claims on the internet, there has been a lot of manual fact checking effort. As this is very time consuming, human fact-checkers (and normal people) can benefit from tools that can support them and help to debunk the non-factual news.

To verify the information in a source (ex: An earthquake hit Greece (metadata: source, date, time)), people tend to search about the same claim in other trusted news sources. If many of those sources agree with the claim, then people tend to consider it as true, and vice versa. Showing relevant news articles (with their source information and stances in one screen) can greatly help people to identify quickly whether the claim is true or false.

The idea of this project is to **cluster** a large number of articles (or tweets) from different news sources based on their **relevance** to the input claim (considering time information), and whether they support or refute it (i.e. considering **stances** towards the input claim). We will show **aggregated results** to users where we combine stances with profiles for each media source.

A pipeline of the working process is highlighted below.



**Project Type:** Research and Development

**Duties/Activities:**

- Conduct literature review
- Collect and prepare dataset
- Build and train ML/DL models to cluster news articles (based on stances) and estimate factuality of the input claim
- Build an interface (web, add-on) to show aggregated results to users

**Required Skills:** Any of the following

- Programming experience in Python
- Experience in NLP/ML (data collection, building classifiers, experimentation and evaluation)
- Web interface development

**Preferred Intern Academic Level:** We take all levels: PhD, MSc, Senior undergrad students enrolled in CS, CSE.

**Learning Opportunities:** You will learn about news clustering, build an ML/DL model to estimate factuality of claims by combining stances and source metadata, and build a user interface to show results.

**Expected Team Size:** 2-4 people.

**Mentors:**

Hamdy Mubarak ([hmubarak@hbku.edu.qa](mailto:hmubarak@hbku.edu.qa))

Firoj Alam ([fialam@hbku.edu.qa](mailto:fialam@hbku.edu.qa))

Samir Abdaljalil ([sabdaljalil@hbku.edu.qa](mailto:sabdaljalil@hbku.edu.qa))

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

**References:**

- Assisting the Human Fact-Checkers: Detecting All Previously Fact-Checked Claims in a Document  
<https://aclanthology.org/2022.findings-emnlp.151/>
- COVIDLIES: Detecting COVID-19 Misinformation on Social Media,  
<https://aclanthology.org/2020.nlpcovid19-2.11.pdf>
- Top news stories over time  
<https://emm.newsbrief.eu/NewsBrief/clusteredition/en/latest.html>