**Explore Data Augmentation Techniques**

**Project Description:** The performance of most ML/DL models depends on the quantity and on the diversity of the training data. However, insufficient data is one of the most common challenges in using machine learning in practice. This is because collecting such data can be costly and time-consuming in many cases.

Data augmentation is a set of techniques to artificially increase the amount of data by generating new data points from existing data. This includes making small changes to data or using deep learning models to generate new data points. Common ways to perform data augmentation in natural language processing (NLP) include:

- Easy Data Augmentation (EDA) operations: synonym replacement, word insertion and swapping, etc.
- Back-translation
- Contextualized word embeddings

Arabic is considered as a low-resource language in many domains (ex: dialectal Arabic used on social media). Thus, applying data augmentation for Arabic NLP is highly needed.

We will explore different data augmentation techniques for text, and we will apply them to generate new training data to detect offensive language, hate speech, spam, fake news, etc.

We will also explore different resources to extract weakly labeled data, e.g., from Wikipedia, Twitter, etc. We will annotate parts of the data to have insights about it, and we will build ML/DL classifiers.

As we build datasets and classifiers for Arabic, we will keep generalization to other languages in mind.

**Project Type:** Research

**Duties/Activities:**

- Conduct literature review
- Annotate and analyze data
- Build and train ML/DL models (using GPU and TPU accelerators)
- Test and benchmark models
Required Skills:

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

Preferred Intern Academic Level:

- Junior/Senior CS and IT majors.

Learning Opportunities: You will learn about data augmentation techniques, and we will build systems to detect offensive language, hate speech, spam, and fake news among others.

Expected Team Size: 3-5 people

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

  Ahmed Abdelali (aabdelali@hbku.edu.qa)

  Samir Abdaljalil (sabdaljalil@hbku.edu.qa)

  Preslav Nakov (pnakov@hbku.edu.qa)