Sentiment Classification on Arabic Texts

Nowadays, millions of users are expressing their opinions and feelings regarding different life aspects or received products and services, resulting in a massive amount of opinionated texts. Such a rich source had encouraged both business and data scientists to try diving in texts and obtain useful insights and predictions.

Question/need:

- The purpose of this project is twofold:
  - Build a model (classifier) to discover the sentiment of Arabic sentences as either positive or negative.
  - Investigate the impact on performance as a result of using different Arabic text types and domains.
- Business, content makers and social bloggers can benefit from building this model by monitoring comments and feedbacks of their customers/followers and take faster reactions whe7n needed.

Data Description:

- Datasets:
  - **Large Multi-Domain Resources for Arabic Sentiment Analysis**
    - Modern Standard Arabic (MSA).
    - 33k reviews.
    - Available: [https://github.com/hadyelsahar/large-arabic-sentiment-analysis-resouces]
  - **ArSenTD-Lev (Arabic Sentiment Twitter Dataset for LEVantine dialect)**
    - Dialectal Arabic (Levantine).
    - 4k Tweets.
    - Available: [http://oma-project.com/]
  - **MARSA: Multi-domain Arabic resources for sentiment analysis**
    - Dialectal Arabic (Gulf).
    - 60k tweets.
    - Available: [https://github.com/imamu-asa/ASA/tree/main/ASA-Dataset]
- Unit of analysis: A sentence in Arabic (review or tweet).
- Prediction target: Polarity (binary).

Tools:

- Data manipulating: Pandas & Numpy
- Modeling: Sklearn & Tenserflow or Pytorch.
- Visualization: Matplotlib & Seaborn.
- Additional tools:
  - Google Collab for cloud processing.
  - Huggingface pretrained language models.

MVP Goal:

- A minimum viable product would be an Exploratory Data Analysis (EDA) and a baseline classifier.