Team7

Pipeline

- Pre-processing
- Feature Extraction
- Model Training
- Testing and Evaluating

Preprocessing

Used regex to:

1. Remove unwanted tokens from the corpus such as: non words, Arabic stop-words and diacritics, links, mentions, tags, English characters, and numbers.
2. Have consistent "ا,ه,ي" throughout the corpus.

This leaves us a clean corpus containing only Arabic words.

Feature Extraction

Out of the corpus we extracted:

1. Bag of words
2. Tf-Idf Transformer
3. Count vectorizer
4. word embeddings and wordvec
5. Contextual word embedding model

Model Training

- Classical machine learning classifiers:
  - SVM
  - Multinomial Naive bayes
  - Logistic Regression

Model Training

- Sequence classifiers:
    - LSTM
    - RNN

Evaluation

1. SVM
2. Naive Bayes
3. Logistic Regression
4. SVM with input word2vec “tockenizer”
5. RNN with input word2vec “tockenizer”
6. LSTM

Submitted model

Logistic Regression model

Highest accuracy beside the macro f1 score,

SVM model may be better concerning the f1 score but significantly worse concerning the accuracy.