Saudi Authority for Data and
Artificial Intelligence
Data Science Bootcamp

Arabic tweets sentiment analysis (MVP)

Done by: Sarah Hamad Alhussiny

1. Intodaction:

In the world of artificial intelligence, which is getting bigger day by day and knowledge of it increases uninterruptedly, Natural Language Processing (Which is part of artificial intelligence and intersects with machine learning and deep learning, as shown in Figure 1)[1] become an important technology and have many applications that are useful in many sides even in our daily life, such as texts translating, speech to text, etc.

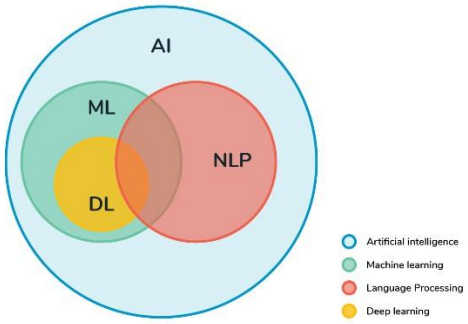

FIGURE 1: AI AND NLP

One of its most important tasks is text classification, which is the process of classifying texts or documents based on the content. Text classification has a variety of applications, such as classification of articles, classification of news, etc.[2].

FIGURE 2: TEXTS CLASSIFICATION

2. Dataset:

To solve the problem, I need the dataset that helps train the machine and build models. I searched for a lot of data and preferred it to be in Arabic to develop my skills in it. I found suitable data on the Kaggle website [3] (dataset link: https://www.kaggle.com/mksaad/arabic-sentiment-twittercorpus?select=test_Arabic_tweets_positive_20190413.tsv).

It initially consists of two features, a tweet, and the category (positive or negative). I needed more features to build well-produced models, so I added many more features which is:

1- Tweet

3- Num of hashtags

4- word density

5- Sentence density

6- char_count

7- Word_count

8- Num of positive words

9- Num of negative words

9 features (columns) and 45275 tweets (rows).

5. Tools:

Regardless of the conventional libraries for storage, statistics, and visualization..., Because I will use NLP libraries, I will do the word embedding, and from that: TF-IDF, N-gram, and AraVec tools. Also using CAMeL Tools tools for pre-processing

6. MVP Goal:

It will be the solution, which is a model that receives a tweet and predict its classification , either positive or negative. This solution will be by building several models using several algorithms and several inputs, then comparing the results and adopting the model that achieved the best result among them.

The algorithms I will use for classification are:

- Logistic Regression
- SVM
- KNN

7. Reference:

[1] https://devopedia.org/natural-language-processing

[2] https://towardsdatascience.com/text-classification-applications-and-use-cases-beab4bfe2e62

[2] www.kaggle.com