



SDAIA
الهيئة السعودية للبيانات
والذكاء الاصطناعي
Saudi Data & AI Authority

Saudi Authority for Data and
Artificial Intelligence
Data Science Bootcamp
Project Proposal

Arabic tweets sentiment analysis



Done by: Sarah Hamad Alhussiny

1. Introduction

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.

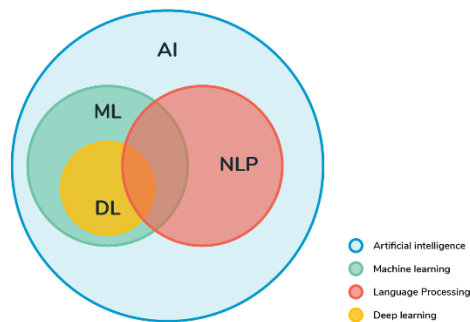


FIGURE 1: AI AND NLP

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. News classification enables specialized websites such as news agencies to classify their content so that the machine can classify any new news directly to its section, which helps the user to easily find the content he/she wants it [2].



FIGURE 2: TEXTS CLASSIFICATION

2. Problem:

Often, it is difficult for a person to **determine the sentiment of others**, whether through their writings or voices. How if these statements are contained in thousands and millions of tweets. This makes it difficult for decision-makers to define them well and quickly. This sometimes leads to the anger of customers and users because of not paying attention to their opinions.

3. Aim:

The project aims to classify the tweets into negative or positive. So that, tweets sentiment analysis helps many people to save time and money as well as develop their businesses. For example, a **company** can find out what people think of their services by analyzing the tweets in their hashtag, A **series director** can know the audience's opinions about his series. It is also possible to know the opinion of the people on the issue of public opinion, like, muting the sounds of mosque loudspeakers.

4. Dataset:

To solve this problem, we 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-twitter-corpus?select=test_Arabic_tweets_positive_20190413.tsv).

It initially consists of two features, a tweet, and the classification (positive or negative). We 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- Num of emojis
- 7- Num of positive words
- 8- Num of negative words

we had 8 features (columns) and 58,000 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

