

Saudi Authority for Data and
Artificial Intelligence
Data Science Bootcamp



**Arabic tweets sentiment
analysis**



Introduction

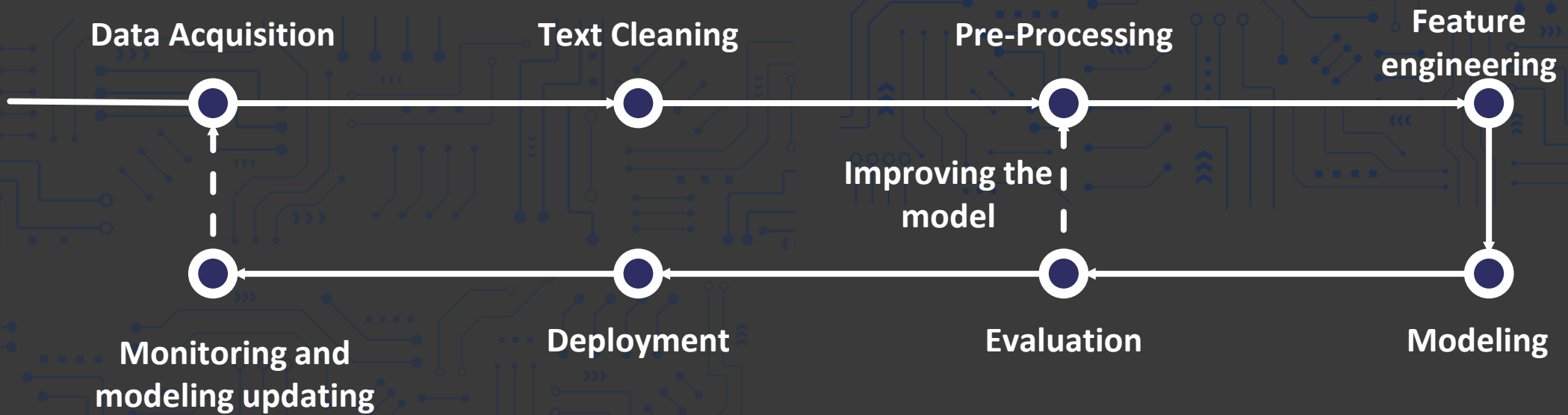


Aim

Classify the Tweets into positive or negative.

Steps

To achieve My goal, I will start with the NLP Pipeline



Dataset

45,275

Tweet in Arabic language

Dataset

The data initially consists of two columns (Tweet and category)

I needed more features to build well-produced models, so I added many more features which is:

Num of hashtags, word density, Sentence density, char_count,

Word_count, Num of positive words, Num of negative words.

Text Cleaning and Preprocessing

01. Text Cleaning

02. Pre-processing

Text Cleaning and Preprocessing

03. Stemming

04. Using CAMEL Tools:

Feature Engineering

Length

Word
density

Char
count

Sentence
density

TF IDF

Word
embedding

N-Gram

AraVec

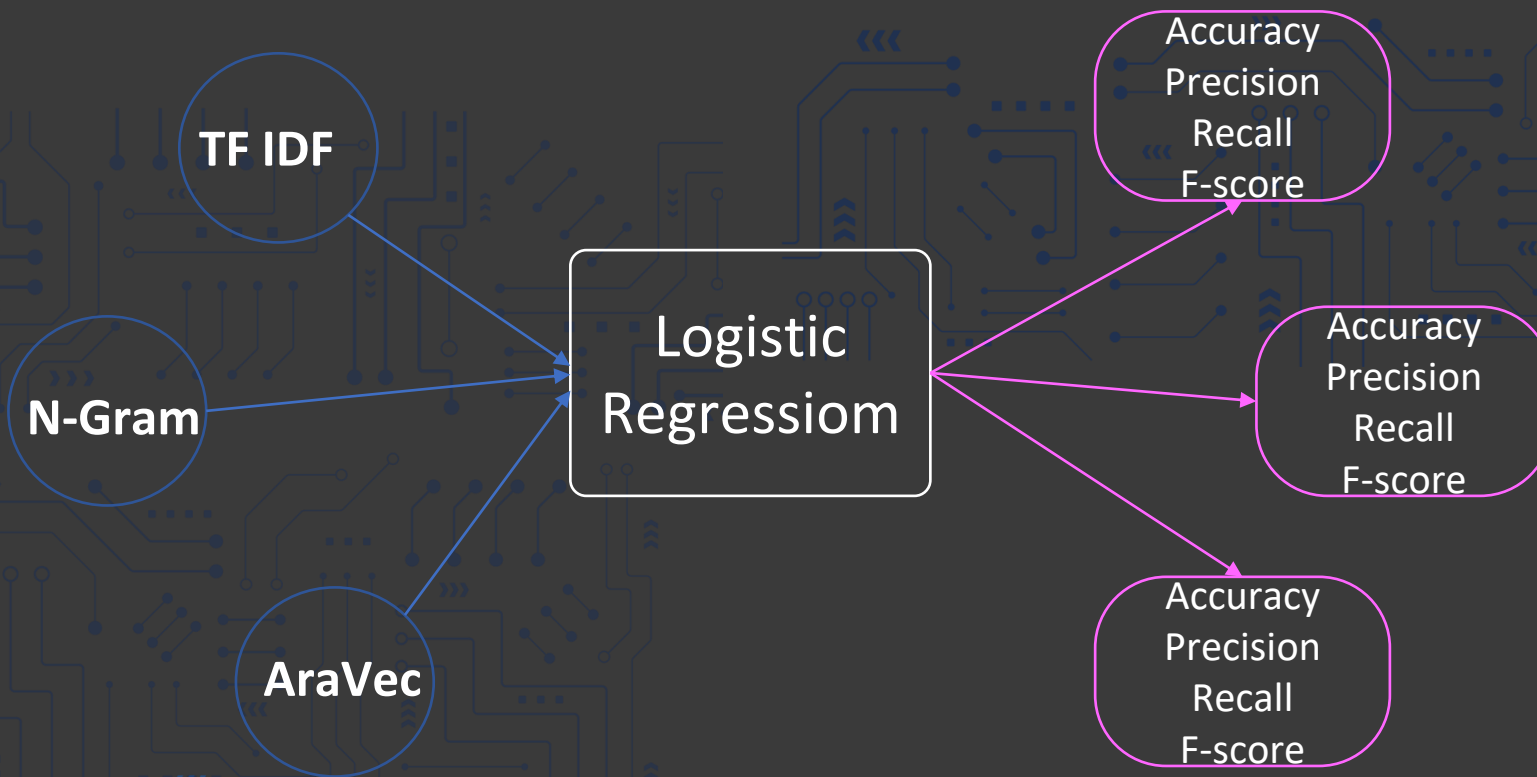
Modeling

01. Logistic Regression

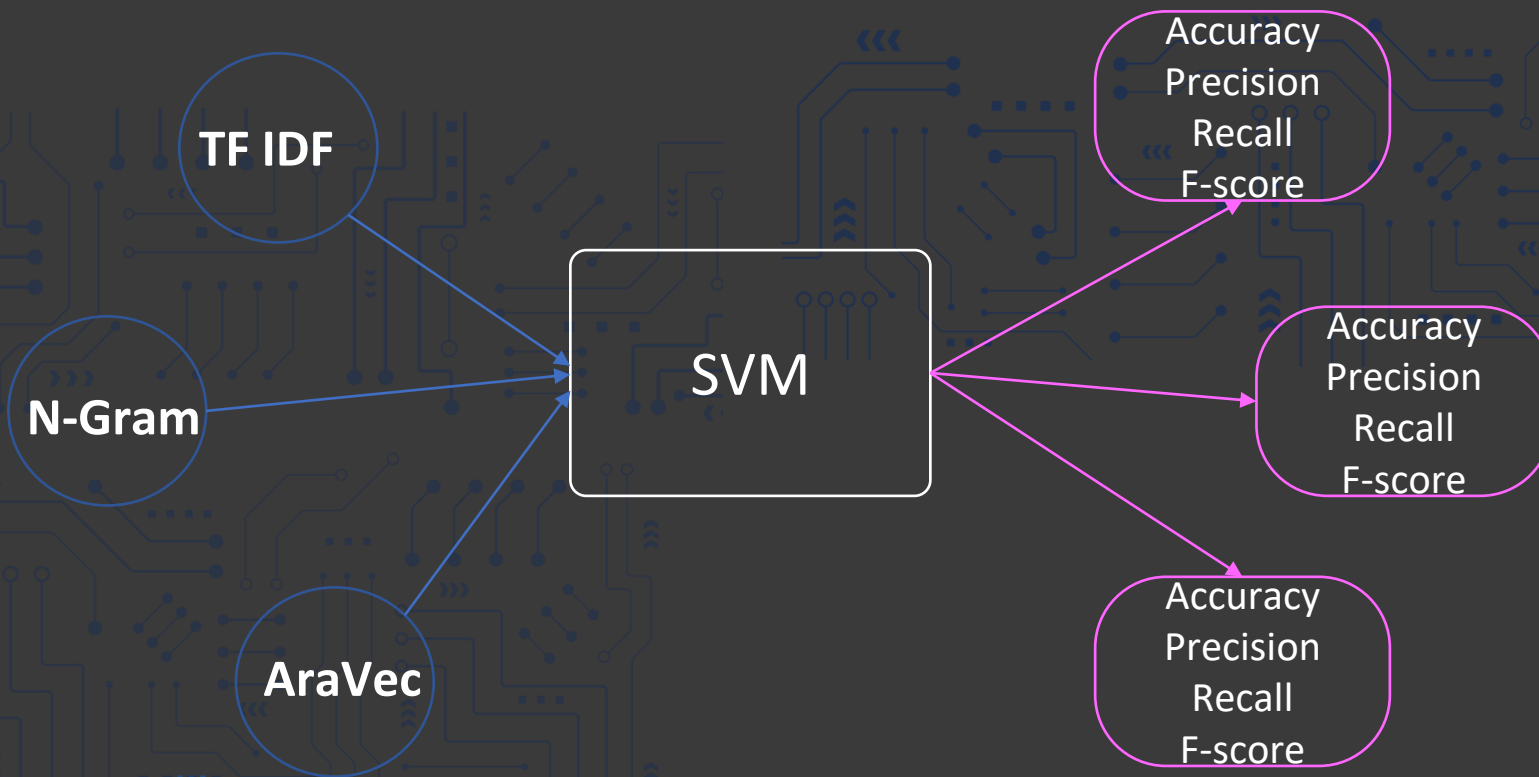
02. SVM(Support-vector machine)

03. KNN(K-nearest neighbors)

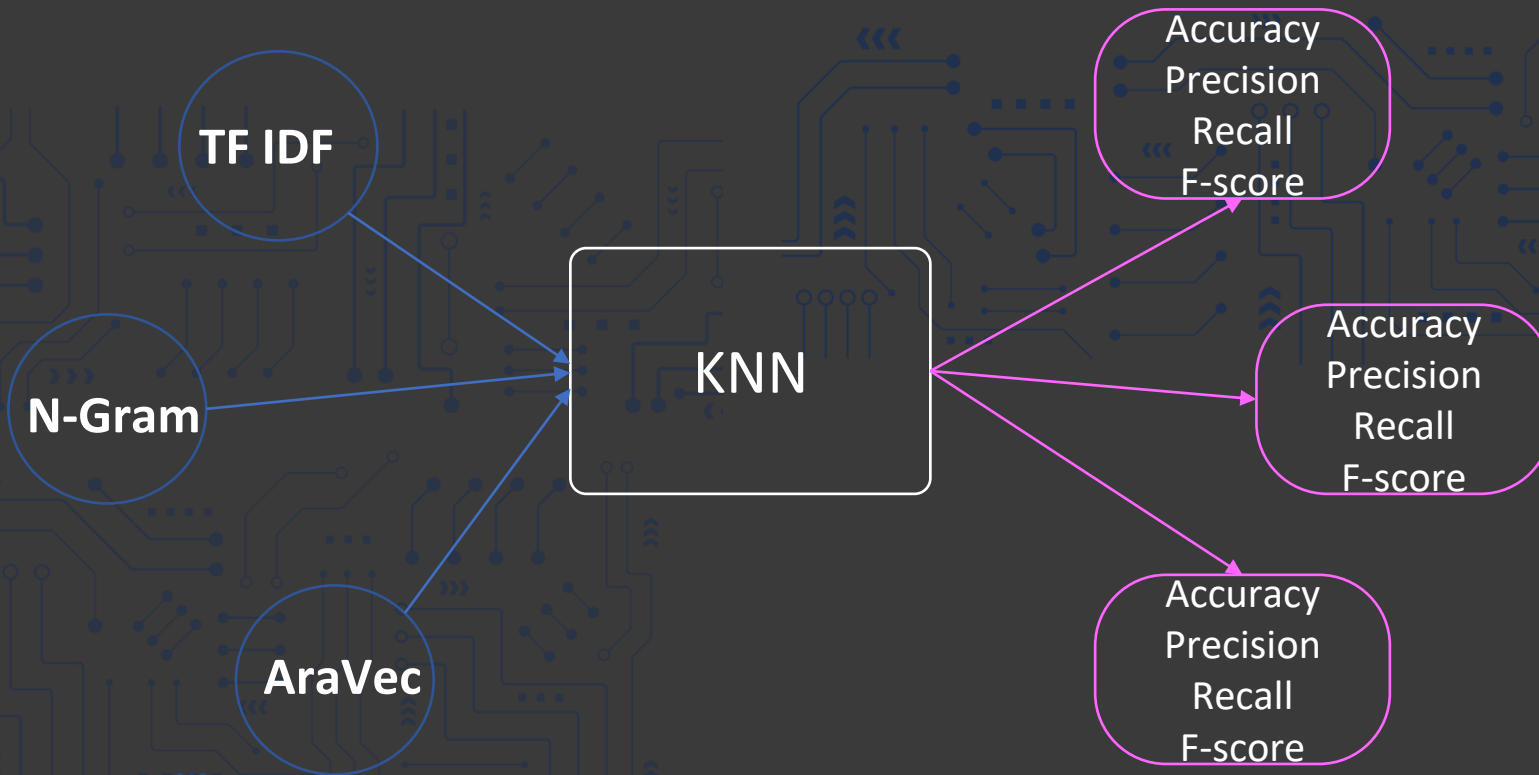
Evaluation Results



Evaluation Results



Evaluation Results



Best model

TF IDF

SVM

Accuracy: 75.85%
Precision: 77.91%
Recall: 77.85%
F-score: 77.85%

The background of the slide is a dark blue-grey color with a complex, light blue circuit board pattern. The pattern consists of numerous thin lines, dots, and small geometric shapes (squares, hexagons) that resemble electronic components and wiring. The lines are interconnected, creating a dense, web-like structure. Some lines are thicker and more prominent, while others are thinner and more delicate. The overall effect is a high-tech, digital aesthetic.

Thanks !

Done by:

Sara Alhussiny