Saudi Authority for Data and
Artificial Intelligence
Data Science Bootcamp

Arabic tweets sentiment analysis

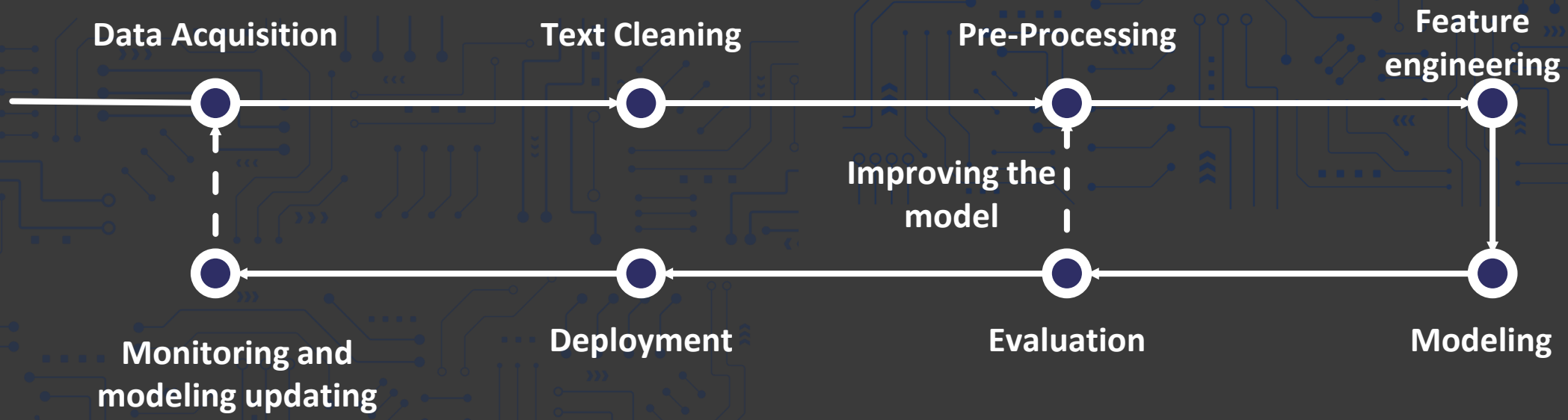
Introduction

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

Sentence density

Char count

TF IDF

Word embedding

AraVec

N-Gram

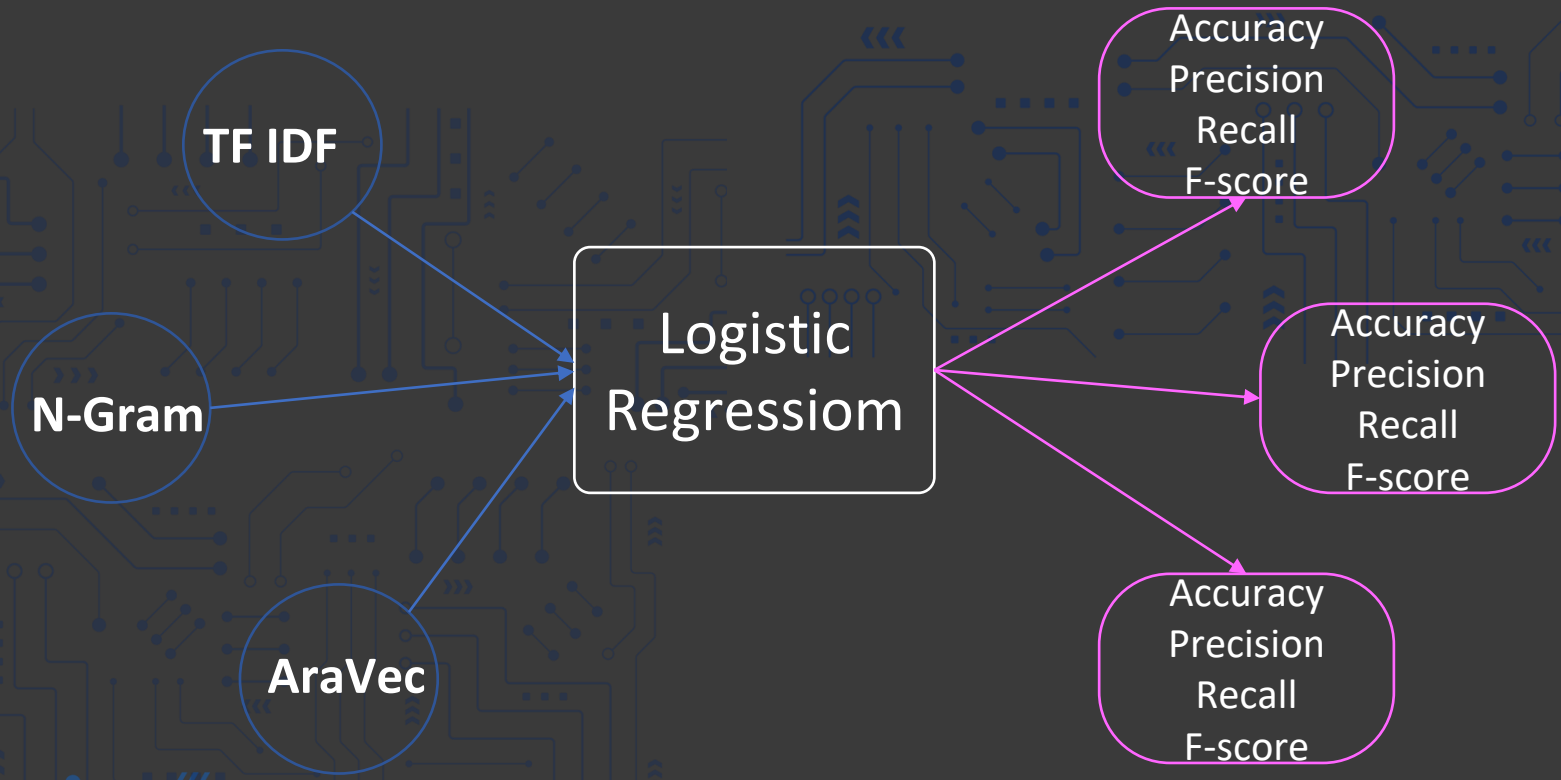
Modeling

01. Logistic Regression

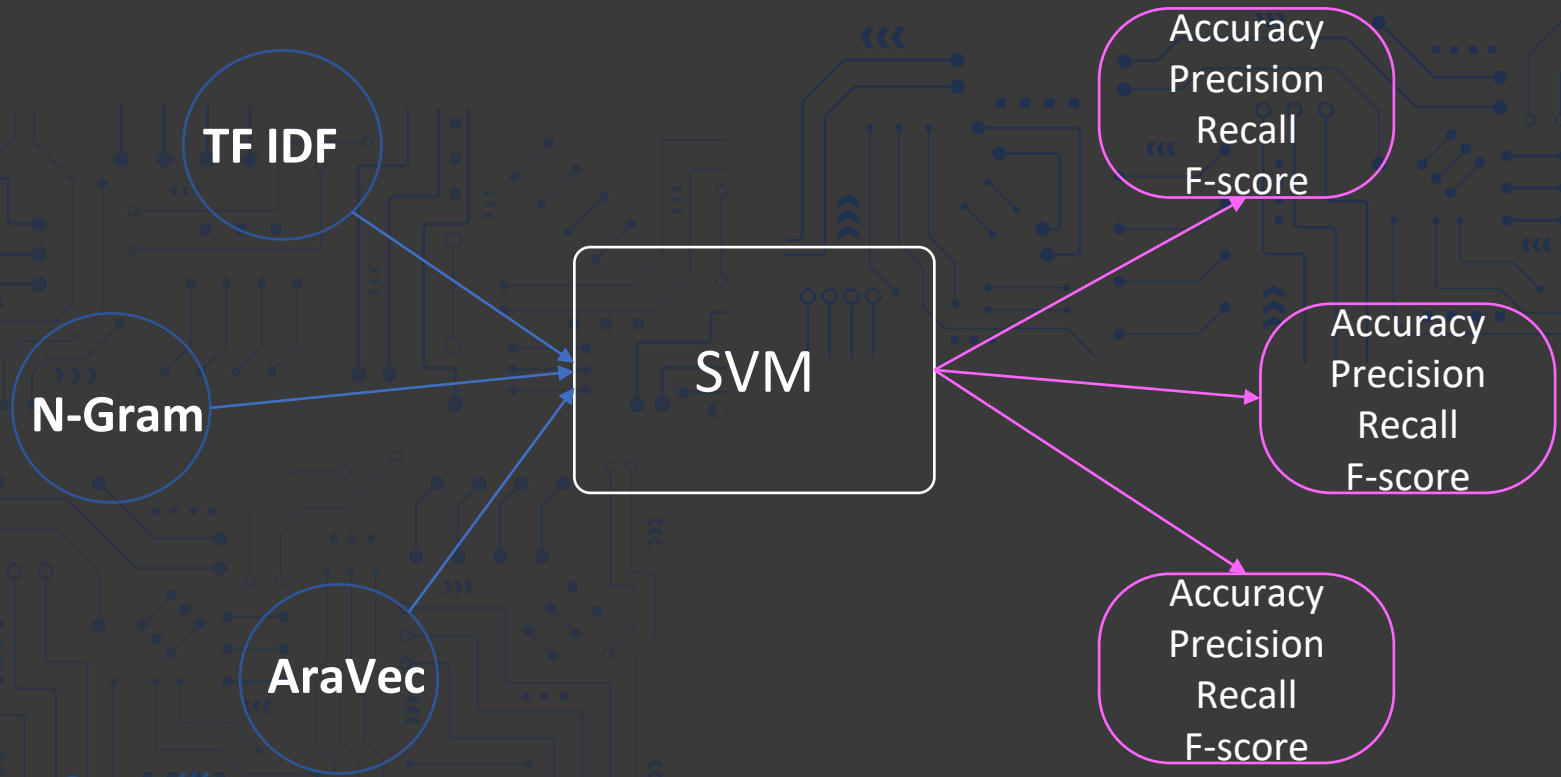
02. SVM(Support-vector machine)

03. KNN(K-nearest neighbors)

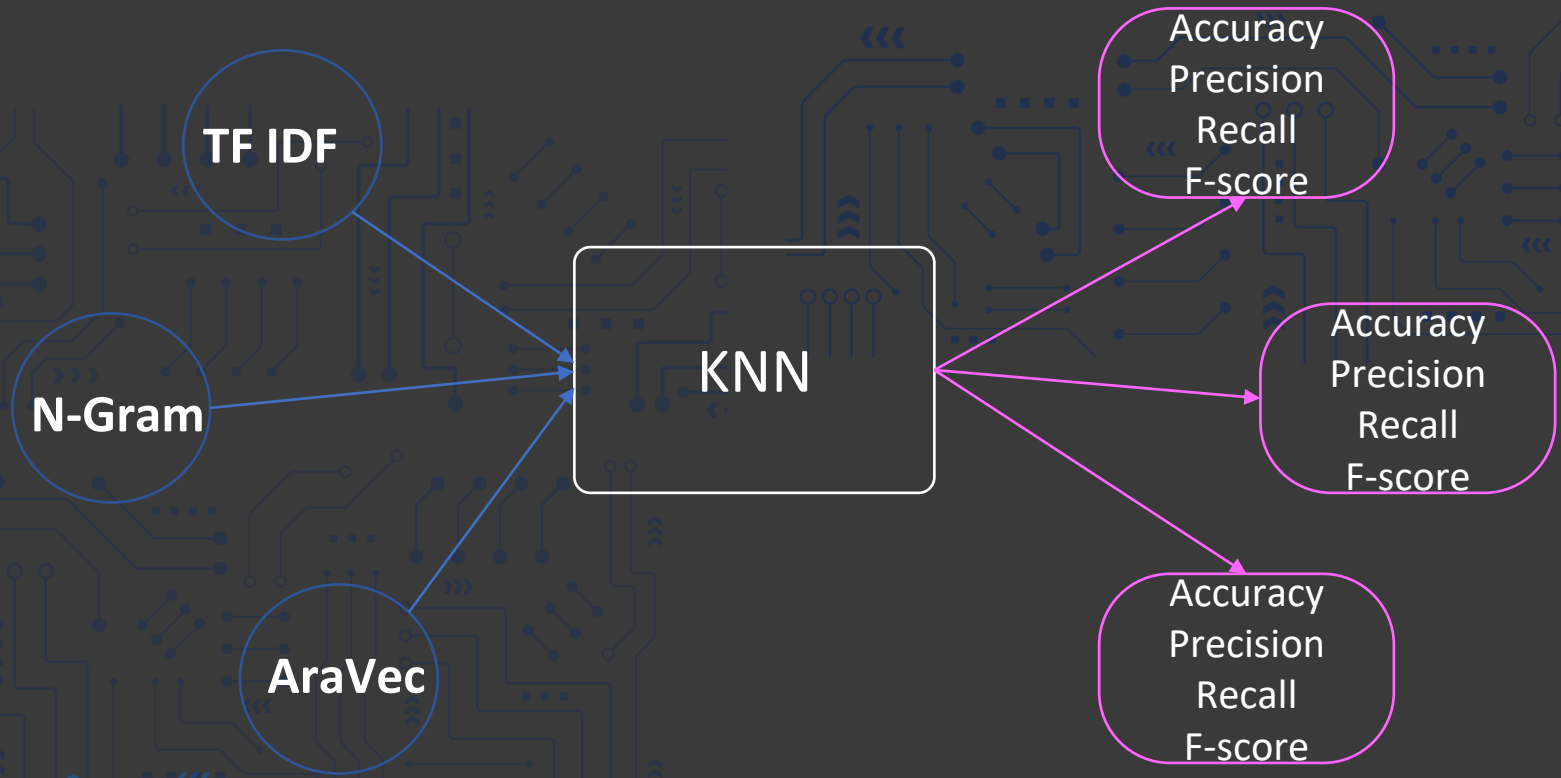
Evaluation Results



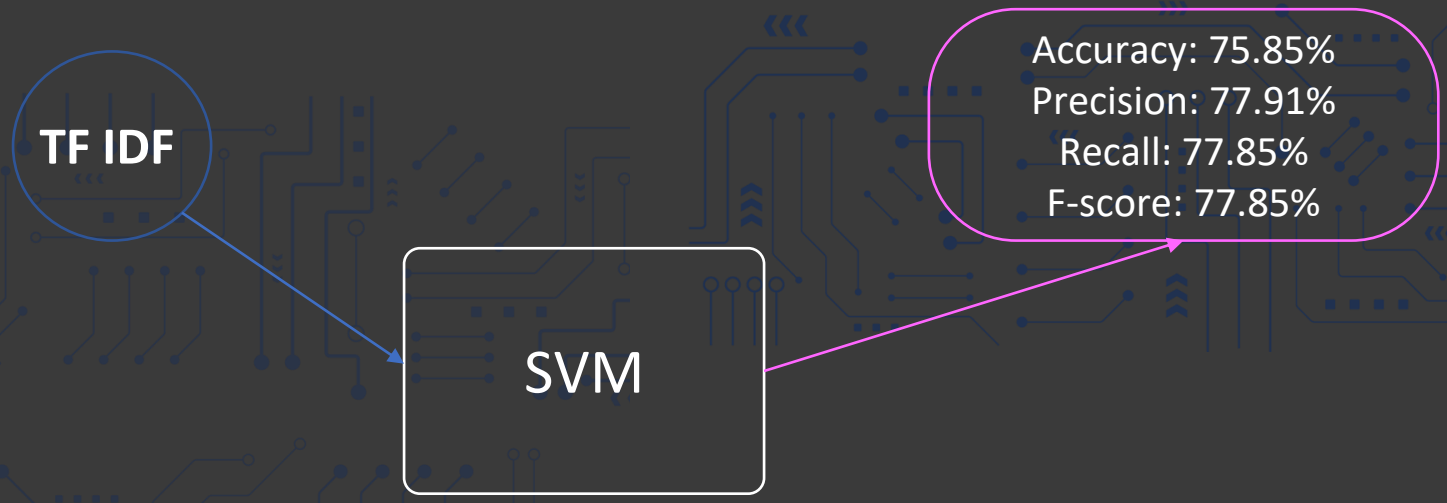
Evaluation Results



Evaluation Results



Best model


TF IDF
SVM
Accuracy: 75.85%
Precision: 77.91%
Recall: 77.85%
F-score: 77.85%

Thanks !

Done by:

Sara Alhussiny