SENTIMENT ANALYSIS APPROACH FOR EVALUATION OF SOCIAL AND ENTERTAINMENT EVENTS
ISY 497

Yasir Almutairi Abdullah Alnoamany Bader Abanmi

Supervisor : Dr.Eslam Almaghayreh
Dr.Mohamad Al Rahhal

Introduction

This is the age of big data. Huge volume of data is being generated daily from different sources. Data science and machine learning techniques introduced many ways to exploit such a huge amount of data to solve different problems in our daily life. Arabic sentiment analysis using natural language processing is a new approach of evaluating the Arabic general public opinion. We apply sentiment analysis on a specific domain which is the entertainment event domain in Saudi Arabia.

Background

Data science goal is to gain insights into data through computation, statistics, and visualization within the respected context of the data domain or problem domain. The aim is to predict, a prediction which at maximum mimic human prediction and to even supersede it and automate it, to do so it requires tools and knowledge from other fields such as machine learning, statistics. Machine learning is a necessary knowledge and skill that data scientists must have where it helps in the process. Because it gives the necessary tools to work with various problems

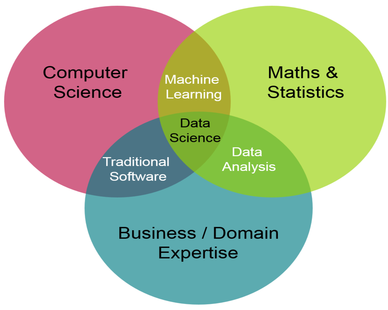

Figure 1: the Intersection of Key Data Science Disciplines

Methodology

Our source of data is twitter. The target is to get the overall general opinion regarding some of the top trending events in KSA. We have collect more than 60,000 tweets by using API. The figure 2 explains the Data Analytics Lifecycle. That we have followed in this project

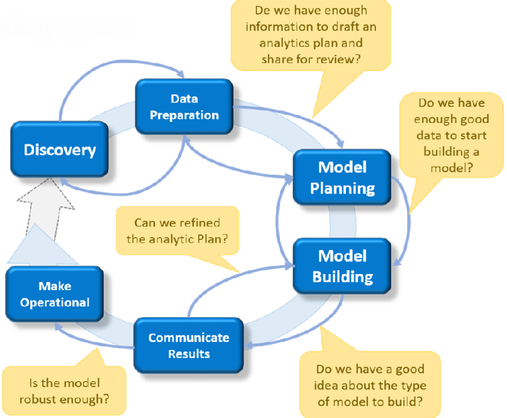

Figure 2: the Data Analytics Lifecycle.

Results

We have applied try 8 models with different combination of features . the figure 3 and 4 shows the accuracy of each model with steaming and without .
(TF-IDF vs CountVect)

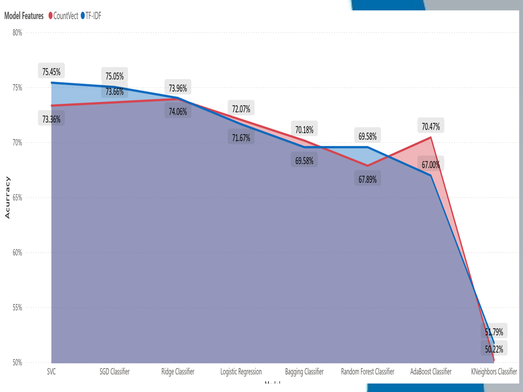

Figure 3: Models accuracy with steaming

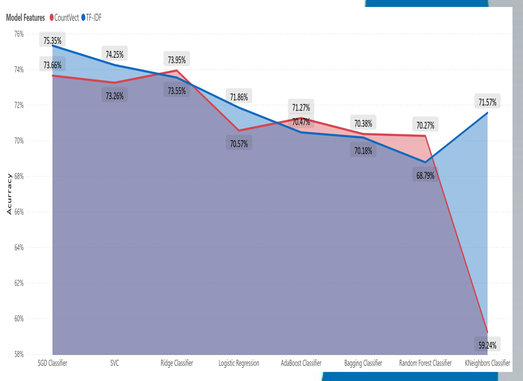

Figure 4: Models accuracy without steaming

The best performance that was achieved was by using the SVC classifier with tf-idf and steaming. The accuracy achieved was 75.45%.Figure 5 shows the accuracy for each category

0 = negative , 1= Positive , 2 = Neutral

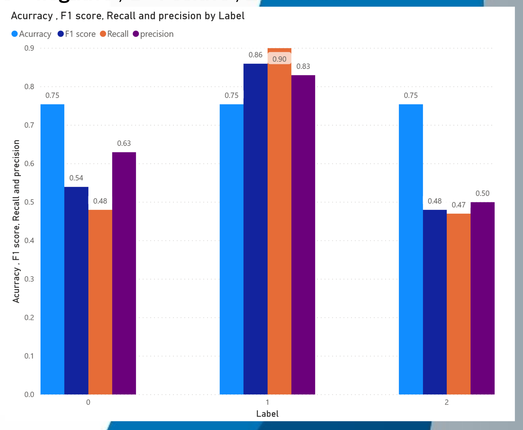

Figure 5: SVC classifier accuracy for each category

Conclusion

In this project we have proposed an events sentiment analysis that contain 60,000 tweets which was labeled manually into 3 categories. We have applied many experiments with different models, the best model was the SVC that classified the tweets with 75% of accuracy.

Future Work

Our future goal is to enhance our results and improve our lexicon content to be more specific, especially with neutral tweets to get more accurate. And publish the lexicon and our project to the public to support the Arabic analysis community