جامعة
الملك سعود
King Saud University

SENTIMENT ANALYSIS APPROACH FOR EVALUATION OF SOCIAL AND ENTERTAINMENT EVENTS

Abdullah Alnoamany

Bader Abanmi

Yasir Almutairi

Supervised By
Dr.Eslam Almaghayreh
Dr.Mohammed Alrahal

Submitted for the partial fulfillment of Bachelor of Applied Computer Sciences in Applied Information Systems degree

College Applied Computer Sciences

KING SAUD UNIVERSITY

April, 2020

TABLE OF CONTENTS

ABSTRACT

Social media generates a huge amount of data reflecting people opinions and feelings about several events. One of the most popular social media platforms used in KSA is twitter. In this project, we will focus on analyzing Arabic tweets to identify people opinions and feelings about the cultural and entertainment events in KSA. The aim of our project is to propose a domain-specific approach for understanding sentiments expressed in tweets related to cultural and entertainment events in KSA. To achieve our goal, we have collected a sentiment dataset consists of a set of tweets related to several events. We have labelled the tweets in the dataset manually. Our final goal is to exploit machine learning to develop a sentiment analysis approach for the evaluation of the cultural and entertainment events in KSA. This approach can help the organizers of these events in identifying any negative sides in these events to avoid it in subsequent events.

.

DECLARATION

We hereby declare that our dissertation is entirely our work and genuine / original. We understand that in case of discovery of any PLAGIARISM at any stage, our group will be assigned an F (FAIL) grade and it may result in withdrawal of our Bachelor's degree.

Group members: 3

Name **Signature**

Abdullah alnoamany ____________________

Yasser almutairy ____________________

Bader abanmi ____________________

ACKNOWLEDGEMENT

We would like to express our thanks to our supervisor (Dr.Eslam Almaghayreh) and (Dr.Mohammed Alrahal), who helped and guided us through this project, SENTIMENT ANALYSIS APPROACH FOR EVALUATION OF SOCIAL AND ENTERTAINMENTEVENTS, and made us learn new things.

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

This is the age of big data. Data is being generated daily, which is getting harder to evaluate and understand. Machine learning techniques introduced many ways that is making this challenge solvable. Arabic sentiment analysis using natural language processing is a new approach of evaluating the Arabic general public opinion. We will apply sentiment analysis on a specific domain which is the entertainment event domain in Saudi Arabia. Our target is to get the overall general opinion regarding some of the top trending events in Saudi. This project will introduce and evaluate machine learning methods which will be applied to try and understand public opinion regarding some major entertainment events. The data will be collected from one of the leading social media platforms in Saudi Arabia.

1.1 Problem Domain

The Saudi Commission for Tourism and National Heritage (SCTH) announced the launching of the "Saudi Seasons 2019" initiative, as an experimental launch this year, where it includes 11 tourism seasons that cover all regions of the Kingdom. The program includes a set of cultural, sports, entertainment and business events. Like any new trend that happens feedback is the method to analyse the success of a new program. The 2019 launch is experimental which emphasize the importance of feedback. There are many ways to analyse the success of an event such as tickets sales and profit. These methods focus on financial success which might not show the overall success. Social media is the prominent method to get an idea about the public opinion about these new trends. We will gather data from twitter to analyse people's opinion about the Riyadh Season. The project will focus on analysing large data from several hashtags that trended during Riyadh Season. This will give us a clear picture of the public's opinion about a certain event.

1.2 Problem Statement

Specific domain Arabic sentiment analysis will be applied in this project to understand the public opinion. The sentiments evaluated are of a specific category. The language used is Arabic (Saudi Dialect) which might be challenging due to the fact that there are no previous research papers in this domain specific field. This project will apply several techniques that will help with tweet classification, such as SVC, Naïve bayes, K-means, and dbscan. And we will see which one provide best classification for the sentiments. There will be three classifications of tweets. A tweet can either be positive (Positive feedback about the event), negative (Negative feedback about the event) or neutral. This will give us an overall view of what the public found entertaining and what failed to entertain. In addition, we will provide lexicon that are specific to our domain in future to help who's want to work in this field.

1.3 Methodology

The data analytics life cycle will be applied to complete this project, As shown in figure 1.1

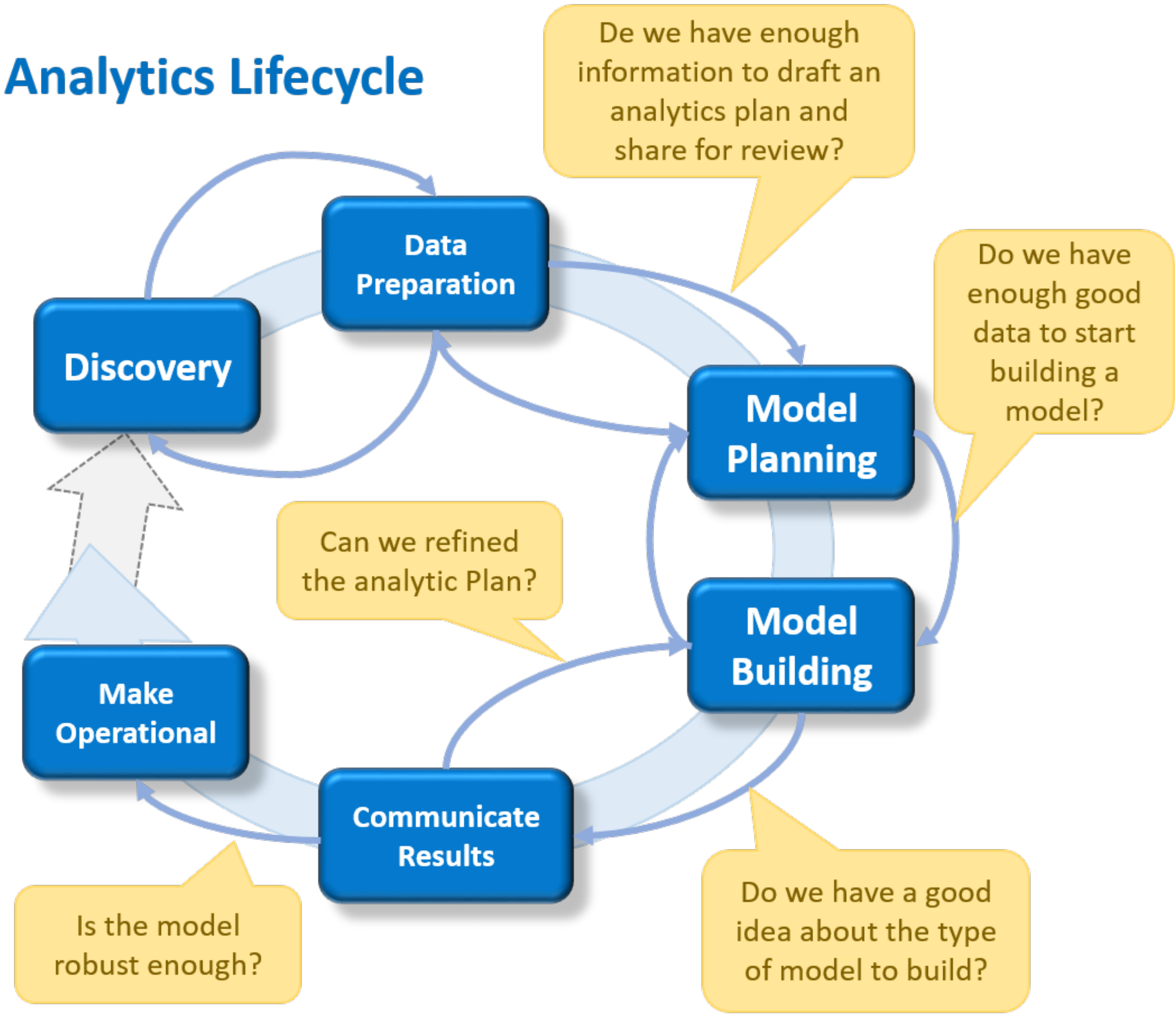

Analytics Lifecycle
De we have enough information to draft an analytics plan and share for review?
Data Preparation
Discovery
Model Planning
Do we have enough good data to start building a model?
Can we refined the analytic Plan?
Model Building
Make Operational
Communicate Results
Is the model robust enough?
Do we have a good idea about the type of model to build?

Figure 1.1: Life Cycle of Data Analytics [1]

1.3.1 Discovery

To clearly observe the performance of Riyadh Season we have to dig in and find effective methods of analysis. First of all, we have to ask the question of “Is this sentiment positive or negative or neutral?”. The answer will give us a clear description of the general opinion about certain events. From the answer we can decide what events did the people enjoy most and what they disliked. Getting information like this can enhance the future of entertainment in Saudi Arabia. There is other valuable information that can be analyzed which, can benefit the entertainment industry of Saudi Arabia.

1.3.2 Data Preparation

The primary source of data used in the project is Twitter. The top trending hashtags on twitter will be targeted. We will grab tweets, understanding them using Eda and clean. After that we will implement a machine learning algorithm which will determine whether the tweet is spam or not. We will try to find other data sources that can help further understand the situation. We will most likely use web scraping methods to obtain more data.

1.3.3 Model Planning

After evaluating the data, we will choose machine learning algorithms that can be applied to our data set. then will look at the candidate algorithms and decide if it shows promise for a classifying the sentiments.

1.3.4 Model Building

In this stage we will implement the chosen models, this will be explained in details in project 2

1.4.5 Communicate Results

We will provide clear visualizations that communicate the result of our findings

1.4 Work Breakdown Structure (WBS)

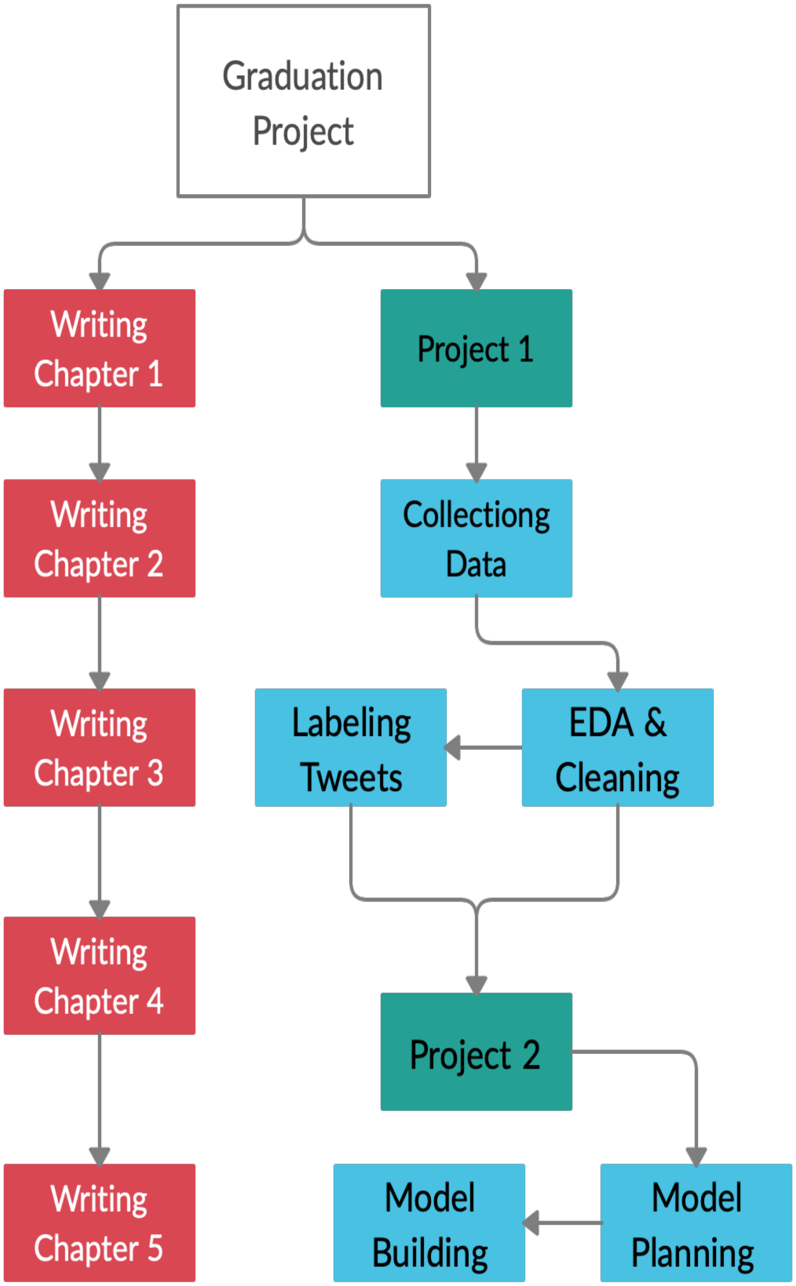
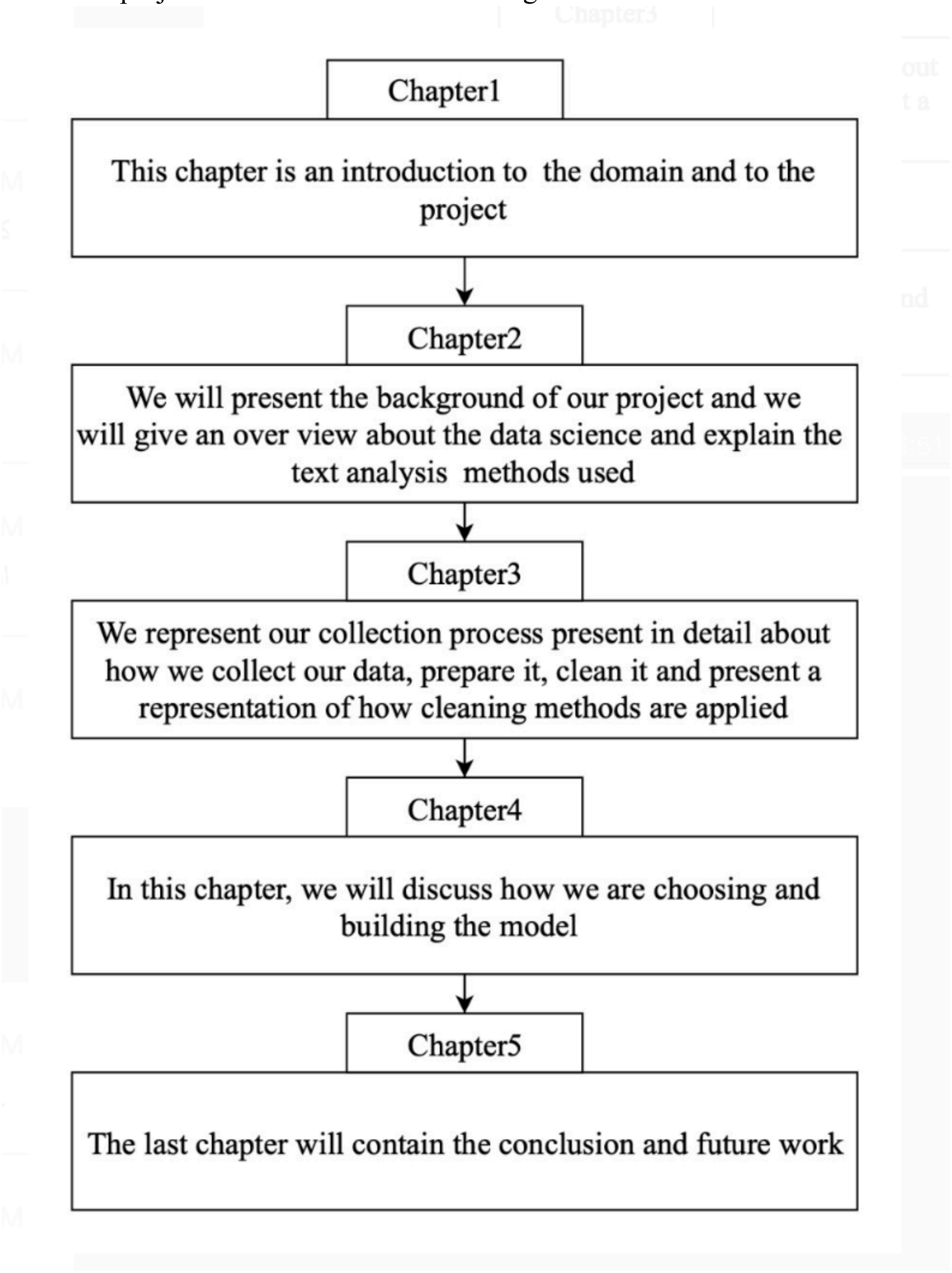


Figure 1.2: Work breakdown structure of project

1.5 Report Layout

The report of this project is structured as shown in Figure 1.2.



Chapter1

This chapter is an introduction to the domain and to the project

Chapter2

We will present the background of our project and we will give an over view about the data science and explain the text analysis methods used

Chapter3

We represent our collection process present in detail about how we collect our data, prepare it, clean it and present a representation of how cleaning methods are applied

Chapter4

In this chapter, we will discuss how we are choosing and building the model

Chapter5

The last chapter will contain the conclusion and future work

Figure 1.3:Project Layout

CHAPTER 2
BACKGROUND/EXISTING WORK

Recently people have been generating massive number of tweets expressing their opinions about the circumstances which occur during entertainment events, these amounts of opinions can be valuable and beneficial for wide range of applications. For example, companies or responsible entities can be tracking the consumer opinions toward their service or product in order to collect information about satisfaction levels of customers and to identify what should they do to improve their product or services. Here we are using Natural Language Process which is subbranch of data science which has recently seen a spike in its popularity among researchers Sentiment analysis is the task of identifying user opinions or sentiment regarding an Entity. This project focus on analyzing these sentiments conveyed in users' tweets during events such as SHM concert or Winter Wonderland and other events to indicating if these events has caused a negative or positive emotion among people. different sentiments within different domains. Additionally, the lack of a sufficient manually labeled events dataset resulted in a limited progress in events-specific sentiment analysis on social media.

2.1 Data Science Fundamentals

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. Data science is "the ability to take data to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it. It is a significant ability in the following decades, at the expert level as well as even at the instructive level for primary school kids, for secondary school kids, for school kids. Since now we truly have basically free and omnipresent information" [2]. The data now is generated in massive scale from tweets to videos and to other texts or non-text and as such data scientists are concerned with taking this data and develop use cases for that data, create a hypothesis, perform an experiment, analyze the results, and finally communicate it. It is an iterative process. "Data science, as it's practiced, is a blend of Red Bull fueled hacking and espresso-inspired statistics. But data science is not merely hacking because when hackers finish debugging their Bash one-liners and Pig scripts, few of them care about non-Euclidean distance metrics. And data science is not merely statistics, because when statisticians finish theorizing the perfect model, few could read a tab-delimited file into R if their job depended on it.” [3].

2.2 Machine Learning in Data Science

Machine learning is a necessary knowledge and skill that data scientists must have. Where it helps in the process of data science as data scientists first initial step is to understand the domain of the problem they are faced with or tasked to solve or research in, it could in a medical domain or a financial domain or any other domains. Their next step is to understand the data they are working with and its limitation and what it does offer of valuable information, and then preparing this data which is the usually the part that consume most of data scientists time. Preparing it involves the removal of noise observations and resolving any misrepresentation it could suffer from. As this step leads to model building where without machine learning algorithms it will not be effectively implemented. This is where data scientists apply machine learning and build models that learns from each observation and make prediction. This feature of prediction that machine learning offers is why it is important to data science as it automate the process of predicting new observation and gains insights from it. Machine learning is valuable to data science because it gives the necessary tools to work with various problems such as supervised and unsupervised problems.

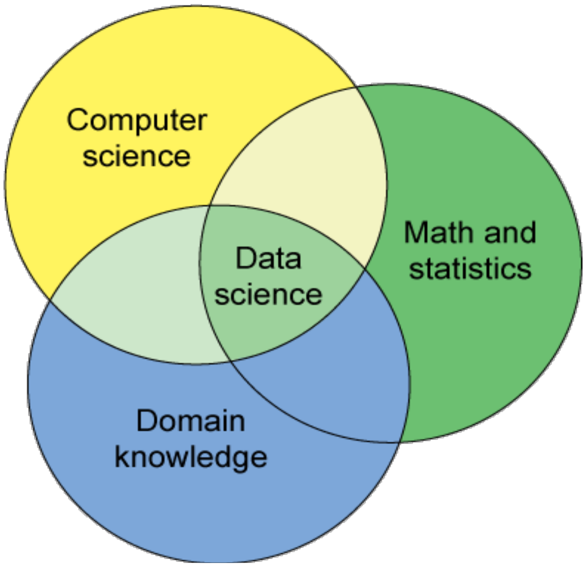


Figure 2.1: Diagram Shows the Intersection of Key Data Science Disciplines [4]

2.3 Text and Sentiment Analysis

The field of sentiment analysis comprises and uses many techniques in other fields to obtain the desired results these fields includes natural language processing which is subfield of linguistics and computer science and it is concerned of programming computers on analyzing large amount of natural language data. Text analysis which also referred to as text analytics and it is concerned of processing texts, structuring and parsing and deriving pattern from the text into data ready for analysis. Sentiment analysis generally aims to identify the feelings of a speaker or writer about a topic or to identify the dominant feelings of a document. As a result, sentiment analysis combines these fields and its techniques and apply it in getting customer sentiment towards a product/service which yields valuable information to companies and governments and marketing campaign managers.

2.4 Related Works

This Section Summarize the result shown in different research papers related to the domain of our project

In [5] the authors focus on analyzing sentiment expressed by football fans through Twitter. The tweets reflect the changes in the fans' sentiment as they watch the game and react to the events of the game. Authors objectives of this paper is to propose approach for understand sentiments in football domain. To achieve their goal, they start by developing a football-specific sentiment dataset which is labeled manually. Then they utilize the dataset to automatically create a football sentiment lexicon. Secondly, they develop a classifier that can capable of recognizing sentiments in football conversation. The results show that approach is effective in recognizing the fans' sentiment during football events. An Arabic Twitter Corpus for Subjectivity and Sentiment Analysis has been present in the paper [6], the authors highlights issues posed by twitter as a genre, such as a mixture of language varieties and topic-shifts. The next step is to extend the current corpus, using online semi-supervised learning. A first sub-corpus will be released via the ELRA repository as part of this submission. The authors of this paper suggests a system for sentiment analysis for Arabic social media data in [7]. Individual settings are required per genre and task.

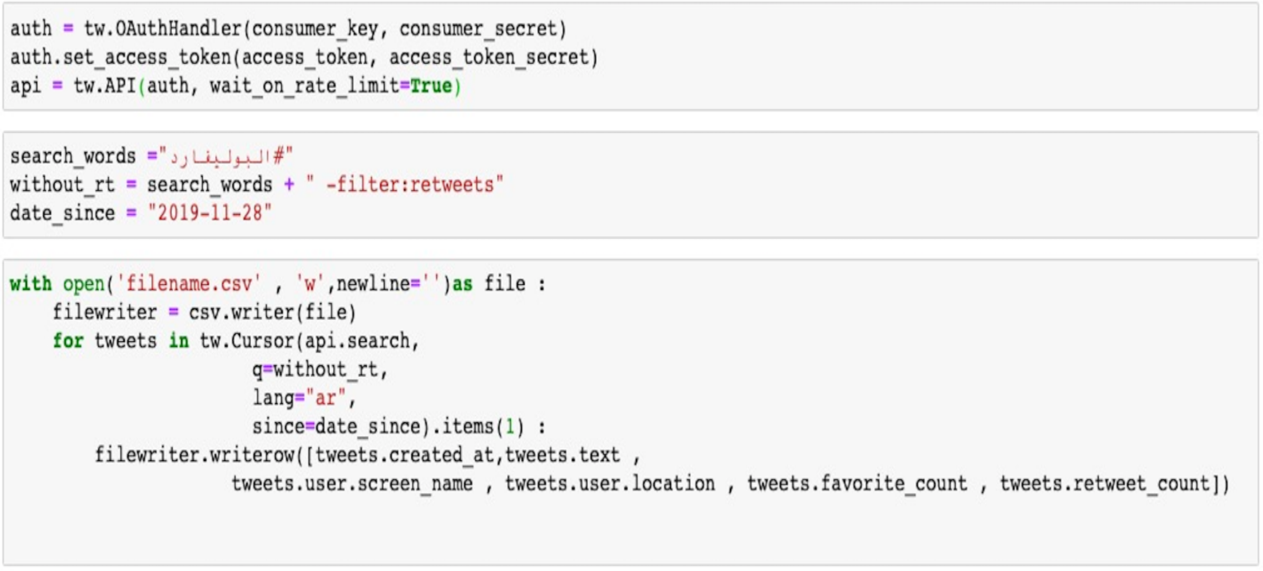
Processing dialects does not improve when it is known which sentences are in dialect. In [8] the authors use Arabic Machine Learning (Classification and Clustering) with social media to discover the polarity or opinion in the contents. Many kinds of classifiers and clusters used with Social Media content detection, like SVM and K-Mean. The paper reviews the literature of the popular ANLP tools with AML software on social media contents toward identifying the best tools in these domains. In this study [9], the authors present the design and implementation of Arabic text classification in regard to university students' opinions through different algorithms such as Support Vector Machine (SVM) and Naive Bayes (NB). The aim of the study is to develop a framework to analyzing Twitter "tweets" as having negative, positive or neutral sentiments in education or, in other words, to illustrate the relationship between the sentiments conveyed in Arabic tweets and the students' learning experiences at universities. Two experiments were carried out, one using negative and positive classes only and the other one with a neutral class. The results show that in Arabic, a sentiments SVM with an n-gram feature achieved higher accuracy than NB both with using negative and positive classes only and with the neutral class. The authors in the paper [10] , detailed the process of collecting the data from Twitter and also the process of filtering, pre-processing and annotating the Arabic text in order to build a big sentiment analysis dataset in Arabic. Several Machine Learning algorithms (Naïve Bayes, Support Vector Machine and Logistic Regression) alongside Deep and Convolutional Neural Networks were utilized in the experiments of sentiment analysis on health dataset. The problem in [11] is how to dealing with dialectical Arabic. Authors of this papers proposed a hybrid approach which combines both machine learning techniques and semantic orientation. Results of the classifier that they have been used as training data for the support vector machine classifier. Their finding shows a hybrid approach has increased the accuracy by 16.41%. And as for the F-measure of a lexical classifier has increased by 5.76%

CHAPTER 3
DATA PREPARATION

Data preparation is the longest process in this project, so we will discuss in details the steps in this process of how we collected the data and what the filters we applied and how we cleaned the dataset until the phase of the data become ready for modeling.

3.1 Data collection

In this section, we provide a detailed description of our data collection process and the data annotation procedure. The goal of this process is to provide a benchmark dataset for the Entertainment event industry. Researchers can utilize this dataset for Arabic sentiment analysis and comparison purposes. The challenge of collecting data firstly is the twitter API, the free version of twitter API have limited futures like we can't collect any data that was before more than 7 days, and limited number of tweets can be gathering. We collected data from twitter by requesting a developer access account therefore, using the twitter API as our main data source. The data collected is related to the event entertainment in Saudi Arabia. We have collected hashtag data related to a specific event during the duration of the event. The filters that applied on the code of API is removing the retweets because we don’t want a duplicates data. We use the code below in figure 3 to collect the data and specify the search word that we want it on the tweets, so we take the trends hashtag for some popular events. The date of collecting in every trend was the same day of event plus a day after.



```
auth = tw.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)
api = tw.API(auth, wait_on_rate_limit=True)
```

```
search_words ="#البوليفارد"
without_rt = search_words + " -filter:retweets"
date_since = "2019-11-28"
```

```
with open('filename.csv' , 'w',newline='')as file :
    filewriter = csv.writer(file)
    for tweets in tw.Cursor(api.search,
                          q=without_rt,
                          lang="ar",
                          since=date_since).items(1) :
        filewriter.writerow([tweets.created_at,tweets.text ,
                         tweets.user.screen_name , tweets.user.location , tweets.favorite_count , tweets.retweet_count])
```

Figure 3.1: Collecting Data API Code

The top trending events collected are:

1) SHM concert 2019

The SHM concert dataset was collected in Nov/5/2019, using the Twitter API. The tweets were filtered by the official hashtag “سهم - ليله# “We obtained 31185 tweets from the hashtag to analyze the event.

2) Winter Wonderland 2019

To collect tweets related to Winter Wonderland 2019, we identified the official hashtag ‘ونتروندرلاند#’ as a filter to retrieve tweets related to this event. The tweets were collected between Oct 19th, 2019 and Oct 26th, 2019. In that period, we collected 3000 tweets, we couldn’t collect more so we tried to collect from another period between 13th Nov to 20th Nov, but we didn't get enough tweets, so we kept the original 3000 tweets.

3) Riyadh Boulevard 2019

Using the official hashtag “الرياض - بوليفارد# “we have collected the tweets related to the event. We collected 5738 tweets between the periods of 17 Oct and 19 Oct. To enrich our data coverage of Riyadh Boulevard event, we noticed another hashtag that appeared in some of our dataset tweets that is also related to same event. Therefore, we began to collect again based on the hashtag “ مسيره الرياض - موسم -#”. From this hashtag we collected around 9550 tweets in the same period of time.

4) Colors Marathon 2019

The Colors Marathon dataset was collected between 26 Oct and 27 Oct 2019, using the Twitter API. The tweets were filtered using the official hashtag “الالوان - سباق# “. We obtained 15454 Arabic tweets from this hashtag. All in all, the overall count of collected tweets is 62,227 tweets from 5 different hashtags. Our dataset was only takes Arabic tweets for sentiment analysis. To ensure data quality, we filtered out duplication and retweets from the dataset.

3.2 Data Exploration

The next phase after data collections is data exploration. It is as crucial as data cleaning, in which it gives us an overview of the data and its structure and helps us design a workflow for data cleaning. In terms of our collected data it consists of Arabic texts as each text is a tweet. The steps involved in the Exploration of collected data include:

- Number of Stop Words present in each tweet: Stop words are words that are commonly used and won’t yield significant information on a particular tweet. Removing it will allow us to focus on the important words.

’الأحياء المجاورة واقترح فرض رسوم رمزية منالازدحام داخل البوليفارد وخارجه فيشكراً لجهودك مبادرة لطيفة للتقليل“

No. Of stop words in this tweet: ‘2’.

- Number of words present in each tweet. Finding the number of words that is not a stop words will give a clear idea on the contribution of a tweet to the feature space.
- Length of tweet. Each tweet abides by twitter policy of tweets max length - 240. This piece of information Find whither a tweet exceeded all of a tweet max length or not.
- Extracting emoticons out of tweets in such a way that these emojis ‘😍💙💙💙’ becomes ‘:smiling_face_with_heart-eyes: :blue_heart: :blue_heart: :blue_heart:’. as some of the tweets we encounter have only an event name with positive or negative emojis.
- The average length of words excluding stop words. As it shows whether a tweet have long words or short. Compared to other tweets.
- Number of tags in a tweet: The total number of tags present in a tweet (#hashtag, @mention)
- Count of numbers present in each tweet. Gives a count of numbers present in a tweet; is a tweet describing a quantitative state (e.g. event size, attendance visitors to an event)
- Frequent words preset in our Arabic tweet’s dataset. Frequent words will yield no valuable information as they’re present in all tweets.

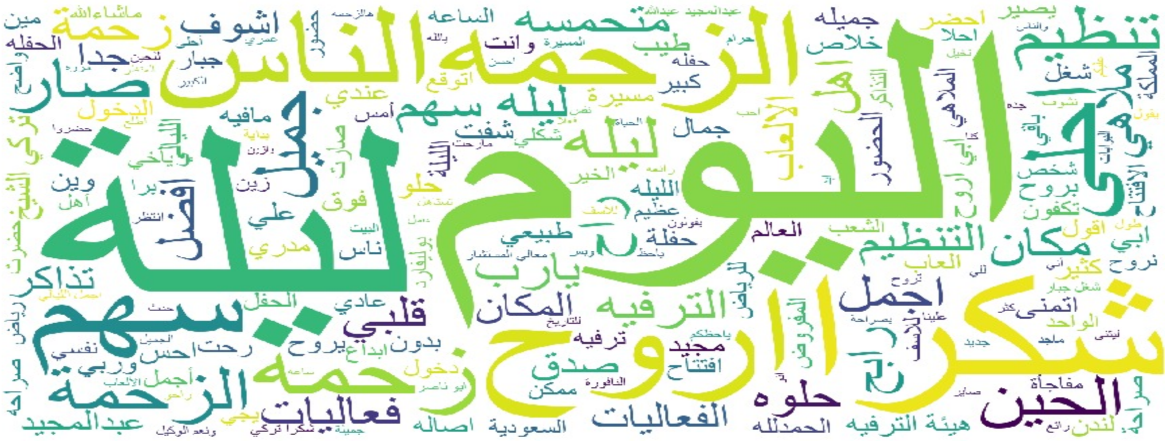


Figure 3.2: Words cloud frequency for saham event

3.3 Data Cleaning

Data cleaning is the process of identifying and removing the errors in the dataset. It's an important step that will affect the final score of our model if we didn’t have high quality data. Our process for cleaning includes:

- Dealing with Arabic stop words:

In python there is already a library can be used for Arabic stop words but it’s to formal Arabic so we added a text that have extra stop words for public Arabic language. we also add some words that are irrelevant to removing it as a stop word.

Table 3.1:Example of Removed Words (Stop Words)

منذ	مع	ليه
وكل	ومن	مابعد
انك	وهب	بعد
هيه	وهذا	بهذي
ايه	وهو	عنها
بأن	وهي	اذا
يااا	وَيْ	اللي
لكنه	وُشْكَانَ	ليبيه

- Removing tags, mentions, url links, username, and hashtag:

When we collect the tweets, the data have to many things we didn’t need it. the cleaning process introduced several difficulties such as in this example.

‘#موسم_الرياض فيه مطاعم ألعاب نارية النافورة، الرعب،حديقة الملز،المربع، ملاهي #ونتر_وندرلاند كل الاشياء حلوه
وانبسطن… https://t.co/jgSi5jR03b’

As in it seen in ‘الملز،المربع‘ ,’الرعب،حديقة’ are separated by a فاصله “faseleh” removing it will result into concatenating the two words together and be treated as a single word which is a result we want to avoid. And the same situation applies to ‘#موسم_الرياض’. By simply removing the punctuation as they are present in a text will introduce these issues. The implemented fix is to examine each step alone and replace it’s with an appropriate replacement. Such as space. The next step will be removing all punctuations and by taking care of ‘faselah’ separated words. The removal of punctuation will not result in joining two words together. Therefore, we can proceed in the removal of all the punctuations in present in our tweets data. The implemented process is summarized by the following flowchart

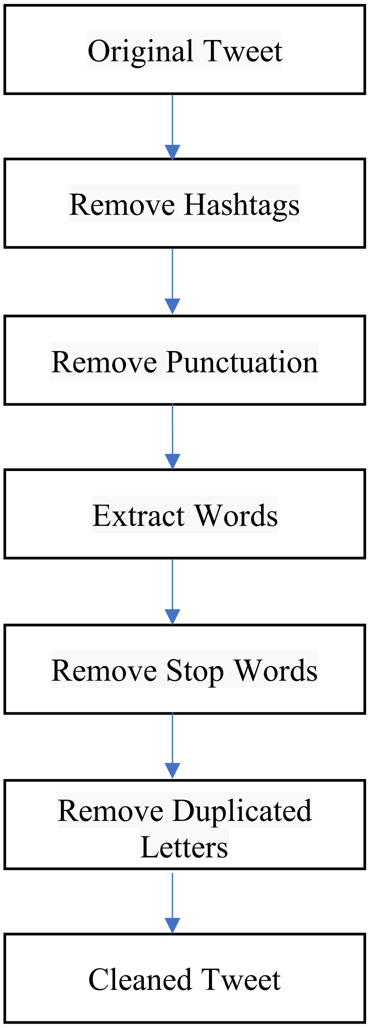


Figure 3.3: Data Cleaning Process

The following table shows examples of five tweets collected from the API. As shown in figure 4 we can see that the left column contains the original tweet. The right side on the other hand contains examples of processed tweets. The flow chart shown in table 4 was applied to clean and process the tweet.:

Table 3.2: Examples of Processed Tweets

Original Tweets	Processed Tweets
حد يعرف سعر الالعاب في ونتر وندر لاند ؟!! #موسم_الرياض #ونتر_وندرلاند	حد يعرف سعر الالعاب ونتر وندر لاند موسم الرياض ونتر وندرلاند
#موسم_الرياض فيه مطاعم العاب نارية النافورة، الرعب،حديقة الملز،المربع، ملاهي #ونتر_وندرلاند كل الاشياء حلوه وانبسطن… https://t.co/jgSi5jR03b	موسم الرياض مطاعم العاب نارية النافورة الرعب حديقة الملز المربع ملاهي ونتر وندرلاند الاشياء حلوه وانبسطن
🌃سينما على الاجواء المفتوحة وياسلام سلم من قدك يا #الرياض#موسم_الرياض #بوليفارد_الرياض #ونتر_وندرلاند 😏 https://t.co/cK8WSZ6MZE	سينما الاجواء المفتوحة وياسلام سلم قدك الرياض موسم الرياض بوليفارد الرياض ونتر وندرلاند
#موسم_الرياض #ونتر_وندرلاند من أحلى الأيام فعلياhttps://t.co/RKeUt3Pdzt ❤️	موسم الرياض ونتر وندرلاند احلى الايام فعليا
نفسيتي تحتاج اروح لـ #ونتر_وندرلاند 💔💙😭	نفسيتي تحتاج اروح ونتر وندرلاند

3.4 Data Labeling

In this project we intend to address the problem of sentiment analysis of entertainment events in KSA domain, we provide an event-specific sentiment dataset, which contain individuals' tweets related to some of large events that has been held in these periods and is annotated manually by human beings. Each tweet in our dataset is manually labeled by sentiment (positive, negative, or neutral). The process of labeling a dataset is not a straightforward step. While it is not science by itself, it is still a difficult problem. And an indispensable step in data preprocessing of supervised learning. Deciding the best fitted approach is an elusive task. Its main challenge which it presents the time it will take to label them correctly. The approach we took is labeling them manually by each member of our team and passing the labels to each other for rechecking and second opinion. Using a standard sentiment classifier with events tweets could lead to learning confusion. For example, some people may writing a song's lyrics that include negative words, these words considered as positive in event domain because that fans is just loving this song that he listened it in concert, even though it's associated with negative sentiment in general domain. the classification performance quality is depending on determining a good set of features. This is especially true for lexical features, since one word or sentence may reflect Secondly social data are written in informal languages which include slang and abbreviations, also affected by spelling and grammar errors.

3.5 Data Representation

In our process of extracting features. We have utilized this type of feature extraction which is popular in the domains of the text classification and sentiment analysis Bag-Of-Words, which will be used in the creation of a vocabulary of unique words in our entertainment-focused dataset. Its extracted vectorized features will be used in training a machine learning model. BOW is a popular textual data extraction method. The implementation of BOW can be categorized in these steps.

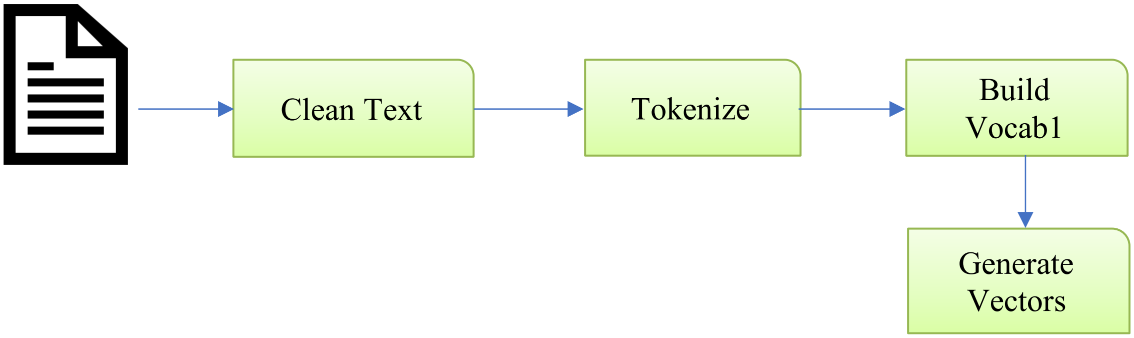


Figure 3.4: BOW Implantation Steps

As it shown in Figure 7 (3.4 7 is not there), BOW process are four steps.

1) Cleaning a text involves removing stop words which are very common in a document such as “في”, and “كذا”.
2) Tokenizing a document as in representing it as a collection of words.
3) Build vocabulary of unique words in all of the documents.
4) Represent a document in vectorized features by Bag-Of-Words or TF-IDF methods.

BOW is simple algorithm that given a document it creates a vocabulary of unique words occurring in the document in our case its Arabic tweets. A generated vocabulary V = {w1, w2, w3, w4, ... , wn }, Where X = { x1, x2. x3, x4, …, xn}, In which each element in xi corresponds to a word in wi. The value of xi is either a binary of a word occurrence in a tweet or a number of occurrences of a word in a tweet which gives it term frequency (TF). Term frequency is a value which tells how frequent a word is our dataset. For example, we have these two tweets in our dataset.

‘حضرو حفلة ليلة سهم الي من كثير غيرانه’

‘اشوف مقاطع حفلة ليلة سهم قلبي يتقطع قطعه قطعه وانا’

A vocabulary would consist of words that are unique but not including stop words.

The bag of words representation of the two tweets above will result into.

Table 3.3:BOW Representation of Tweet 1 and Tweet 2

Vocabulary	Tweet 1	Tweet 2
كثير	1	0
غيرانه	1	0
حضرو	1	0
حفلة	1	1
ليلة	1	1
سهم	1	1
قلبي	0	1
يتقطع	0	1

Term frequency representation is going to result in the number of occurrences of a word in tweets.

To represent the two tweets demonstrated above collected from an event of ‘ليلة سهم’ (Layla shm) in terms of term frequency in which it will tell the number of occurrences of a word in the document.

Table 3.4: TF Representation of Tweet 1 and Tweet

Vocabulary	**Term Frequency (TF)**
كثير	1
غيرانه	1
حضرو	1
حفلة	2
ليلة	2
سهم	2
قلبي	1
يتقطع	1
اشوف	1
مقاطع	1

Another feature representation is Term Frequency-Inverse Document Frequency (TF-IDF) In which it reduces the weight assigned to most frequent words. As we want more weight to rare words as their more informative than frequent words. TF – IDF can be calculated by as shown in the equation below:

$$TF - IDF_{(w_i)} = TF_{(w_i)} \times log \frac{|D|}{DF_{(w_i)}}$$

Equation 3.1 TF-IDF Formula

In where TFwi is the number of occurrences of wi, and |D| is the total number of documents, and DFwi is the number of documents contains the word wi

CHAPTER 4 (Project 2)
Model Planning and Building

4.0 Introduction

In this chapter, will provide our model planning and building, and explain the experiments with evaluation for each model.

4.1 Model Selection and Building

We choose 8 types of classification models and do the experiments on them which are as follow:

1) Ridge Classifier
2) Logistic Regression
3) KNeighbors Classifier
4) SGD Classifier
5) AdaBoost Classifier
6) SVC
7) Bagging Classifier
8) Random Forest Classifier

- Ridge regression: as the name suggests, is a method for regression rather than classification. Presumably you are using a threshold to turn it into a classifier. In any case, you are simply learning a linear classifier that is defined by a hyperplane.[12]

- Logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.[13]

- **KNeighbors Classifier** is a basic calculation that stores every accessible case and groups new cases dependent on a similitude measure (e.g., separation capacities). KNN has been utilized in measurable estimation and example acknowledgment as of now toward the start of the 1970s as a non-parametric system.[14]

- **Stochastic gradient descent** is a simple yet very efficient approach to fit linear models. It is particularly useful when the number of samples (and the number of features) is very large. The classes SGDClassifier and SGDRegressor provide functionality to fit linear models for classification and regression using different (convex) loss functions and different penalties .[15]

- **AdaBoost**, short for “Adaptive Boosting”, is the first practical boosting algorithm proposed by Freund and Schapire in 1996. It focuses on classification problems and aims to convert a set of weak classifiers into a strong one.[16]

- **SVC (Support Vector Classifier)** is to fit to the data you provide, returning a "best fit" hyperplane that divides, or categorizes, your data. From there, after getting the hyperplane, you can then feed some features to your classifier to see what the "predicted" class is. This makes this specific algorithm rather suitable for our uses, though you can use this for many situations.[17]

- **Bagging classifier** is an ensemble meta-estimator that fits base classifiers each on random subsets of the original dataset and then aggregate their individual predictions (either by voting or by averaging) to form a final prediction. Such a meta-estimator can typically be used as a way to reduce the variance of a black-box estimator [18]

- **Random Forest** is an ensemble learning technique for arrangement, relapse and different errands that work by building a large number of choice trees at preparing time and yielding the class that is the method of the classes (grouping) or mean forecast (relapse) of the individual trees[19]

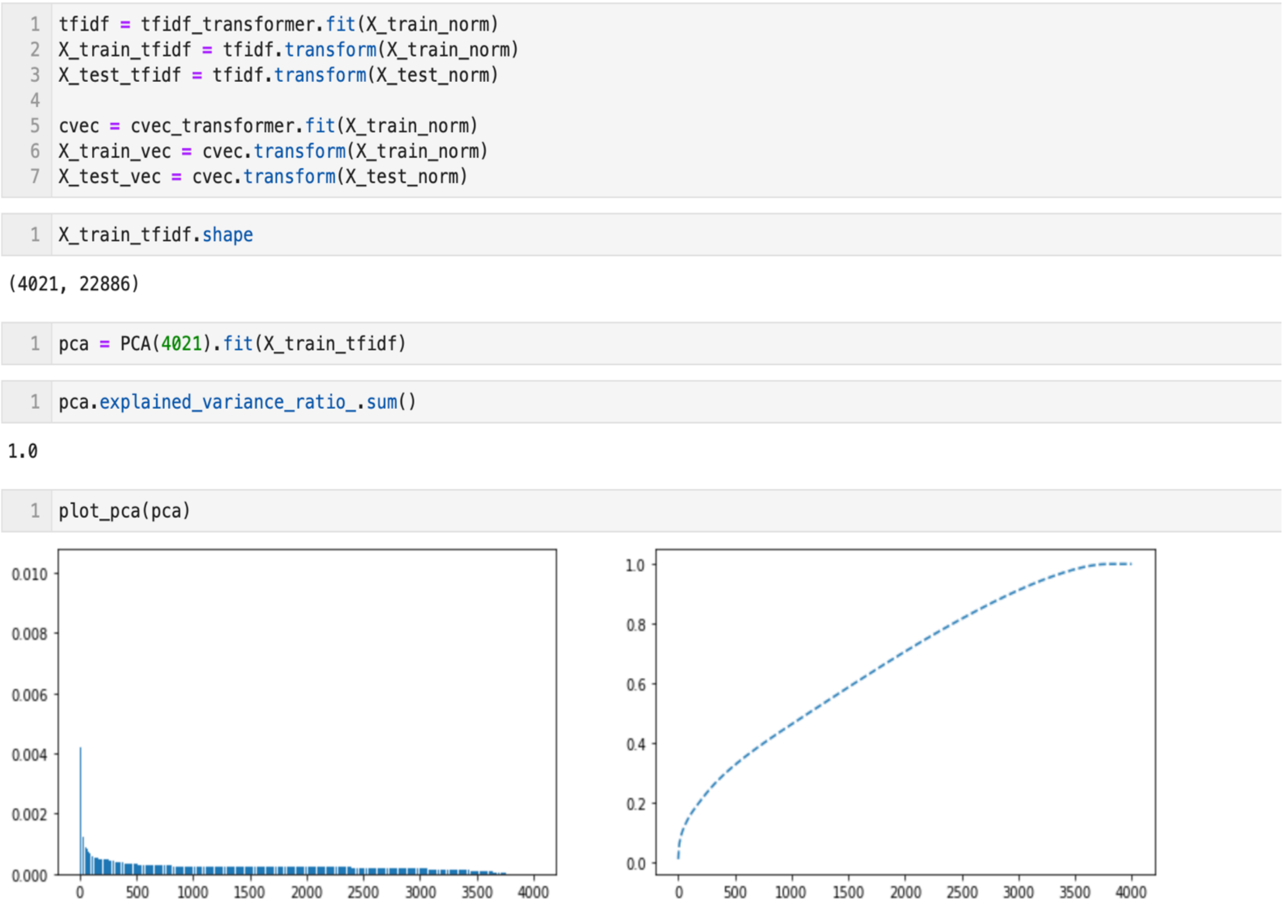
4.2 Crowdsourcing

Because the labeling phase is important and affects the result, we should be doing it with different people's perspectives.

The labeling process was distributed among 30 students, we have shared files through cloud drive. each student has taken a file with 500 tweets. the process has been done by choosing specific indexes from each file as a golden question to measure the accuracy of each student. we write simple code in python that chooses these indexes to measure the accuracy and compared it with the rights label. so any student that has accuracy under 80% percent were rejected and ask him to relabel. most of the students' accuracy was above 80%.

4.3 Features Engineering

. The total amount of features that we have is 32,000 so we try to reduce them by using Principal Component Analysis (PCA) which is "an unsupervised, non-parametric statistical technique primarily used for dimensionality reduction in machine learning"[20]. We get 4,000 components that represent 95% of all features as shown at the next page.



```
tfidf = tfidf_transformer.fit(X_train_norm)
X_train_tfidf = tfidf.transform(X_train_norm)
X_test_tfidf = tfidf.transform(X_test_norm)

cvec = cvec_transformer.fit(X_train_norm)
X_train_vec = cvec.transform(X_train_norm)
X_test_vec = cvec.transform(X_test_norm)
```

```
X_train_tfidf.shape
```

```
(4021, 22886)
```

```
pca = PCA(4021).fit(X_train_tfidf)
```

```
pca.explained_variance_ratio_.sum()
```

```
1.0
```

```
plot_pca(pca)
```

Figure 4.3.1: no of components that represent most of the data and pca code

After that each model in the previous section has been used with a different combination of features as shown below:

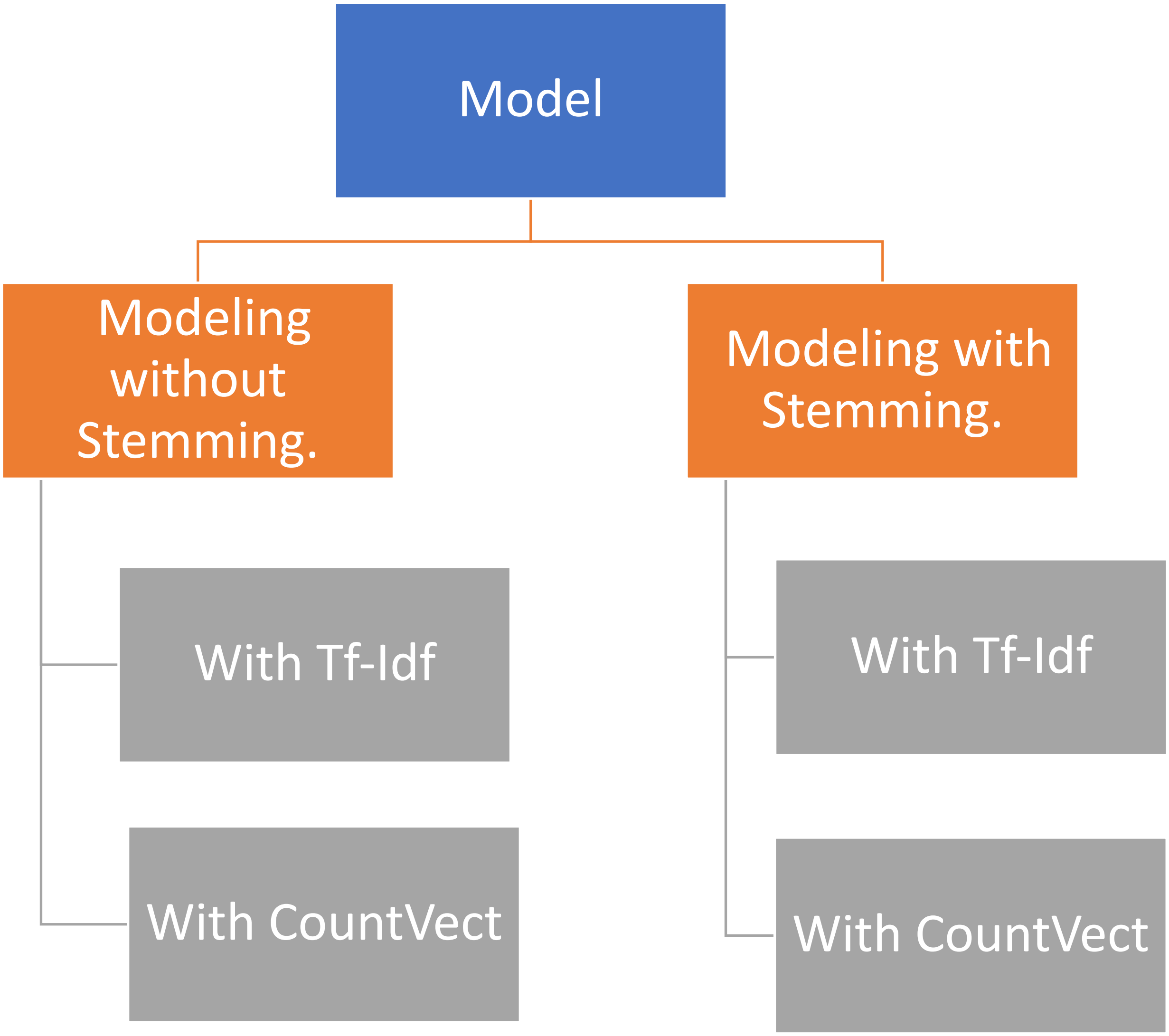


Figure 4.3.2: Combinations of features

The stemming process has done by using the **Tashaphyne** library, which is an Arabic light stemmer and segment. It mainly supports light stemming (removing prefixes and suffixes) and gives all possible segmentations. We tried each model with stemming one time with **TF-IDF** and one with **CountVect** and the same process without stemming as shown in the figure 4.3.

4.4 Model Evaluation

Figure 4.4.1 and 4.4.3show the performance for each model with different combinations:

- With Stemming

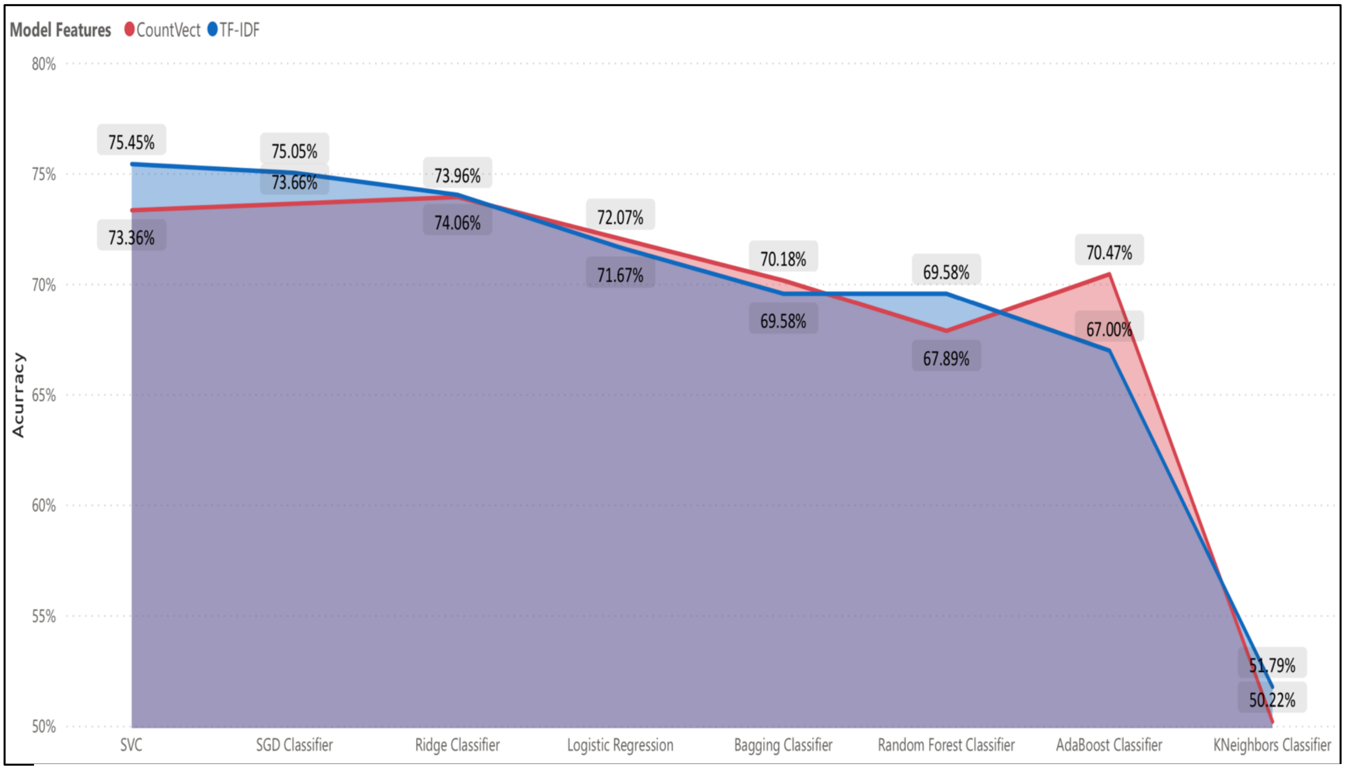
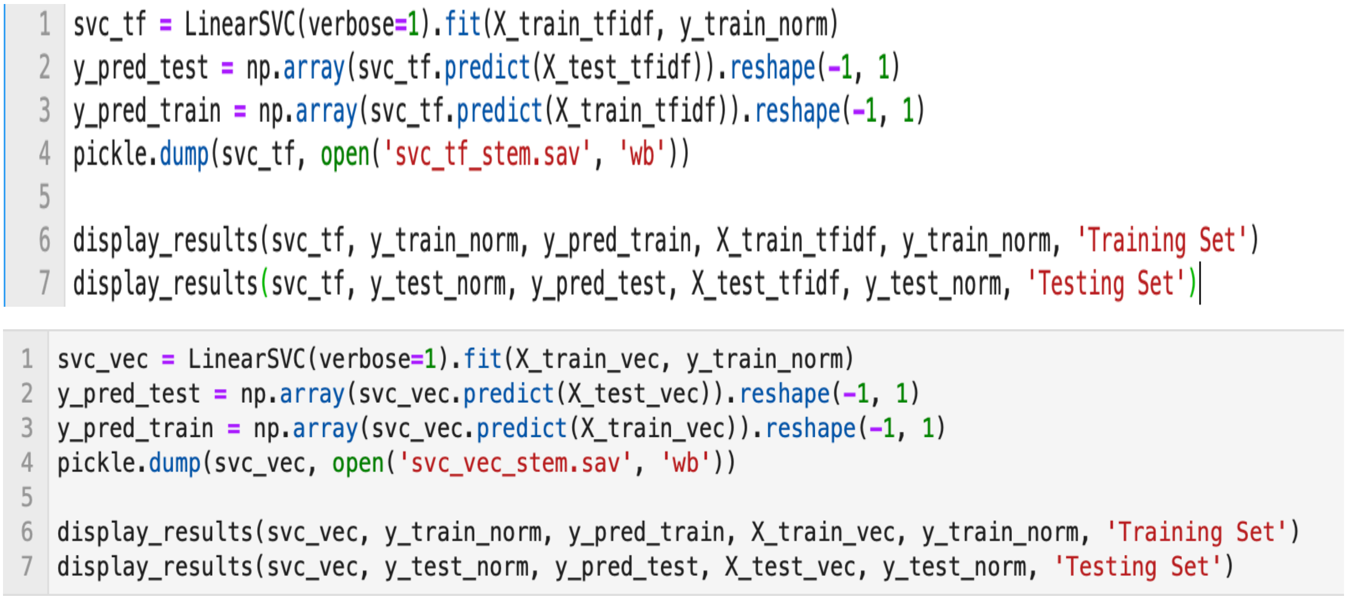


Figure 4.4.1: Models performance with Stemming

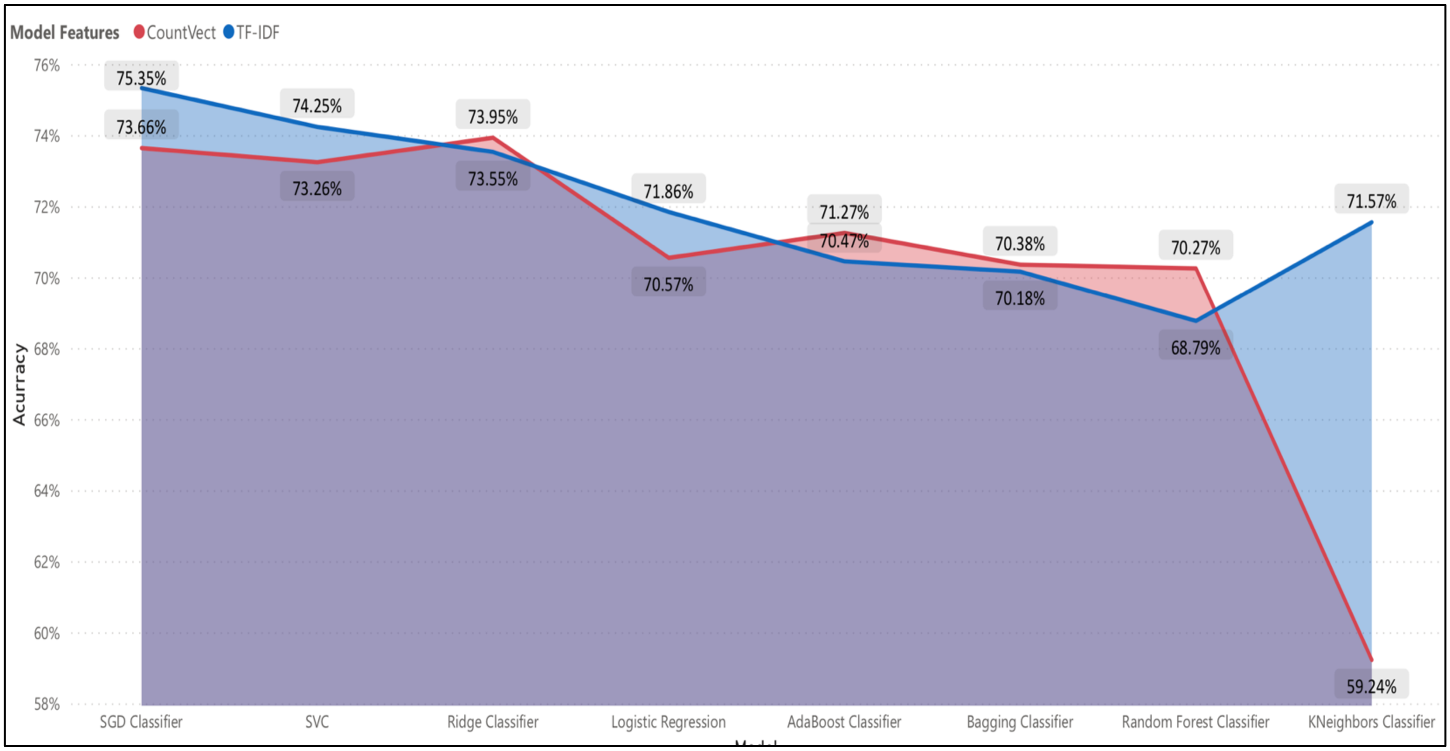


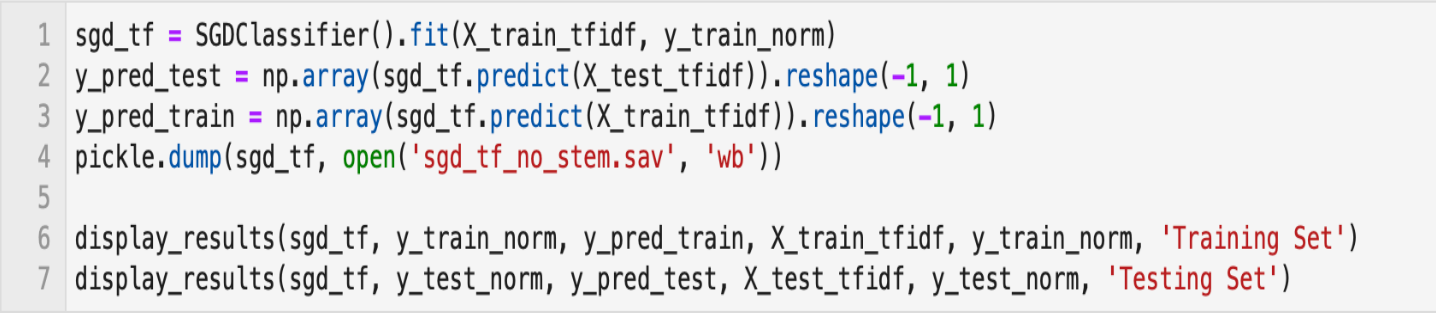
```python
svc_tf = LinearSVC(verbose=1).fit(X_train_tfidf, y_train_norm)
y_pred_test = np.array(svc_tf.predict(X_test_tfidf)).reshape(-1, 1)
y_pred_train = np.array(svc_tf.predict(X_train_tfidf)).reshape(-1, 1)
pickle.dump(svc_tf, open('svc_tf_stem.sav', 'wb'))

display_results(svc_tf, y_train_norm, y_pred_train, X_train_tfidf, y_train_norm, 'Training Set')
display_results(svc_tf, y_test_norm, y_pred_test, X_test_tfidf, y_test_norm, 'Testing Set')
```

```python
svc_vec = LinearSVC(verbose=1).fit(X_train_vec, y_train_norm)
y_pred_test = np.array(svc_vec.predict(X_test_vec)).reshape(-1, 1)
y_pred_train = np.array(svc_vec.predict(X_train_vec)).reshape(-1, 1)
pickle.dump(svc_vec, open('svc_vec_stem.sav', 'wb'))

display_results(svc_vec, y_train_norm, y_pred_train, X_train_vec, y_train_norm, 'Training Set')
display_results(svc_vec, y_test_norm, y_pred_test, X_test_vec, y_test_norm, 'Testing Set')
```

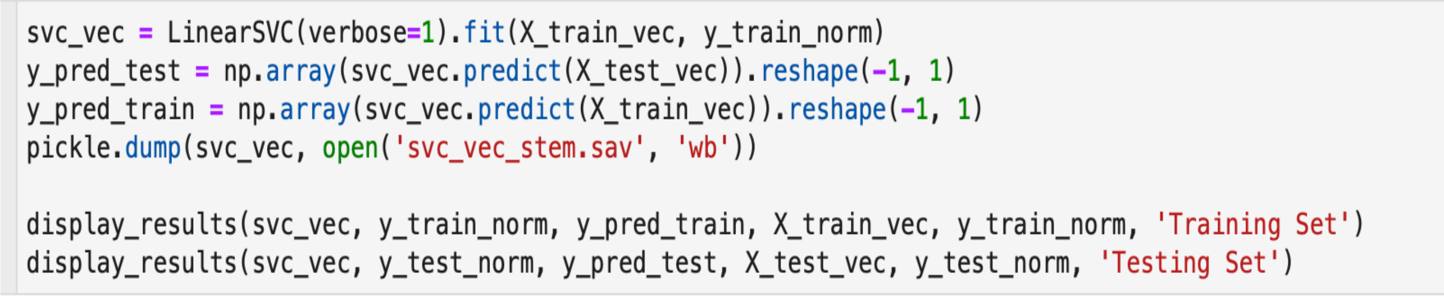
Figure 4.4.2: svc modeling code

- Without Stemming





```
sgd_tf = SGDClassifier().fit(X_train_tfidf, y_train_norm)
y_pred_test = np.array(sgd_tf.predict(X_test_tfidf)).reshape(-1, 1)
y_pred_train = np.array(sgd_tf.predict(X_train_tfidf)).reshape(-1, 1)
pickle.dump(sgd_tf, open('sgd_tf_no_stem.sav', 'wb'))

display_results(sgd_tf, y_train_norm, y_pred_train, X_train_tfidf, y_train_norm, 'Training Set')
display_results(sgd_tf, y_test_norm, y_pred_test, X_test_tfidf, y_test_norm, 'Testing Set')
```



```
svc_vec = LinearSVC(verbose=1).fit(X_train_vec, y_train_norm)
y_pred_test = np.array(svc_vec.predict(X_test_vec)).reshape(-1, 1)
y_pred_train = np.array(svc_vec.predict(X_train_vec)).reshape(-1, 1)
pickle.dump(svc_vec, open('svc_vec_stem.sav', 'wb'))

display_results(svc_vec, y_train_norm, y_pred_train, X_train_vec, y_train_norm, 'Training Set')
display_results(svc_vec, y_test_norm, y_pred_test, X_test_vec, y_test_norm, 'Testing Set')
```

Figure 4.4.3: Models performance without Stemming and the code for best one

So, the best performance was the SVC model with the combination of TF-IDF and Stemming with 75.45% of accuracy, the figure 4.4.3 shows the details of SVC performance with Recall and F1 score for each label. Considering the labels below stands for

0 = negative

1 = positive

2 = Neutral

Accuracy = TP+TN/TP+FP+FN+TN

Precision = TP/TP+FP

Recall = TP/TP+FN

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

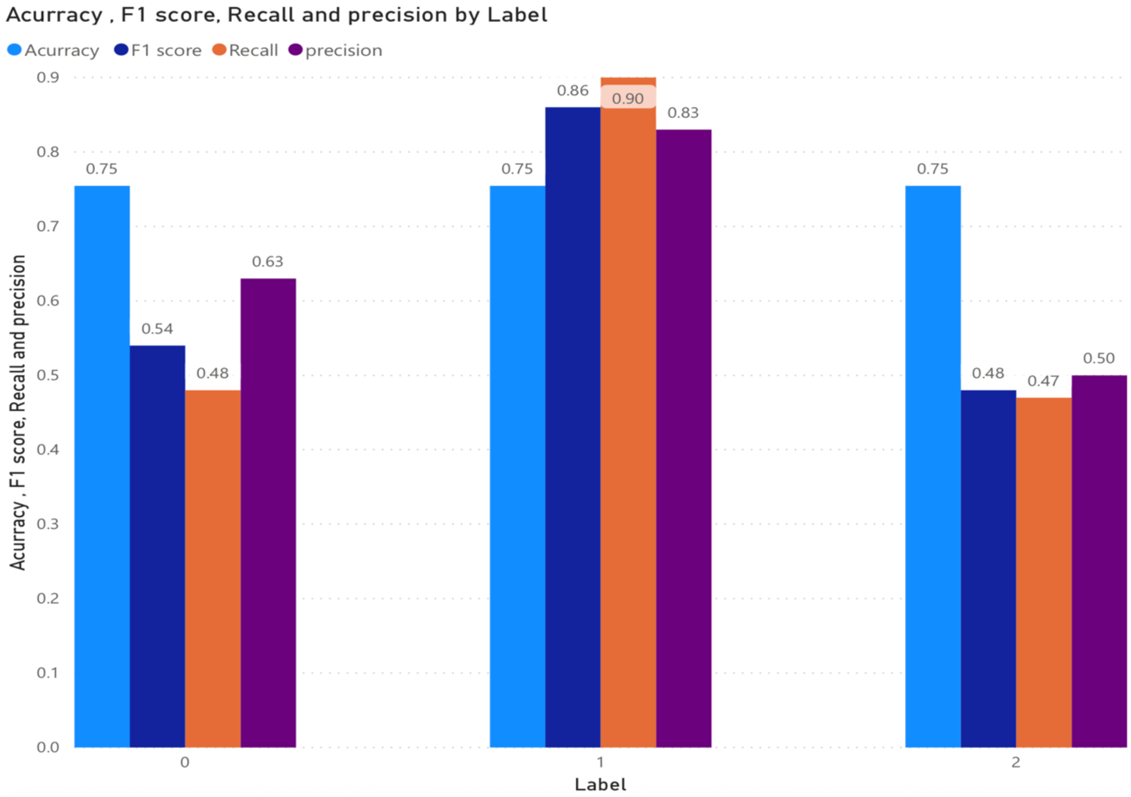


Figure 4.4.4: Details of the svc model performance

4.4 Sentiment Lexicon

Sentiment lexicon is one of the two main approaches to sentiment analysis. With this approach, a dictionary of positive and negative words is required, with a positive or negative sentiment value assigned to each of the words. Lexicon-based approaches is a piece of a text message is represented as a bag of words. Following this representation of the message, sentiment values from the dictionary are assigned to all positive and negative words or phrases within the message. Apart from a sentiment value, the aspect of the local context of a word and even the domain of the lexicon is usually taken into consideration, so our lexicon will help anyone who wants to work on the entertainment domain that will save their time for labeling and speed up the process of the work. The table below shows samples of our sentiment lexicon:

Table 4.4: lexicon sample

Positive	Negative
انجاززززز عظيم	وين وصلنا
عظيم صدقوني	وصلنا الأنحلال
صدقوني باقي	الأنحلال والتخلف
باقي ونتر	والتخلف والكبت
وندرلاند يقدرني	والكبت والشفاحه
يقدرني فعل	والشفاحه ؟؟؟

CONCLUSIONS AND FUTURE WORK

5.0 Conclusion

In this project, we have proposed a Sentiment Analysis Approach for the Entertainment Events domain which contains 62,000 tweets. Our data was labeled manually into 3 categories. We have applied many experiments with different models and combinations of features trying to get the most accurate model, most of the dataset was a positive tweets more than the other categories. The best two models were

- SVC classifier with steaming and tf-idf with 75.45% of accuracy.
- SGD classifier with tf-idf and without steaming with the accuracy 75.35%.

5.1 Future Work

Our future goal is to enhance our results and improve our sentiment lexicon content to be more specific and efficient, especially with neutral tweets to get more accurate results. And we want to expand the experiments and try to implement the different features and combinations such as part of speech and the sentiment lexicon to improve the accuracy of sentiment classifying. We are also planning to publish the lexicon and our project to the public to support the Arabic analysis community Due to the limited number of published works in the field of entertainment and events in particular

REFRENCES

[1] J. Strickl, “Analytics: is it more than a buzzword?,” *BI Corner*, 16-Sep-2015. .

[2] “Hal Varian on the Need for Data Interpreters,” *aLittleCode*, 08-Oct-2017. .

[3] C. O’Neil and R. Schutt, *Doing Data Science: Straight Talk from the Frontline*. O’Reilly Media, Inc., 2013.

[4] “Data science and open source,” 09-Aug-2013. [Online]. Available: http://www.ibm.com/developerworks/library/os-datascience/index.html. [Accessed: 01-Dec-2019].

[5] S. Aloufi and A. E. Saddik, “Sentiment Identification in Football-Specific Tweets,” *IEEE Access*, vol. 6, pp. 78609–78621, 2018.

[6] E. Refaee and V. Rieser, “An Arabic Twitter Corpus for Subjectivity and Sentiment Analysis,” p. 6.

[7] M. Abdul-Mageed, M. Diab, and S. Kübler, “SAMAR: Subjectivity and sentiment analysis for Arabic social media,” *Comput. Speech Lang.*, vol. 28, pp. 20–37, Jan. 2014.

[8] T. Kanan *et al.*, “A Review of Natural Language Processing and Machine Learning Tools Used to Analyze Arabic Social Media,” in *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*, 2019, pp. 622–628.

[9] H. AL-Rubaiee, R. Qiu, K. Alomar, and D. Li, “Sentiment Analysis of Arabic Tweets in e-Learning,” *J. Comput. Sci.*, vol. 12, no. 11, pp. 553–563, Feb. 2017.

[10] A. M. Alayba, V. Palade, M. England, and R. Iqbal, “Arabic language sentiment analysis on health services,” in *2017 1st International Workshop on Arabic Script Analysis and Recognition (ASAR)*, 2017, pp. 114–118.

[11] “(PDF) Arabic tweets sentiment analysis - A hybrid scheme.” [Online]. Available: https://www.researchgate.net/publication/283664122_Arabic_tweets_sentiment_analysis_-_A_hybrid_scheme. [Accessed: 01-Dec-2019].

[12] “machine learning - Why does ridge regression classifier work quite well for text classification? - Cross Validated.” https://stats.stackexchange.com/questions/17711/why-

does-ridge-regression-classifier-work-quite-well-for-text-classification (accessed Apr. 08, 2020).

[13] “What is Logistic Regression?,” *Statistics Solutions*. https://www.statisticssolutions.com/what-is-logistic-regression/ (accessed Apr. 08, 2020).

[14] O. Harrison, “Machine Learning Basics with the K-Nearest Neighbors Algorithm,” *Medium*, Jul. 14, 2019. https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761 (accessed Apr. 08, 2020).

[15] “LinearRegression vs. SGDRegressor,” *Evening Session*, 13:30:00-04:00. ./week2-andrew-ng-machine-learning-with-python.html (accessed Apr. 08, 2020).

[16] SauceCat, “Boosting algorithm: AdaBoost,” *Medium*, Apr. 30, 2017. https://towardsdatascience.com/boosting-algorithm-adaboost-b6737a9ee60c (accessed Apr. 08, 2020).

[17] “Python Programming Tutorials.” https://pythonprogramming.net/linear-svc-example-scikit-learn-svm-python/ (accessed Apr. 08, 2020).

[18] “sklearn.ensemble.BaggingClassifier — scikit-learn 0.22.2 documentation.” https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.BaggingClassifier.html (accessed Apr. 08, 2020).

[19] “Random forest,” *Wikipedia*. Mar. 12, 2020, Accessed: Apr. 08, 2020. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Random_forest&oldid=945233128.

[20] H. Goonewardana, “PCA: Application in Machine Learning,” *Medium*, Feb. 28, 2019. https://medium.com/apprentice-journal/pca-application-in-machine-learning-4827c07a61db (accessed Apr. 09, 2020).