

**Arabic Sentiment Analysis**

|  |
| --- |
| Faris Adel Alahmadi |
| Sultan Fadel Fatani |
| Ramzi Yousef Almazroe |
| Faisal Majed Aljuhani |

Supervisor: Dr. Ali Alkhathlan

Faculty of Computing and Information Technology King Abdulaziz University

2021-2nd semester

Table of Contents

[**Arabic Sentiment Analysis** 3](#_Toc68891360)

[**Abstract** 3](#_Toc68891361)

[**Introduction** 4](#_Toc68891362)

[**Problem statement** 5](#_Toc68891363)

[**Language challenge** 6](#_Toc68891364)

[**Solution** 6](#_Toc68891365)

[**Datasets** 7](#_Toc68891366)

[**Hotel reviews** 7](#_Toc68891367)

[**Attraction reviews** 8](#_Toc68891368)

[**Hotel reviews** 8](#_Toc68891369)

[**Movie reviews** 8](#_Toc68891370)

[**Products reviews** 9](#_Toc68891371)

[**Restaurants reviews** 9](#_Toc68891372)

[**Data preparation methods** 9](#_Toc68891373)

[**What is Data Preparation?** 9](#_Toc68891374)

[**Benefits of Data Preparation** 9](#_Toc68891375)

[**Data Preparation Steps** 10](#_Toc68891376)

[**Text conversion to numbers** 10](#_Toc68891377)

[**Machine learning algorithms** 13](#_Toc68891378)

[**Naïve bayas** 13](#_Toc68891379)

[**Pros and Cons of Naive Bayes** 15](#_Toc68891380)

[**Applications of Naive Bayes Algorithms** 15](#_Toc68891381)

[**Logistic regression** 15](#_Toc68891382)

[**Support vector machine (SVM)** 16](#_Toc68891383)

[**Random forest** 18](#_Toc68891384)

[**Flow chart of the project** 19](#_Toc68891385)

[**Data preprocessing flow chart** 20](#_Toc68891386)

[**Training & testing flow chart** 21](#_Toc68891387)

[**GUI app flow chart** 22](#_Toc68891388)

[**Timeline of the project** 23](#_Toc68891389)

[**Project timeline** 23](#_Toc68891390)

[**Equipment or software will be used** 25](#_Toc68891391)

[Figure 1 Sentiment analysis approach 6](#_Toc68867415)

[Figure 2: Booking Hotal Reviews Datasets 7](file:///C:\Users\SPC\Downloads\Document.docx#_Toc68867416)

[Figure 3 Tripadvisor: Attraction Reviews Dataset 7](#_Toc68867417)

[Figure 4 Tripadvisor: Hotel Reviews Dataset 7](#_Toc68867418)

[Figure 5 Elcinema: Movie Reviews Dataset 8](#_Toc68867419)

[Figure 6 Souq: Products Reviews Dataset 8](#_Toc68867420)

[Figure 7 Qaym & tripadvisor: Restaurants Reviews Dataset 9](#_Toc68867421)

[Figure 8: Naïve bayas 13](#_Toc68867422)

[Figure 9: Sigmoid function 15](#_Toc68867423)

[Figure 10: Support vector machine SVM 15](#_Toc68867424)

[Figure 11: Random forest 17](#_Toc68867425)

[Figure 12: Data preprocessing 19](#_Toc68867426)

[Figure 13:Training & testing 20](#_Toc68867427)

[Figure 14: GUI app 21](#_Toc68867428)

[Figure 15:Milestones 22](#_Toc68867429)

[Figure 16: Project timeline 22](#_Toc68867430)

[Table 1 Bag of words 11](#_Toc68867431)

[Table 2: BOW Eg:1 12](#_Toc68867432)

[Table 3: BOW Eg:2 12](#_Toc68867433)

[Table 4: BOW Eg:3 12](#_Toc68867434)

[Table 5: 1-gram 12](#_Toc68867435)

[Table 6: 2-gram 12](#_Toc68867436)

[Table 7: 3 gram 13](#_Toc68867437)

[Table 8: Example of naïve bayas 14](#_Toc68867438)

[Table 9: Project timeline 24](#_Toc68867439)

[Table 10: List of references: 25](#_Toc68867440)

# **Arabic Sentiment Analysis**

# **Abstract**

Our project is about the Arabic Sentiment analysis which is a task of natural language processing that has recently attracted increasing attention. However, sentiment analysis research has mainly been administered for the English language. Although Arabic is one of the most used languages on the Internet. In this report, we will Introduce the Arabic Sentiment Analysis topic and know more about its importance, tasks, uses and we will discuss the (SA) approaches, the software, tools also libraries we are going to use to complete our project.

# **Introduction**

Sentiment Analysis indicates the use of Natural Language Processing (NLP) and Machine Learning to specify and extract subjective information in a piece of text. Sentiment Analysis is powerfully useful as it enables us to obtain an overview of the wider public opinions (e.g., happy, sad, and anger) towards certain topics, products, or services. It is the method of determining the emotional tone behind a series of words, used to have an understanding of the attitudes, opinions, and emotions expressed within a web review. In general, Sentiment Analysis requires identifying four elements comprising entity, its aspect, opinion holder, and his sentiment. The extracted opinions can be classified as either objective or subjective text. The subjective text also can be classified as positive or negative or natural sentiments. Sentiment Analysis is important for the Arabic language, which suffers from a lack of interest in terms of applications of artificial intelligence, including machine learning and deep learning. As there is no scientific research so far has monitored and obtained findings on the processing levels of the Arabic language at a high and comprehensive level.

It has become an important element for decision-makers and business managers also as for the general public users to know sentiment and opinions. Because users increasingly tend to consult one another before making purchasing choices, decision-makers and business managers are now making important investments in measuring public opinion about their products and services. They invest in sentiment analysis not only to keep their clients satisfied but also to further improve new products and services and attract prospective customers. This fact motivates the NLP community to extend its efforts towards sentiment analysis research. Recently, the Arab world has become the main target of many multinational analysts because it represents an important player within the global economy and international politics. Mining opinions about issues, like oil and gas prices, market movements, and politics, has received extensive attention, particularly after the Arab spring. Despite this, Arabic sentiment analysis (ASA) research has not witnessed significant developments yet, mostly due to the lack of sentiment resources in Arabic.

Thus, many studies have tried to use machine translation on English sentiment resources. This bilingual approach, however, is just inept because of Arabic linguistic features, which are essentially different from the English language in terms of structure and grammar. Arabic studies must consider the various dialects and also the free writing forms that make research in this field more complicated. Most Arabic interactions in social media are produced in local dialects. However, monolingual studies, which treat Arabic texts, commonly ignore dialects and Arabic texts containing Latin letters. Most of the available resources and tools consider only modern standard Arabic (MSA). Thus, their reproduction isn't relevant because it would cause lower accuracy in real applications.

Our project seeks to enhance Arabic Sentiment Analysis progressing and with the exponential growth of Arabic online content and the increasing number of Arabic internet users, the spread of social media and microblogs like Twitter, Facebook, etc...,

The focus on Sentiment Analysis is deeply being studied. Sentiment Analysis in this language has gained the eye of many researchers in the last decade. However, in sentiment analysis, there are many research studies deal with the English language, but the Arabic sentiment analysis is still in its early stages and there are a few kinds of research that have been conducted in this direction. Also, one of the six official languages of the United Nations is the Arabic language. It is the official language of 27 countries and spoken by more than 422 million people within the Arabic world. The Arabic language is the fastest growing language during the last five years with a growth rate of 6091.9% in the number of Internet users, has ranked as the fourth most used language on the web, with internet availability and the exponential increase in it is usage, the web has become a platform to read and write information. This textual information is growing day by day on the internet, so there is an abundance of web forums, social media, personal blogs, etc.

The users of the internet are not just consumers of web content, but also producers of it. Their contributions have become very important to enrich the web content. These users express their opinion and share information on different fields and topics. Due to the increase of the opinions, reviews, feedbacks, and emotions on the web, exploring and analyzing this information becomes very important in determining the best decision making by other users ranging from products, movies, education, health, and politics to hotels and different services that help people, companies, institutions, and states making their decisions or prediction. So, the aim of this project is to develop an Arabic Sentiments Analysis software system that analyzes a huge number of texts from different resources by analyzing each review sentence with focusing on specific aspects or specific keywords it will help people in determining the best decision making towards certain topics, products, or services.

# **Problem statement**

The problem is that we need to make a machine recognize the compliments reviews and the criticism reviews so that we can use these classified reviews to get many benefits, the store funder needs to know how many users like a product and how many user dislikes it (it shouldn’t be product maybe service), so if the reviews classified into two classes its can be use the negative classes to see what is the thing that the user was obsessed about.

## **Language challenge**

Sentiment Analysis is highly dependent on the morphology of the language being analyzed. Unlike Latin language Arabic has a very complex and rich morphology compared to other languages. Most Arabic word can have several morphological aspects such asDerivational morphology(1) which is the mechanism of creating a new word based on an existing one for example the word (kataba, كَتَبَ) which means write and the word (Katib, كاتِب) means writer. Another morphology is **Inflectional morphology** (2) which Inflectional morphemes change what a word does in terms of grammar but does not create a new word like (eat, ate). finally, **Agglutinative morphology** (3) this is a unique morphology for Arabic language for example “and with his work” is written in Arabic as “وبعمله” This word can be split into four parts (و + ب + عمل + ه). Therefore, a single word in Arabic can contain more information than any other languages.

# **Solution**

We will tackle the problem of Sentiment analysis by using machine learning algorithms. The input of the model is a sequence of words and the output is weather the input has a positive or negative sentiment; thus, it is a classification problem. The pipeline of our approach is as follow: we first need to look for datasets for ASA that are ready to use. Second, we must prepare/clean the input text, cleaning data involves in removing stop words, punctuation and diacritics. Third, since machine learning algorithms do not work on raw texts directly, we must convert texts to numeric values (vectors), that is called *feature extraction.* A simple and effective feature extraction model for vectorizing texts is called Bag of Words (BOW). Two algorithms follow the BOW concept that are: CountVectorizer and Term Frequency-Inverse Document Frequency (TF-IDF). Fourth, after vectorizing our text input, we will train our machine learning algorithms and develop our classifiers. Algorithms such as Support Vector Machine, Logistic regression, and more will be used. We finally test our model based on the performance metrics found on the literature.

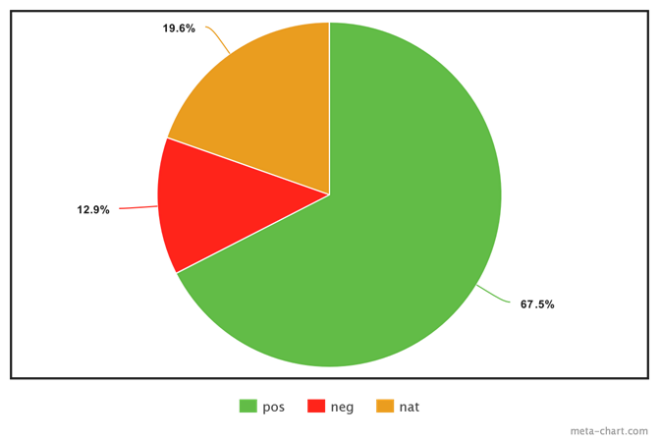
Sentiment analysis is usually solved using one of those approaches. Supervised approach, unsupervised approach, and hybrid approach. The approach used in this project is the supervised approach based on Support Vector Machine (SVM), Logistic regression, Random forest, and Naïve bayas. This architecture has two phases for reaching to sentiment score. The input is online reviews. The first phase is called "Analyze Data" which includes some functions as Reading and extracting data, Text analysis, and Text conversion to numbers using (BOW). Reading and extracting data which extracts the information and their parameters data and create arrays of records. Text analysis contains that splits sentiment reviews into sentences and tokenizes each review into some words. Natural language processing (NLP) Linguistics makes normalization functions and reformat data. Our proposed technique introduces an improved version of Bag-of-words (BOW) algorithm which is based on a word weight (TF-IDF) we may use Countvectorizer. The phase two called " Sentiment analysis score and polarity " the proposed technique can detect the polarity based on three levels of classification. At document level which evaluate the review as whole and assign it to a positive, negative, or neutral class. Second at sentence level Sentiment Analysis tries to identify whether the sentence contain an opinion or not. Finally, at word level Sentiment Analysis evaluate each word with positive, negative, or neutral value. Although the proposed technique is based on the word-by-word evaluation, it can handle expressions, wish words, some special cases.

# **Datasets**

We used different data set in our project that include different reviews:

* Attraction Reviews
* Hotel Reviews
* Movie Reviews
* Product reviews
* Restaurant reviews

Consist of number of examples for each class: Positive, Negative, and Neutral.



## **Hotel reviews**

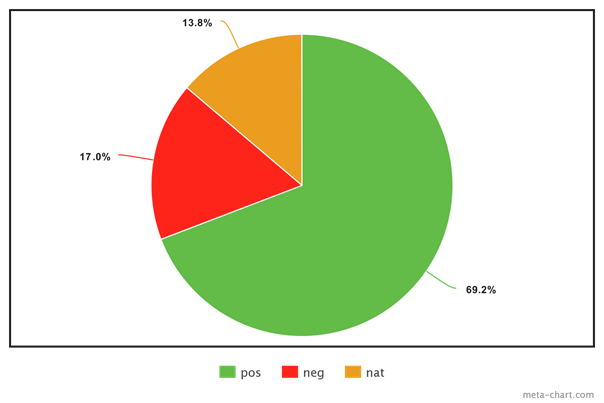
* Source: booking.com
* Positive: 276387
* Negative: 52849
* Natural: 80326
* Total 409562 record

Figure : Booking Hotel Reviews Datasets

## Chart, pie chart Description automatically generated**Attraction reviews**

* Source: tripadvisor.com
* Positive: 2073
* Negative: 81
* Natural: 0
* Total 2154 record

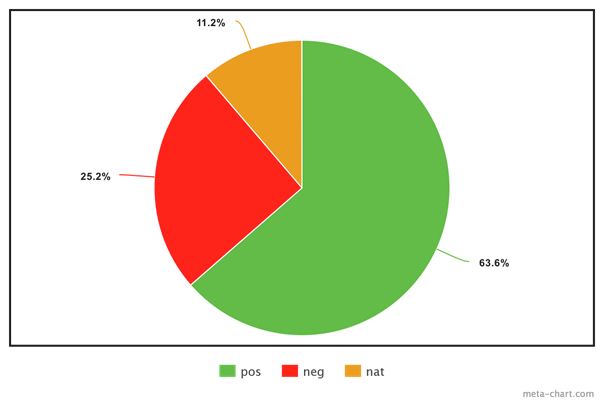
Figure 3 Tripadvisor: Attraction Reviews Dataset



## **Hotel reviews**

* Hotel reviews data set from tripadvisor.com
* Positive: 10775
* Negative: 2647
* Natural: 2150
* Total 15572 record

Figure 4 Tripadvisor: Hotel Reviews Dataset



## **Movie reviews**

* Movie reviews data set from elcinema.com
* Positive: 969
* Negative: 384
* Natural: 171
* Total 1524 record

Figure 5 Elcinema: Movie Reviews Dataset

## **Products reviews**

* Products reviews data set from souq.com
* Positive: 3101
* Negative: 863
* Natural: 308
* Total 4272 record

Figure 6 Souq: Products Reviews Dataset

## **Restaurants reviews**

* Dataset from qaym.com & tripadvisor.com
* Positive: 8030
* Negative: 2675
* Natural: 265
* Total 10970 record

Figure 7 Qaym & tripadvisor: Restaurants Reviews Dataset

# **Data preparation methods**

## **What is Data Preparation?**

is the technique of gathering, sorting, and consolidating into one file or data table for evaluation. It is a key stage before processing that often require reformatting data, making data corrections, and merging data sets to improve data.

## **Benefits of Data Preparation**

**Fix errors quickly**: Data preparation helps to handle errors before processing. After data beaning removed from its source, these errors become more difficult to be understood and corrected.

**Produce top-quality data**: Cleaning and reformatting datasets ensures that all data used in analysis will be of the highest possible quality.

**Make better business decisions:** Higher quality data that can be processed and analyzed more quickly and efficiently as a result of more timely, efficient, and high-quality business decisions.

## **Data Preparation Steps**

The specifics of the data preparation process vary by industry, organization and need, but the framework remains largely the same.

**1. Gather data.**

The data preparation process begins with selecting the appropriate data.

**2. Discover and assess data.**

After collecting the data, it is quite crucial to learn more about each dataset. This step is about getting to know the data and understanding what must be done before the data becomes useful in a particular context.

**3. Cleanse and validate data.**

Cleaning up the data is traditionally the longest process part of the data preparation process; However, it is critical for deleting erroneous details and filling in gaps. Important tasks here include:

* Removing extraneous data and outliers.
* Filling in missing values.
* Conforming data to a standardized pattern.
* Masking private or sensitive data entries.

After the data has been cleaned, it must be validated by testing for errors in the data preparation process up to this point. Often, during this process, a system error will become obvious. and will need to be resolved before continuing on.

**4. Transform and enrich data.**

Transforming data is the process of reformatting or value entries to reach a well-defined outcome, or to make the data more understandable by a wider audience. *Enriching* data refers to adding and connecting data with other related information to provide more details.

**5. Store data**

When all is in order, the data can be stored or channeled into a third-party application—such as a business intelligence tool—clearing the way for processing and analysis to take place.

# **Text conversion to numbers**

Machine learning algorithms cannot work with text data type, only with numeric data types, so how we can make the text readable by the machine learning algorithms? We just must find a way to convert the text to numbers or we can call it mapping not converting. There are a lot of algorithms to do that like word2vec, GLOVE, word embedding, Bag of Words.

In our project we will use bag of words approach, so what is bag of words (BOW) and what is the methods that belong to BOW:

BOW is a way of extracting features from text for use in modeling, to simplify the definition its bagging all words in the data set in one bag and assign all words to zero, but only the words that appear in the text assign it to one.

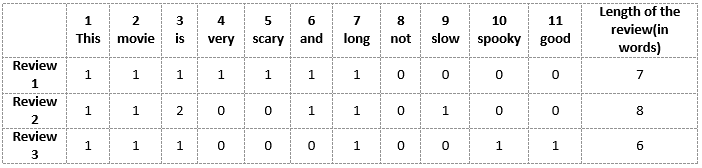
Example:

Review 1: This movie is very scary and long.

Review 2: This movie is not scary and is slow.

Review 3: This movie is spooky and good.

Table Bag of words



The BOW approach is too simple and easy to implement but also quite effective.

BOW approach has a lot of methods and techniques but here we will only use two of them:

* Countvectorizer
* TF-IDF (Term Frequency-Inverse Document Frequency)

Countvectorizer:

It is a BOW approach its work by count the word that appear in the text, so for example if the text has 3 of 4 “good” word this text has more probability to be a positive text and the same with word “bad”.

Example:

“this book is good”, “this movie is bad”, “the app is not good”.

“this book is good”

Table : BOW Eg:1



“this movie is bad”

Table : BOW Eg:2



“the app is not good”

Table : BOW Eg:3



TF-IDF

It’s like CountVectorizer but instead of Count the number of times each word appears in the document its TF\*IDF score.

TF-IDF is an information retrieval technique that weighs a term’s frequency (TF) and its inverse document frequency (IDF). Each word or term that occurs in the text has its respective TF and IDF score. The product of the TF and IDF scores of a term is called the TF-IDF weight of that term.

**Put simply, the higher the TF-IDF score (weight), the rarer the term and vice versa.**

Limit of CountVectorizer and TF-IDF is not always “good” word mean a positive sentiment sometimes the word “not” comes before the “good” word, so for that we will perform 1-gram ,2-gram and 3-gram method, so what is a N-gram method?

In the example above we pick every word and put it in a list and give every word its count, but what if we pick 2 words? Or 3 words? For example, instead of “good” we pick “not good”, “it’s good”, “not bad’ and “not too bad”.

As you see if we choose N window and pass it through the text then the performance will be better for sure.

Example:

“the app is not good.”

Table : 1-gram



Table : 2-gram



Table : 3 gram



Not always more (N) means better performance sometimes when you increase (N) the performance will go down, so we must try and find the best performance.

In our project we will do Countvectorizer and TF-IDF for each 1,2 and 3 grams.

# **Machine learning algorithms**

There are a lot of machine learning algorithms that’s work will specialty in the few decades the algorithms perform well in many problems.

Examples of machine learning algorithms:

* Naïve bayas.
* Neural network.
* Decision tree.
* Ada boost.
* Logistic regression.
* Support vector machine.
* Random forest.

And more.

In our project we will use these four algorithms:

* Naïve bayas.
* Logistic regression.
* Support vector machine.
* Random forest.

We choose these algorithms because it’s simple and does not include complex math, also its good algorithms and sometimes it gets good results with small data sets.

## **Naïve bayas**

It is a [classification algorithm](https://courses.analyticsvidhya.com/courses/introduction-to-data-science-2/?utm_source=blog&utm_medium=6stepsnaivebayesarticle) based on Bayes’ Theorem with an assumption of independence among predictors.

In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

Bayas’ theorem:

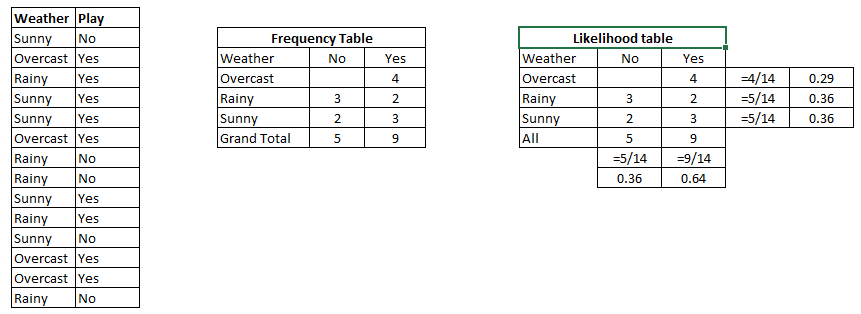


Figure : Naïve bayas

Example of how naïve bayas work:

Now, we need to classify whether players will play or not based on weather condition. The steps below explain how to do it.

Table : Example of naïve bayas



1. Convert the data set into frequency table.
2. Create likelihood table by calculate the probabilities.
3. Then use naïve bayas to make prediction.

The prediction: Players will play if weather is sunny. Is this statement being correct? Yes/no

We can solve it using above discussed method of posterior probability.

P (Yes | Sunny) = P(Sunny | Yes) \* P(Yes) / P (Sunny)

Here we have P (Sunny |Yes) = 3/9 = 0.33, P(Sunny) = 5/14 = 0.36, P(Yes)= 9/14 = 0.64

Now, P (Yes | Sunny) = 0.33 \* 0.64 / 0.36 = 0.60, which has higher probability.

The prediction is YES.

### **Pros and Cons of Naive Bayes**

**Pros:**

* easy and fast to predict class of test data set. It also performs well in multi class prediction.
* It performs well in case of categorical input variables compared to numerical variable(s).

**Cons:**

* On the other side naive Bayes is also known as a bad estimator, so the probability outputs from predict\_proba are not to be taken too seriously.
* Another limitation of [Naive Bayes](https://courses.analyticsvidhya.com/courses/naive-bayes?utm_source=blog&utm_medium=naive-bayes-explained) is the assumption of independent predictors. In real life, it is almost impossible that we get a set of predictors which are completely independent.

### **Applications of Naive Bayes Algorithms**

* **Real time Prediction.**
* **Multi class Prediction.**
* **Text classification/ Spam Filtering/ Sentiment Analysis.**
* **Recommendation System.**

# **Logistic regression**

Logistic Regression is a classification Machine Learning algorithm, it is a predictive analysis algorithm and based on the concept of probability.

Logistic regression algorithm uses hypothesis and a sigmoid function to make a prediction, so what is hypothesis and sigmoid function.

Hypothesis: is a linear function that its output is the dot product of weights and the features. Like this example:

sigmoid function: is a nonlinear function that its output between 0 and 1.

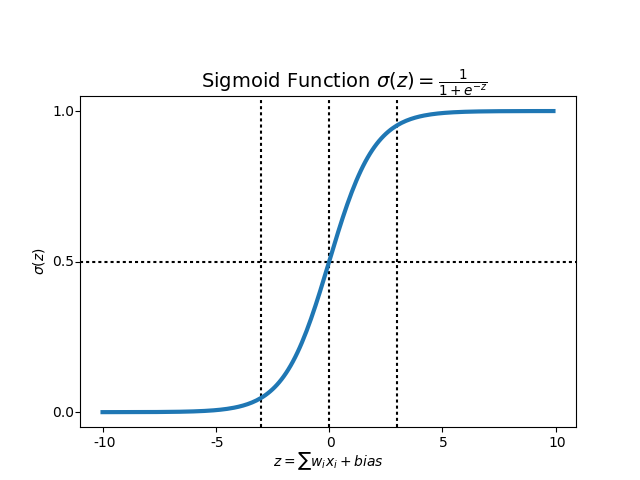


Figure : Sigmoid function

We will not go further in the mathematics behind this algorithm but at the end these algorithms use a probability function to make prediction, the prediction is a probability.

# **Support vector machine (SVM)**

Support Vector Machine (SVM) [machine learning algorithm](https://courses.analyticsvidhya.com/courses/introduction-to-data-science-2?utm_source=blog&utm_medium=understandingsupportvectormachinearticle) which can be used for both classification or regression challenges. In the SVM algorithm, we plot each data item as a point, and we perform classification by finding the best decision boundary that differentiates the two classes very well.



Figure : Support vector machine SVM

So, assume that we have 3 decision boundaries:



First, we make it a right boundary.



Then we chose the best and equally difference between two classes.

In this example we will choose the boundary C.

# **Random forest**

Random forest is a classification algorithm, that use bagging method by combining a lot of decision trees together, that is why they call it forest.

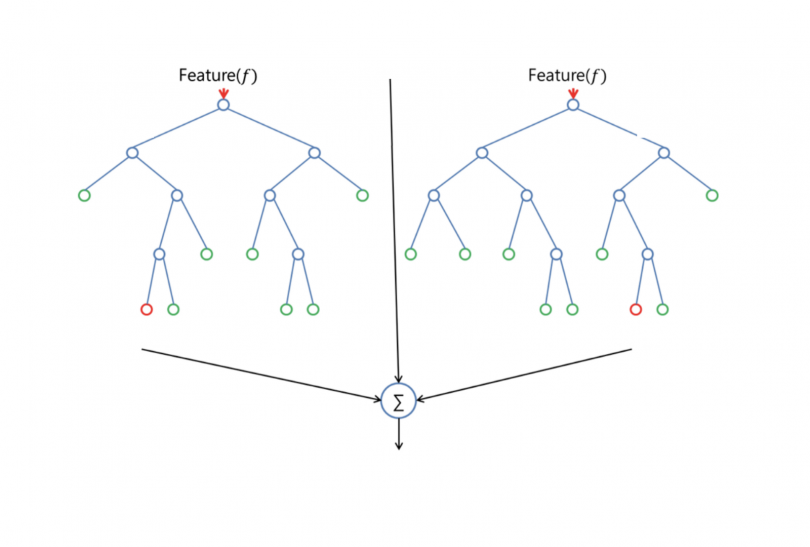


Figure : Random forest

This is the basic idea of random forest, by make prediction from multiple decision trees and then merging the results and make more accurate and more suitable prediction.

AT THE END

We will not perform these algorithms from scratch, we will use some APIs to make it easy and fast. We will make 24 experiment and we will choose the best accuracy to use it in a GUI app.

Diagram

Description automatically generated

Figure : Experiment flow chart

# **Flow chart of the project**

The project will go through 3 phases.

* Data preprocessing.
* Training & testing
* GUI app

There is a flow chart for every phase. And the phases one and two are performing just one time after that we will extract the model and use the model in phase 3 when we are using GUI app.

## **Data preprocessing flow chart**

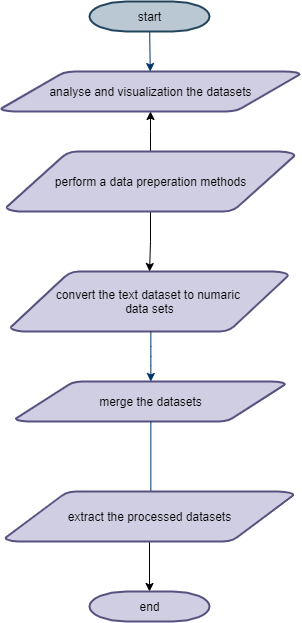


Figure : Data preprocessing

## **Training & testing flow chart**

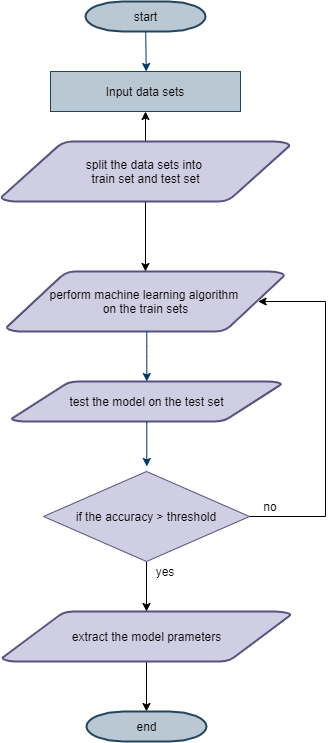


Figure :Training & testing

## **GUI app flow chart**

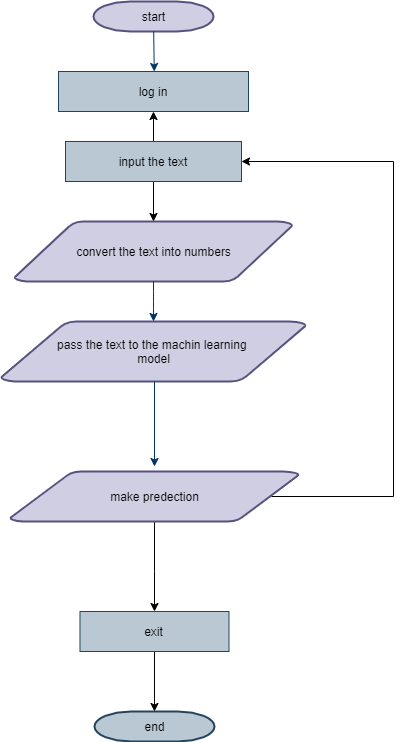


Figure : GUI app

## **Timeline of the project**

We designed the milestones to be that the planning and preprocessing phase will be in this semester and the executing & testing, GUI designing and closing phases will be in the next semester.

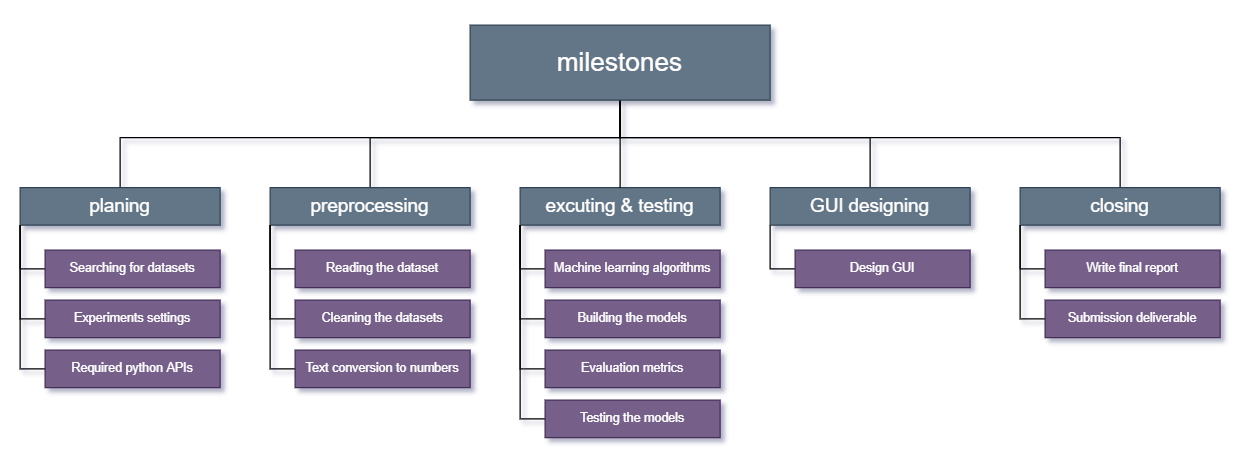


Figure :Milestones

## **Project timeline**

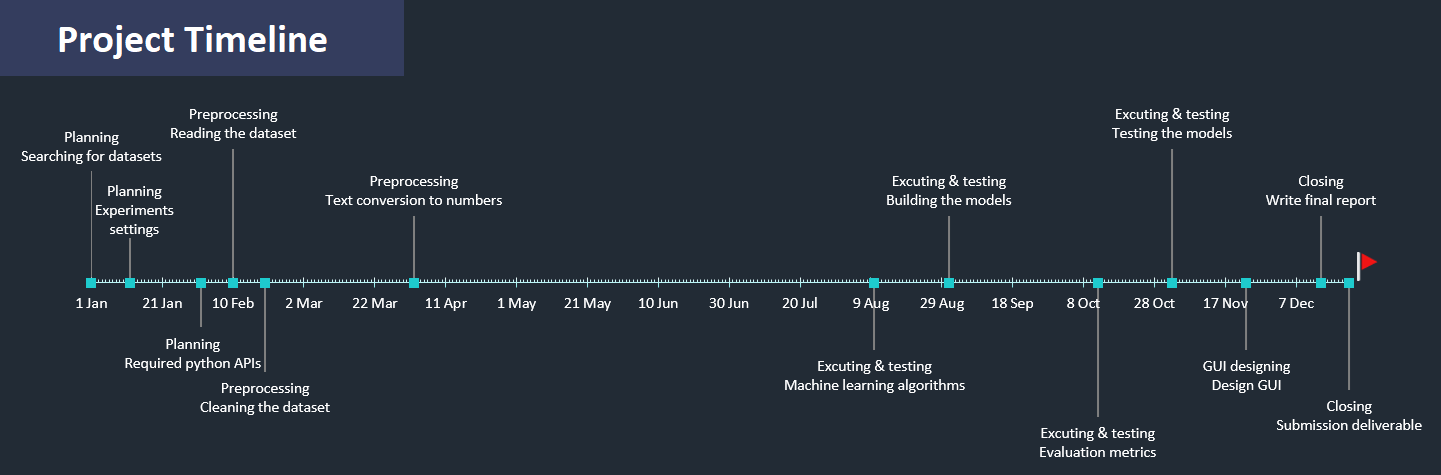


Figure : Project timeline

Table : Project timeline

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Phases | Phase No | TASK NAME | description | START  DATE | END  DATE | DURATION  in days |
| Planning | 1 | Searching for datasets | Searching for sentiment Arabic datasets. | 1-1-21 | 11-1-21 | 7 |
| Experiment’s settings | Set a plan for experiments. | 12-1-21 | 29-1-21 | 14 |
| Required python APIs. | Define a required python APIs. | 1-2-21 | 9-2-21 | 7 |
| preprocessing | 2 | Reading the dataset | Reading datasets to choose the best techniques to analyze it. | 10-2-21 | 18-2-21 | 7 |
| Cleaning the dataset | Clean the datasets and analyze it. | 19-2-21 | 1-4-21 | 30 |
| Text conversion to numbers | Define the algorithms to convert text to number and execute it. | 2-4-21 | 22-4-21 | 15 |
| Executing & testing | 3 | Machine learning algorithms | Determine the machine learning algorithms and determine the parameters. | 10-8-21 | 30-8-21 | 15 |
| Building the models | Write python code for models | 31-8-21 | 11-10-21 | 30 |
| Evaluation metrics | Evaluate the accuracy by the confusion metrics. | 12-10-21 | 1-11-21 | 15 |
| Testing the models | Test the model and fix bugs. | 2-11-21 | 22-11-21 | 15 |
| Designing GUI | 4 | Design GUI | Design a Graphical user interface and export the app. | 23-11-21 | 13-12-21 | 15 |
| Closing | 5 | Write final report | Write the final project report | 14-12-21 | 22-12-21 | 7 |
| Submission deliverable | Submit the app and the report | 22-12-21 | 22-12-21 | 0 |

# **Equipment or software will be used**

The equipment we will used:

* Aziz supercomputer.

Software we will used:

* Python programing language.
* NumPy API, for linear algebra operation.
* Scikit-learn API, for machine learning and conversion text to number algorithms.
* Pandas API, for data preprocessing.
* PyCharm IDE, for python programing language.
* IntelliJ IDE, for GUI and UX/UI designing.

Those software and equipment can change in the future.

**List of references:**

Table : List of references:0

|  |  |
| --- | --- |
| Type of material | Work cited |
| Research paper | Mohammad S. Hasan, Shahnoor C. Eshan, (2017) An application of Machine Learning to Detect Abusive Bengali Text, |
| Research paper | Anjuman Prabhat, Vikas Khullar, (2017) Sentiment classification on Big Data using Naïve Bayes and Logistic Regression. |
| Research paper | Sarah O. Alhumoud, Mawaheb I. Altuwaijri, Tarfa M. Albuhairi, Wejdan M. Alohaideb, (2015) Survey on Arabic Sentiment Analysis in Twitter. |
| Research paper | Doaa Mohey El-Din, (2016) Enhancement Bag-of-Words Model for Solving the Challenges of Sentiment Analysis |
| Research paper | Ibrahim Awajan, Mumtazimah Mohamad, (2018) A Review on Sentiment Analysis in Arabic Using Document Level |
| Article | [SUNIL RAY](https://www.analyticsvidhya.com/blog/author/sunil-ray/), 6 Easy Steps to Learn Naive Bayes Algorithm with codes in Python and R artical, 2017 |
| Article | Niklas Donges, A COMPLETE GUIDE TO THE RANDOM FOREST ALGORITHM, June 16, 2019 |
| Article | SUNIL RAY, Understanding Support Vector Machine (SVM) algorithm from examples (along with code), SEPTEMBER 13, 2017 |
| Article | Ayush Pant, Introduction to Logistic Regression, [Jan 22, 2019](https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148?source=post_page-----66248243c148--------------------------------) |
| Article | Jason Brownlee, A Gentle Introduction to the Bag-of-Words Model, October 9, 2017 |

**Conclusion**

In this paper, we reviewed the Arabic Sentiment Analysis importance, tasks, uses, and some problems related to this topic and the obtained solution to these problems and presented the data set used which consist of a number of examples for each class: Positive, Negative, and Neutral. And we discussed the (SA) approaches, the software, tools also libraries we are going to use also we dived more into the methods like data preparation which include (gathering data, discovering and assessing data, cleaning and validating data, transforming and enriching data, storing the data) and text conversion to numbers using Bag of words (BOW) which has a lot of methods and the chosen ones in this project which is (CountVectorizer, TF-IDF)Term Frequency-Inverse Document Frequency) ) and we went through different machine learning algorithms and the chosen ones in this project which is (naïve bayas, logistic regression, support vector machine, random forest) implementation and how it works, also some applications example related to it. In addition experiments we will conduct on 1-gram,2-gram,3-gram for each method using the four chosen machine learning algorithms in this project. Finally, we come to a flow chart of the project which has three phases (data preprocessing, training & testing, GUI app) and reviewed the timeline of the project, and show a list of references related to the Arabic Sentiment Analysis topic. In the end, we hope that this paper can provide researchers with a comprehensive overview of Arabic sentiment analysis in terms of resources, approaches, and open challenges to further advance the state of the art in this domain.

**Declaration of Originality**

We hereby declare that this project report is based on our original work except for citations and quotations, which had been duly acknowledged. We also declare that it has not been previously and concurrently submitted for any other degree or award at KAU or other institutions.

|  |  |
| --- | --- |
| Name | Student ID |
| Faris Adel Alahmadi | 1845562 |
| Sultan Fadel Fatani | 1945396 |
| Ramzi Yousef Al Mazroe | 1743719 |
| Faisal Majed Aljuhani | 1846023 |
| Date:4/9/2021 | |