

King Abdelaziz University

Faculty of Computing & Information Technology

Computer Science Department

CPCS-498 – (Senior project 1)

**Arabic Sentiment Analysis**

**Title and aim(s) for the project:**

Our project seeks to enhance Arabic Sentiment Analysis progressing only a few studies have addressed Sentiment Analysis in the Arabic language 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.

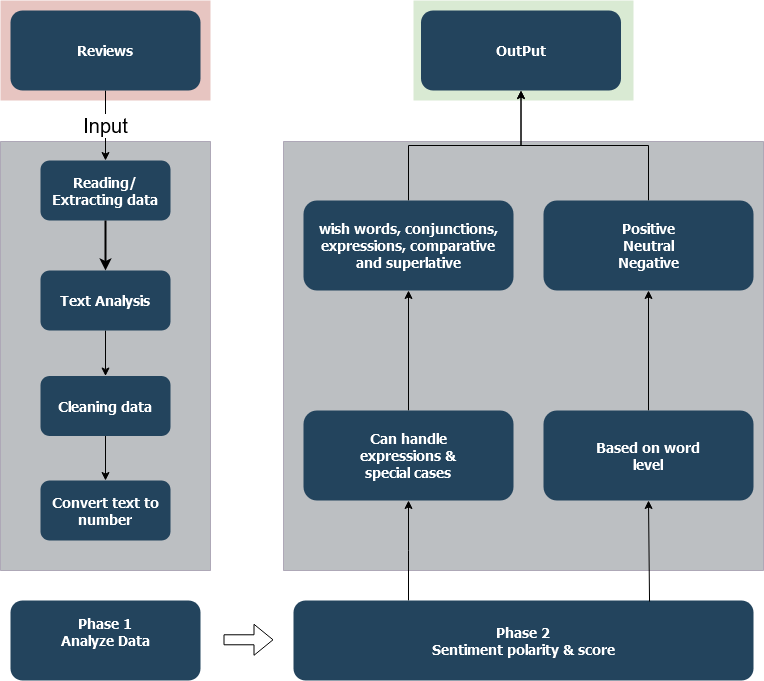
The users of the internet are not just consumers of web content, but also producers of it. Their contributions have become very important to enrich the web content. These users express their opinion and share information on different fields and topics. Due to the increase of the opinions, reviews, feedbacks, 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 definition:**

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.

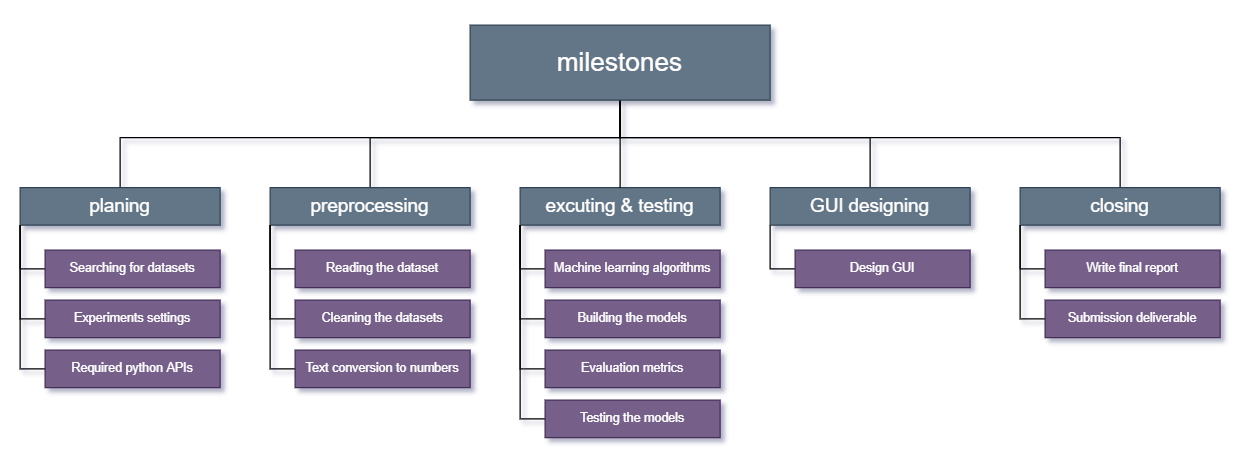


Sentiment analysis is usually solved using one of those approaches. Supervised approach, unsupervised approach and hybrid approach. The approach used in this project is the supervised approach based on Support Vector Machine (SVM), Logistic regression, Random forest and Naïve bayas. The sentiment analysis architecture is presented here which illustrated in the figure. 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 count vectorizer. 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.

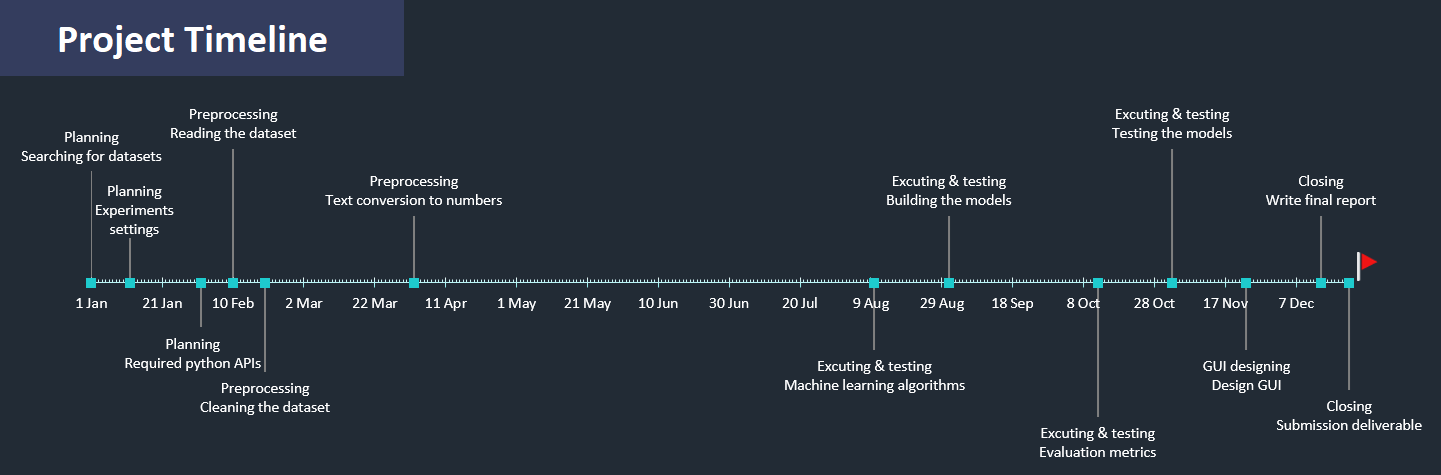
**Schedule for the project:**

Milestones diagram

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.



**Project timeline:**



|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Phases | Phase No | TASK NAME | description | START  DATE | END  DATE | DURATION  in days | STATUS |
| Planning | 1 | Searching for datasets | Searching for sentiment Arabic datasets. | 1-1-21 | 11-1-21 | 7 | Done |
| Experiments settings | Set a plan for experiments. | 12-1-21 | 29-1-21 | 14 | Done |
| Required python APIs. | Define a required python APIs. | 1-2-21 | 9-2-21 | 7 | Done |
| preprocessing | 2 | Reading the dataset | Reading datasets to choose the best techniques to analyze it. | 10-2-21 | 18-2-21 | 7 | Done |
| 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 |  |

**Literature search and bibliography:**

We have prepared for our project by searching in google scholar and watch YouTube videos and some courses that we have taken in FCIT.

List of references:

|  |  |
| --- | --- |
| Type of material | Work cited |
| Research paper | **[ Mohammad S. Hasan, Shahnoor C. Eshan]**, An application of Machine Learning to Detect Abusive Bengali Text. |
| Research paper | **[Anjuman Prabhat, Vikas Khullar]**, 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],** Survey on Arabic Sentiment Analysis in Twitter**.** |
| Research paper | **[Doaa Mohey El-Din]**, Enhancement Bag-of-Words Model for Solving the Challenges of Sentiment Analysis |
| Research paper | **[Ibrahim Awajan, Mumtazimah Mohamad ],** A Review on Sentiment Analysis in Arabic Using Document Level |

[hint] these references can change or add more references to it.

**Summarize papers and why we chose them:**

**[ Mohammad S. Hasan, Shahnoor C. Eshan]:**

An application of Machine Learning to Detect Abusive Bengali Text, the problem is detecting Abusive Bengali Text by using binary classification supervised learning, in this paper the researchers use SVM, C4.5, naïve bayas algorithm for training and testing phase, and use unigram, bigram and trigram to convert text to numbers.

We chose this paper because its problem simi like our problem and its use all algorithms we intend to use except logistic regression.

**[Anjuman Prabhat, Vikas Khullar]:**

Sentiment classification on Big Data using Naïve Bayes and Logistic Regression, the problem is classifying the text into positive, negative, or natural class, in this paper the researchers use naïve bayas and logistic regression algorithms for training and testing phase and use *countvectorizer*, *fidf* to covert text to numbers.

We chose this paper because our problem is sentiment classification problem and using logistic regression algorithm.

**[Sarah O. Alhumoud, Mawaheb I. Altuwaijri, Tarfa M. Albuhairi, Wejdan M. Alohaideb]:**

Survey on Arabic Sentiment Analysis in Twitter, in this paper the researchers do some data preprocessing techniques like filtering, misspelling, repeated letters and normalization on Arabic tweets from twitter.com.

We choose this research paper because the researchers do some data cleaning techniques that we need for our Arabic datasets and some filtering methods.

**[Doaa Mohey El-Din]:**

Enhancement Bag-of-Words Model for Solving the Challenges of Sentiment Analysis, in this research paper the researcher gives a lot of challenges and some techniques to enhancement the bag of word method (bag of word method is and conversion text to number method).

We choose this paper because our conversion text to number algorithms is bag of words algorithms and we will try to enhance it as possible to increase accuracy.

**[Ibrahim Awajan, Mumtazimah Mohamad]:**

A Review on Sentiment Analysis in Arabic Using Document Level, in this paper the researchers explain briefly the machine learning path in the document level and some data preprocessing methods, gives some statistical information about Arabic language and Arabic internet users.

We choose this research paper because the researcher explain the ML path better then other research paper and it was simple and easy to read and understandable.

**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.

**What do we intend to carry out in the future?**

We intend to explore deeper and more advanced approaches to enhance the quality of text analysis regarding the sentiment analysis topic.

State of the art approaches such as neural networks are the base approaches for lots of AI systems. Many architectures that based on neural networks concepts have been applied in NLP, which have led to huge improvements in almost all NLP specific tasks. Our directions will be to go in line with those approaches by exploring, applying and investigating possible directions of improvements.