|  |  |
| --- | --- |
| **Faculty of Computer Science and Information Technology** |  |

****

**Twitter Sentiment Analysis**

**A Report Submitted in Partial Fulfillment of the Requirements for the Degree**

**of Bachelor of Computer Science and Information Technology**

***Prepared by***

Abdel Rahman Mahmoud Rashad (41510157)  
 Abdel Rahman Sobhy Saleh (41510144)  
 Abdel Rahman Tarek Mahmoud (41510089)  
 Bishoy Alper Aziz (41510137)  
 Khaled Mohamed Abdel Hamid (41510123)  
 Yasser El-Sayed Mohamed (41510072)

***Supervised by***

Dr. Diaa El-Din Mohamed Fayed

Cairo, 2018/1019

**Acknowledgement**

It is a matter of great pleasure to present this progress report on development of the graduation project “Twitter Sentiment Analysis”. We are grateful to College of Computer Science for providing us with this great opportunity to develop a web application for this project.

We are also grateful to our project supervisor **Dr. Diaa El-Din Fayed** for his valuable advices and suggestions.

In addition, we would like to convey our thanks to the teaching and non-teaching staffs of the Department of Computer Science, for their invaluable help and support throughout the period of the project hours. We will not miss to express our gratitude to all our friends and everyone who has been the part of this project by providing their comments and suggestions.

**ABSTRACT**  
  
  
Analysis of public information from social media could yield interesting results and insights into the world of public opinions about almost any product, service or personality. Social network data is one of the most effective and accurate indicators of public sentiment. The explosion of Web 2.0 has led to increased activity in Podcasting, Blogging, Tagging, Contributing to RSS, Social Bookmarking, and Social Networking. As a result, there has been an eruption of interest in people to mine these vast resources of data for opinions. Sentiment Analysis or Opinion Mining is the computational treatment of opinions, sentiments and subjectivity of text.

In project, we will develop a program for sentiment analysis is an approach to be used to

computationally measure customers' perceptions. This project reports on the design of a sentiment analysis, extracting and training a vast amount of tweets. Results classify customers' perspective via tweets into positive and negative, which is represented in fully web page.

**Table of contents**

LIST OF FIGURE…………………………………………………………………... v

List of abbreviations and acronyms…………………………………………………. vi

**Chapter 1: Introduction**

1.1 Overview.…………………………………………………...…………………... 1

1.2 Motivation……………………………………………………………………….. 1

1.3 Aims and Objectives……...……………………………………………………... 1

1.4 Scope……………...……………………………………………………………... 2

1.5 General constraints……...…………………………………………………...…... 2

**Chapter 2: Background and Previous work**

2.1 Background……………………...……………………………………………..... 3

2.1.1 Data Mining…….…………………………………………………….... 3  
2.1.1 Natural Language Processing.………………………………………….. 3

2.2.1 Machine Learning Algorithms….………………………………….…... 4

2.2.2 Naive Bayes……….……………………………………….…… 4  
2.2.2 Support Vector Machine.……………………………………….. 7

2.2.2 Sentiwordnet…………………………………………………….. 8

2.3 Previous work……………………………………………………….……………. 10

**Chapter 3: Planning and Analysis**

3.1 Planning………………………………………………………………..……..……11

3.1.1 Feasibility study and estimated cost……………………………...……...11

3.1.2 Gantt chart and Communication Plan……………………………………12

3.2 Analysis and limitations of existing system…………………………………..…...13

3.3 Need for new system……………………………………………….…………..….13

3.4 Analysis of new system…………………………………….…………..………….14

3.4.1 User Requirements……………………………………………………....14

3.4.2 System Requirements………………………..…………..……………...14

3.4.3 Domain Requirements…………………………..………….………….. 14

3.4.4 Functional Requirements & Non- Functional Requirement……….…... 14

3.5 Advantages of new system………………………………………………………...16

3.6 User characteristics………………………………………………………………...16

**Chapter 4: Design**

4.1 Design and Implementation Constraints…………………………….………….... 18

4.2 Risks and risk management………………………………………………………. 18

4.3 Design of database ERD……………………………………………………….…. 19

4.3.1 Entity Relationship Diagram…………………………………………..... 19

4.3.2 Mapping of Entity Relationship Diagram………………………………. 19

4.4 Class diagram……………………………………………………………………... 20

4.5 Use case diagram………………………………………………………………….. 21

4.5.1 Primary Use-case Diagram……………………………………….……... 21

4.5.2 Use-case Scenarios…………………………………………………...…. 21

4.6 Activity diagram…………………………………………………………………... 22

4.7 Sequence diagram…………………………………………………………………. 23

4.8 State diagram………………………………………………………………………. 24

4.9 Software Architecture……………………………………………………………... 25

4.10 User Interface………………………………………………………………….…. 26

LIST OF FIGURES

**Figure 2.1 Natural Language Processing……………………..….…………...…. 3**

**Figure 4.1 Entity Relationship Diagram……………………..….…………...…. 19**

**Figure 4.4 Class diagram……………………………………………………...…. 20**

**Figure 4.5** **Use Case Diagram……………………………………………………... 21**

**Figure 4.6 Activity Diagram……………………………………………………... 22**

**Figure 4.7 Sequence Diagram…………………………………………………… 23**

**Figure 4.8 State Diagram…………………………………………………........... 24**

**Figure 4.9 System architecture………………………………………………….. 25**

**Figure 4.10.1 Interface (1)……………………………………………………….. 26**

**Figure 4.10.2 Interface (2)……………………………………………………….. 27**

**Figure 4.10.3 Interface (3)……………………………………………………….. 28**

**Figure 4.10.4 Interface (4)……………………………………………………….. 29**

**Figure 4.10.5 Interface (5)……………………………………………………….. 30**

LIST OF ABBREVIATIONS AND ACRONYMS

**API:** Application Programming Interface.

**NLP:** Natural Language Processing.

**NLTK:** Natural Language Toolkit. **SVM:** Support Vector Machine

1. **INTRODUCTION**

**1.1 Overview**   
This proposal entitled “Twitter Sentiment Analysis” is a web application, which is used to analyze the tweets. We will be performing sentiment analysis in tweets and determine where it is positive, negative or neutral.  
This web application can be used by any organization office to review their works or by political leaders or by any others company to review about their products or brands. [1]

**1.2 Motivation**  
The emergence in the last decade of social media platforms such as Twitter, Facebook, and Instagram, enabled people to engage in social activities to express their opinions, thoughts, and emotions on a variety of topics. On such platforms, large amounts of data are produced  
(e.g.: 6000 tweets per second), this representing an opportunity for companies to assess their social influence and people opinions towards their products.   
Consequently, a computational framework is desirable to perform opinion mining and sentiment analysis which can adapt to the activity domain of the user. [1]

|  |  |  |
| --- | --- | --- |
| **No** | **Objective** | **Priority** |
| 1 | Build two infrastructures: real time sentiment analysis and long-term sentiment analysis, employing an engine capable to adapt to both infrastructures | Highest |
| 2 | Understand and implement natural language processing techniques | Highest |
| 3 | Implement a machine-learning algorithm to perform sentiment analysis | High |
| 4 | Achieve 80% or more in classification accuracy | High |
| 5 | Build a web application graphical user interface for visualization purposes | Medium |

**1.3 Aims and Objectives**   
The project aims to produce real time sentiment analysis associated with a range of brands, products and topics. The project’s scope is not only to have static sentiment analysis for past data, but also sentiment classification and reporting in real time.   
As such, the system should automatically collect and analyze data from Twitter, the primary data source for this project. For example, the sentence "Brand A is awesome" has positive sentiment for Brand A. More sophisticated structures can be built,  
for example, the sentence "Brand A is okay but Brand B is great" has neutral sentiment for Brand A, and positive sentiment for Brand B. By the end of the project the goal is to produce up to the minute sentiment values for brands and topics. As such, a system which determines the polarity of tweets (Twitter messages) by using machine learning algorithms and natural language processing techniques is proposed, Table 1.1 presents the objectives set, prioritized by the level of contribution to the scope of the project. [3]  
  
 Table 1.1: Objectives set for the project

**1.4 Scope**   
For this project, only data regarding Twitter tweet that are in Arabic and English languages.  
It will be helpful to the companies, political parties as well as to the common people. It will be helpful for political party for reviewing about the program that they are going to do or the program that they have performed.

Similarly, companies also can get review about their new product on newly released hardware or software.

Also the movie maker can take review on the currently running movie By analyzing the “tweets” analyzer can get result on how positive or negative or neutral are peoples about it.

**1.5 General constraints**  
As we are planning to launch a prototype for testing purposes then a beta version for more advanced validation process then launching the final version.

1. **BACKGROUND AND PREVIOUS WORK**

**2.1 Background**

**2.1.1 Data Mining**

Data mining is the computational process of finding patterns in large datasets and its methods are at the intersection between Artificial Intelligence, Machine Learning, computer science, data base technologies, and statistics. The objective of data mining is to extract information or knowledge from a dataset and transform it into a structure that can be understood.

Data preparation is an important part of any data analysis. To properly prepare data it is necessary to understand the application domain, this is important as the researcher must be able to identify pertinent data and cleansing the dataset removing any data which is deemed as unimportant to the analysis.

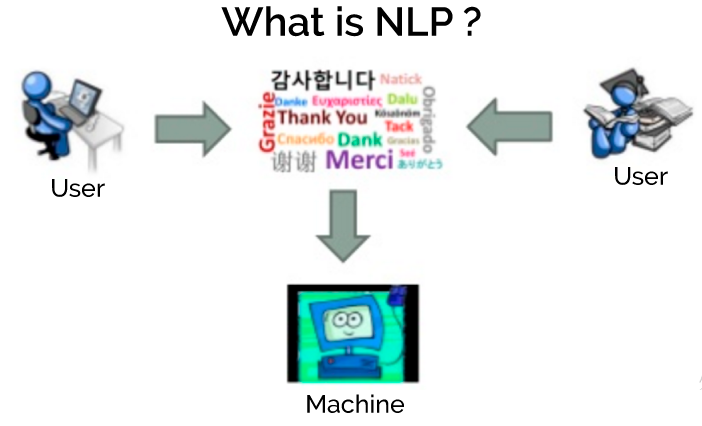
Some of these pre-processing techniques include; Data cleaning, Noise treatment, sampling, strategy’s to deal with missing values, Normalization, and feature extraction. Many of these pre-processing techniques will be examined in more detail in the implementation section of this report.

Data mining focuses on discovering patterns in data. Sentiment analysis which is also known as opinion mining focuses on discovering patterns in text that can be analyzed to classify the sentiment in that text. [6]

**2.1.1 Natural Language Processing (NLP)**

Natural Language Processing (NLP) is ability of machines to understand and interpret human language the way it is written or spoken. [5]

# **What Natural Language Processing (NLP) is using for?**

Natural language processing (NLP) has used for tasks such as sentiment analysis, topic detection, language detection, key phrase extraction. The objective of NIP is to make computer/machines as intelligent as human beings in understanding language do. [4]

**- Figure (2.1) – [5]**

**2.2.1 Machine Learning Algorithms**

Machine learning is a branch of artificial intelligence which focuses on building models that have the ability to learn from data. As it is such an enormous field that encompasses many areas there is no standardized definition for it, but Arthur Samuels’s general definition which describes it well:

In general, there are two types of machine learning algorithms supervised, and unsupervised, there are variations on these algorithms such hybrid types (semi-supervised learning) but for this report they will be classified into one of the two categories.

The method of supervised learning consists of presenting an algorithm with a training dataset; this dataset consists of training examples and the corresponding expected output for each example. The expected output is general known as the target. A supervised learning algorithm uses this dataset so that it can learn to map the input examples to their expected target. If the training process is implemented correctly, the machine-learning algorithm should be able to generalize the training data so that it can correctly map new data that it has never seen before.

Unsupervised machine learning algorithms do not require training data, they operate on data where the output is unknown. The object of this form of learning is usually to discover patterns in the data that may not be known by the researcher. An example of an unsupervised method would be clustering where the algorithm uses a distance function to group similar data points together.

This project focuses exclusively on supervised learning algorithms for the task of text classification, however there are two major applications for supervised methods

**Classification:**

The target output is a class or label, the simplest case is a choice between zero ore one, although there can also be multiple alternative classes. Classification is used for many applications such as test classification, object recognition, and voice recognition software.

**Regression**:

In this case the target is a real number or vector of real numbers. Regression is mostly used for prediction. Supervised regression algorithms are used mostly for prediction. Example applications are stock market prediction, in power systems analysis it can be to predict spikes in a network, and most recently they have been used in memory caching to complement the ‘locality of reference’ method.

This project focus on supervised classification algorithm the models that were used are described below. [7]  
  
2.2.2 **Naive Bayes Algorithm**

What is the naive Bayes algorithm?

It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a NaiveBayesclassifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

Naive Bayes is a classification algorithm for binary (two-class) and multi-class classification problems. The technique is easiest to understand when described using binary or categorical input values.

It is called naive Bayes or idiot Bayes because the calculation of the probabilities for each hypothesis are simplified to make their calculation tractable. Rather than attempting to calculate the values of each attribute value P(d1, d2, d3|h), they are assumed to be conditionally independent given the target value and calculated as P(d1|h) \* P(d2|H) and so on.

This is a very strong assumption that is most unlikely in real data, i.e. that the attributes do not interact. Nevertheless, the approach performs surprisingly well on data where this assumption does not hold.

**What is naive Bayes used for?**

It has been successfully used for many purposes, but it works particularly well with natural language processing (NLP) problems. Naive Bayes is a family of probabilistic algorithms that take advantage of probability theory and Bayes' Theorem to predict the tag of a text.

What is naive Bayes in machine learning?

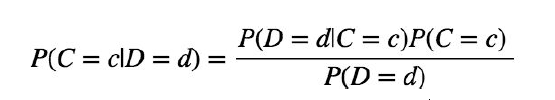
In machine learning, naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive)independence assumptions between the features. Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem.

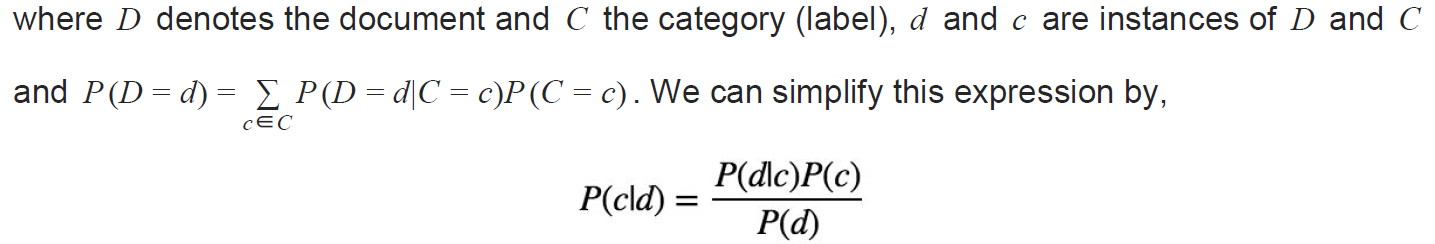
Maximum likelihood training can be done by evaluating a closed form expression (mathematical

expression that can be evaluated in a finite number of operations), which takes linear time.

It is based on the application of the Baye’s rule given by the following formula:

*Formula: Baye’s rule*



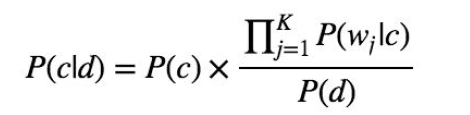


In our case, a tweet *d* is represented by a vector of *K* attributes such as *d* = (*w*1, *w*2, ..., *wk*) .

Computing *P*(*d*|*c*) is not trivial and that's why the Naive Bayes introduces the assumption that

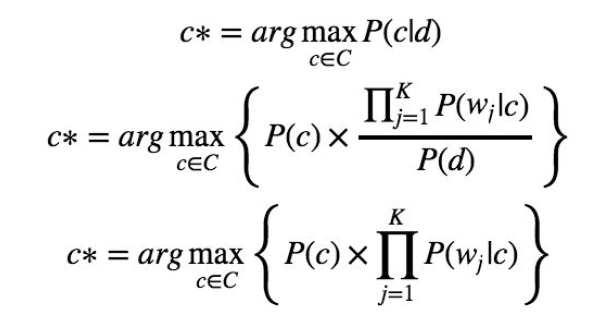
all of the feature values *wj* are independent given the category label *c* . That is, for *i* ≠ *j* , *wi* and

*wj* are conditionally independent given the category label *c*. So the Baye's rule can be rewritten



Based on this equation, maximum a posterior (MAP) classifier can be constructing by seeking

the optimal category which maximizes the posterior P(c|d) :



What is naive Bayes in data mining?

Naive Bayes classification. The Naive Bayes classification algorithm is a probabilistic classifier. It is based on probability models that incorporate strong independence assumptions. ... Datamining in Info Sphere Warehouse is based on the maximum likelihood for parameter estimation for Naive Bayes models.

Why is naive Bayes so naive?

In simple terms, a naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature, given the class variable. ... Basically, it's "naive" because it makes assumptions that may or may not turn out to be correct

**Advantages**

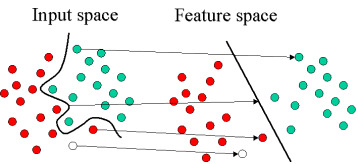
Advantages of some particular algorithms. Advantages of Naive Bayes: Super simple, you're just doing a bunch of counts. If the NB conditional independence assumption actually holds, a Naive Bayes classifier will converge quicker than discriminative models like logistic regression, so you need less training data. [8]

**{\displaystyle p({\text{height}}\mid {\text{male}})={\frac {1}{\sqrt {2\pi \sigma ^{2}}}}\exp \left({\frac {-(6-\mu )^{2}}{2\sigma ^{2}}}\right)\approx 1.5789}{\displaystyle p({\text{weight}}\mid {\text{male}})=5.9881\cdot 10^{-6}}2.2.2 Support Vector Machine**

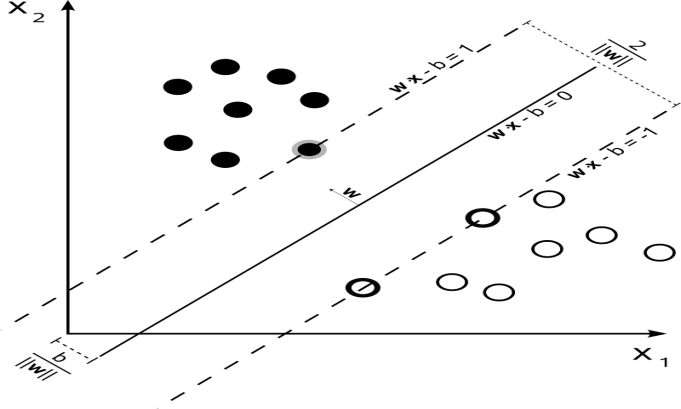
Support Vector Machine is another example of supervised algorithm. While Naïve Bayes uses a probabilistic model and it is becoming an increasingly common method for text classification, Its increased popularity is largely due to the high classification accuracy that is associated with its use.

The support vector machine is classed as a non-probabilistic binary linear classifier. It works by plotting the training data in multidimensional space, it then tries to separate the classes with a hyperplane. If the classes are not immediately linearly separable in the multidimensional space the algorithm will add a new dimension in an attempt to further separate the classes. It will continue this process until it is able to separate the training data into its two separate classes using a hyperplane.

A basic representation of how it splits the data is shown in figure below.



One of the main areas where this method differs from other linear classifiers such as the perceptron is in the way it selects the hyperplane. In most cases there may be multiple hyperplanes or in some cases an infinite number of hyperplanes that could separate that classes. The SVM algorithm chooses the hyperplane which provides the maximum separation between the classes has the greatest margin or the maximal margin hyperlane which minimizes the upper bound of the classification errors. A standard method for finding the optimum way of separating the classes is to plot two hyperplanes in a way that there are no data points between them, and then by using these planes the final hyperplane can be calculated. This process is shown in figure below. The data points that fall on these planes are known as the supports. [9]



Now that the algorithm has calculated the hyperplane that provides the maximum level of separation between the classes, new data can be classified. New instances are mapped into the feature space and are classified by which side of the hyperplane they fall onto. [10]

**{\displaystyle p({\text{height}}\mid {\text{male}})={\frac {1}{\sqrt {2\pi \sigma ^{2}}}}\exp \left({\frac {-(6-\mu )^{2}}{2\sigma ^{2}}}\right)\approx 1.5789}{\displaystyle p({\text{weight}}\mid {\text{male}})=5.9881\cdot 10^{-6}}2.2.2 SentiWordNet**

**The purpose of SentiWordNet: -**

The aim of SentiWordNet is to provide an extension for WordNet, such that all synsets can be associated with a value concerning the negative, positive or objective connotation. SentiWordNet 3.0 is the improved version of SentiWordNet 1.0 and publicly freely available for research purpose with a webinterface.4 This extension labels each synset with a value for each category between 0.0 and 1.0. The sum of the three values is always 1.0, so each synset can have a nonzero value for each sentiment,

because some synsets can be positive, negative or objective depending on the context in which they are used. The web interface allows the user to search for any synset belonging to WordNet with its associated SentiWordNet scores. Additionally, the user is able to see a visualisation of those scores. Each category is linked to a color, which is red for negativity, blue for objectivity

and green for positivity. The visualization of the synset good#5 can be seen in figure 1.

The advantage of using synsets instead of terms is to offer different sentiment scores for eachsense of one word, because the connotations can differ in one word depending on the sense.

**The methods of building SentiWordNet**

* **Creation and purpose of WordNet**

The idea behind WordNet is to create a “dictionary of meaning” integrating the functions of dictionaries and thesauruses. Lexical information is not organized in word forms, but in word meanings which is consistent

with the human representations of meaning and their processing in the brain. Besides creating a innovatively organized lexical semantic resource, the researchers aim

furthermore, to support and promote automatic text analysis for applications in the field of artificial intelligence.

**Structure and contents**

WordNet contains English nouns, verbs, adjectives and adverbs. They form so called “synsets”, i.e. sets of distinctive cognitive synonyms, which glosses, i.e.

descriptions of the synsets with sample expressions or sentences, are attached to. What constitutes the “net “like

structure of WordNet are the links between the synsets. Synsets that have a certain lexical or conceptual relation are linked.

● Nouns can be connected through hyponymy/hyponymy and metonymy/homonymy

relations which can also be inherited. They form a hierarchy which all goes back up to one

root. There is also a differentiation between types (common nouns) and instances(persons, entities).

● Verbs are organized via troponin, hyponym and entailment relations.

● Adjectives are linked to their antonyms, and relational adjectives point to their related.

**Explanation**

The system was built in two main steps. In

the **semi supervised**

**learning step** 8 classifiers were established to decide for each synset

belonging to WordNet if it is negative, positive or objective.

This provides on the one hand a higher generalization factor and a low risk of over fitting, on the

other hand the different classification results help to give the synsets a tendency of being more

positive or negative, rather than just one opportunity: namely that a synset can positive, negative

and objective to a certain extend.

In the second step, the **random walk**

**step** the scores for the positive and negative scores are

due to the “defiensdefiendum” relationship.Through averaging of all the classification results a value between 0.0 and 1.0 can be obtained

for each category for each synset. If all classifiers will decide on the same category, this

sentiment will have the maximum value, which is 1.0.

**2.3 Previous work:**

**2.3.1 Expedia Canada**

Around Christmastime, Expedia Canada ran a classic “escape winter” marketing campaign. All was well, except for their choice of screeching violin as background music. Understandably, people took to social media, blogs, and forums. Expedia noticed that and removed the ad. Then, they created a series of follow-up spin-off videos: one showed the original actor smashing the violin, and in another one, they invited a real follower who had complained on Twitter to come in and rip the violin away. Though their original product was far from flawless, they were able to redeem themselves by incorporating real customer feedback into continued iterations [1].

**2.3.2 City Voices project**

In Brazil, federal public spending rose by 156% from 2007 to 2015 while people’s satisfaction with public services steadily decreased. Unhappy with this counterproductive progress, the Urban-planning Department recruited McKinsey to help them work on a series of new projects that would focus first on user experience, or citizen journeys, when delivering services. This citizen-centric style of governance has led to the rise of what we call Smart Cities.

McKinsey developed a tool called City Voices, which conducts citizen (customer) surveys across more than 150 different metrics, and then runs sentiment analysis to help leaders understand how constituents live and what they need, in order to better inform public policy. By using this tool, the Brazilian government was able to surface urgent needs –a safer bus system, for instance– and improve them first.

If even whole cities and countries, famous for their red tape and slow pace, are incorporating customer journeys and sentiment analysis into their decision making processes, then innovative companies better be far ahead.

**2.3.3 Analyzing customer support interaction on Twitter**

programmers makes some analysis on how the four biggest US phone carriers (AT&T, Verizon, Sprint, and T-Mobile) handled customer support interactions on Twitter. they downloaded tens of thousands of tweets mentioning the companies (by name or by handle), and ran them through a Monkey Learn sentiment model to categorize each tweet as positive, neutral, or negative. We then used our new Insight Extractor, which reads all text as one unit, extracts the most relevant keywords, and returns the most relevant sentences including each keyword.

**2.3.4 Hotel reviews on TripAdvisior**

how people feel about hotels in several major cities around the world, so some programmers scraped and analyzed more than one million reviews from TripAdvisor. they looked at hotels in London, Paris, New York, Bangkok, Madrid, Beijing, and Rio de Janeiro. [12]

**3- PLANNING AND ANALYSIS**

**3.1 Planning**

**3.1.1 Feasibility study and estimated cost**

A feasibility study is an analysis used in measuring the ability and likelihood to complete a project successfully including all relevant factors. It must account for factors that affect it such as economic, technological, legal and scheduling factors.

Project managers use feasibility studies to determine potential positive and negative outcomes of a project before investing a considerable amount of time and money into it.

A feasibility study tests the viability of an idea, a project or even a new business.

**Importance of Feasibility Studies**

Feasibility studies allow companies to determine and organize all the details to make a business work. A feasibility study helps identify logistical problems, and nearly all business-related problems and their solutions. Components of a Feasibility Study includes: Technical feasibility, Economic feasibility, Operational feasibility, Schedule feasibility

### **Technical feasibility**

Evaluating the technical feasibility is the trickiest part of a feasibility study. This because, at the point in time there is no any detailed designed of the system, making, it difficult to access issues like performance, costs (on account of the kind of technology to be deployed) etc. A number of issues have to be considered while doing a technical analysis; understand the different technologies involved in the proposed system. Before commencing the project, we have to be very clear about

what are the technologies that are to be required for the development of the new

### system. Is the required technology available?

### Our project "Twitter Sentiment Analysis" is technically feasible since all the required tools are easily available. Python and Php with JavaScript can be easily handled. Although all tools seems to be easily available.

### The cost of the project is high because we bought an expensive laptop to work on it to decrease the time that we need to finish the project and increase the speed and performance of any program or ide that we using it in our project and to run programs fast.

### The issues that we faced in our project are:

### We don’t know what the idea of the sentiment analysis.

### We don’t have enough information about the data analysis especially sentiment analysis.

### We take a lot of time to bring API of Twitter.

**Operational feasibility**

Proposed project is beneficial only if it can be turned into information systems that will meet the operating requirements. Simply stated, this test of feasibility asks if the system will work when it is developed and installed. Are there major barriers to Implementation?

### The proposed was to make a simplified web application. It is simpler to operate and can be used in any webpages. It is free and not costly to operate.

### **Economic feasibility**

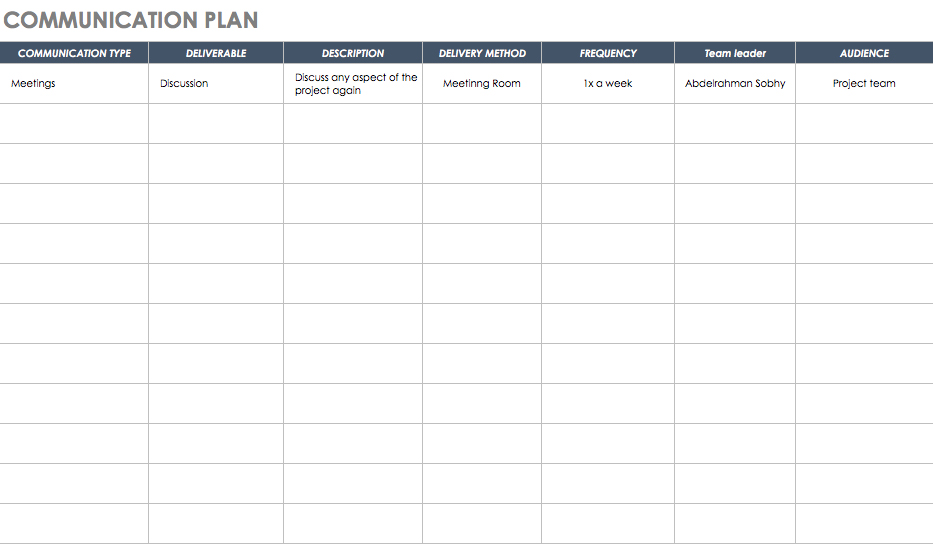
Economic feasibility attempts to weigh the costs of developing and implementing a new system, against the benefits that would accrue from having the new system in place. This feasibility study gives the top management the economic justification for the new system. A simple economic analysis, which gives the actual comparison of costs and benefits, are much more meaningful in this case. In addition, this proves to be useful point of reference to compare actual costs as the project progresses. There could be various types of intangible benefits on account of automation. These could increase improvement in product quality, better decision making, and timeliness of information, expediting activities, improved accuracy of operations, better documentation and record keeping, faster retrieval of information. This is a web-based application. Creation of application is not costly.

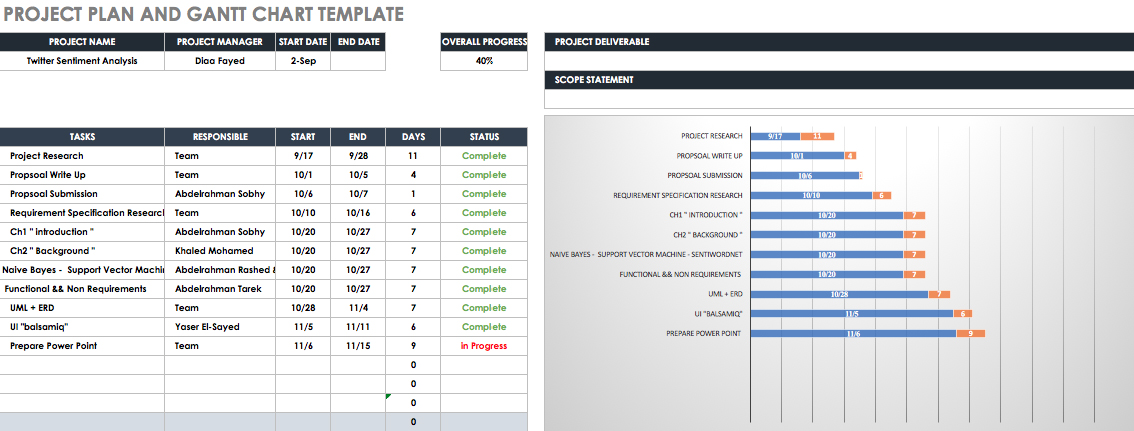
### **Schedule feasibility**

A project will fail if it takes too long to be completed before it is useful. Typically, this means estimating how long the system will take to develop, and if it can be completed in a given period of time.

Schedule feasibility is a measure how reasonable the project timetable is. Given our technical expertise, are the project deadlines reasonable? Some project is initiated with specific deadlines.

### It is necessary to determine whether the deadlines are mandatory or desirable.

**3.1.2 Gantt Chart && Communication Plan:**



**3.2 Analysis and limitations of existing system**

We collected dataset containing from Twitter which are positive and negative and neutral data. Those datasets were trained data and was classified using Naïve Bayes and SVM Classifier. Before training the classifier unnecessary words, punctuations, meaning less words were cleaned to get

pure data. To determine positivity and negativity of tweets we collected data using

twitter API. Those data were stored in database and then retrieved back to remove

those unnecessary word and punctuations for pure data. To check polarity of test

tweet we train the classifier with the help of trained data. Those results were stored

in database and then retrieved back using PHP, HTML, JavaScript and CSS.

**3.3 Need for new system**

The main feature of our web application is that it helps to determine the opinion

about the peoples on products, government work, politics or any other by analyzing

the tweets. Our system is capable of training the new tweets taking reference to

previously trained tweets.

The computed or analyzed data will be represented in various diagram such as Pie chart,

Bar graph and Scatter Plot

Most importantly, you didn't get enough the analysis of the feelings of the ratio of true.  
But you will now get 99%

**3.4 Analysis of new system**

**3.4.1 User Requirements:**

1- A place to search to do a search for what he/she wants and place to put in it the comment

or file to make sentiment analysis on it.

2- The search is based on a large number of tweets so that the result is closer to correct.

3- The result appears clearly and easily.

**3.4.2 System Requirements:**

1- make a user interface have a search place to search.

2- make a user interface to have a place to put a file or comment on it.

3- the search place have connected to twitter API.

4- we take a tweet or file or comment and make a classification on it and know if it positive, natural or negative if it positive we collect it with positives tweets and if it negative collect it with negatives tweets and if it natural collect it to the naturals groups.

5- finally we will display the results on the graph or animations.

**3.4.3 Domain Requirements:**

we have to make search or put comment or file to make sentiment analysis in English language if not in English we have to display message to user to know him.

can user search on words are not a concept or unknown words.

**3.4.4 Functional Requirements & Non- Functional Requirement**

**Requirement Definition:**

After the extensive analysis of the problems in the system, we are familiarized with the requirement that the current system needs. The requirement that the system needs is categorized into the functional and non-functional requirements. These requirements are listed below:

**1-Functional Requirements:**

Functional requirement are the functions or features that must be included in any system to satisfy the business needs and be acceptable to the users. Based on this, the functional requirements that the system must require are as follows: • System should be able to process new tweets stored in database after retrieval • System should be able to analyze data and classify each tweet polarity

**2- Non-Functional Requirements:**

Non-functional requirements is a description of features, characteristics and attribute of the system as well as any constraints that may limit the boundaries of the proposed system. The non-functional requirements are essentially based on the performance, information, economy, control and security efficiency and services. Based on these the non-functional requirements are as follows:

• User friendly

• To perform with efficient throughput and response time

**Functional Requirements:**

A set of functional requirements was developed in the GUI as a Table 1:

|  |  |  |
| --- | --- | --- |
| # | ​Functional Requirements | Priority |
| 1 | Include textual input field on which sentiment analysis is performed | Very High |
| 2 | Output stream chart in real time | Very High |
| 3 | Output and store the model after the training phase of the classification algorithm | Very High |
| 4 | Test the classification algorithm proposed | Very High |
| 5 | Clean the noise from tweets retrieved and address orthography issues | High |
| 6 | Store the acquired data depending on the component in which the engine is used | High |

**Non-Functional Requirements:**

The non-functional requirements for the graphical user interface were set following three principles: responsiveness, readability, and usability.

non-functional requirements developed for the graphical user interface.

|  |  |  |
| --- | --- | --- |
| **#** | **Type** | **Non-functional requirements** |
| 1 | Responsiveness | The graphical user interface should have a quick response time for user interactions |
| 2 | Usability | The graphical user interface should update in real time the aggregate sentiment score of the Tweets analysis |
| 3 | Readability | The graphical user interface should present an intuitive design |

|  |  |  |
| --- | --- | --- |
| # | Non-functional requirements | Priority |
| 1 | Efficiency - the engine should be able to support large intakes of data without affecting its performance | High |
| 2 | Speed - The engine should perform sentiment analysis in seconds from the time the request was made | Very High |
| 3 | Scalability - the engine should scale the output size based on the size of the input | Very High |

* **Usability**:

The GUI for the website must be easy to navigate around and provide simplistic design

complying with related standards and also must be uniform throughout the site to allow the

users to quickly familiarize with the website

* **Reliability**:

The system must be available 24 hours a day 7 days a week as it may be used at any time,

throughout many various locations to access all available options

* **Performance**:

The system must be able to support multiple users. There’s no initial target user base but the

system should be able to support more than 100’s of users and allow for expansion to different social networks such as Facebook & Google+. The response time of the system must allow for all actions to be performed in real time, the results must be available for review instantly.

* **Supportability**:

The website must be compatible with all major desktop browsers and mobile devices such as smartphones and tablets. The overall look and feel as well as usage must be done the same way across platforms.

* **Security**:

The website uses OAuth 2.0 protocol to contact and interact with the Twitter API using application-only authentication therefore user accounts and login facilities are not required. [18]

**{\displaystyle {\text{posterior numerator (male)}}={\text{their product}}=6.1984\cdot 10^{-9}}3.5 Advantages of new system**

Reduce time and effort. New system will help us to reduce time and efforts by doing some operation automatically. New system will reduce the cost If you don’t have repeat activities several times to finish them in the most quality way, you will dramatically cut costs due to fewer mistakes in your company. All your tasks must be done with the higher possible quality for the first time .

replace paper processes if we have a new system we will not need paper any more because we can do any things by programs and machine. New system Will Meet and Surpass All Customers Expectations communicate more effectively with customers, suppliers or partners

improve the efficiency of staff and the company

**3.6 User characteristics**

This part is to identify various user classes that we anticipate will use the web application. User classes will be differentiated based on the use, product functions and features, technical expertise, security and privilege levels and educational level. The solution is intended to be used by three main different user classes; system administrators, system operators and customers or regular users.

No special knowledge or skills should be assumed for the part of the regular users. Users are not expected to learn or remember a set of commands in order to start using the application. The prototype application will be only a web based and then for the product versions there will be a desktop versions, smart phones and smart Tablets. The following clearly describes a visionary role for each participant.

* Users:

users with no particular knowledge needed, users who are interested to use the tool looking for knowing people’s thoughts about a desired topic..

* System Operators:

Maintains for the functional interface of the application and troubleshooting issues.

Suggest possible updates and identifying renewal application needs.

Coordinate with service providers and infrastructure vendors.

Coordinate and communicate with system administrators.

* System Administrators:

Develop and maintain installation and configuration procedures and operational requirements.

Perform weekly/monthly backup operations, ensuring all required files and data are successfully backed up.

Repair and recover from hardware or software failures.

Coordinate and communicate with system operators.

**4-DESIGN**

**4.1 Design and Implementation Constraints**

Sentiment analysis of short texts such as single sentences and Twitter messages is challenging and have some Constraints because of the limited contextual information that they normally contain. Effectively solving this task requires strategies that combine the small text content with prior knowledge and use more than just bag-of-words, Sentiment analysis tools can identify and analyze many pieces of text automatically and quickly.

But computer programs have problems recognizing things like sarcasm and irony, negations, jokes, and exaggerations - the sorts of things a person would have little trouble identifying. And failing to recognize these can skew the results. [6]

'Disappointed' may be classified as a negative word for the purposes of sentiment analysis, but within the phrase “I wasn't disappointed", it should be classified as positive.

We would find it easy to recognize as sarcasm the statement "I'm really loving the enormous pool at my hotel!", if this statement is accompanied by a photo of a tiny swimming pool, whereas an automated sentiment analysis tool probably would not, and would most likely classify it as an example of positive sentiment.

With short sentences and pieces of text, for example like those you find on Twitter especially, and sometimes on Facebook, there might not be enough context for a reliable sentiment analysis. However, in general, Twitter has a reputation for being a good source of information for sentiment analysis, and with the new increased word count for tweets it's likely it will become even more useful.

So, automated sentiment analysis tools do a really great job of analyzing text for opinion and attitude. The result of Sentiment analysis must to be accurate and fast, To make it more accurate, we must use more than one algorithm but often make the process takes more time and become more slowly. [7]

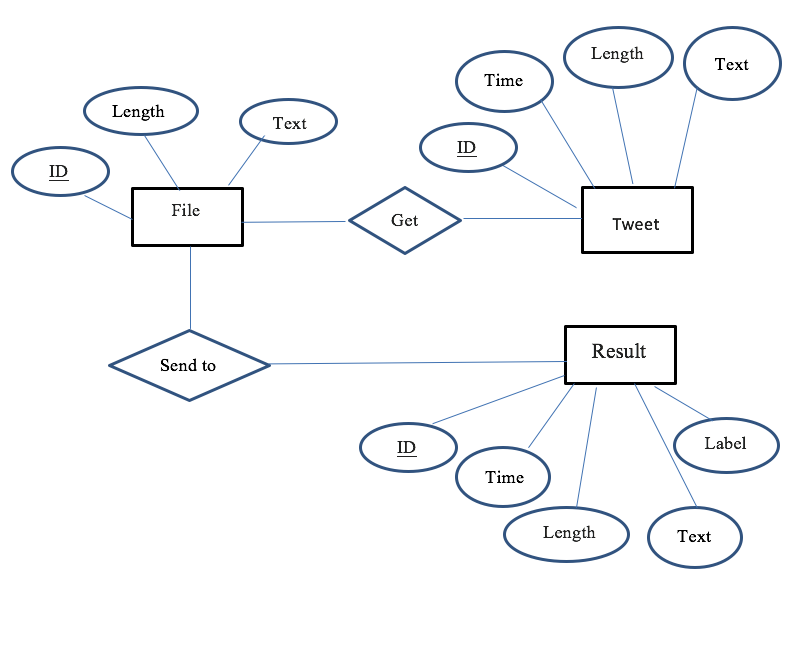
**4.3 Risks and risk management**

1. Poor Resources we didn’t find many open resources to help us in our project while document project .
2. API’s find many difficulties to choose the API will Work on.
3. Algorithm’s we take more time to search on the algorithm to find what will work in our project to be perfect program.
4. We work as a team that is make us problem to find time to all the team to meeting .
5. Working in parallel with studying semester. [11]

**4.4 Design of database ERD**

**4.4.1 Entity Relationship Diagram**

4.4.2 Mapping of Entity Relationship Diagram



**- Figure (4.1) -**

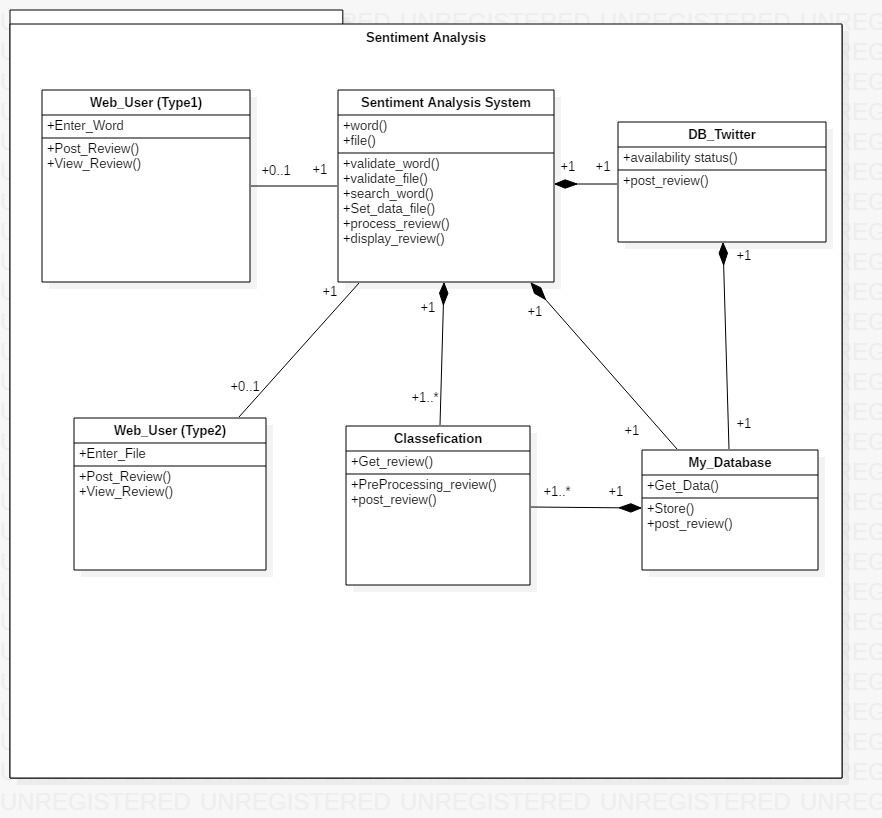
**4.4 Class diagram**

­­­if The user enters through a web page a 1-'word' he wants to search for or 2- enters a 'file' he wants to analyze

1-If the user has entered 'word' the Sentiment Analysis System search into DB\_Twiteer It then sends out this data to my database “My Database" to store Then it happens to her Classification "preprocessing " and then send new data to Sentiment Analysis System To happen "processing new data " and Display review

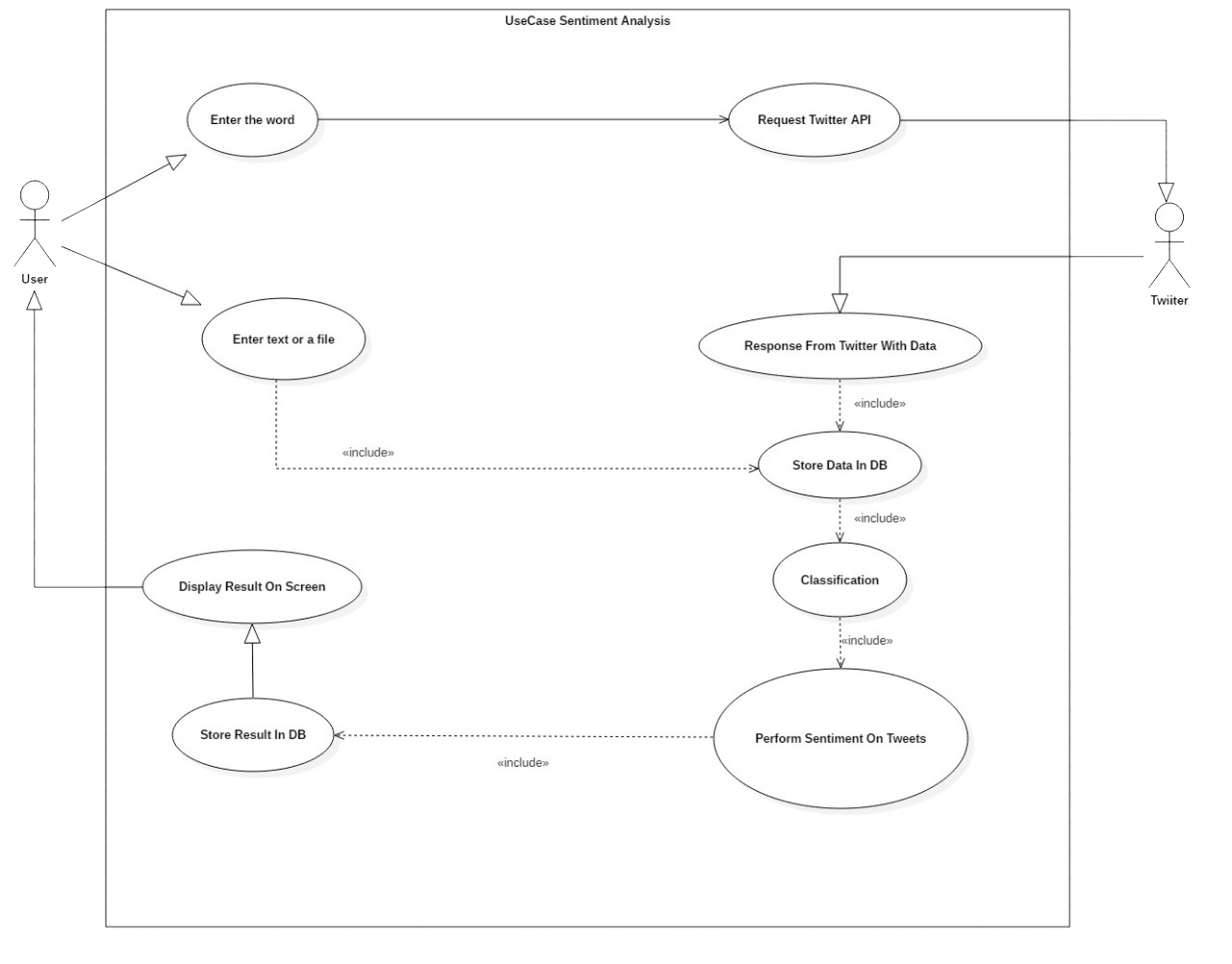
2-If the user has entered "file" the Sentiment Analysis System send data to "My Database" to store

Then it happens to her Classification "preprocessing " and then send new data to Sentiment Analysis System To happen "processing new data " and Display review



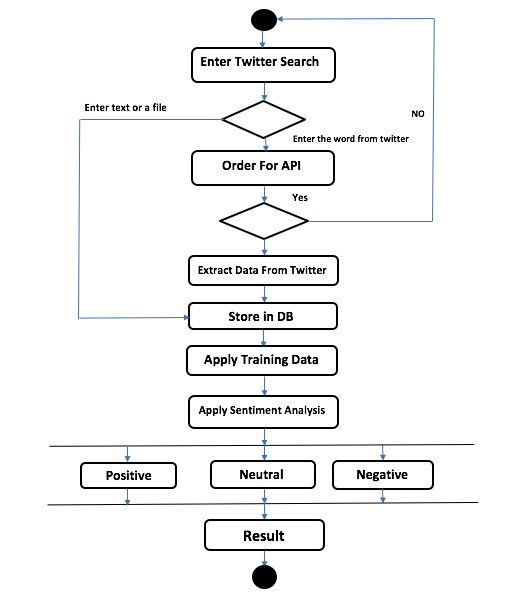
**- Figure (4.4) -**

**4.5 Use Case Diagram**



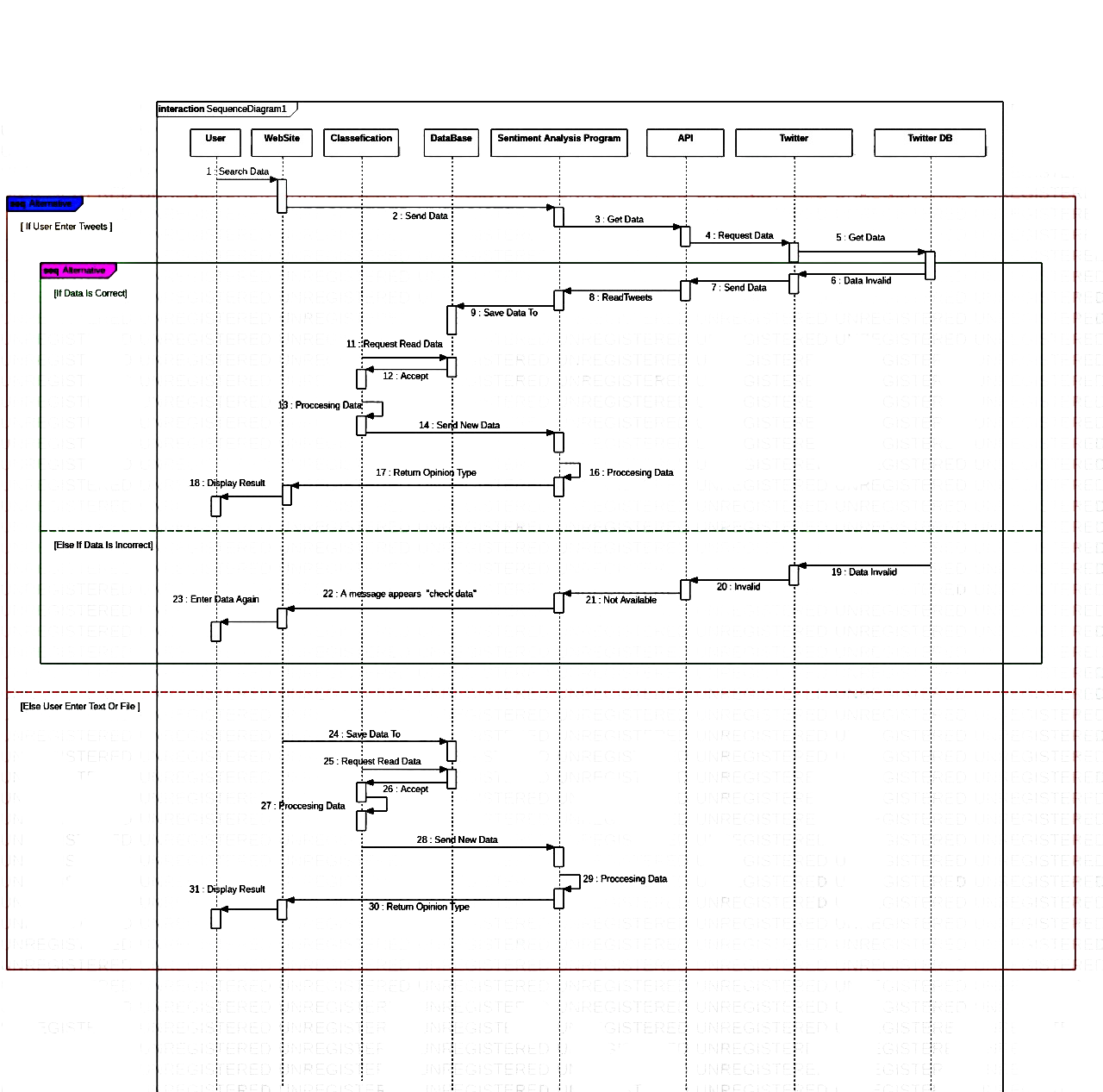
**- Figure (4.5) –**

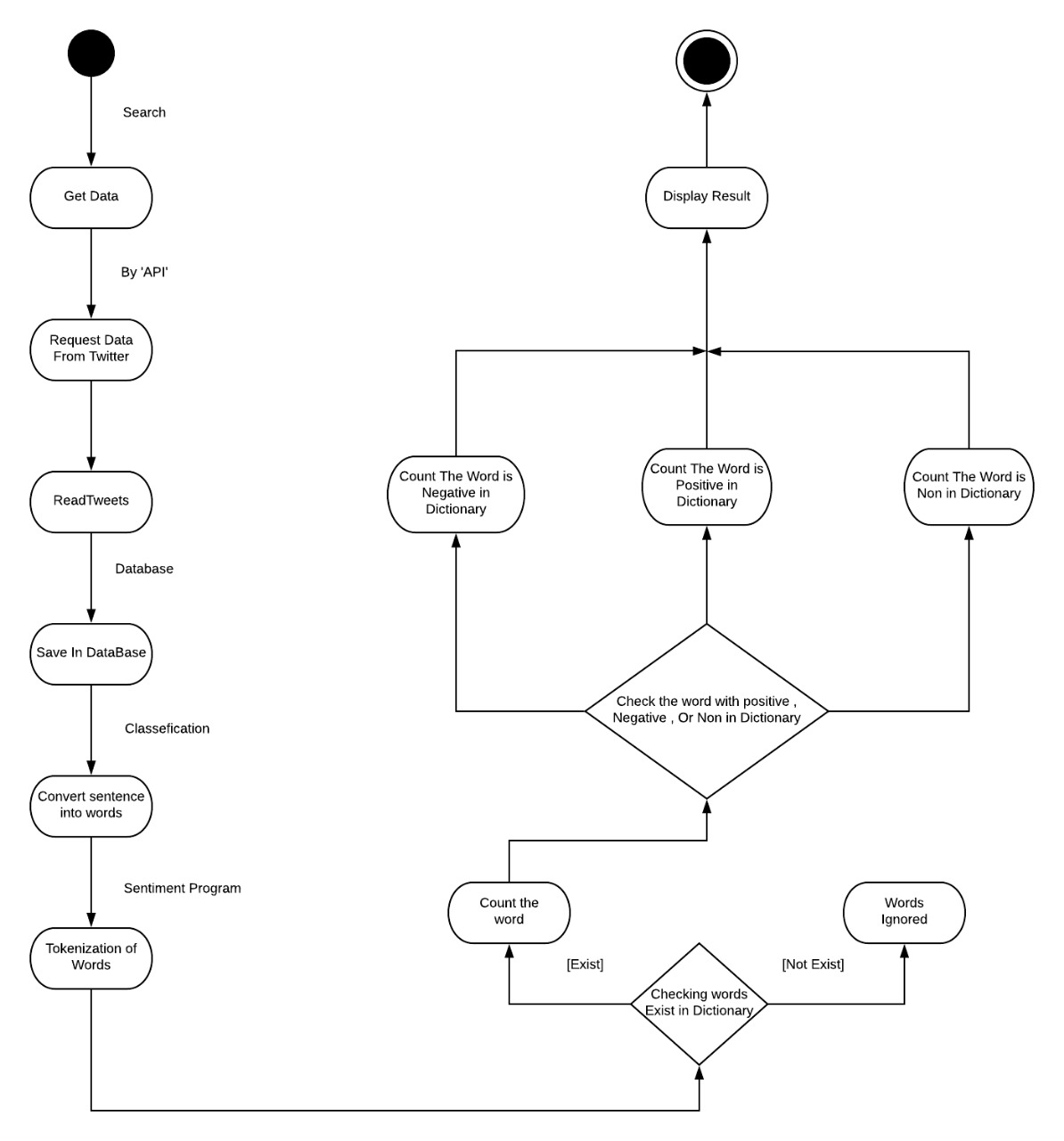
**4.6 Activity Diagram**

****

**- Figure (4.6) -**

**4.7 Sequence diagram**

Search for text within the Twitter database with 'API' Then it happens save to my database then happens to process and send the results to the client

**4.8 State diagram]**

**- Figure (4.8) -**

**4.9 Software Architecture**

is mean software systems are constructed to satisfy organizations business goals. The architecture is a bridge between those (often abstract) business goals and the final (concrete) resulting system.

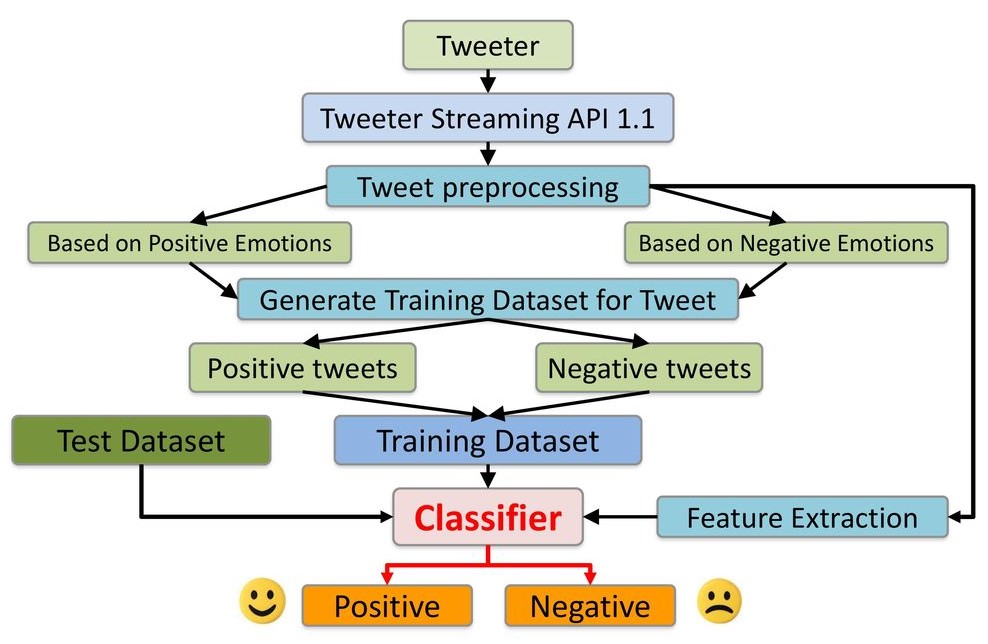
**System architecture**

A system's architecture is a representation of a system in which there is a mapping of functionality onto hardware and software components, a mapping of the software

architecture onto the hardware architecture, and a concern for the human interaction with these components.

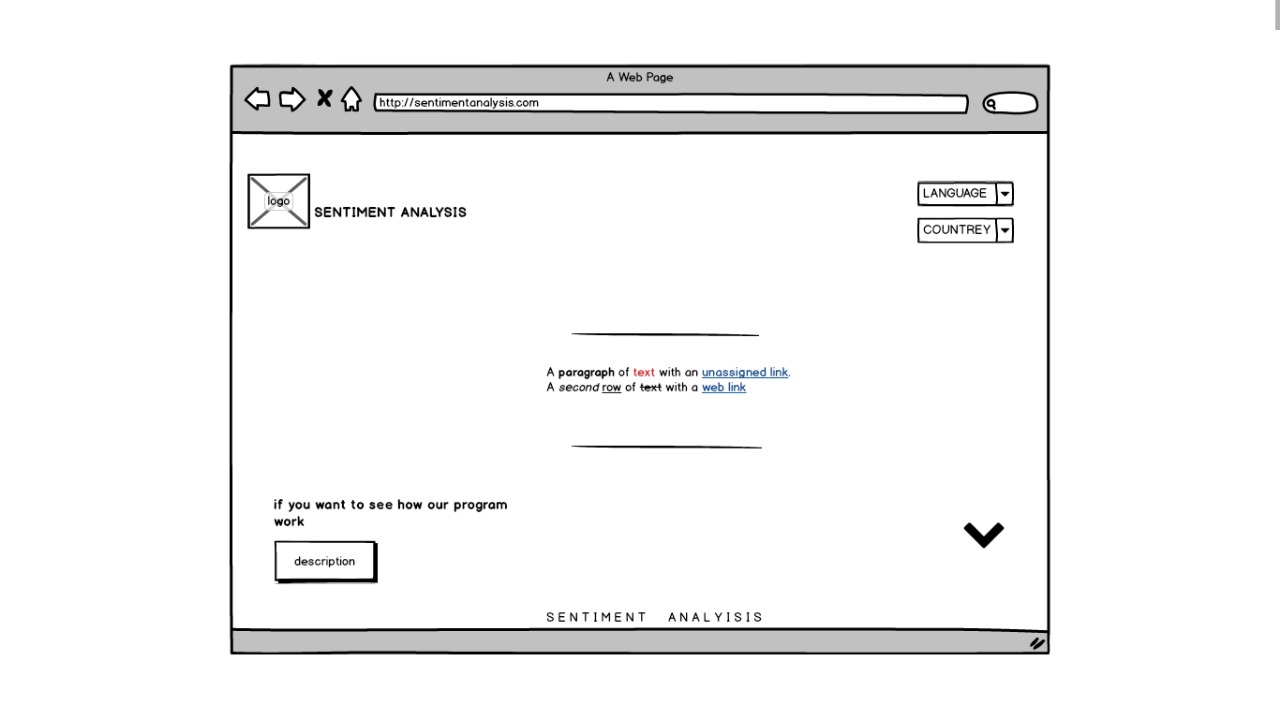
That is, system architecture is concerned with a total system, including hardware, software, and humans.

How the system is works?

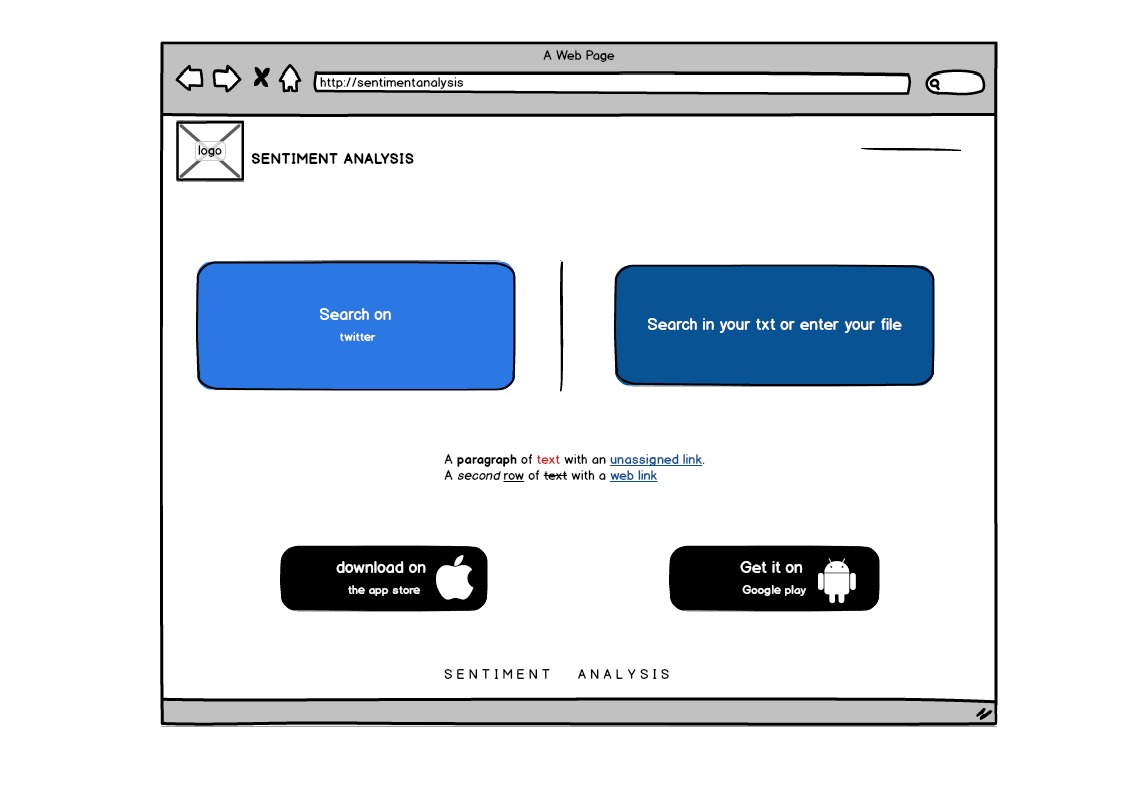


**- Figure (4.9) -**

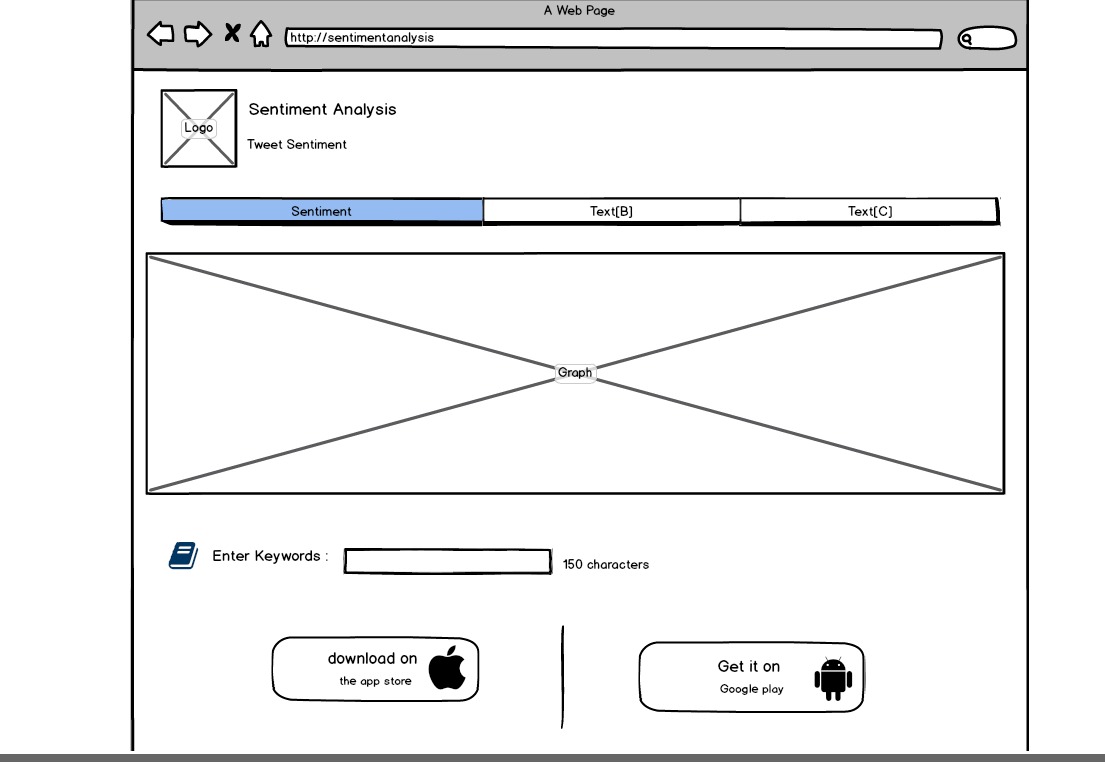
**4.10 User Interface**

****

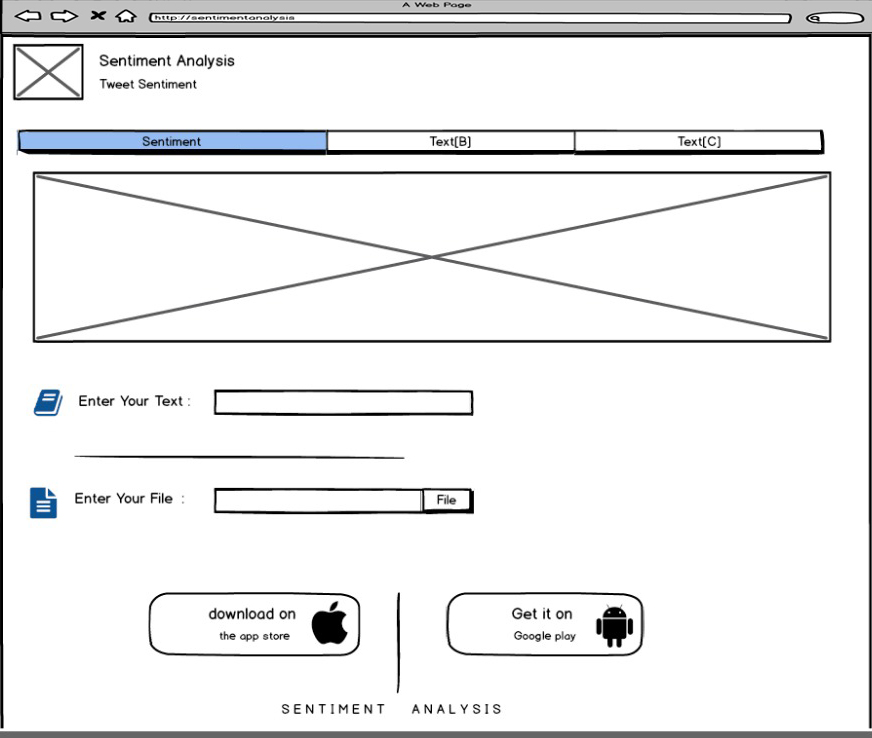
**- Figure (4.10.1) -**

****

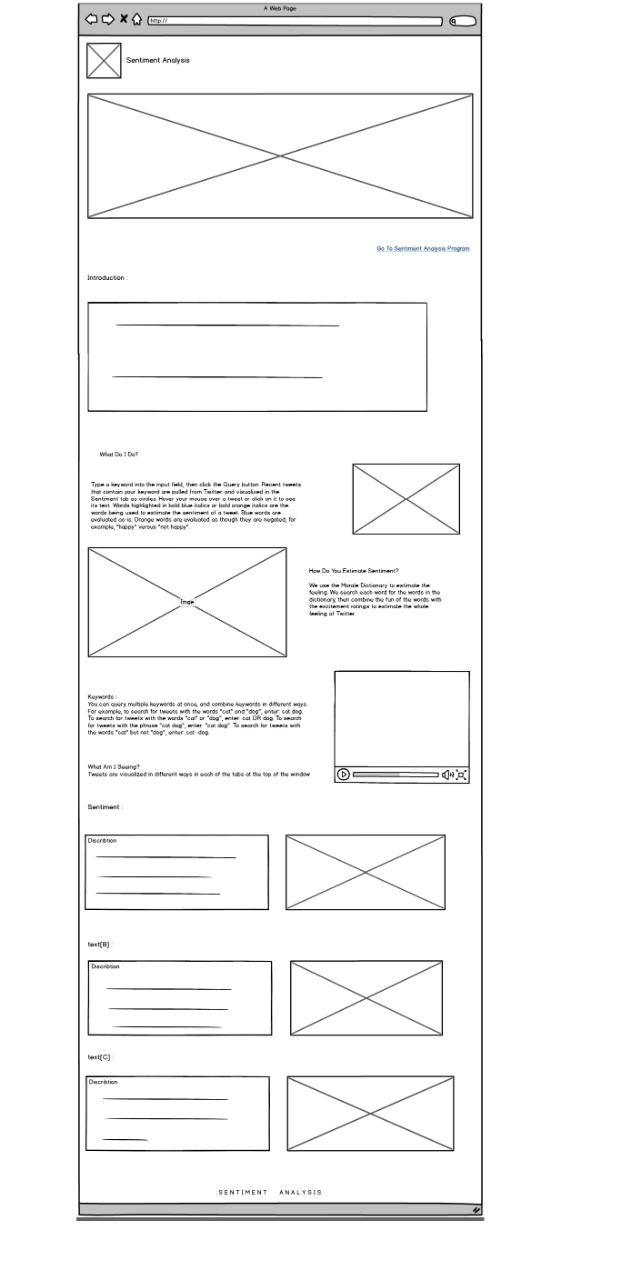
**- Figure (4.10.2) -**

****

**- Figure (4.10.3) -**



**- Figure (4.10.4) -**

****

**- Figure (4.10.5) -**

**REFERENCES:**

1. <https://monkeylearn.com/sentiment-analysis>
2. Kim S-M, Hovy E (2004) Determining the sentiment of opinions In: Proceedings of the 20th international conference on Computational Linguistics, page 1367.

Association for Computational Linguistics, Stroudsburg, PA, USA.

1. Twitter (2018) Twitter API. <https://dev.twitter.com/start>.
2. https://docs.microsoft.com/en-us/azure/architecture/data-guide/scenarios/natural-language-processing
3. https://www.upwork.com/hiring/for-clients/artificial-intelligence-and-natural-language-processing-in-big-data
4. https://searchsqlserver.techtarget.com/definition/data-mining
5. https://searchenterpriseai.techtarget.com/definition/machine-learning-ML
6. https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained
7. http://aclweb.org/anthology/W04-3253
8. https://medium.com/nlpython/sentiment-analysis-analysis-part-2-support-vector-machines-31f78baeee09
9. https://www.investopedia.com/terms/r/riskmanagement.asp
10. https://arxiv.org/ftp/arxiv/papers/1711/1711.10377.pdf
11. <https://www.entrepreneurshipinabox.com/1062/business-systems-7-benefits>
12. <https://www.nibusinessinfo.co.uk/content/business-benefits-new-software>
13. [https://www.dataversity.net/the-possibilities-and-limitations-of-sentiment-analysis#](https://www.dataversity.net/the-possibilities-and-limitations-of-sentiment-analysis)-
14. <https://typely.com/blog/making-use-of-sentiment-analysis>
15. <https://www.researchgate.net/publication/304024274_LIMITATIONS_OF_SENTIMENT_ANALYSIS_ON_FACEBOOK_DATA/download>
16. https://www.slideshare.net/anilsth91/tweet-sentiment-analysis-78718965