**A Project on**

**SENTIMENT ANALYSIS OF IMDB MOVIE DATA SET**

**Master of Tecnology**

**In**

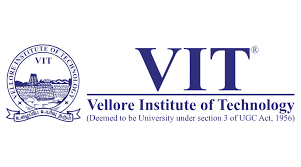
**Software Engineerning (Integrated)**

**By**

P. YAMINI 17MIS1123

**Under the guidance of**

PROF. TULASI PRASAD SARIKI



**Department of Computing Science and Engineering**

**Vellore Institute of Technology**

**VIT UNIVERSITY, CHENNAI CAMPUS**

**FALL SEM -2019-20**

**Table of Contents**

1. [Introduction 3](#_bookmark0)
   1. [Purpose 3](#_bookmark1)
   2. [Intended Audience 3](#_bookmark2)
   3. [Project Scope 3](#_bookmark3)
2. [General Description 4](#_bookmark4)
   1. [Product Perspective 4](#_bookmark5)
   2. [Product Functions 4](#_bookmark6)
   3. [User Classes and Characteristics](#_bookmark7) 6
   4. [Product Documentation](#_bookmark8) 7
      1. [User Documentation](#_bookmark9) 7
      2. [Technical Documentation](#_bookmark10) 7
3. [Interface Requirements](#_bookmark12) 8
   1. [User Interfaces 8](#_bookmark13)
   2. [Hardware Interfaces](#_bookmark14) 8
   3. [Communications Interfaces 8](#_bookmark15)
   4. [Software Interfaces](#_bookmark16) 9
4. [System Features](#_bookmark17) 9
5. [Non-functional Requirements](#_bookmark18) 12
   1. [Performance Requirements](#_bookmark19) 12
   2. [Safety Requirements](#_bookmark20) 12
   3. [Security and Privacy Requirements](#_bookmark21) 12
   4. [Software Quality Attributes](#_bookmark22) 12
6. [Use Case Diagrams](#_bookmark23) 13
7. [Research requirements](#_bookmark23) 15

# Introduction

## **Purpose:**

This Software Requirements Specification (SRS) documents key specifications, describes a prototype in terms of functional and non-functional requirements for Sentiment Analysis of IMDB movie data set. The information documented, helps the intended audience to design and develop the product.

## **Intended Audience:**

Primary readers of this document are the web researches, movie audience, Directors, actors, producers and developers. This document is intended for the following:

**Developers:**

in order to be sure that they are developing the right project that fulfil requirements provided in this document.

**Directors**

In order to have an exact list of the view about the director’s audience tweets.

**Producers:** To know about the popularity of the director and actors in the public. The type of movie that is being accepted by the public.

**Actors:** To choose the director based on his success rates in the past and his craze in the public.

## **Project Scope**

The scope of the project is to provide a public, actors, directors a complete view about their movie and the expected outcomes based on the previous results.

#### The project aims to:

1. Provide an accurate sentiment analysis result.
2. Get a past history of director, actor to the public.
3. Get the likes and craze of the actors in public.
4. Smooth, fast, efficient, reliable and easy way to analyse a movie.
5. Providing a user-friendly menu and good entertainment visualization capabilities.
6. Having a plenty of options in term of filtering and viewing information according to user’s needs.

# General Description

## **Product Perspective**

Due to the world’s massive growth of social networks and the rapid flow of news over the internet; Link Development and we came up with the sentiment analysis of imdb movie dataset. The main aim of this project is that it can allow users to analyse a movie based on collecting all comments, reviews, tweets or even notes in blogs related to the user’s search keyword.

Then perform an intelligent processing technique to extract the true meanings of the people’s comments and to decide and classify them in terms of positive, negative or neutral thus to know the majority of people like or dislike the desired topic. More specifically providing people's feelings regarding certain topics with high accuracy will lead to a better decision making.

## **Product Functions**

The architecture diagram of the tool is shown in figure 1. This tool will provide the following functions:

**Topic Extraction**

This part is considered a key stone in the project as it detects and extracts topics titles from the tweets. Using hash tags is not informative enough about the topic of sentiment the author mentions in his/her tweet. Our approach goes as follows, first we do pre-process which includes removal of stop words that occur frequently in the tweets but have no relevant meaning, then generate the feature vector. The features used are n-grams, unigram, bigram, and trigrams, and some named entities that are extracted from the crawled tweets. The main step is to cluster related tweets together using similarity measures so we can have multiple clusters each has one topic. Afterwards key-phrase extraction is used on each cluster to extract the key-phrases that are candidates to be title topics. Clusters that result in similar key phrase are merged together and this key phrase has higher weight to be the topic title.

**Sentiment Classification**

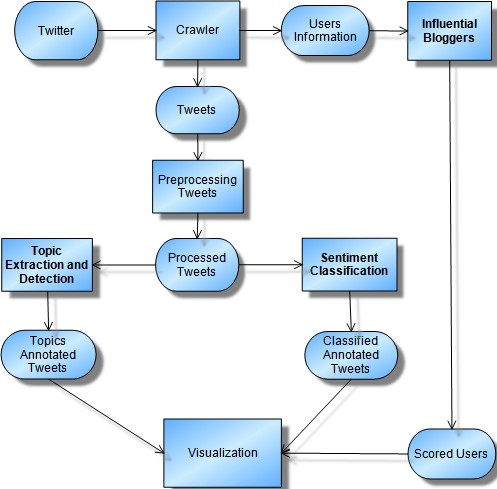
Sentiment classification is the primary module of the product. The objective of this part is to provide as much as possible an accurate classification for opinions embedded in certain sentences like tweets or micro-blogs written in Egyptian dialect as positive, negative or neutral. In addition to counting the total numbers of positive, negative and neutral tweets found in the data source with regards to specified topic.

**Determining Influential Bloggers**

Since influential members in a social network can be responsible for starting a buzz or getting the community to notice a new trend, product, or even adopt an opinion, we are interested in the problem of identifying which users are leaders. For Directors, actors, audience it is of great importance to learn about opinions in order to assess

chances and risks. A manual analysis is only possible on a very limited scale. An automated computer supported

analysis is necessary given the large number of virtual communities with huge amounts of posting



## **User Classes and 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 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.
* **Advanced end users:** advanced users are those who have valuable input and feedbacks. Users who are more familiar with informative sites and can use our features efficiently. These valuable feeds will lead to enhancement of users’ satisfaction.

#### 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

## **Product Documentation**

### User Documentation

User manual and CD will be made available for troubleshooting and help. Also, this will represent as a full backup of the system. The user manual will contain detailed information about the usage of the product from a layman perspective to an advanced network/system administrator. The manual may also be made available online however this manual will be made for the product version but not for the prototype

### Technical Documentation

Technical manual will be made for the purpose of current and future audience involved in watching the movie. The document will also include the development of technical requirements and the functional specifications components for the sake of verifying the technical accuracy of all procedural steps included in the document to help in annual reviews process for developers over the product. Also, as the user documentation this technical manual will be for the product version and not for the prototype.

# Interface Requirements

**User Interfaces**

User interface includes various forms and windows. The main window will consist of the main search bar and a main menu bar with file, edit, view, tools and help. The interface will visualize the features and functionalities listed in this document for this prototype as the included below not limited to:

* Drop down menu for various option selection
* Selection list for filtering results
* Push buttons for users’ feedback and reclassifying tweets
* Visual graphs to show results
* Help button

**Hardware Interfaces**

The solution makes extensive use of several hardware devices. These devices include;

* RapidMiner database server.
* Windows and Linux users’ computers.

**Communications Interfaces**

Internet connection and a web browser are required in order to make use of several functions and to be executed such as searching, viewing and downloading.

**Softw****are Interfaces**

For the data gathering twitter is the only source and using Streaming API that offers high throughput. Using this API is perfect because we can retrieve real time information and also this continuous stream will be retrieved with no end and capturing all the messages in the stream without missing any information. The information retrieved in JSON format.

# System Features

This section illustrates the functional features using the following template:

**System Feature:** Name of the feature.

**Priority:** Indicate the priority of the feature to the user whether it is of High, Medium, or Low.

**Description:** Provide a short description of the feature

**Action/ Response Sequences:** List the sequences of actions required to be done in order to use this feature.

**Result:** List the system responses of this feature.

**Functional Requirements:** List the software modules required to carry out the function provided by the feature.

|  |  |
| --- | --- |
| System Feature | Sentiment Classification |
| Priority | high |
| Description | Identifying the sentiment polarity (positive, negative or neutral) of tweets on certain topics from twitter. |
| Action | This module is activated after the user provides a query (topic, service or a product) or following the activation of the hot topic module. |
| Result | The system shows the results of the search of a query or the output of the hot topic module associated with the sentiment polarity of each item retrieved together with the percentage of Positive, Negative and Neutral sentiment of the whole result. |
| Functional requirements | A focused crawler, pre-processing module, sentiment classifier module, hot topic module and sentiment visualization module. |

|  |  |
| --- | --- |
| System Feature | User Feedback |
| Priority | Medium |
| Description | The user can give feedback by correcting the polarity of the classified retrieved tweets, and save the results |
| Action | The user selects a result and suggests a better annotation for it. |
| Result | The suggested correction by the user is stored in a system database to be handled by an administrator, and it is applied for future training and modifications to the system. |
| Functional requirements | A feedback interaction module |

|  |  |
| --- | --- |
| System Feature | Influential Bloggers Identification |
| Priority | medium |
| Description | Identifying the influential users on social media in certain topics. |
| Action | This module is activated after the user provides a query (topic, service or a product) or following the activation of the hot topic module. |
| Result | The system shows a list of all influential users on Twitter platform in certain topic, with indications on the level of influence. |
| Functional requirements | Influential bloggers identification module, hot topic module, and influential blogger visualization module. |

|  |  |
| --- | --- |
| System Feature | Hot Topics Identification |
| Priority | High |
| Description | Identifying the Hot topics and Trending topics in Twitter according time period. |
| Action | This module is activated after the user provides a date interval. The default interval is the last week using the system date. |
| Result | The system shows the hot and trending topics, putting them in order from high trending topics to lower and the user can browse the tweets related to any of the topic. |

|  |  |
| --- | --- |
| Functional requirements | A focused Crawling, and the topic extraction module |

|  |  |
| --- | --- |
| System Feature | Results Visualization of the tweets |
| Priority | medium |
| Description | Visualizing the results of sentiment classification and topic extraction modules into clear and interesting form. |
| Action | The proper modules will be activated by the user using a bottom included in the output screen of each of modules. |
| Result | The system shows the results in the visualization form selected. |
| Functional requirements | Sentiment classification and topic extraction Visualization modules. |

|  |  |
| --- | --- |
| System Feature | Statistics and info-graphics |
| Priority | Low |
| Description | Viewing different collected statistics about retrieved classified tweets, hot topics tweets, and influential bloggers in a good visualized form such as info-graphics. |
| Action | The proper modules will be activated by the user using a bottom included in the output screen of each of SATA modules. |
| Result | Reports and Graphs shows the statistics required |
| Functional requirements | Sentiment classification and topic extraction Statistics modules. |

# 5. Non-functional Requirements

**Performance Requirements**

**As for this prototype version we will** keep on detecting if the system crashed, hanged or an operating system error occurred. Also detecting the performance of the system in terms of the efficiency of integration of the different components



**Safety Requirements**

For the safety requirements nothing but an operation of weekly backups for the data base should take place.

**Security and Privacy Requirements**

There are no specific security requirements, anyone can access and use the portal but only authorized persons who are allowed to use and access the database, web pages and the product engine.

**Software Quality Attributes**

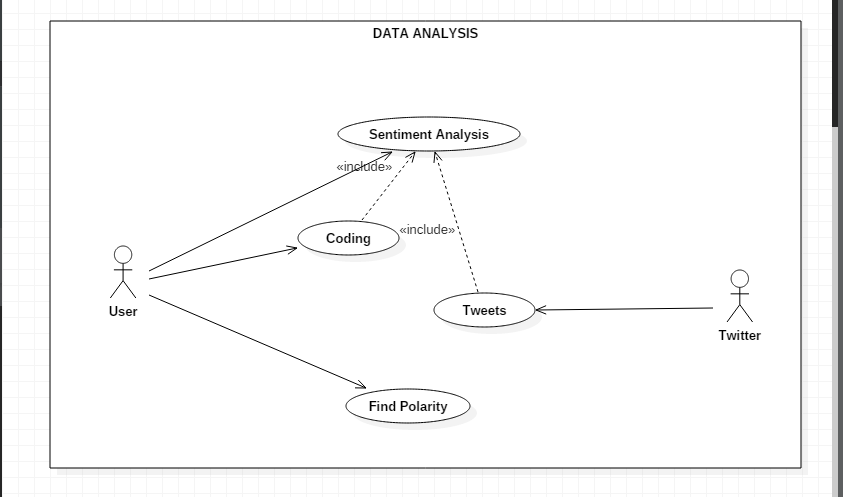
#### Reliability

The solution should provide reliability to the user that the product will run with all the features mentioned in this document are available and executing perfectly. It should be tested and debugged completely. All exceptions should be well handled.

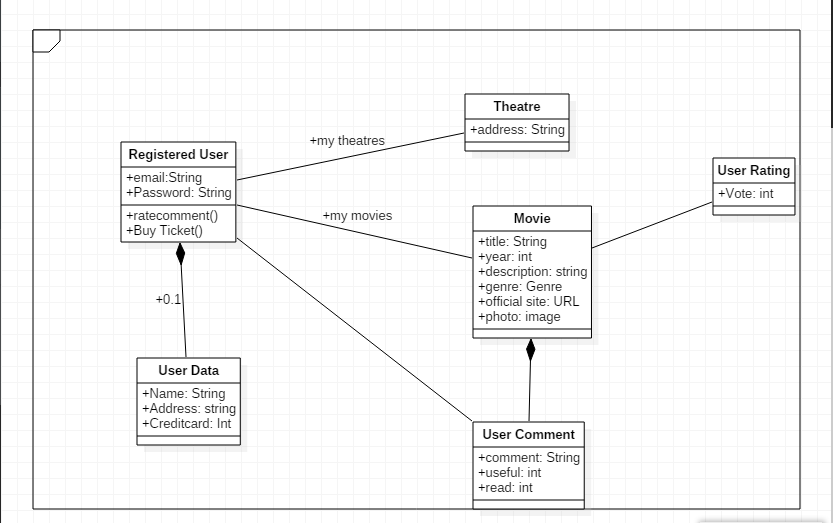
#### Accuracy

The solution should be able to reach the desired level of accuracy. But also keeping in mind that this prototype version is for proving the concept of the project.

# 6. Use Case Diagrams



CLASS DIAGRAM:



# 7. Research Requirements

|  |  |
| --- | --- |
| **Module Name** | Sentiment Classification |
| **Research Objective** | Extract the sentiment words in the tweets for the aim of creating a hybrid  approach which combines the benefits of the ML approach and the SO approach. |
| **Description** | Given the limited work done for Arabic text in the field of sentiment analysis, especially for the Egyptian dialect, two lists of sentiment words will be built manually one for the most occurring positive sentiment words, and one for the most occurring negative sentiment words. Then for each word in these lists a  weight is given to it based on its frequency in the positively labelled tweets, and the negatively labelled tweets in the corpus. |
| **Expected Outcome** | Weighted lists of the positive and the negative sentiment words mostly used by  the Egyptian bloggers. |

|  |  |
| --- | --- |
| **Module Name** | Detecting Influential Users and Opinion Leaders |
| **Research Objective** | Determine the method for retrieving User Information |
| **Description** | These two methods will be investigated:  1. Get the user information using the Twitter API |

|  |  |
| --- | --- |
|  | The Twitter REST API enables developers to access user information. However, the API is rate limited; it only allows clients to make a limited number of calls in a given hour. Also, there are limitations to the information retrieved, for example, it does not return more than 5000 followers per users even though the number of followers may exceed that, and for information such as retweets and mentions, it only returns the 20 most recent retweets or mentions, which may also exceed that. I find that such limitations may exclude information that could be of value to determine which users are influential.  2. Get the user information from the user profile page  Develop a crawler to access a user’s twitter profile page source code, and retrieve the user information available, such as the number of tweets posted by that user, the number of followers, friends and list, and other information available that may be useful. However, there are a couple of issues in regard to this approach. First, Twitter has recently changed the layout of its user profile pages more than once. Such changes require adjustment to the crawler code that extracts the user information. Second, this approach limits the amount of information we may have access to what is only available on  the page. |
| **Expected Outcome** | Users information retrieval tool |