Software Requirements Specification

Of

Sentiment Analysis

Team Members:

Tirth Patel (170020107041)
Ravi Sahani (170020107049)

Internal Guide:

Prof. Bansari Thakkar

Table of Contents

Software Requirement Specification

1. Introduction

Sentiment Analysis is contextual mining of text which identifies and extracts subjective information in source material, and helping a business to understand the social sentiment of their brand, product or service while monitoring online conversations.

1.1 Purpose:-

This Software Requirements Specification (SRS) documents key specifications, describes a prototype in terms of functional and nonfunctional requirements for Sentiment Analysis Tool for Arabic (SATA). The information documented, helps the intended audience to design and develop the product. There will be a need for future updates of this document as we are planning to launch a prototype version for testing then start officially the beta version then the final version.

1.2 Scope:-

The scope of the project is to provide a user friendly web based product that extracts people's sentiment feelings toward certain services, products, organizations, political or nonpolitical topics and any influential people on social media. In this project phase which aims at developing a filed prototype, emphasis will be put on Arabic tweets from Twitter in the political domain.

The project aims to:

1. Provide an accurate sentiment analysis results.
2. Achieve a wide range of users in Egypt and the MENA region.
3. Support Arabic Egyptian dialect in the first run and English will be considered later.
4. Smooth, fast, efficient, reliable and easy to use web-based tool.
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.

1.3 Definitions, Acronyms and Abbreviations:-

The algorithm proposed works on Twitter Data, primarily it collects the tweets and then study it with the help of different statistical computing procedures. In the age of artificial intelligence and machine learning, competition is between best and best. So inorder to gain control over market, it is essential to understand market condition especially during covid-19 situation. For that sentiment of market is very important and sentiment of market is what consumer think of certain product.

1.4 References:-

Kaggle.com
SRS from various universities
Coursera.org
Udemy.com
Github.com

1.5 Overview:-

Due to the world's massive growth of social networks and the rapid flow of news over the internet; Link Development and AUC came up with the sentiment analysis tool for Arabic (SATA) research project. The SRS contains product perspective, functional and non-functional requirements, product documentation, product constraint, interface requirement, user classes and characteristics, research requirements.

2. Overall Description

2.1 Product Perspective:-

The main aim of SATA is that to develop a tool that can allow users to use a simple search bar to search for any services, products or any political topics and the engine of that tool is to crawl over the internet 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.
The purpose of the prototype is to demonstrate the concept and to deliver operational and functional services for testing purposes. As for initial Twitter will be the only source of data for the prototype and then integration will be needed to include more sources like facebook, news websites and blogs.

2.2 Product Functions:-

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

Topic Extraction:

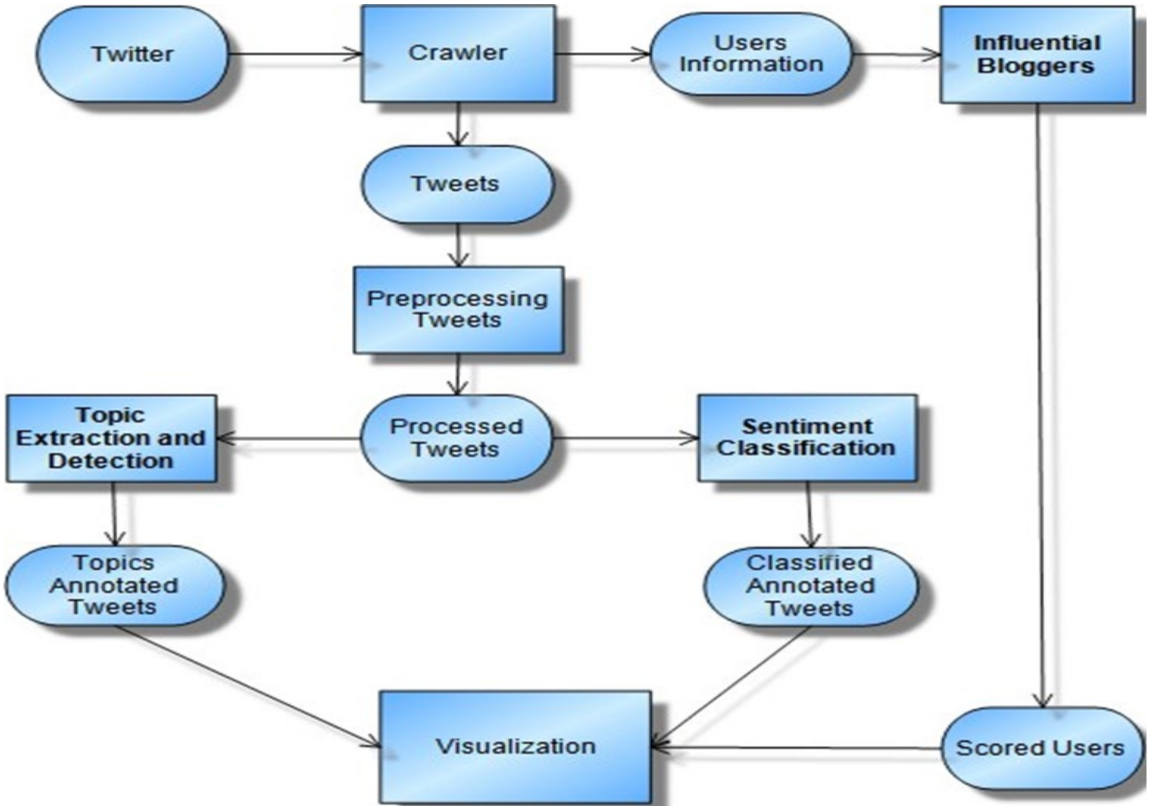
This part is considered a key stone in the project as it detects and extracts topics titles from the tweets. Using hash tags is not informative enough about the topic of sentiment the author mentions in his/her tweet. Our approach goes as follows, first we do preprocessing which includes removal of stop 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 companies, organizations and governments, 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 postings.



Architectural diagram for the product

2.3 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 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.

- **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 o Suggest possible updates and identifying renewal application needs o Coordinate with service providers and infrastructure vendors o 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 o Coordinate and communicate with system operators

2.4 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 developers involved in the product to understand and follow the solution at the level of coding and the programing languages used. 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.

2.5 Product 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. The following constraints will apply to for both the prototype and the different live service solution versions.

Processing Power: SATA requires high speed machine for data capturing from various sources, classifying the sentiment polarity of large data and extracting topics.

Deployment Point: SATA is built to be deployed as internet services. High bandwidth of the portal is required to fulfill the large number of concurrent users.

Operating Platform: SATA may work for several distributions of Linux and Windows PCs, also smart phones and smart tablets.

3. Specific Requirements

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.

3.1 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, preprocessing 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	A focused crawler, 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 SATA components
Priority	medium
Description	Visualizing the results of sentiment classification, influential blogger 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 SATA modules.
Result	The system shows the results in the visualization form selected.
Functional requirements	Sentiment classification, influential blogger 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 Info-graphics that shows the statistics required
Functional requirements	Sentiment classification, influential blogger and topic extraction Statistics modules.

3.2 Nonfunctional 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.

3.3 Research Requirements:-

This section describes the needed research and experiments work efforts to develop each module: hot topic detection and extraction, sentiment classification, and detection of influential bloggers and opinion leaders.

Module(s) Name	Hot Topic Detection and Sentiment Classification
Research Objective	To find a list of Arabic stop words to be removed from tweets to enhance clustering and classification results.
Description	Finding a proper list of stop words is not an easy task specially when dealing with the Arabic dialect. Different spelling of the same word by users makes it difficult to include all the word forms in the list. Using natural language processing tools like stemmer to detect different forms of the same word is not just difficult but also gives bad results as some dialect words do not follow the inflection rules of modern standard Arabic. We will develop a list by getting frequent unigrams that occur more than a certain threshold from the total crawled tweets that reached about 20,000 tweets. Named entities are removed from this list as it’s relevant to our work and being repeated that frequent gives it more weight not the opposite.
Expected Outcome	A list of Egyptian dialect stop words

Module Name	Hot Topic Detection
Research Objective	Select proper features that will achieve accurate clustering
Description	The features are the words or phrases that are relevant to our domain, and help clustering the tweets properly. In our work we are using n-grams and named entities. N-grams are unigram, bigram, and trigram. Determining the threshold of each n-gram that will lead to get better clustering is what we are targeting. The Named entities will be also considered as features will be also investigated.
Expected Outcome	Thresholds for all features

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 labeled tweets, and the negatively labeled tweets in the corpus.
Expected Outcome	Weighted lists of the positive and the negative sentiment words mostly used by the Egyptian bloggers.

Module Name	Sentiment Classification
Research Objective	Compare the performance of the Machine Learning and the Semantic Orientation methodologies and choose the one which produces the best result.
Description	Although the ML approach was used extensively in the sentiment analysis process throughout the literature, it was still very important to test the SO approach with respect to our case which is dealing with the Egyptian dialect. Thus, we need to test both methodologies for the aim of comparing their performance and interpret the results obtained in each methodology.
Expected Outcome	The methodology which is most suitable to our case which is dealing with the Egyptian dialect.

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

Module Name	Choosing the Machine Learning classification algorithm
Research Objective	Compare the performance of the Support Vector Machine and the Naïve Bayes algorithms when used in the machine learning methodology and choose the algorithm which produces the best result.
Description	Although it was observed in more than one study that the Support Vector Machine algorithm produce higher result than the Naïve Bayes algorithm, it is still important to test the performance of the Naïve Bayes classification algorithm. The Support Vector Machine algorithm is believed to have some principle advantages over the Naïve Bayes algorithm. Some of these advantages are robustness in high dimensional spaces, any feature is relevant, robustness when there is a sparse set of samples and, finally, most text categorization problems are linearly separable. On the other hand, Naïve Bayes algorithms are also most suitable for classification problems with high dimensionality. That is why we need to try both algorithms and choose the one which produces the highest accuracy.
Expected Outcome	The classification algorithm which produces the highest accuracy.

3.4 Interfaces:-

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;

- MySQL database server with intensive use of memory space.
- PHP server with high performance and intensive use for CPU usage.
- 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.

Software Interfaces

For the prototype we will launch the portal over the internet and other than the hardware specified in the hardware interface section, the software requirements are to support windows operating system with support to MySQL, apache and PHP servers.

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.