**Software Requirements Specification**

For

**Automated System for Sentimental Analysis of Streaming Data**

**31/10/2021**

Prepared by

|  |  |  |
| --- | --- | --- |
| **Specialization** | **SAP ID** | **Name** |
| Big Data | 500067177 | Lakshay Vasuja |
| Big Data | 500068760 | Lakshay Sharma |
| Big Data | 500068095 | Divyansh Chandna |
| Big Data | 500068030 | Janvi Arya |



Department of Informatics

School Of Computer Science

UNIVERSITY OF PETROLEUM & ENERGY STUDIES,

DEHRADUN- 248007. Uttarakhand

Table of Contents

|  |  |  |
| --- | --- | --- |
| **Topic** | | **Page No** |
| Table of Content | | 2 |
| Revision History | | 3 |
| 1 | Introduction | 3 |
|  | 1.1 Purpose of the Project | 3 |
|  | 1.2 Target Beneficiary | 3 |
|  | 1.3 Project Scope | 3 |
|  | 1.4 References | 3-4 |
| 2 | Project Description | 4 |
|  | 2.1 Reference Algorithm | 4 |
|  | 2.2 Data/ Data structure | 4-5 |
|  | 2.3 SWOT Analysis | 5 |
|  | 2.4 Project Features | 6 |
|  | 2.5 User Classes and Characteristics | 6-7 |
|  | 2.6 Design and Implementation Constraints | 7 |
|  | 2.7 Design diagrams | 8-9 |
|  | 2.8 Assumption and Dependencies | 9 |
| 3 | System Requirements | 9 |
|  | 3.1 User Interface | 9 |
|  | 3.2 Software Interface | 10 |
|  | 3.3 Database Interface | 10 |
|  | 3.4 Protocols | 10 |
| 4 | Non-functional Requirements | 10-11 |
|  | 4.1 Performance requirements | 11 |
|  | 4.2 Security requirements | 11 |
|  | 4.3 Software Quality Attributes | 11-12 |
| 5 | Other Requirements | 12 |
| Appendix A: Glossary | | 12 |
| Appendix C: Issues List | | 13 |

1. INTRODUCTION
   1. Purpose of the Project

The main purpose of this project is to create a general-purpose twitter sentimental analysis system where a user can analyze both tweets posted by a particular user and tweets related to a specific hashtag or keyword. It will help to overcome the challenges of identifying the sentiments of the tweets.

* 1. Target Beneficiary

The main target audience for this project are those companies who want to build their own sentimental analysis system for various purposes and those people who want to do research related to sentiments of a particular user and hashtag.

* 1. Project Scope

With the aid of multiple features like automating the whole process, gathering of streaming data or fastening the whole process by collectively performing all the steps together (gathering of data, pre-processing of data, sentimental analysis of data, graphical analysis), companies can improve their own sentimental analysis systems and save time.

* 1. References
* Hossam ELzayady, Khaled M.Badran, Gouda I. Salama,(2018). Sentiment Analysis on Twitter Data using Apache Spark Framework, 10.1109/ICCES.2018.8639195.
* Dr. Khalid N. Alhayyan & Dr. Imran Ahmad (2017). “Discovering and Analyzing Important Real-Time Trends in Noisy Twitter Stream” n.p
* Shashank Gupta (2018, January 7). Sentiment Analysis: Concept, Analysis and Applications. Retrieved From: <https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17>
* Sandeep Dayananda (2019, May 22). Spark Streaming Tutorial – Sentiment Analysis Using Apache Spark. Retrieved From: <https://www.edureka.co/blog/spark-streaming/>
* Nikhil Raj (2021, June 15). Starters Guide to Sentiment Analysis using Natural Language Processing. Retrieved From: <https://www.analyticsvidhya.com/blog/2021/06/nlp-sentiment-analysis/>

1. PROJECT DESCRIPTION
   1. Reference Algorithm
2. Rational Unified Process (RUP) methodology will be followed throughout this project for developing the system.
3. The inception phase will comprise of fetching the live tweets with a keyword, using the streaming API’s and data streaming pipeline will be created.
4. Moving towards the elaboration phase, pre-processing of fetched tweets will done using the Apache Spark transformations.
5. Further, in the construction phase, applying the sentimental analysis using Natural Language Processing concepts through which sentiment analysis score will be generated based upon the fetched tweets.
6. Adding to this, we can visualize the extracted result by plotting it. Finally, in the transition phase, the system is tested and modified according to the requirements.
   1. Characteristic of data



Fig 1: Unstructured form of fetched tweets (username: PMOIndia)

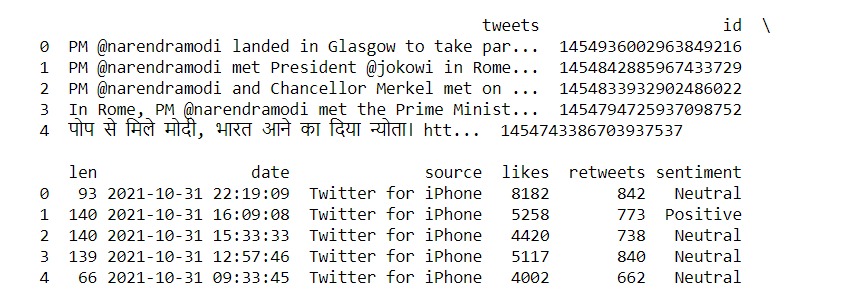


Fig 2: Structured form of fetched tweets (username: PMOIndia)

The source of dataset is twitter user’s timeline which consists of unstructured tweets. (Username entered by the user during runtime).

Then fetched tweets were preprocessed by pandas library and converted them into structured form.

* 1. SWOT Analysis



Fig 3: SWOT Analysis

* 1. Project Features

Project gives two functionalities for the user which are user timeline sentiment analysis and the second one is hashtag sentiment analysis. In user timeline analysis, the major features are: -

1. User interface model-asks the user to give username or twitter handle of any individual as an input to perform further analysis on his/her tweets.

2. Fetching live tweets

3. Tweets are then preprocessed into positive, negative and neutral reviews of any particular username.

4. Visualization of output on a graph.

5. It helps audience to judge the tweets of any particular username.

In hashtag sentiment analysis, the major features are: -

1. User interface model-asks the user to enter any hashtag as an input to perform further analysis on the tweets based on this hashtag.

2. Fetching live tweets

3. Tweets are then preprocessed into positive, negative and neutral reviews of any particular hashtag.

4. Visualization of output on a graph.

5. It helps audience to get an idea about any trending hashtag or current news.

* 1. User classes and characteristics
* TwitterClient

This class contains five different functions (\_\_inti\_\_, get\_twitter\_client\_api, get\_user\_timeline\_tweets, get\_friend\_list, get\_home\_timeline\_tweets). This class works on the API module of tweepy library to generate a connection between entered keyword and a user’s timeline.

* TwitterAuthenticator

This class contains one function(authenticate\_twitter\_app). This class works on the OAuthHandler module of tweepy library to authenticate the twitter user by authenticating their unique keys and tokens (these keys and tokens were developed at the time of twitter developer account creation).

* TwitterStreamer

This class contains two functions (\_\_init\_\_, stream\_tweets). This class works on Stream module of tweepy library to fetch the live tweets.

* TwitterListener

This class contains three functions (\_\_inti\_\_, on\_data,on\_error). This class deals with StreamListener module of tweepy.streaming to collect the fetched tweets by twitter streamer class.

* TweetAnalyzer

This class contains four different function(clean\_tweet, analyse\_sentiment, tweets\_to\_data\_frame, count\_values\_in\_column). This class works with textBlob, matplot and pandas libraries of python. First of all cleaning of fetched tweets is done, then sentimental analysis using textBlob which in turn converts this into a structured form (dataframe) and then generates graph as the output.

* 1. Design and Implementation constraints

1. Python 3 is the language used and import libraries tweepy and textBlob.

2. A twitter developer account is needed.

3. Further, a connection is to be made for fetching tweets and for that you need to use tweepy.

4. Matplot library is required to plot the graph of the output.

* 1. Design diagrams
* Activity Diagram

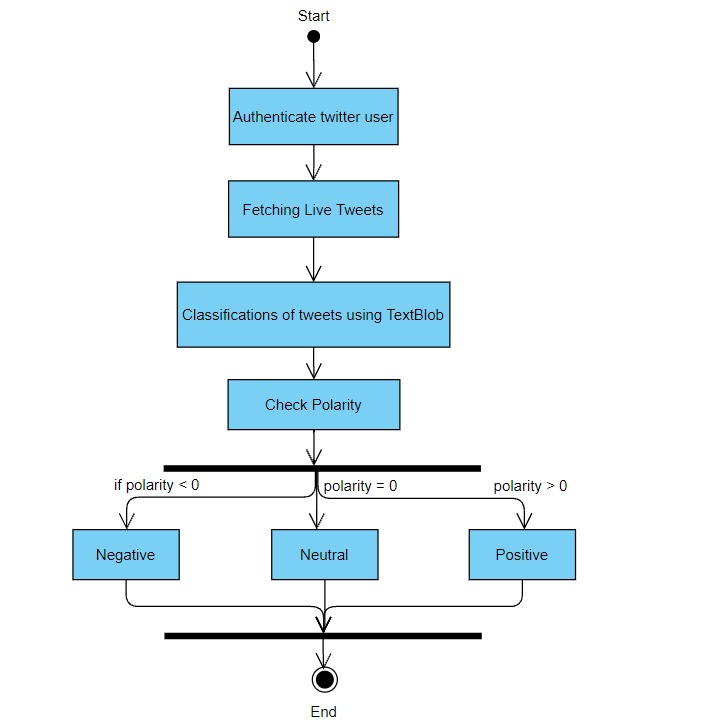


Fig 4: Represents the Activity Diagram of the System

* Collaboration Diagram

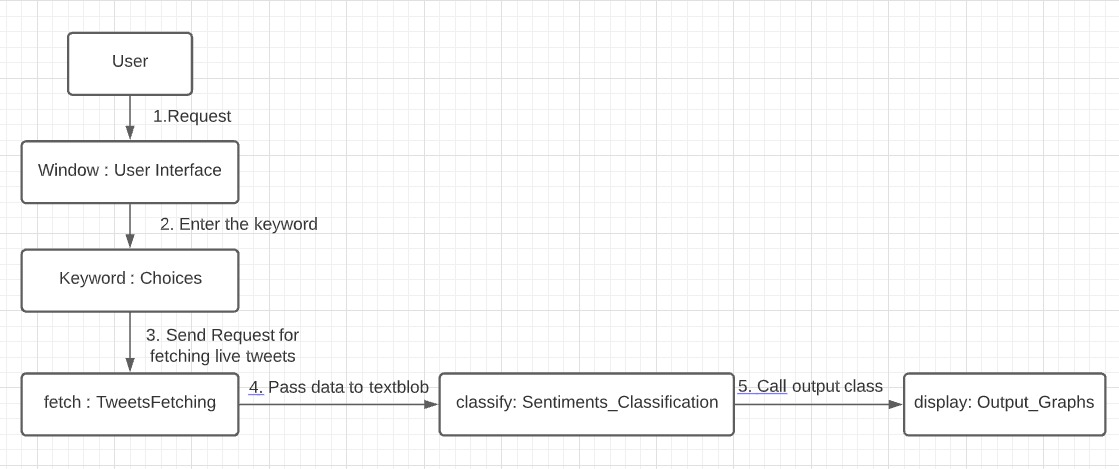


Fig 5: Represents the Collaboration Diagram of the System

* Data Flow Diagram

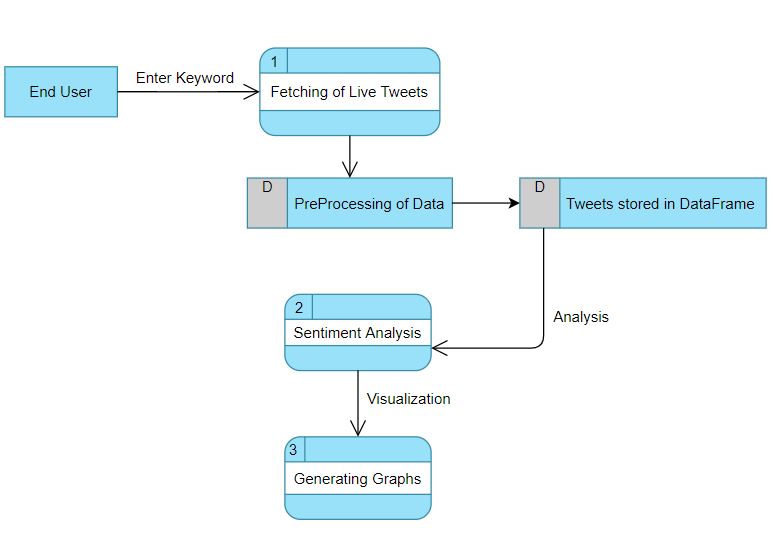


Fig 6: Represents the Data Flow Diagram of the System

* 1. Assumptions and Dependencies

The assumption in our project is that the tweets are being fetched in real time, although it is near to real time.

The dependency is the accuracy of the library used for classification of tweets (textBlob). Since the accuracy is 52%, some tweets might get classified in the wrong category.

1. FUNCTIONAL REQUIREMENTS
   1. User Interface

Front end Software- Anaconda/Jupyter Notebook

Back end software- Python

Operating System- Windows 10

Libraries- Numpy, Pandas, tkinter for GUI

* 1. Software Interface

The connection between different modules is described as follows:

A connection is established using twitterclient using function get\_twitter\_client\_api that returns a reference to the twitter application. User timeline function is used pass the keyword or username and also specify the size of data batch we need to extract. When creating an object of tweet analyzer class, we converted the above fetched tweets into dataframe. Now we have all the requirements and references ready to be passed for sentimental analysis.

* 1. Database Interface

Although we are fetching live twitter data but we are using temporary storage(dataframe) to store only during the program runtime which displays the result because there is no point in storing the data since it is changing live.

No database is currently used in our project.

* 1. Protocols

A user needs to have a twitter developer account so that he/she can create an app in the UI provided by the twitter developer website and then they can generate keys which will be used in the authentication process while fetching the data. Only these protocols need to be followed. The synchronization of dataset is not required as well because the fetching of data is done from the latest tweet of the user or in case of hashtag the latest tweet about that hashtag so synchronization is already performed by the program.

1. NON-FUNCTIONAL REQUIREMENTS
   1. Performance requirements

|  |  |  |
| --- | --- | --- |
| S.NO | Entities | Minimum Requirement |
| 1. | OS | 1. Windows 7, Recommended: Windows 10 |
| 2. | CPU | Intel or AMD processor with 64-bit support; Recommended: 2.8 GHz or faster processor |
| 3. | GPU | Not required to run this. Recommended: Nvidia GeForce GTX 1660 or Quadro T1000 |
| 4. | Disk Storage | 4 GB of free disk space |
| 5. | Monitor Resolution | 1280x800; Recommended: 1920x1080 |
| 6. | Internet | Internet connection required for software activation |

4.2 Security requirements

* In our project, user is requested to create a developer account on Twitter to fetch tweets from Twitter.
* After creating an application o the developer twitter website, twitter develops keys and tokens for the user.
* The keys and tokens developed by Twitter woks as an authenticator in fetching the tweets or working of our project.
  1. Software Quality Attributes
* Adaptability: - Definition of adaptability is the degree to which a product framework adjusts to change in its current circumstance. A versatile programming framework can endure changes in its environment. Our project works on different versions of python.
* Availability: - It is defined as the possibility that the system is going to properly operate when it is requested for use. As in our project, user can easily fetch tweets from Twitter anytime and have the sentiment analysis done.
* Correctness: - It is defined as the behavior of the user or the product interaction with each other. As our project is GUI-based which eases the user to request for the output they want.
* Flexibility: - It is defined as the ability of the product to change according to the user or environment requirements. Our project accepts different user inputs and give different outputs according to their requirements.
* Interoperability: - It is defined as the ability of the systems to interact or communicate freely. Like in our project, we are fetching the latest tweets for the analysis.
* Maintainability: - This attribute judges the product based on the improvisation in the software according to latest trends or environment changes.
* Portability: - It is the ability to move software from one machine platform to another. As our project outcome i.e., analysis can be done on any device meeting the above-mentioned performance requirements of the system.
* Reliability: - It is defined as the working of software failure-free in any specified environment for specified time. As in our project, we are using TextBlob library to maintain the accuracy of the classification done which ensures that the output which user gets is reliable.
* Reusability: - It is defined as the ability of the product to be reused when in need. Our project can be used by multiple user multiple times i.e., user can do sentiment analysis of tweets as many times as they want.
* Robustness: - It is defined as the ability to cope with any failures or errors at any given time.
* Testability: - This aspect covers the ease in testing of the product in any environment.
* Usability: - Defined as the degree to which a product will be used by consumers to achieve specific objectives with satisfaction and efficiency.

5. OTHER REQUIREMENTS

There are no other requirements left to be discussed. All the essential requirements have been thoroughly discussed above.

APPENDIX A: GLOSSARY

|  |  |
| --- | --- |
| Terms | Definition |
| Tweets | The tweets is referring to the data which is fetched from twitter. |
| DFD | It stands for Data Flow Diagram |
| GUI | It stands for graphical user interface by which end users easily interact with the system. |
| Dataframe | It is a structured form representation of the twitter data. |
| textBlob | It is the library used for classifying the fetched data. |
| RUP | It stands for Rational Unified process. It uses object oriented approach to regulate/maintain the software lifecycle. |
| Keyword | A keyword refers to the specific username on twitter |
| Hashtag | A hashtag is a social media term that starts with “#”. These hashtags are used to signify a particular event or thing. |

APPENDIX B: ISSUES LIST

|  |  |  |  |
| --- | --- | --- | --- |
| S.No | Issues | Priority | Status |
| 1. | Convert batch stream into live stream | P1 | Ongoing |
| 2. | Creation of GUI | P2 | Creation of different modules is completed. Combining them is remaining, |
| 3. | Building our own algorithm for preprocessing | P2 | Right now we are using inbuilt algorithm |