Software Requirements Specification

for

Sentiment Analysis Based on tweets during COVID-19 using Twitter

Version 1.0 approved

Prepared by Tanisha Nazare(59)

Ketaki Mankar(61)

Somu Sharma(50)

Table of Contents

1. Introduction

1.1 Purpose

The Coronavirus outbreak has been a severe disruption to the global economy and this has affected all nations. In this pandemic the impact of social media platforms is becoming more noticeable than ever before. Social networking has a remarkable impact and is one of the most increasingly growing social information structures. Analyzing tweets during and after Coronavirus could be worthy as the condition and people's reactions are changing every instant during this critical period. The contribution of this work is to analyze the COVID-19 tweets dataset on news to aid the understanding of sentiment trends.

1.2 Document Conventions

This project of analyzing sentiments of tweets comes under the domain of "Pattern Classification" and "Data Mining". Both of these terms are very closely related and intertwined, and they can be formally defined as the process of discovering "useful" patterns in a large set of data, either automatically (unsupervised) or semi automatically (supervised). The project would heavily rely on techniques of "Natural Language Processing" in extracting significant patterns and features from the large data set of covid19 tweets and on "Machine Learning" techniques for accurately classifying individual unlabeled data samples (tweets) according to whichever pattern model best describes them.

1.3 Product Scope

This model can be used for different regions, countries and social groups to understand their behavior. It can be extended to understand reactions towards vaccinations with the rise of anti-vaccine sentiments given fear, insecurity and unpredictability of COVID-19. Sentiment analysis which can provide more details for emerging topics during the rise of COVID-19 cases in relation to various government protocols such as lockdowns and vaccination plans.

1.4 References

- https://www.analyticsvidhya.com/blog/2021/02/sentiment-analysis-predicting-sentiment-of-covid-19-tweets/
- https://towardsdatascience.com/twitter-sentiment-analysis-based-on-news-topics-during-covid-19-c3d738005b55
- https://drive.google.com/file/d/1-bOQ_NQXq2CQsGP-FwMqwE7MoxZxuC2i/view?usp=sharing

2. Overall Description

2.1 Product Perspective

The recent explosion in data pertaining to users on social media has created a great interest in performing sentiment analysis on this data using Machine Learning principles to understand people's interests. This project intends to perform the same tasks. The difference between this project and other sentiment analysis tools is that it will perform real time analysis of tweets based on hashtags and not on a stored archive. SRS defines a component of a larger system, relates the requirements of the larger system to the functionality of this software and identifies interfaces between the two. A simple diagram that shows the major components of the overall system, subsystem interconnections, and external interfaces can be helpful.

2.2 Product Functions

• Collect tweets in a real time fashion i.e. , from the twitter live stream based on specified hash tags in News

• Remove redundant information from these collected tweets.

• Store the formatted tweets in dataset

• Perform Sentiment Analysis on the tweets stored in the database to classify their nature viz. positive, negative and so on.

• Use a machine learning algorithm which will predict the human emotions and concerns with respect to COVID-19.

It is a Natural Language Processing Problem where Sentiment Analysis is done by Classifying the Positive tweets from negative tweets by machine learning models for classification, text mining, text analysis, data analysis and data visualization

2.3 User Classes and Characteristics

The model can be used for different regions, countries and social groups to understand their behavior. It can be used by officials for better Covid-19 management through policies and projects, such as support for depression and mental health issues. Our analysis provides a potential approach to reveal the public's sentiment status and help institutions respond timely to it.

2.4 Operating Environment

Language Used: Python

This application will run on an Android device as this application is a stand alone single user system.

Processor - Pentium or above.

RAM - 128 MB or above.

Tools used: Google Colab/Jupyter Notebook

Inputs will be received from two sources

1) User Interface - Supply the keywords and analyze the session duration.

2) Twitter API - Supplies the Tweet text.

Outputs will display the current mood of the Twitter on the Covid19 related tweets and the historical data would be displayed in the form of graphs.

2.5 Design and Implementation Constraints

This project is based on the Natural Language Processing Problem where Sentiment Analysis is done by Classifying the Positive tweets from negative tweets by machine learning models for classification, text mining, text analysis, data analysis and data visualization. The preprocessing of the text data is an essential step as it makes the raw text ready for mining. In one of the later stages, we will be extracting numeric features from our Twitter text data. Exploring and visualizing data.

3. External Interface Requirements

3.1 User Interfaces

It meets the requirements to conform to the user's needs.It controls which allow the user to interact with the application as well as to imply their functionality within the application. This interface includes user inputs as well as graphs for a visual representation of the output produced.

User Inputs - For controlling the sentiment analysis first by adding,removing keywords for each Covid-19 topic and second by specifying the duration of each analysis session.

Graphs - Displaying the percentage of Twitter users who are currently for or against the topic will be analyzed.Also display the total number of Tweets which have been processed as well as the user should be able to interpret the trend of sentiment toward the topic related to Covid19.

3.2 Hardware Interfaces

This application will run on an Android device as this application is a stand alone single user system.
Processor - Pentium or above.
RAM - 128 MB or above.

3.3 Software Interfaces

The software will run on an Operating system in which the inputs will be received from two sources
1) User Interface - Supply the keywords and analyze the session duration.
2) Twitter API - Supplies the Tweet text.

Outputs will display the current mood of the Twitter on the Covid19 related tweets and the historical data would be displayed in the form of graphs.

Tools : Google collab

3.4 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 for downloading purposes of the tweets dataset for Twitter.

4. System Features

4.1 Data extraction:

4.1.1 Description

Our main priority of this model is the dataset which contains data of covid19 tweets. These tweets are collected using Twitter API and a Python script. A query for this high-frequency hashtag (#covid19) is run on a daily basis for a certain time period, to collect a larger number of tweets samples.

4.1.2 Functional Requirements

Data should be extracted through tweets from twitter based on covid19.

4.2 Sentiment Analysis:

4.2.1 Description

Sentiment analysis, one of the most promising methods for content analysis in social media, known as emotion AI or opinion mining, leads to natural language processing(NLP) and text analysis to systematically, quantify, extract, identify, and study effective states and personal information. As tweet is widely used and according to the surveys users tweet are more likely to express their feelings towards their post. So by using this sentiment analysis feature we are going to predict the sentiment of that particular user's twitter.

There are three main steps to show how sentiment analysis works:

- Data collection: Data collected from twitter.
- Prepossessing: The collected information is processed during this step in order to prepare the data for the next phase. This phase includes three main stages.

    1. The cleaning stage contains the removal of repeated letters, text correction, normalization, stop word removal, and language detection.

2. The Tokenization method focuses on converting text into tokens until it becomes vectors.
3. The extraction of features such as grammatical structures and mining characteristics.

- Data exploration: extracted data is represented in such a way that a common consumer can also read and understand our dataset which we are using.
- Data analysis: In this stage, all data should be processed and then identified based on the main purpose of research, such as polarity identification, sentiment analysis, or frequency analysis.

4.2.2 Functional Requirements

Model should be able to process new tweets stored in the database after retrieval and also be able to analyze data and classify each tweet polarity.

4.3 Topic Modeling:

4.3.1 Description

Topic modeling analyzes “bags” or groups of words together—instead of counting them individually–in order to capture how the meaning of words is dependent upon the broader context in which they are used in natural language. Topic modeling is not the only method that does this cluster analysis, latent semantic analysis, and other techniques have also been used to identify clustering within texts. One of the techniques is BERT. It is an open source machine learning framework for natural language processing (NLP) and is designed to help computers understand the meaning of ambiguous language in text by using surrounding text to establish context.

4.3.2 Functional Requirements

The collected tweets containing the different words which are related to covid 19 are segregated and bags of words are formed. This would help our NLTK model to process its further for training and testing.

FUNCTIONALITY	GET DATA
INPUT	.csv files consisting of Dataset
REQUIREMENT	Panda library to convert dataset into Dataframe
OUTPUT	Tabular format
SEQUENCE	1

FUNCTIONALITY	PROCESSING DATA
INPUT	Raw Data
REQUIREMENT	Data Preprocessing & Tokenization Algorithm
OUTPUT	Cleaned & Processed Data
SEQUENCE	2

FUNCTIONALITY	TRAIN MODEL
INPUT	Cleaned & Processed Data

REQUIREMENT	Natural Language Processing toolkit(NLTK) and BERT Algorithm
OUTPUT	Training Model Created to test the new Data
SEQUENCE	3

FUNCTIONALITY	TEST MODEL
INPUT	Input in the form of text from user
REQUIREMENT	Training Model
OUTPUT	Result of entered text for analyzing the sentiments
SEQUENCE	4

5. Other Nonfunctional Requirements

5.1 Performance Requirements

It would provide up-to-date information limited only by the rate of Twitter input.The output should display the latest results at all times and if lags the user should probably be notified.

5.2 Safety Requirements

System should not cause any harm to human users.The system is able to avoid or tackle disastrous action i.e. it should be foul proof and robust.

5.3 Security Requirements

It should never disclose any personal information of Twitter users and should collect no personal information from its own users.

5.4 Software Quality Attributes

Reliability - It would meet all of the functional requirements without any unexpected behavior. The output shouldn't display incorrect or outdated information without alerting the user to errors.

Availability - It will be available at all times on the user's Android device and its functionality depends on any external services such as Internet access.

Maintainability - The code should be written clearly and should be documented properly for ensuring its maintenance.

Portability - It will be designed to run on any operating system including all the latest system versions. It should not be architecture specific.

Appendix A: Analysis Models