AUTOMATED PROCESSING FOR SOCIAL MEDIA DATA IN A MASS EMERGENCY

Project ID - 18-007

System Requirement Specification

Sri Lanka Institute of Information Technology

Special Honors Degree of Bachelor of Science in Information Technology

Specialized in Software Engineering

May, 2018

Table of Content

Table of Figures

List of Tables

1 INTRODUCTION

1.1 Purpose

The purpose of the document is to give a detailed description of requirement for the "Automated Processing System for Social Media data in a Mass Emergency". The document will give brief idea of the detailed descriptions of the functional requirements, non-functional requirements, hardware and software requirements, user characteristics and user interfaces. This document will give detail overview of the product and its parameters. In addition, the document outlines, constraints which may limits the developers options and assumptions and dependencies which made by developer while implementing the component. This document is primarily intended for the Supervisor, Co – Supervisor and research team members to refer as a reference document while developing the system, but also will be of interest to researchers who are interested in implementing this kind of product. The document is written in a form that any person can read and understand the content

1.2 Scope

This document includes the details relevant to the basic functionalities in the "Automated Processing System for Social Media data in a Mass Emergency" project. To describe requirements of the system, different diagrams and technical challenges which should overcome are also included to this document along with overview of the system, goals, tasks, benefits, users and research areas. This document will provide a comprehensible design of the system.

1.2.1 Main Objectives

Main goal of this research work is to develop an open source application programming interface (API) for processing social media textual data at presence of a natural disaster to support individuals of natural disaster supporting teams. The end product would be composed with four modules or components which will be focusing on major or foremost and priorities aspects which can be expected from automated processing of social media data in a mass emergency. The four principal components are,

1. Developing an automatic text summarization component for processing social media posts in an emergency and generating related summaries.
2. Categorizing the information identified and prioritizing the information of social media posts in order to obtain filtered information.
3. Semantic analysis of information to measure how critical, the corresponding situation could be.
4. Validating the accuracy of each social media post by analyzing the follow up comments.

1.2.2 Benefits

· Stop spreading false information
· Give accurate and up to date information
· Avoid hearsay and rumors.

1.3 Definitions, Acronyms, and Abbreviations

SRS	Software Requirement Specification
API	Application Programming Interface
NLP	Natural Language Processing
SQL	Structured Query Language
DB	Database
RAM	Random Accesses Memory
GPRS	General Packet Radio Service
EDGE	Enhanced Data rates for GSM Evolution
GSM (G)	Global System for Mobile

DMC	Disaster Management Center

Table 1: Definitions, Acronyms, and Abbreviations

1.5 Overview

1.5.1 Software Overview

The world is full of emergencies caused by natural disasters. In such situations, vast amount of information exchange through social media like Facebook, Twitter, official websites and applications that are dedicated to natural disaster management. In countries where natural disasters are frequent, the disaster management centers have employed teams to monitor and analyze information to get a closer insight into the situation. It may be helpful to identify areas that have suffered the most in an emergency, the type of emergency, and the value of the information that has been trusted. Manually analyzing the overwhelming amount of information is difficult, error prone, and tedious. Real-time disaster information is critical for rapid decision-making in response to emergencies. This research work aims to introduce an effective and productive automated tool for analyze the information generated on social media using modern concepts such as, Semantic Analysis, Natural Language Processing, Machine Learning and Artificial Intelligence.

1.5.2 Document Overview

Chapter 1: Explain purpose of preparing this document. In scope, describes what this system will do. It further describe objectives, benefits and the goals of the system. In overview it describes how the SRS is organized and a brief description of what the rest of the document contains.

Chapter 2: Describe the user-understandable overall description in a non-technical way. It includes product perspectives, product functions, user characteristics, constraints, assumptions and dependencies, and the apportioning of requirements. In product perspective, compares with existing systems and other competing products to provide perspective of the system. Product functions give summary of the major functions of the application. User characteristics indicate what kind of people the typical user is likely to be. Constraints describe all conditions that may

limit developer's options. The Assumptions and Dependencies section describes any assumptions made when developing the component.

Chapter 3: In this chapter, it describe external interface requirement, performance requirement, design constraints, software system attributes and classes and object that specific system. In other words, this chapter describes the developer point of view of the component.

2 OVERALL DESCRIPTION

Social media become key source that people go for help and information in disaster situation. Some organizations and government agencies have identified the use of social media as an important role in emergency response. The task of processing social media entries requires new means of information filtering, classifying and summarization. The lacking feature of most current systems available is the accuracy and the dependability of a given entry. Hybrid systems highly depend on crowdsourcing which requires volunteers so called digital volunteers. This affects the latency of the process. Existing systems are highly dependent on the Twitter. Extracting data from numerous sources other than Twitter streaming API is a challenging task to be completed. The unstructured data needs to be cleaned in order to be used in other stages. Finding appropriate optimal number of categories to match the requirements of different parties (Organization, Government agencies etc.), identifying ways of calculating accuracy levels for entries, defining thresholds and finding the criticality of situation are major research areas which would be covered throughout the research project.

Figure 1: Overall system diagram

2.1 Product perspective

Tough social media is practically and widely used in financial business-oriented scenarios applications for other purposes are scarce increasing widespread use, popularity and large user base of social media had lead the way for researchers to identify various other uses of social media platforms. In fact, there is a lot of work to be done for the context of social media usage in an emergency.

Some organizations and government agencies have identified the use of social media as an important role in emergency response. For example, American Red Cross has deployed so called Digital Response Center in order to provide situational awareness information and help who are in need. Due to the lack of manpower, lack of funds to conduct proper research and criticality of a situation stakeholders believe that it is resource wasting unachievable task

The task of processing social media entries requires new means of information filtering, classifying and summarization. The lacking feature of most current systems available is the accuracy and the dependability of a given entry. Hybrid systems highly depend on crowdsourcing which requires volunteers so called digital volunteers. This affects the latency of the process. Existing systems are highly dependent on the Twitter. Extracting data from numerous sources other than Twitter streaming API is a challenging task to be completed. The

unstructured data needs to be cleaned in order to be used in other stages. Finding appropriate optimal number of categories to match the requirements of different parties (Organization, Government agencies etc.), identifying ways of calculating accuracy levels for entries, defining thresholds and finding the criticality of situation are major research areas which would be covered throughout the research project. Here is a comparison of existing systems which can be included under the domain of proposed system.

Figure 2: Comparison table with existing systems

2.1.1 System interfaces

- Twitter API
- WordLift API
- Indata lab API

2.1.2 User interfaces

Figure 3: System interface design 1

Figure 4: System interface design 2

Figure 5: System interface design 3

2.1.3 Hardware interfaces

- Compatible smartphone / tablet pc.
- Personal computer / laptop with required specifications.

2.1.4 Software interfaces

- Virtual servers (Amazon EC2) in the cloud by Amazon Web Services (AWS)
- Docker
- Neo4J graph platforms
- MongoDB NoSQL database
- Firebase realtime database
- NLP Tool Kit

2.1.5 Communication interfaces

- Modem (Built in GPRS/EDGE/3G/4G)

- Wi-Fi Router

2.1.6 Memory constraints

The system will be developed as microservices for each module or component, deployed on Amazon Web Services (AWS). Memory and storage usages will be allocated according to the consumption of resources of proposed system. Virtual server instances for microservices will be initiated with 8GB of RAM and 1TB of storage.

2.1.7 Operations

Operations of the system and subsystems can be carried out at three different levels. There will be parameters associated with each level of operation which define the status of the system and / or control the system.

Observing Level

This is the normal operational mode. It allows a user to access the API through a Web at a fairly high level. Monitoring is also done at this level. It is anticipated that all user categories have access to this level.

Maintenance Level

This allows the system to be maintained in a more efficient way. New releases would be planned depending on this level. Limited access would be given to the system administrators.

Test Level

Sub-component level testing would be carried out in this level. This would also be given restricted access

2.1.8 Site adaptation requirements

The user interfaces of proposed system are expected to be designed only in English language. Multilingual support will be available in future releases depending on end user requirements.

2.2 Product functions

There are four major components in this system. Under those component there are major sub functions which will carryout by developers.

1. Text Summarization - From summarizing social media posts will attempt to get on identification of the core meaning of a particular post and extracting the summarized content over the bulk of social media posts. Below points are the major functions of this component.
    - Upload text corpus to be processed through automatic text summarization.
    - Generate textual graph by uploaded text corpus.
    - Generate word cloud from a bulk of social media posts which are related to a particular context.
    - Compare and contrast generated summarized text output over input text corpus.
    - View generated summarized text output by means of graphs in a diagramatic way.
2. Categorizing and prioritizing - Categorizing each post content in order to assign into corresponding buckets which are priorities are labeled on.Below points are the major functions of this component.
    - Extract data from social media to identify the relevant data.
    - Identify the relevant data from the extracted data.
    - Categorize (Classify) the relevant data into meaningful categories.
    - Prioritize the categorized (classified) entries.
3. Define Criticality - Analysis of social media post to disclose the criticality of the disaster situation. Below points are the major functions of this component.
    - Call API
    - Critically level prediction
    - Monitor details
4. Validate accuracy - Analysis of the comments for the social media posts to validate accuracy of the main post.
    - Get comments posted for the post want to validate
    - Filter comments to separate relevant comments

- Generate word cloud from relevant comments to particular context.
- Generated accuracy by counting words
- View generated accuracy as graph in percentage.

2.3 User characteristics

The application is intended to be used by any personal who is interested in an emergency. Although DMC, Humanitarian Organizations, Victims, General Public and Journalists are the main benefactors of the system.

User does not need to have a special training in other words user doesn't have to be a expert in technology. Anyone with basic knowledge and understanding of domain should be able to operate the system.

2.4 Constraints

- All the tools and technologies should be open source
- Limitation of available time will be major constraint to development.
- Due to the restrictions in Facebook Graph API. The analysis is limited only to posts / comments that are posted to groups which are owned by the users. In the case of twitter, it doesn’t limit to the user’s followers any public tweet can be obtained.
- Code base must be open source.
- Data usage should strictly follow the policies mention by the data source providers.
- Git must be used as the VCS for the project.

2.5 Assumptions and dependencies

- User must have internet connection
- User Should have basic knowledge of using computer and internet
- Server must be up and running
- There should be sufficient processing power to run the system

2.6 Apportioning of requirements

The requirements described in sections 1 and 2 of this document are referred to as primary specifications. In the section 1.4 describe the overview of the system component and it further describe under section 2. The way system is been implemented can be change in system design phase from the content of the document. However, functional and nonfunctional requirements describe in this document will not change any point.

3 SPECIFIC REQUIREMENTS

3.1 External interface requirements

3.1.1 User interfaces

Figure 3: System interface design 1- Validate accuracy UI – Describe the information about validity of the post. Display disaster information with map and accuracy level of the post with graph and pie chart.

Figure 4:System interface design 2 - View Criticality Level UI - Describe the information about criticality level of the disaster. Display disaster information using map and graphs.

Figure 5: System interface design 3 - Dashboard user interface - This UI shows the state of the application of dashboard that the end user interacts with. The end user will be able to upload an input text corpus and view the generated summarized text output side by side and also this will give opportunity to the end user to access other user interfaces as well.

Note: The main outcome of the system is a publicly available API which is user can get the data processed by the system. This API is publicly available for outsiders. They can do their work interdependently by interacting with the API.

3.1.2 Hardware interfaces

Compatible smartphone / tablet pc and personal computer / laptop with required specifications - function as means of displaying frontend which is more specific to the service providers who are intended to consume the use of proposed application programming interface (API). The latter mentioned hardware interfaces should have the capability of connecting to the internet which is supposed to be high speed and bandwidth while required web browsers are installed.

3.1.3 Software interfaces.

- NLP Tool Kit – Use for NLP needs such as tokenizing, parsing, identifying named entities.
- Virtual servers (Amazon EC2) in the cloud by Amazon Web Services (AWS) - Amazon Elastic Compute Cloud (Amazon EC2) provides scalable computing capacity in the Amazon Web Services (AWS) cloud. Using Amazon EC2 eliminates the user need to invest in hardware upfront, so the user can develop and deploy applications faster. The user can use Amazon EC2 to launch as many or as few virtual servers as the user need, configure security and networking, and manage storage. Amazon EC2 enables the user to scale up or down to handle changes in requirements or spikes in popularity, reducing the user need to forecast traffic.
- Docker - Docker containers will be used for implementing microservices architecture
- Neo4J graph platforms - Neo4J graph platforms will be used for automatic summarization module.
- MongoDB NoSQL and Firebase realtime databases - NoSQL databases will be used for store training dataset of social media posts.

3.1.4 Communication interfaces

- Wi – Fi Router – Use to connect to the internet.

3.3 Performance requirements

Virtual server instance for microservices should not exceed 8GB RAM and 1TB storage capacity since the size of system effect to the overall performance.

The network calls and API calls should not take more than 5 seconds as the user might not wait any longer than that..

3.4 Design constraints

There are no specific design constraints for this system.

3.5 Software system attributes

3.5.1 Reliability

The system should provide reliable information to the end user and perform constantly well when the user needs the system. The system must be capable of recovering any crashes.

3.5.2 Availability

The system must be available for the user whenever user want to accesses the system. Especially during a disaster time the system must available 24/7.

3.5.3 Security

Since the system going to be hosted in AWS, there are many security features to protect the user.

3.5.4 Maintainability

Maintainability is defined as the probability of performing a successful repair action within a given time. In other words, maintainability measures the ease and speed with which a system can be restored to operational status after a failure occurs. In this system we use component based architecture. Each main function we develop as a separate component. So if any change to be done, it will easy to maintain.

3.6 Other requirements

- Use of IEEE standard colors when designing the interfaces, font sizes/styles
- Positioning of buttons and screen components
- Maximize use of open source tools/technologies

4 Supporting information

4.1 Appendices

REFERENCES

[1] A.T.M Shahjahan and Kutub Uddin Chisty "Social Media Research and Its Effect on Our Society " World Academy of Science, Engineering and Technology International Journal of Information and Communication Engineering Vol:8, No:6, 2014

[2] Muhammad Imran, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. 2015. Processing social media messages in mass emergency: A survey. ACM Comput. Surv. 47, 4, Article 67 (June 2015), 38 pages.DOI: http://dx.doi.org/10.1145/2771588

[3] Bing Liu. Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2012.

[4] Yelena Mejova, Ingmar Weber, and Michael W Macy. Twitter: A Digital Socioscope. Cambridge University Press, 2015.

[5] Ahmed Nagy and Jeannie Stamberger. Crowd sentiment detection during disasters and crises. In Proceedings of the 9th International ISCRAM Conference, pages 1–9, 2012

[6] Dragomir R Radev, Eduard Hovy, and Kathleen McKeown. 2002. Introduction to the special issue on summarization. Computational linguistics 28, 4 (2002), 399–408

[7] Ganesan, K., C. Zhai and J. Han, 2010. Opinosis: A graph-based approach to abstractive summarization of highly redundant opinions. Proceedings of the 23rd International Conference on Computational Linguistics

[8] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D., J. B., and K. Kochut, "Text Summarization Techniques: A Brief Survey," International Journal of Advanced Computer Science and Applications, vol. 8, no. 10, 2017.

[9] Y. J. Kumar, O. S. Goh, H. Basiron, N. H. Choon, and P. C. Suppiah, "A Review on Automatic Text Summarization Approaches," Journal of Computer Science, vol. 12, no. 4, pp. 178–190, Jan. 2016.

[10]https://www.google.lk/search?q=semantic+analysis+social+media&rlz=1C1CHBD_enLK771LK771&source=lnms&tbm=isch&sa=X&ved=0ahUKEwiU98SQ4JPaAhXFtI8KHUUoDD8Q_AUICigB&biw=1366&bih=613#imgrc=sT1wUMlUtiK_GM: