AUTOMATED PROCESSING FOR SOCIAL MEDIA DATA IN A MASS EMERGENCY

Final Report

Perera P.A.D – IT14093210

Project ID: **18-007**

B.Sc. Special (Honors) Degree in Information Technolog Specializing

in Software Engineering

Department of Information Technology

Sri Lanka Institute of Information Technology
Sri Lanka

September 2018

DECLARATION

"I declare that this is my own work and that this dissertation does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning, and to the best of my knowledge and belief, it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to Sri Lanka Institute of Information Technology the nonexclusive right to reproduce and distribute my dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books)."

Signature: ……..... Date: ……....

Perera P.A.D

The above Person has carried out research for the B.Sc. Special (Hons) degree in IT Dissertation under my supervision.

Signature: ……..... Date: ……....

Nuwan Kuruwitarachchi

Senior Lecturer – Department of IT.

ACKNOWLEDGEMENT

We are sincerely thankful to our supervisor Mr. Nuwan Kuruwitarachchi and external supervisor Mr. Raj Prasanna and all the staff members of Sri Lanka Institute of Information Technology for their kind encouragement, dedication constant support given in development of this research study. Also, we would like to remember and grateful to Mr. Jayantha Amaraarachchi Senior Lecturer, including other lecturers who conduct the CDAP module.

The Research study was Developed and carried out under the Comprehensive Design Analysis Project module. Project is done by all the group members, who work hard and higher level of commitment, have given in every single possible manner.

ABSTRACT

The world is full of emergencies caused by natural disasters. In such situations, vast amount of information will be exchanged via social media networks (Facebook, Twitter, etc), official websites and public forums which are dedicated for management of natural disasters. In countries where natural disasters are frequent, disaster management centers and disaster management coordinating units have employed teams to monitor and analyze information to obtain a closer insight into a particular situation. It helps to identify areas that have suffered the most in an emergency, the type of emergency, immediate needs of victims, casualties and infrastructure damages. Manually analysis of overwhelming amount of information is difficult and time consuming. Real-time disaster information is critical for rapid decision-making in response to emergencies. Rest of the document contains overall summary the working progress of research which aims to introduce an effective and productive automated tool to analyze the information generated on social media using modern concepts such as, Semantic Analysis, Natural Language Processing, Machine Learning and Artificial Intelligence. Currently the research work is at preliminary level and has been completed with formation of training labeled datasets and selecting candidate algorithms and methods to be followed. The evaluation stage of candidate algorithms and concepts are yet to be completed.

Table of Contents

LIST OF FIGURES

1. INTRODUCTION

1.1 Background Context

The term “social media” refers to Internet-based applications that enable people to communicate and share resources and information [1]. Due to natural disasters there is an increased communication since people seek to contact family and friends in the disasters zone, and seek information regarding food, shelter and transportation. Social media has played a significant role in disseminating information about these disasters by allowing people to share information and ask for help. Social media are also becoming vital to recovery efforts of crises, when infrastructure must be rebuilt and stress management is critical.

The extensive reach of social networks allows people who are suffering from disasters to rapidly connect with needed resources. There are lot of groups in the most popular social networking sites, allowing individuals involved in various aspects of emergency awareness and preparedness to connect, discuss, and share knowledge in specific fields.

Sentiment analysis of disaster related posts in social media is one of the techniques that could gear up detecting posts for situational awareness. In particular, it is useful to better understand the dynamics of the network including users’ feelings, panics and concerns as it is used to identify polarity of sentiments expressed by users during disaster events to improve decision making. It helps authorities to find answers to their questions and make better decisions regarding the event assistance without paying the cost as the traditional public surveys. Sentiment information could also be used to project the information regarding the devastation and recovery situation and donation requests to the crowd in better ways. Using the results obtained from sentiment analysis, authorities can figure out where they should look for particular information regarding the disaster such as the most affected areas, types of emergency needs [2].

The evaluation of sentiment and the extent of effect of a post or the level of criticality will be important when referring to this kind of scenario. For example if a post contains a request for immediate help that can be considered as a high priority. Intention of this component is to help the first responders with their critical decision making. Most of the time they have to make important decisions rapidly during an emergency. They have to decide what needs to be done, which issue needs to be attended first. So by giving a level of importance high, medium or low, they will be able to get a chance to attend high priorities by spending less time.

1.2 Background Literature

There are several applications around the world which are performing disaster management in social media. They also used NLP techniques, and semantic analysis algorithms. But in those systems, they have use sentiment analysis and semantic analysis algorithm to show the feeling of the peoples. Those systems haven't built mechanism to predict the criticality level of the situation going on.

In this component I used sentiment analysis and semantic algorithms for predict the how critical the situation is. It will really helpful for supporting teams to get the idea about criticality of the situation and to get the correct decisions. Because if the supporting teams know the critical level of the situation and who are the people need help first, they can be ready for the situation before they go to the area where the disaster is happening. That will really helpful for saving people's lives which is a very important thing in disaster management. [2], [3], [4]

Features	Twitris	Senseplace 2	EMERSE	AIDR	Proposed System
Automated Classification	✓	✓	✓	✓	✓
Prioritizing	✗	✗	✗	✗	✓
Criticality Analysis	✗	✗	✗	✗	✓
Accuracy Validation	✗	✗	✗	✗	✓
Text Summarization	✗	✗	✗	✗	✓

Figure 1: Over view of Literature survey

1.3 Research gap

Recently it has given a lot of attention for the usage of social media in various aspects of industries. Since the beginning of the "Information Era" social media are a proven method in digital marketing and advertising it has helped the small businesses to grow. Social media monitoring and regulation is another topic that come to life time which is taking limelight at a slow pace.

Usage of social media for different reasons other than financial benefits is somewhat ignored comparingly. One might think that there are no other goals that can be achieved but simply that is not true. Information is powerful. Proper use of information will result in valuable outcomes. In 2018 March it has revealed that millions of Facebook user information has been used illegally to generate manipulative messages to influence voters in America during elections (Cambridge Analytical incident). User behavior analysis which is used in e-commerce sites to suggest products is another example which shows the power of information and big data.

Most researches conducted in social media usage related to disasters or in other words emergencies have used the popular micro blogging platform Twitter which provides a streaming API to collect publicly available entries (Posts) each maximum length of 140 characters in real time. There is so much information generated elsewhere other than Twitter. For instance, forums blogs dedicated channels for disaster response. Among the techniques used to filter the entries that are related to some specific event provided hash tags (for example #earthquakes) keyword filtering are more common. When identifying a trend (Trend analysis through social media) some systems use word count mechanisms and give the most repeated words as an output. Limitations in streaming API (Maximum number of requests per minute) slows down the process increasing the Latency.

Although mechanisms have provided to filter out entries for a given event, the ability for the existing systems to evaluate the accuracy or the dependability of an entry is limited. Systems that use mix of human interaction and computation power are called "Hybrid systems". They use crowdsourcing to create a model to filter the future entries.

1.4 Research Problem

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.

1.5 Research Objectives

Main Objective

Main goal of this research work is to develop an open source application programming interface for processing social media textual data at presence of a natural disaster to support individuals of natural disaster supporting teams.

Specific Objective

As per reason literature, almost all the disaster management systems used the sentiment analysis technique to show the emotions of the people who face to the disaster situation going on. But In this system we use sentiment analysis algorithm to predict that how critical the situation is. Because if they is a mechanism to predict criticality level of the situation disaster management teams can prioritize the people who need help.

The component to be produced called “Criticality analyzer”. This component predicts the critical level of the situation by analyzing social media post. Basically in disaster management one of main problem is to identify and prioritize the people how need help. Because there may be a people who are in a very critical situation and they need help before others. The main objective of having this component in the system is to identify the people who need help first.

The main purpose of this component is help supporting teams to categorize and prioritize the people who need help first. It will really helpful for supporting teams to get the idea about criticality of the situation and to get the correct decisions.

2 Methodology

Criticality Analyzer

This component will provide the criticality level of the situation going on by analyzing social media post. As an input of this component it will talk the social media posts which are filtered from the upper component of the system. This component developed using NLP techniques with natural language processing toolkit of python language.

There will be a new classification algorithm which can predict criticality level of the situation going on. For develop that algorithm I am maintaining a data set to train the model. The data set is previous social media posts which are describing the situations of the disasters. All post has the criticality level of the situation. By using that data set model will be train. To train the model I use the NLTK Naivebayes algorithm which is very common algorithm use in sentiment analysis. It will consider the features of the words to train the model in each class (high, medium, low)

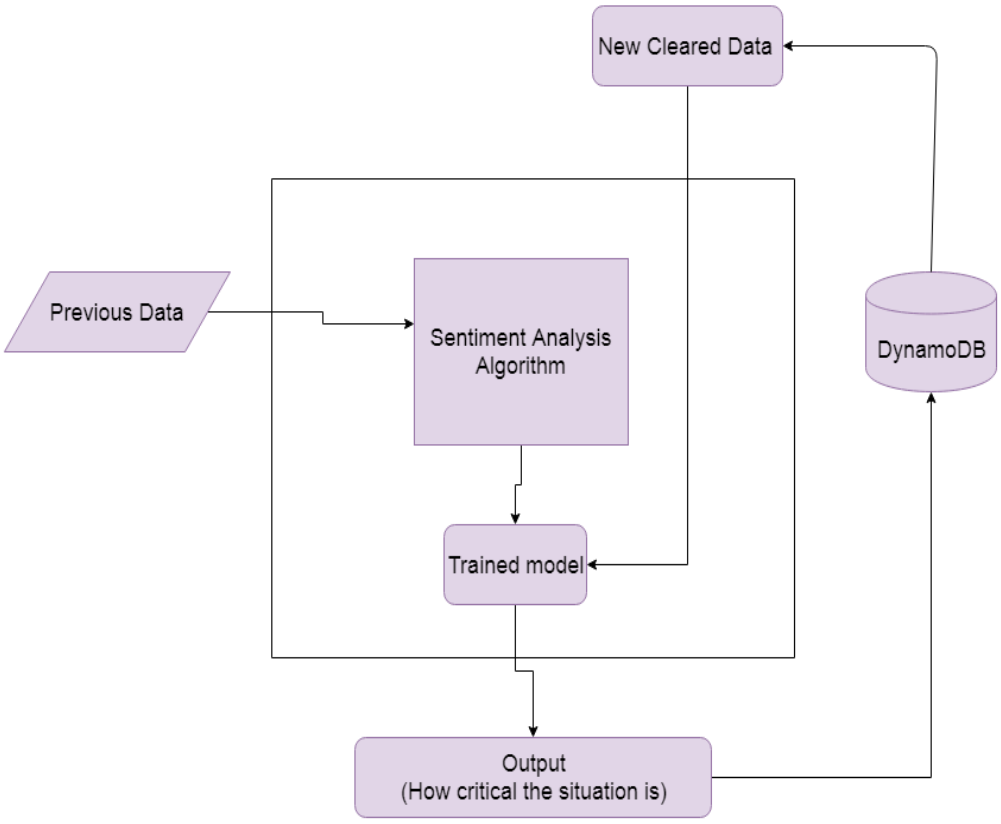

Figure 2: Criticality analysis component

According to the diagram component has sentiment analysis algorithm which is train the model to predict criticality of the disaster situation. It uses previous data which has the label of criticality level to train the model. Inside the algorithm it analysis word features and train the model according to the classes. Then save the model for future use. Using the saved model component will predict the criticality level of new social media posts and update the data base with predicted value.

2.1 Testing & Implementation

Implementation

Basically python NLTK tool kit and naivebayse library used for implementation. Also tried a probability base algorithm and realize that it giving bit low accuracy level. Because of that moved in to naïvebayse sentiment analysis algorithm. In that methodology, according to the data set get large accuracy is also getting higher, which mean accuracy depend on the data set.

While the implementation it can be see what are the word features used to train the model by using word cloud.

Word feature used to train ‘high’ class

Figure 3: word features for ‘high’ class

Word feature used to train ‘medium’ class

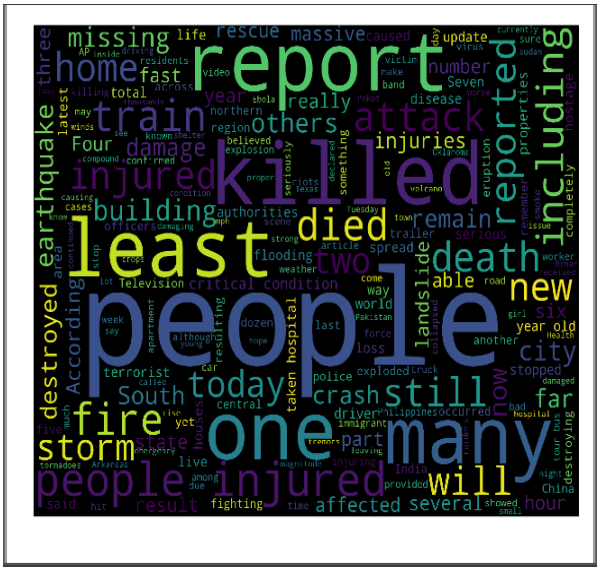

Figure 4: word features for ‘medium’ class

Word feature used to train ‘low’ class

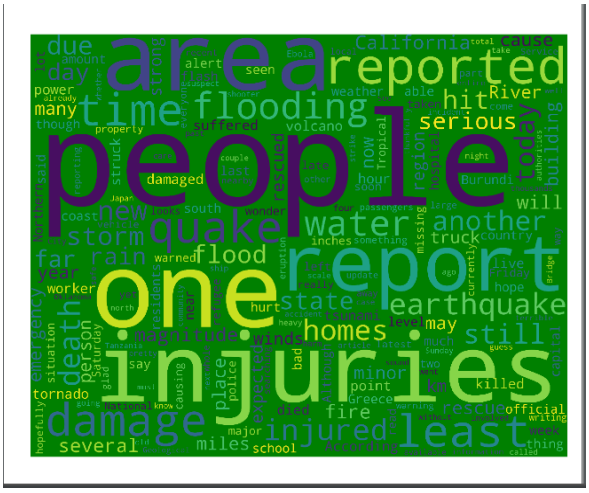

Figure 5: word features for ‘low’ class

Testing

Static testing

- Review
- Inspection
- Checking the documents manually
- walkthrough

Dynamic testing

- Component testing
- Integration testing
- System testing

Unit Testing

Each unit individually to test whether it's fit for use. This used to identify smallest part of problems earlier stages of testing, and most important thing in unit testing is identify the bug than correcting it.

Component Testing

Each component testing done in the application separately also its known as program testing here it found the bugs or defect and take the actions to correct it.

Integration Testing

Each module of the software combined and tested as group.it must be test after unit testing.

System Testing

This is the level of testing where complete software and integrated software is tested. It verified as system whether it meets the requirements. This will ensure the quality level of the system.

3. Results & Discussion

The result of this component is a class label. It can be ‘high’, ‘medium’ or ‘low’. According to the testing accuracy of the results is 60-70 %. The accuracy is depending on the amount of the data in the training data set. By using big data set we can train a more accurate model. As well has it good to have similar amount of data for all the classed in the training data set.

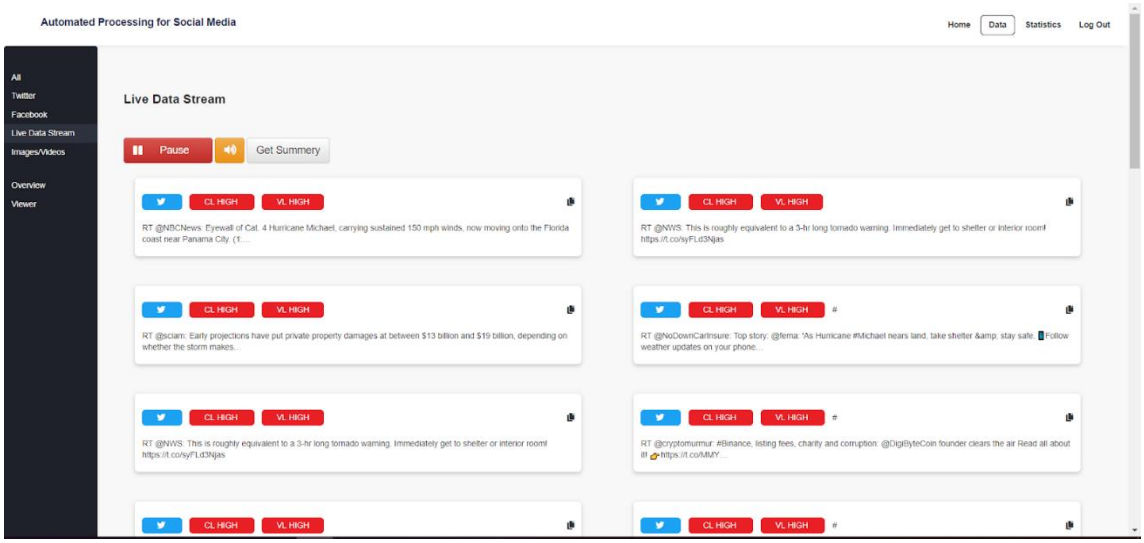

Figure 6: Criticality analyzed live data

CONCLUSIONS

Social media receives overwhelming number of posts during an emergency situation. This research paper proposes a novel process to capable of real time analysis of social media data during mass emergency and generate useful meaning. Upon implementing the process it would allow the decision makers, first responders with actionable information with higher accuracy. Semantic analysis would give overall perspective for the status of the affected society. Post ranking will be focused on identifying reliable social media posts through huge collection of them. Automatic text summarization generates a shorter and concise form of a particular social media post to make it easy for the supporting teams to go through massive datasets of social media posts easily.

Although there exists a potential need for such a system which automates the process most of the organizations are hesitant to appreciate the value. This research project proposes a novel process capable of near real time analysis of social media data during mass emergency and generate useful meaning. Upon implementing the process, it would allow the decision makers, first responders with actionable information with higher dependability. Semantic analysis would give overall perspective for the status of the affected society.

Reference

[1] Bruce R. Lindsay, Social Media and Disasters: Current Uses, Future Options, and Policy Considerations
Sep 6, 2011

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

[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.