AUTOMATED PROCESSING FOR SOCIAL MEDIA DATA IN A MASS EMERGENCY

Gunarathna T.M.T.A

(IT14145476)

Special Honors Degree of Bachelor of Science in Information Technology

Specialized in Software Engineering

Department of Information Technology

Sri Lanka Institute of Information Technology

Sri Lanka

August, 2018

DECLARATION

"I declare that this is my own work and 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 non-exclusive 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: 02.08.2018

The above candidate has carried out research for the Special Honors Degree of Bachelor of Science in Information Technology Dissertation under my supervision.

Signature of the supervisor: Date: 02.08.2018
(Nuwan Kuruwita Arachchi)

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.

ACKNOWLEDGEMENT

The work described in this research paper was carried out as our 4th year research project for the subject Comprehensive Design Analysis Project. The completed final project is the result of combining all the hard work of the group members and the encouragement, support and guidance given by many others. Therefore, it is our duty to express our gratitude to all who gave us the support to complete this major task.

We are deeply indebted to our supervisor Mr. Nuwan Kuruwita Arachchi and our external supervisor Prof. Raj Prasanna, Lecturers of Sri Lanka Institute of Information Technology whose suggestions, constant encouragement and support in the development of this research, particularly for the many stimulating and instructive discussions. We are also extremely grateful to Mr. Jayantha Amararachchi, Senior Lecturer/ Head-SLIIT Centre for Research who gave and confirmed the permission to carry out this research and for all the encouragement and guidance given.

We also wish to thank all our colleagues and friends for all their help, support, interest and valuable advices. Finally, we would like to thank all others whose names are not listed particularly but have given their support in many ways and encouraged us to make this a success.

Table of contents

List of Figures

This will include list of figure which are in the body of content of this document.

List of Tables

List of abbreviations

POS	Part of Speech
NLP	Natural Language Processing
NLG	Natural Language Generation
AWS	Amazon Web Services
API	Application Programming Interface

Table 1.0 Definitions, Acronyms, and Abbreviations

1. Introduction

1.1. Background context

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 focussing 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.

This document will be mainly focusing on developing an automatic text summarization component for processing social media posts which are emerging dramatically by the time and going parallel to an event of mass emergency where the current or ongoing status and critical informations are hidden inside of the respective event occured and generating respective summaries by identifying the core meaning of a particular post and extracting the summarized content over the bulk of social media posts while preserving or maintaining the core meaning of a particular post without damaging the actual core meaning of it.

1.3. Research problem

In an emergency, the social media which are dedicated for posting the current or on-going status of natural disasters publish emerging number of huge datasets which are incapable to process them manually. Because of that the research aspect of summarizing social media posts will be focused on identifying the core meaning of a particular post and extracting the summarized content over the bulk of social media posts. It will be crucial to be responsible to

maintain the core meaning of a particular post without damaging the actual meaning of it. This volume of text is an invaluable source of information and knowledge which needs to be effectively summarized to be useful [8] for the natural disaster supporting teams to take actions timely, effectively and efficiently. Summarization helps to gain required information in less time.

Automatic text summarization is very challenging, because when we as humans summarize a piece of text, we usually read it entirely to develop our understanding, and then write a summary highlighting its main points. Since computers lack human knowledge and language capability, it makes automatic text summarization a very difficult and non-trivial task[8].According to Radef et al. [6] a summary is defined as "a text that is produced from one or more texts, that conveys important information in the original text(s), and that is no longer than half of the original text(s) and usually, significantly less than that". Automatic text summarization is the task of producing a concise and fluent summary while preserving key information content and overall meaning. [8]

Text summarization approaches can be broadly divided into two groups: extractive summarization and abstractive summarization. Extractive summarizations extract important sentences or phrases from the original documents and group them to produce a summary without changing the original text. Abstractive summarization consists of understanding the source text by using linguistic method to interpret and examine the text. Abstractive methods need a deeper analysis of the text. These methods have the ability to generate new sentences, which improves the focus of a summary, reduce its redundancy and keeps a good compression rate.

Summaries produced by extractive summarization techniques are constructed by choosing a subset of sentences in the original text which is being the input for the text summarizer. The chosen sentences are supposed to be most important sentences of the input text corpus. According to the context of the research, the input text could be a social medial post with follow up comments. Extractive methods tend to be verbose and this is especially problematic as produced summaries should not be lengthy and be readable for natural disaster supporting teams. Thus, an informative and concise abstractive summary would be a better solution.

Existing work in abstractive summarization has been quite limited and can be categorized into two categories: (1) approaches using prior knowledge and (2) approaches using Natural Language Generation (NLG) systems. The first category of work requires considerable amount of manual effort to define schema such as frames and templates that can be filled with the use of information extraction techniques. These systems were mainly used to summarize news articles. The second category of work uses deeper NLP analysis with special techniques for text regeneration. Both approaches either heavily rely on manual effort or are domain dependent. [7]

Because of the latter mention failures of using extractive and abstractive summarization for automatic text summarization of social media posts, a novel flexible summarization framework, Opinosis, can be proposed. That uses graphs to produce abstractive summaries of highly redundant opinions.

1.3. Research gap

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.

System name Data; example capabilities	Reference and URL
Twitris Twitter; semantic enrichment, classify automatically, geotag	[Sheth et al. 2010; Purohit and Sheth 2013] http://twitris.knoesis.org/
SensePlace2 Twitter; geotag, visualize heat-maps based on geotags	[MacEachren et al. 2011] http://www.geovista.psu.edu/SensePlace2/
EMERSE: Enhanced Messaging for the Emergency Response Sector Twitter and SMS; machine-translate, classify automatically, alerts	[Caragea et al. 2011] http://emerse.ist.psu.edu/
ESA: Emergency Situation Awareness Twitter; detect bursts, classify, cluster, geotag	[Yin et al. 2012; Power et al. 2014] https://esa.csiro.au/
Twitcident Twitter and TwitPic; semantic enrichment, classify	[Abel et al. 2012] http://wis.ewi.tudelft.nl/twitcident/
CrisisTracker Twitter; cluster, annotate manually	[Rogstadius et al. 2013] https://github.com/jakobrogstadius/crisistracker
Tweedr Twitter; classify automatically, extract information, geotag	[Ashktorab et al. 2014] https://github.com/dssg/tweedr
AIDR: Artificial Intelligence for Disaster Response Twitter; annotate manually, classify automatically	[Imran et al. 2014a] http://aidr.qcri.org/

Table 1.1 **Available Systems and where to find them Source [2]**

System/tool	Approach	Event types	Real-time	Query type	Spatio-temporal	Sub-events	Reference
Twitter Monitor	burst detection	open domain	yes	open	no	no	[Mathioudakis and Koudas 2010]
TwitInfo	burst detection	earthquakes+	yes	kw	spatial	yes	[Marcus et al. 2011]
Twevent	burst segment detection	open domain	yes	open	no	no	[Li et al. 2012b]
TEDAS	supervised classification	crime/disasters	no	kw	yes	no	[Li et al. 2012a]
LeadLine	burst detection	open domain	no	kw	yes	no	[Dou et al. 2012]
TwiCal	supervised classification	conflicts/politics	no	open	temporal	no	[Ritter et al. 2012]
Tweet4act	dictionaries	disasters	yes	kw	no	no	[Chowdhury et al. 2013]
ESA	burst detection	open domain	yes	kw	spatial	no	[Robinson et al. 2013a]

The table includes the types of events for which the tool is built (open domain or specific), Whether detection is performed in real time, the type of query (open or "kw" = keyword-based), and whether it has spatio-temporal or subevent detection capabilities. Sorted by publication year.

Table 1.2 **Features of existing systems, Source [2]**

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

Table 1.3 **Features of existing systems and proposed system comparison.**

1.4. Research objectives

The general idea behind the project is to help organizations like DMC, humanitarian organizations, volunteers and services to get situational information more quickly and more accurately. Finding informative and salient posts among a large amount of posts with noise is hard and time consuming, due to the fact that they are scarce.The ability to analyze these massive amounts of information to be useful, it must be faster. (Low Latency). Although there exists many solutions which provides trends as word clouds through finding the frequency of recurring words they are not useful in disaster response.To produce a viable

solution for the lack of information during an emergency, the proposed system would be equipped with four main components.

2. Methodology

2.1. Methodology

Automatic text summarization module will consume a set of social media posts as input and they will be tokenized by sentences and obtain a collection of sentences. Then the later mentioned collection of sentences will be further iterate through each sentence one by one and gone through the process of word tokenizing to retrieve a collection of word tokens for each sentence respectively. As the next step of data preprocessing, the collection of word tokens taken from the previous immediate phase will be gone through of adding POS tags accordingly and each word will be tagged with the sentence identifier and position identifier for identifying which sentence a particular word was in and for identifying where a word was placed in the sentence body. By the end of this process, data preprocessing phase will be finished.

In the next phase of process, all the sentences which have to be summarized will be taken and map on graph database to build a word graph. The built word graph will have only one node for a particular word even though a word has been repeated multiple times in the same sentence or multiple sentences in the input text. The ventex are referred as nodes in the context of Neo4J graph databases. The edges will represent as the relationships in between two words. For developing word graph, a common relationship variable will be used and each relationship will be placed in between two nodes according to the pattern of nodes have been distributed with the direction from the start node to end node. The relationship variable value will be used only for identifying a collection of sentences which are referred to a particular input dataset.

Methodology part will discuss the methodology which has been used for developing automatic text summarization module as above.

2.2. Testing & implementation

Testing and implementation part will discuss about test cases which will be used for testing and corresponding results and the implementation details. This would contain the details of supporting libraries and software solutions, apis and frameworks which have been used. This will not be added to this document as this module is under development.

2.3. Research findings

Research findings part will discuss about facts that has been found while doing the research. This will not be added to this document as this module is under development. This will not be added to this document as this module is under development.

3. Results & Discussion

The below result and discussion part will discuss about the final result which will be given by the module for the user and how much it will be accurate enough for taking better decisions on them. This will not be added to this document as this module is under development.

3.1. Results

3.2. Discussion

4. Conclusion

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.

5. 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:

6. Glossary

7. Appendices