AUTOMATED PROCESSING FOR SOCIAL MEDIA DATA IN A MASS EMERGENCY

Kavindu Chinthana Kodithuwakku

(IT14136252)

Degree of Bachelor of Science

Department of Information Technology

Sri Lanka Institute of Information Technology
Sri Lanka

September 2018

AUTOMATED PROCESSING FOR SOCIAL MEDIA DATA IN A MASS EMERGENCY

Kavindu Chinthana Kodithuwakku

(IT14136252)

Dissertation submitted in partial fulfillment of the requirements for the degree of Science

Department of Information Technology

Sri Lanka Institute of Information Technology

September 2018

Declaration

I declare that this is my own work and this dissertation1 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:

The above candidate has carried out research for the B.Sc. Dissertation under my supervision.

Signature of the supervisor: Date

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 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.

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

I would like to express my deep gratitude to supervisor Mr. Nuwan Kuruwitaarachchi and external supervisor Dr Raj Prasanna for their patient guidance, enthusiastic encouragement and useful critiques of this research work. I would also like to thank my fellow colleagues for their dedication and assistance in keeping the research on schedule.

My grateful thanks are also extended 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. Finally, I would like to thank all others whose names are not listed particularly but have given their support in many ways to make this a success

Table of Contents

Definitions, Acronyms, and Abbreviations

Term	Definition
API	Application Programming Interface.
JSON	JavaScript Object Notation.
AWS	Amazon Web Services.
DMC	Disaster Management Center.
REST	Representational State Transfer.
RDS	Relational Database Service.
S3	Simple Storage Service.
Crowdsourcing	The practice of obtaining information or input into a task or project by enlisting the services of many people, either paid or unpaid, typically via the Internet.
VCS	Version Control System.
Stakeholder	Any person with an interest in the project.
Open-source software	Software for which the code is freely available for use and research
Tweet	Tweet is a buzz word used to refer to a text message which contains maximum 140 characters.

1 Introduction

Social media has become a vital role in current society. It is the streamline communication medium to share not only textual but also pictorial and vocal information. Millions of users have made it a daily practice to surf through social media to find and share information. Researchers have found a burst of information generated during and aftermath of a mass emergency through social media platforms as a useful resource to get an insight into the situation which is also known as situational awareness.

In an emergency such as earthquake there can be entries which contain facts related to the extent of the affect, infrastructure damages, casualties, donations available or requested, the kinds of necessities required, for instance food or water. Trustworthiness and reliability of incoming or extracted information is a matter that is yet to be solved. Formal first responders, disaster managers humanitarian organizations, NGO's, general public, local police, area firefighters are few of the stakeholders who benefit from such information. Moreover, different types of information sought by different stakeholders. For example, humanitarian organizations might be interested in potential donators and the requirements of the victims while first responders and disaster managers are interested in the infrastructure damages and casualties.

To make use of this information generated in some developed countries have employed teams to analyses and generate reports which are useful for first responders to make time critical life and death decisions. Reading and manually analyzing continuous streams of unorganized textual information generated at an increased rate is a tedious and stressful task. Responding to each incoming message timely might require increase of employees. Therefore, processing social media data during a mass emergency is a matter that requires new ways of data processing reducing the amount of information examined by the humans.

Although there exists a potential need to use social media platforms to serve this kind of specific purposes as mentioned above, some countries, organizations have failed to or hesitant to appreciate the benefits and the widespread usage of social media to improve situational awareness. One of main objective of this research is to provide evidence for the usage of a system which uses social media as its base source of information.

1.1 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 Analytica 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 microblogging platform Twitter (Figure 2.2) 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 hashtags (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.

1.2 Research Problem

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. In recent literature (refer APPENDIX B) it was clear that one way of getting information quickly is to analyze the huge throughput of social media activity during an emergency. But there are major problems with this approach.

- Scarcity of entries (Messages, Posts, Comments) with relevant information (Noise factor). A lot of entries show emotions or prayers.

- The ability to analyze these massive amounts of information to be useful, it must be faster. (Low Latency)

Although 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 led 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.

1.3 Objectives

Finding trending topics.

Identifying trending topics helps in detecting a disaster happening near real time. It is commonplace to witness a high traffic in social media during an disaster.

Extract data from social media to identify the relevant data.

The amount of social media post threads would be dramatically increased up to tens of thousands throughout the duration and aftermath of a natural disaster. Processing these entries in a timely manner is a reoccurring matter in the disaster management aspect as this information get obsolete quickly. Although finding disaster relevant posts is important. It works similar to an email client for instance Gmail, which clusters the similar emails as spams, social media, promotions.

Categorize (Classify) the relevant data into meaningful categories.

Categorizing and prioritizing is a critical sub component in the flow of automated processing for social media entries. It works as a middle layer between other components and social media, providing only the information that are useful. Entries (Posts) are categorized into main categories, namely informative, personal, infrastructure damages, donations, requests for help. While understanding the meaning of a text, identifying meaningful (relevant) text is more important. During a natural disaster relevant information are generated by eyewitnesses, people who are affected (victims), government and humanitarian groups/ agencies, people who wants to donate and people who are willing to take part in voluntary services.

1.4 Goals

- Provide a better tool for disaster management.
- Reduce the disaster response time.
- Filter the relevant entries and provide them to the main flow of the process.

1.5 Benefits

- Ability to extract information quickly.
- Identify trends.
- Ability to visualize data.
- Improved response time.
- Open source API for others to develop their own GUIs.

1.6 Background

During the Background research similar systems were found which were using twitter as their main data source. (Figure 2.1). [1] The most popular one among those existing system is known as the AIDR (Artificial Intelligence for Disaster Response). It uses few categories to classify the incoming stream of data namely casualties, infrastructure damages and donations. Furthermore, it uses Crowdsourcing to train a model during an emergency which leads to response latency.

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/

Figure 2.1 **Example Systems Available**

Proposed tool /system has

- Extended number of categories

- Prioritization of entries in each category.

And it doesn't use crowdsourcing to train a model.to minimize the response latency.

3 Methodology

Intention of this section is to present the steps followed to build the proposed tool. Rest of the section will provide an overview of the fundamental approach which was followed throughout the research. Subsection 3.1 General Approach describes how the research was conducted in much more general manner.

3.1 General Approach

Defining the problem.

Discussions were subsequently held with members and supervisors to properly define and assess problem to identify the scope. Initial problem was reduced to match with the time given and added extra functionality to match the required complexity. Next step was to identify the potential stakeholders.

Background Research / Literature Review

Once the problem was defined and stakeholders were identified, background research was required to distinguish what exists already, what are the ongoing projects and what they are lacking. This step was important as it gave us an overview of how other researchers has conducted their and what their outcome was.

Identifying the outcomes, defining objectives and goals.

After analyzing what needs to be implemented outcomes objectives and goals were defined. Next schedule was developed which contains the major milestones.

Time to Time Verification

Regular discussions with members and supervisors were conducted to verify the followed path and the methodology was valid as decided in prior stages. When unexpected problems occurred during the process they were discussed and resolved as they came.

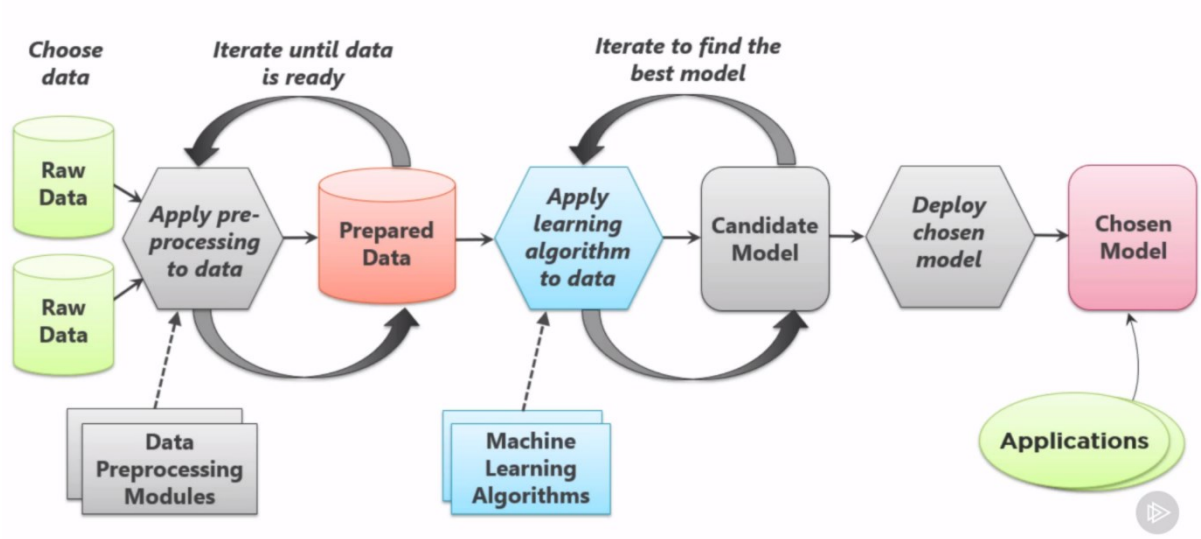
Documentation

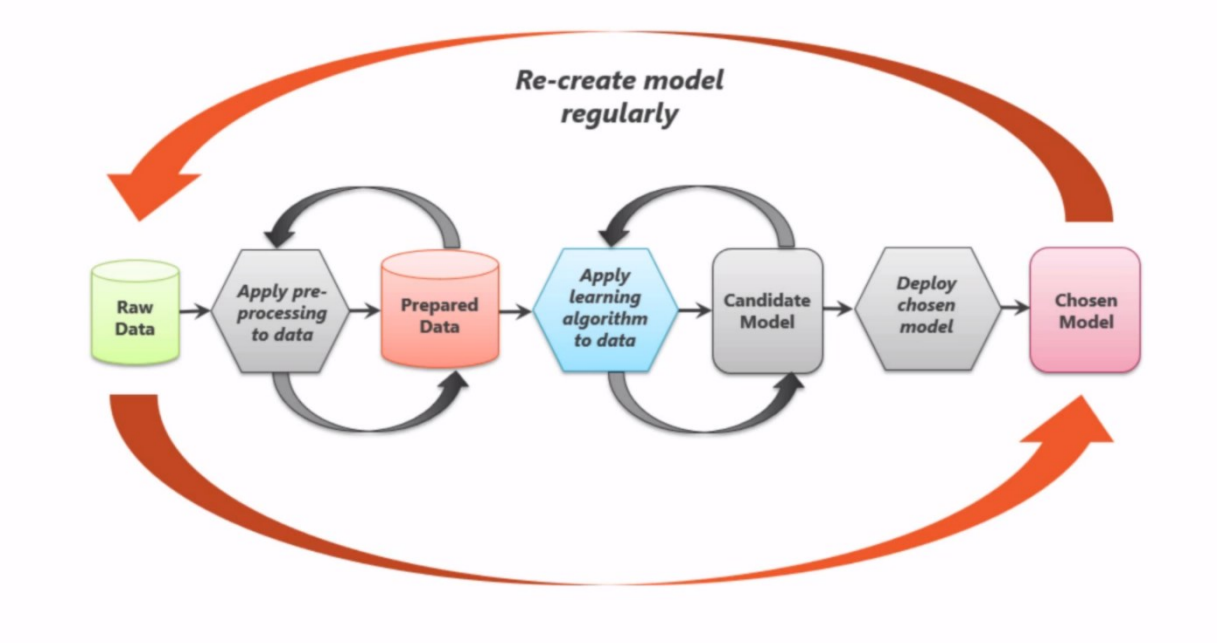
More high-level journal was maintained throughout the research including daily work done. SRS document containing all the requirements and a proposal with the proposed solution was produced as main design documents.

3.2 Specific Approach

3.1.1 Machine Learning

Machine learning is the process of finding patterns in data to predict potential outcome. Following process was followed when applying machine learning algorithms to the given problem. For machine learning it requires a lot of data, computing power and effective machine learning algorithms. All of these are now available more than ever thanks to the information era. Applying machine learning techniques to Business purpose is a common real-world use case. Some problems require domain knowledge. But in this case the required domain knowledge was less.





3.1.2 Choosing the right data /Collecting data

Since machine learning is so much dependent on data. Usually the data collected in the csv format. A good clean dataset usually results in a better trained model with higher accuracy. There are several methods which can be used to preprocess data.

Real world data are generally

- **Incomplete**: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
- **Noisy**: containing errors or outliers
- **Inconsistent**: containing discrepancies in codes or names

Tasks in data preprocessing

- **Data cleaning**: fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies.
- **Data integration**: using multiple databases, data cubes, or files.
- **Data transformation**: normalization and aggregation.
- **Data reduction**: reducing the volume but producing the same or similar analytical results.
- **Data discretization**: part of data reduction, replacing numerical attributes with nominal ones.

Data transformation

1. Normalization:

- Scaling attribute values to fall within a specified range. Example: to transform V in [min, max] to V' in [0,1], apply V'=(V-Min)/(Max-Min)
- Scaling by using mean and standard deviation (useful when min and max are unknown or when there are outliers): V'=(V-Mean)/Standard Deviation

2. Aggregation: moving up in the concept hierarchy on numeric attributes.
3. Generalization: moving up in the concept hierarchy on nominal attributes.
4. Attribute construction: replacing or adding new attributes inferred by existing attributes

Data reduction

1. **Reducing the number of attributes**

- Data cube aggregation: applying roll-up, slice or dice operations.
- Removing irrelevant attributes: attribute selection (filtering and wrapper methods), searching the attribute space.
- Principle component analysis (numeric attributes only): searching for a lower dimensional space that can best represent the data

2. **Reducing the number of attribute values**

3.1.3 Developing a hypothesis

3.1.4 Training a model

Learning requires

- Identifying patterns
- Recognizing those patterns when you see them again

Same applies for a machine to learn it requires past data. Depending on the historical data algorithm builds a model which will be used to label (in supervised learning) unseen new data.

3.1.5 Evaluating the model

Cross Validation

Cross validation attempts to avoid overfitting (training on and predicting the same datapoint) while still producing a prediction for each observation dataset. This is accomplished by systematically hiding different subsets of the data while training a set of models. After training, each model predicts on the subset that had been hidden to it, emulating multiple train-test splits. When done correctly, every observation will have a 'fair' corresponding prediction.

4 Implementation

This section will provide an overview of the steps taken during the implementation of the system in developer perspective.

4.1 Dataset

	choose_one	text
2	Relevant	Just happened a terrible car crash
3	Relevant	Our Deeds are the Reason of this #earthquake May ALLAH Forgive us all
4	Relevant	Heard about #earthquake is different cities, stay safe everyone.
5	Relevant	there is a forest fire at spot pond, geese are fleeing across the street, I cannot save them all
6	Relevant	Forest fire near La Ronge Sask. Canada
7	Relevant	All residents asked to 'shelter in place' are being notified by officers. No other evacuation or shelter in place orders are expected
8	Relevant	13,000 people receive #wildfires evacuation orders in California
9	Relevant	Just got sent this photo from Ruby #Alaska as smoke from #wildfires pours into a school
10	Relevant	#RockyFire Update => California Hwy. 20 closed in both directions due to Lake County fire - #CAfire #wildfires
11	Relevant	Apocalypse lighting. #Spokane #wildfires
12	Relevant	#flood #disaster Heavy rain causes flash flooding of streets in Manitou, Colorado Springs areas
13	Relevant	Typhoon Soudelor kills 28 in China and Taiwan
14	Relevant	We're shaking...It's an earthquake
15	Relevant	I'm on top of the hill and I can see a fire in the woods...
16	Relevant	There's an emergency evacuation happening now in the building across the street
17	Relevant	I'm afraid that the tornado is coming to our area...
18	Relevant	Three people died from the heat wave so far
19	Relevant	Haha South Tampa is getting flooded hah- WAIT A SECOND I LIVE IN SOUTH TAMPA WHAT AM I GONNA DO WHAT AM I GONNA DO FVCK #flooding
20	Relevant	#raining #flooding #Florida #TampaBay #Tampa 18 or 19 days. I've lost count
21	Relevant	#Flood in Bago Myanmar #We arrived Bago
22	Relevant	Damage to school bus on 80 in multi car crash #BREAKING
23	Not Relevant	They'd probably still show more life than Arsenal did yesterday, eh? EH?
24	Not Relevant	Hey! How are you?
25	Not Relevant	What's up man?
26	Not Relevant	I love fruits
27	Not Relevant	Summer is lovely
28	Not Relevant	My car is so fast

4.2 Algorithm

Naive Bayes classifier

Naive Bayes is a family of algorithms based on applying Bayes theorem with a strong(naive) assumption, that every feature is independent of the others, in order to predict the category of

a given sample. They are probabilistic classifiers, therefore will calculate the probability of each category using Bayes theorem, and the category with the highest probability will be output. Naive Bayes classifiers have been successfully applied to many domains, particularly Natural Language Processing (NLP).

Latent Dirichlet Allocation (LDA)

LDA or latent Dirichlet allocation is a "generative probabilistic model" of a collection of composites made up of parts. In terms of topic modeling, the composites are documents and the parts are words and/or phrases (n-grams).

If you view the number of topics as number of clusters and the probabilities as the proportion of cluster membership, then using LDA is a way of soft clustering your composites and parts. Contrast this with say k-means where each entity can only belong to one cluster. These fuzzy memberships provide a more nuanced way of recommending similar items, finding duplicates, or discovering user profiles/personas.

4.3 Tools and Technologies

Tools

- Notepad++
- Aws Developer Console
- Putty
- Visual Studio Code (IDE)
- PyCharm (Python IDE)

Technologies

- Python
- NoSQL Database - DynamoDB
- Angular 5
- JavaScript

- HTML 5
- SCSS/CSS

5 Product/ Solution Overview

5.1 Functionality

- **Extract data from social media to identify the relevant data.**

The system would use APIs mentioned in section 2.1.4 as its main data source. This is first step which other functions would depend on.

- **Identify the relevant data from the extracted data.**

This would extract/ filter only the entries that are relevant to an emergency.

- **Categorize (Classify) the relevant data into meaningful categories.**

This is where the entries that were filtered in the previous step would be categorized into groups with similar content.

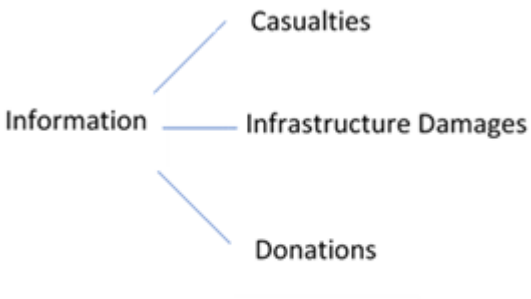


Figure 2.1.0 **Basic types of categories**

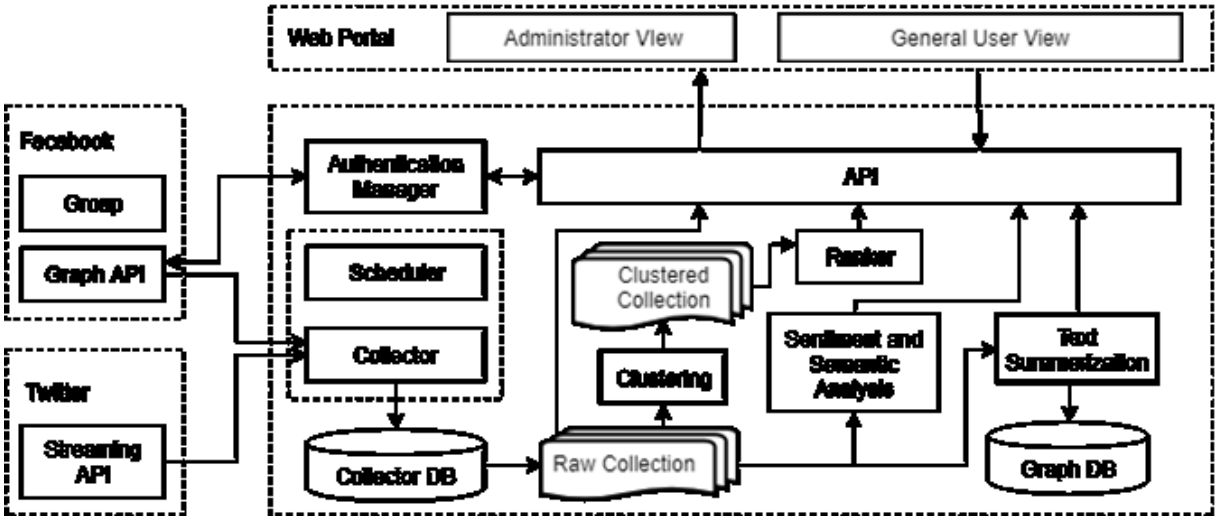
- **Prioritize the categorized (classified) entries.**

Response time is a critical factor when it comes to an emergency. Responsible parties must know where to start and where to end in other words how to prioritize events/ actions during an emergency.

- **Allow users to link their Facebook pages to the system to retrieve data.**

For example, if a certain DMC has an official Facebook page then they should be able to connect their page to the system easily to allow the system to retrieve data.

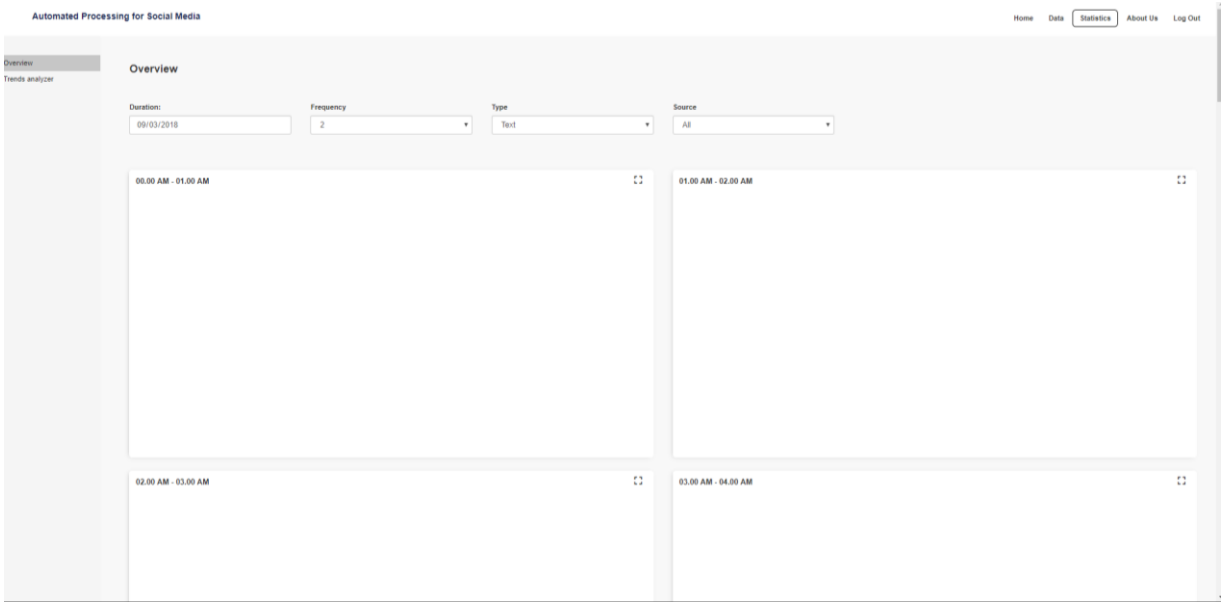
5.2 Architecture

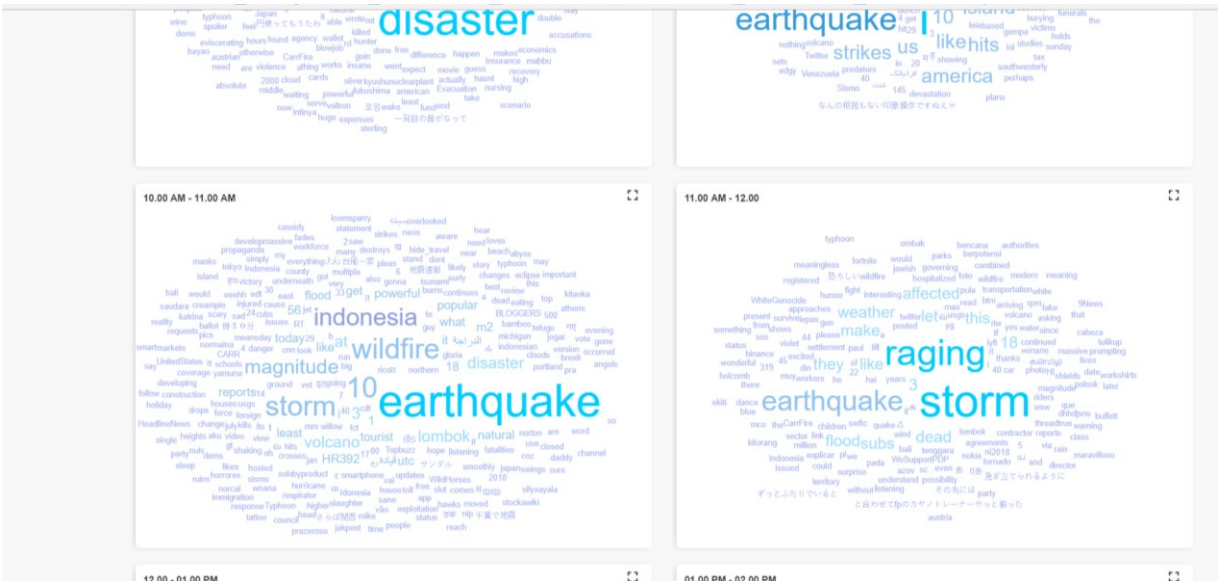


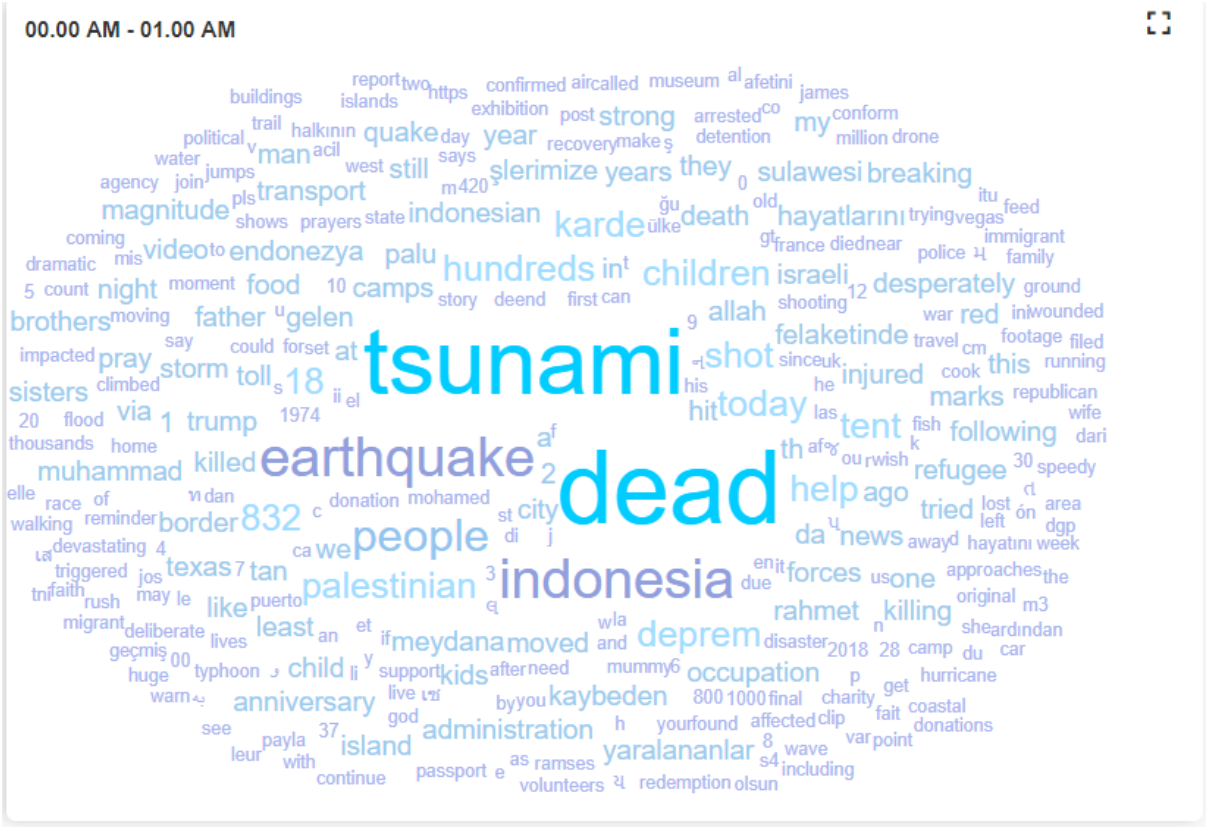
5.3 User interfaces

Provided user interfaces are only for the purpose of giving the stakeholders a basic idea how the features would be organized. All the user interfaces are bound to change as the requirements change. Going through several iterations of demonstrations / prototypes final interfaces would be designed.

Figure 3.1.1 Categorized data view.







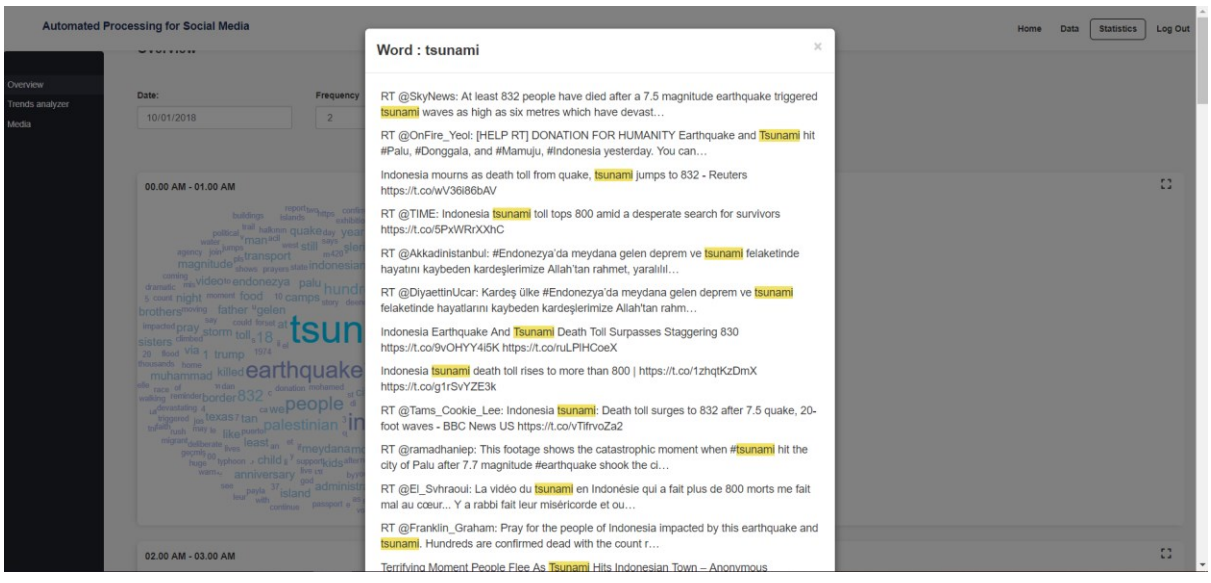


Figure 3.1.2 Charts view

User interfaces are an optional feature of the subcomponent since the research/project is conducted on underlining functionality. Although for commercial purposes and to improve user interaction. Interfaces would be designed to serve the purpose of data visualization.
The purpose of providing an open API is to provide interested parties to develop their own tools to visualize the data published through the API. Basic user interfaces will be provided to match the needs of user's data visualization requirements. All user interfaces are described in detail in section 3.

5.4 Testing

Yet to be done

5.5 External Software interfaces

Twitter Streaming API

Twitter is a free microblogging social networking platform which allows its registered users to publish 140 characters of maximum length messages. To weave tweets into a conversation thread or connect them to a general topic, members can add hashtags to a keyword in their post. The hashtag which acts like a meta tag, is expressed as #keyword. It uses special tags called Hashtags to annotate/ filter text messages (which are called Tweets).
Twitter has exposed an API for developers/businesses to create applications on top of the public information available in Twitter. It provides and extensive documentation which can be found at https://developer.twitter.com/en/docs. There is some limitation to the API when it come to free usage. Part of the API called Streaming API which will be used in the project as a source of data. For more information about the Twitter API refer to APPENDIX A.

Facebook Graph API

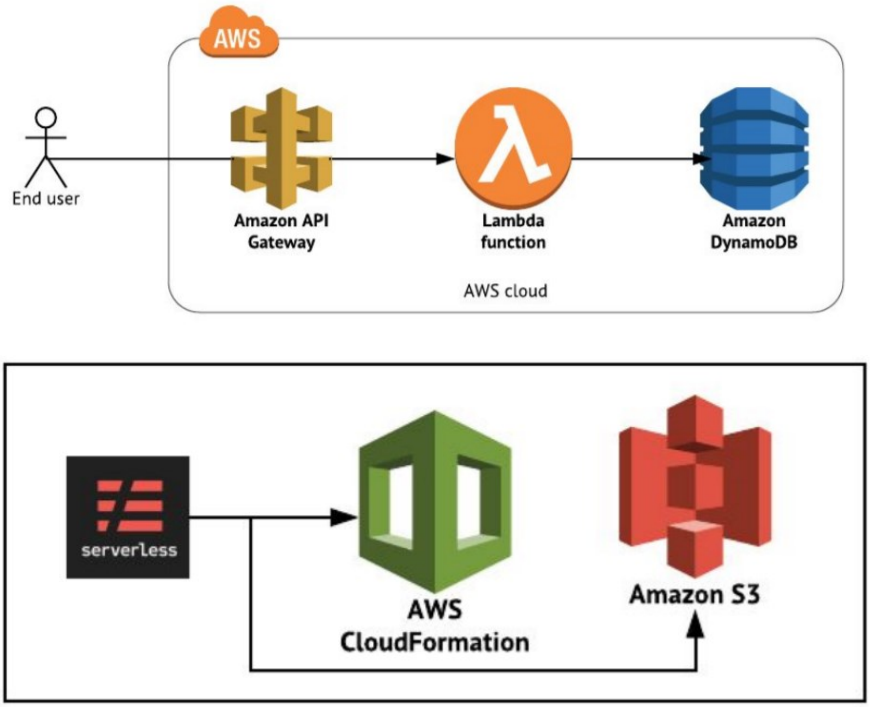
Facebook Graph API is used to extract data only from official Facebook groups and chat messages which are owned by DMC's and other organizations due to the restrictions on Facebook Policy.

Amazon Web Services

To provide backend infrastructure AWS will be used. It is collection of cloud services which provides the facility of pay as you go. Using AWS, the cost to maintain infrastructure physically can be reduced drastically. These will be required when the final product is ready to be used in real life situations.

Specifically.

- EC2 instances: to be used as a cloud server.
- Amazon S3: As a storage to store images/ videos and to host web pages.
- Lambda functions: computations to be done when necessary.
- DynamoDB.



5.6 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 an expert in technology. Anyone with basic knowledge and understanding of domain should be able to operate the system.

2.4.1 Data Usage policies.

It is important to realize that Data Usage Policies are vital when it comes to applications build upon user data. The proposed system has to comply to the terms and agreements provided in the API usage policies by Twitter and Facebook.

Twitter API usage policy.
https://developer.twitter.com/en/developer-terms/agreement-and-policy
Facebook Graph API usage policy.
https://developers.facebook.com/policy

In addition, internal data usage policy contains

- Data should not be shared across different users.
- User sensitive data should be encrypted.

6 Results/ Discussion

	precision	recall	f1-score	support
Not Relevant	0.90	0.98	0.94	4118
Relevant	0.97	0.85	0.91	3156
avg / total	0.93	0.93	0.92	7274

7 Conclusion

Social media receives overwhelming number of posts during an emergency. 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

8 Future Works

Validating information for accuracy can be achieved by more advanced. For example, by acquiring geo location of posts (where it was generated) and comparing it with the location of the actual incident took place. Taking data from more than one source is another way. Optimizing algorithms for the lack of data and finding proper solutions for domain specific challenges such as identifying duplicated information, fake rumors. Other than that, we wish to image filter for identifying disaster related images.

Appendices

APPENDIX A – Twitter JSON object

Figure 4.1 show a sample twitter post, more commonly known as a tweet. Figure 4.2 shows the relevant JSON object for it.

Figure 4.1. Sample Tweet [2]

```
"tweet": {
  "created_at": "Thu Apr 06 15:24:15 +0000 2017",
  "id_str": "850006245121695744",
  "text": "1\/ Today we\u2019re sharing our vision for the future of the
          Twitter API platform!\nhttps:\/\/t.co\/XweGngmxlP",
  "user": {
    "id": 2244994945,
    "name": "Twitter Dev",
    "screen_name": "TwitterDev",
    "location": "Internet",
    "url": "https:\/\/dev.twitter.com\/",
    "description": "Your official source for Twitter Platform news, updates & events.
                    Need technical help? Visit https:\/\/twittercommunity.com\/ \u2328\ufe0f #TapIntoTwitter"
  },
  "place": {

  },
  "entities": {
    "hashtags": [

    ],
    "urls": [
      {
        "url": "https:\/\/t.co\/XweGngmxlP",
        "unwound": {
          "url": "https:\/\/cards.twitter.com\/cards\/18ce53wgo4h\/3xo1c",
          "title": "Building the Future of the Twitter API Platform"
        }
      }
    ],
    "user_mentions": [

    ]
  }
}
```


Figure 4.2 Json Object returned by the Twitter Streaming API for the sample Tweet Figure 4.1 [2]

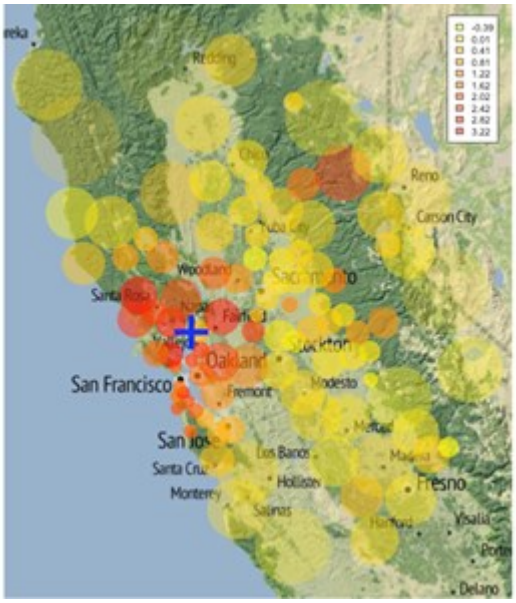
Key	**Description**
tweet	The root of the JSON object.
created_at	Timestamp of the post.
id_str	Tweet id.
text	Body of the tweet.

user	Contains details about the user who created the tweet.

Table 4.1 Description for Twitter API streaming JSON object keys

APPENDIX B – Literature

By Ariel Evnine, Andreas Gros and Aude Hofleitner - Facebook Data Science team [6]



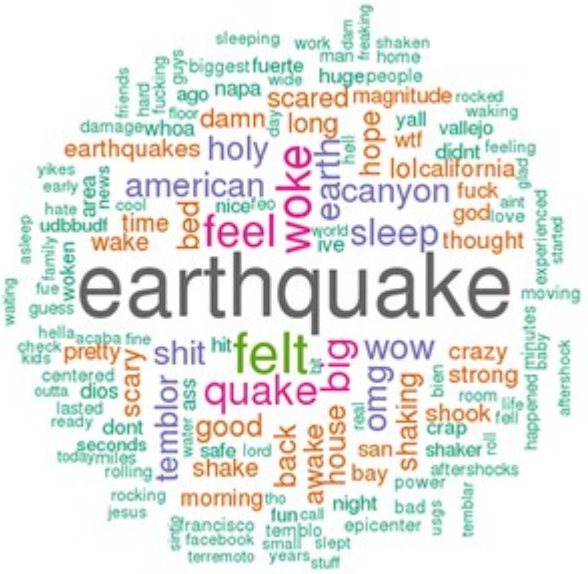
On Sunday August 24th, 3:20 a.m Pacific time, an earthquake of magnitude 6.0 occurred in the Bay Area, 3.7 miles (6.0 km) northwest of American Canyon near the West Napa Fault. It was the largest earthquake in the Bay Area since the 1989 Loma Prieta earthquake.

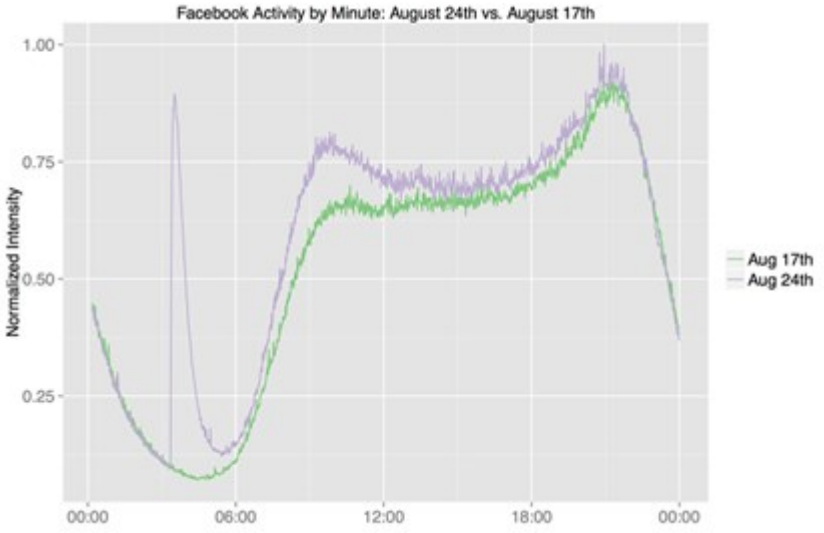
During a crisis, people turn to Facebook to stay connected to their friends and family. They use it to receive social support and keep the people they care about informed on how they are doing.

The map above shows the relative difference in activity on Facebook between the 24th of August, 3:21 a.m. and 3:26 a.m., and the same time period one week earlier.
For visualization, we cluster together nearby cities which showed similar changes in activity. The color represents the percent variation in activity (red: largest activity, yellow: lowest activity). The size represents the area covered by the cluster. The blue cross indicates the location of the epicenter.

We looked at all public posts from people within 300 km from the epicenter during the hour following the earthquake on August 24th. The following word cloud shows the frequency of words used. “Earthquake” comes as the most commonly used word, also very common are

"American Canyon", which is the location where the earthquake occurred. Happening in the middle of the night, the earthquake had a strong effect on people's sleep ("wake", "sleep"). People also express their fear and general feelings and enquire about friends and family.

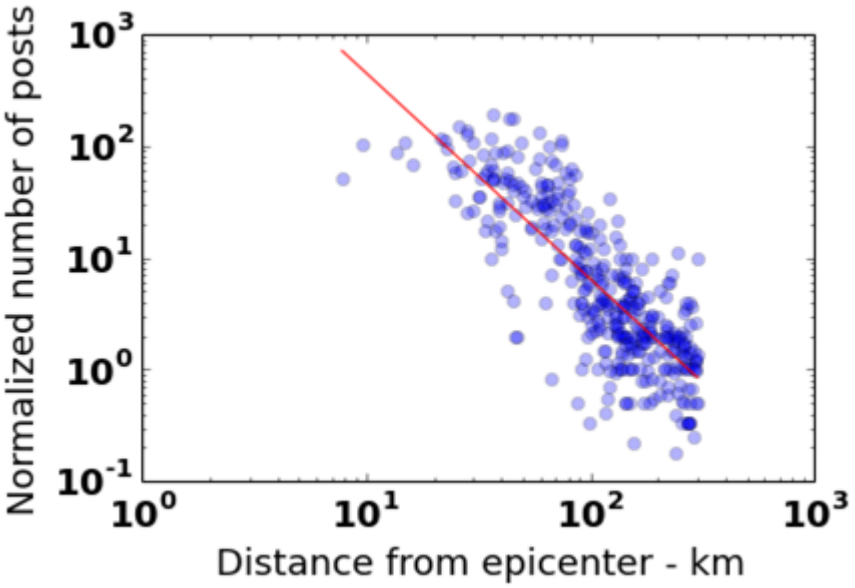




We see very significant spikes in Facebook activity for people located in a 300 km radius of the earthquake. In the beginning of the night, the activity is very similar on both August 17th and 24th. The difference spikes at 3:21a.m., just following the shake. We notice people staying more active than usual throughout the night. The difference decreases in the early morning (more than two hours after the earthquake) but never to the usual level of activity. In the morning, the number of posts increases above normal number of posts and the additional activity remains for the entire day.

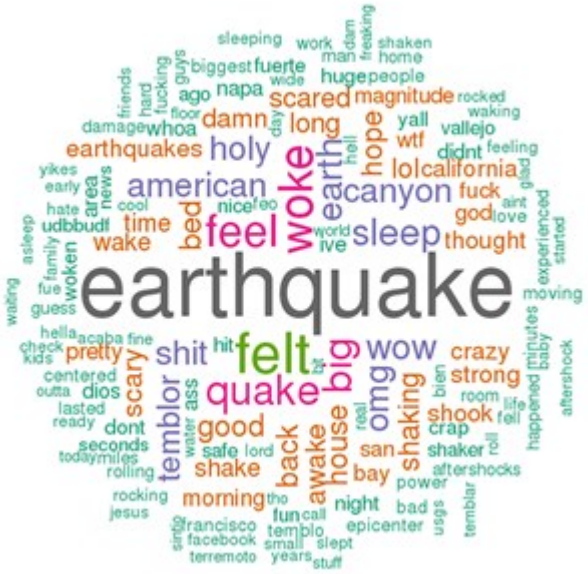
Similarly, we compare the variation in the number of posts in a city to the city's distance from the epicenter. The variation in the number of posts is computed as the ratio between the number of posts in a city within one hour following the earthquake to the number of posts on August 17th at the same time.

We ran a linear regression between the distance to the earthquake and the posting variation (in log scale).

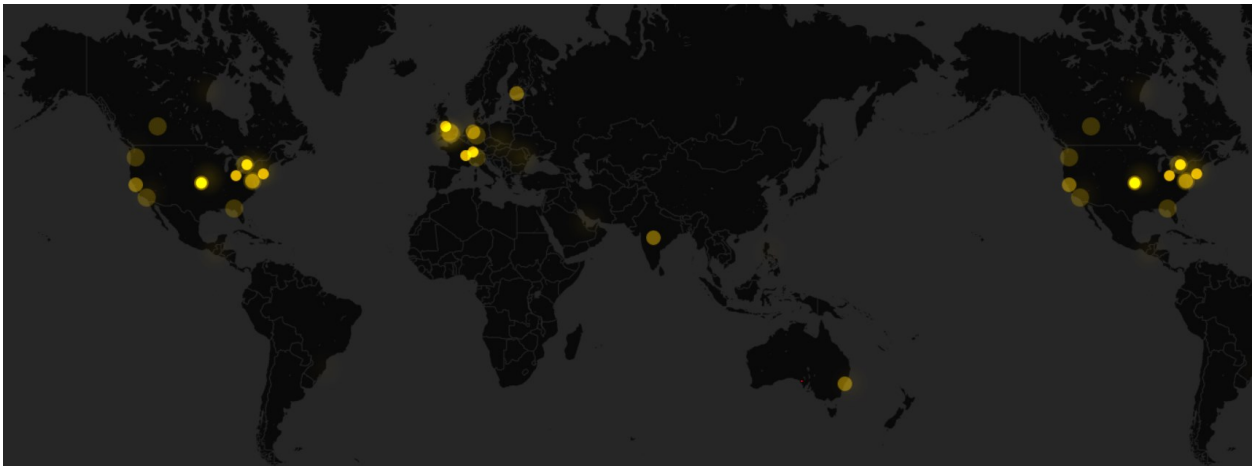

Normalized number of posts
10^3
10^2
10^1
10^0
10^{-1}
10^0
10^1
10^2
10^3
Distance from epicenter - km

APPENDIX C - Visual Tools

Word Clouds



Map showing the locations of live generated data.



REFERENCE

[1] Muhammad Imran, Carlos Castillo, Ji Lucas, Patrick Meier, and Jakob Rogstadius: Coordinating Human and Machine Intelligence to Classify Microblog Communications in Crises. In proceedings of the ISCRAM'14. Pennsylvania, USA.

[2] Bruns, J.E.Burgess, K. Crawford, and F. Shaw. # qldfloods and@ qpsmedia: Crisis communication on twitter in the 2011 south east queensland floods, 2012

[3] Sweetser, K. D., & Metzgar, E. (2007). Communicating during crisis: Use of blogs as a relationship management tool. Public Relations Review, 33, 340–342.

[4] Aude Hofleitner On Facebook when the earth shakes Retrieved from https://www.facebook.com/notes/facebook-data-science/on-facebook-when-the-earth-shakes/10152488877538859

[5] Fraustino, Julia Daisy, Brooke Liu and Yan Jin. "Social Media Use during Disasters: A Review of the Knowledge Base and Gaps," Final Report to Human Factors/Behavioral Sciences Division, Science and Technology Directorate, U.S. Department of Homeland Security.College Park, MD: START, 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] Tim Tinker and Elaine Vaughan, Risk and Crisis Communications: Best Practices for Government Agencies and Non-Profit Organizations, Booz Allen Hamilton, 2010, p.30,

[8] Adam Acar and Yuya Muraki, Twitter for Crisis Communication: Lessons Learned from Japan's Tsunami Disaster, International Journal of Web Based Communities, 2011 (forthcoming), p. 5.

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