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

List of Tables

List of Figures

Definitions, Acronyms, and Abbreviations

Table 1 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 plays a vital role current society where smart phones are commonplace. During disasters researchers have found a surge of social media activity. People looking for their loved ones, prayers from all over the world and volunteers who are willing to work are few of the contributors for this serge of activity. There are three main stages before, during and aftermath of a disaster. In those stages thousands of data pass through social media. There are insightful data as well as noise like prayers and news entries. Insightful data can be information from eyewitnesses, actual victims. Finding this information in a large amount of data is tedious and hard task to be achieved within a smaller time frame. This research aims to find a solution for finding related data from massive amount of data generated though social media during and aftermath of a disaster.

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 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)

- Duplicate data.

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 to be used in other stages. Finding appropriate optimal number of categories to match the requirements of different parties (Organization, Government agencies etc.),

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 a 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 like 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

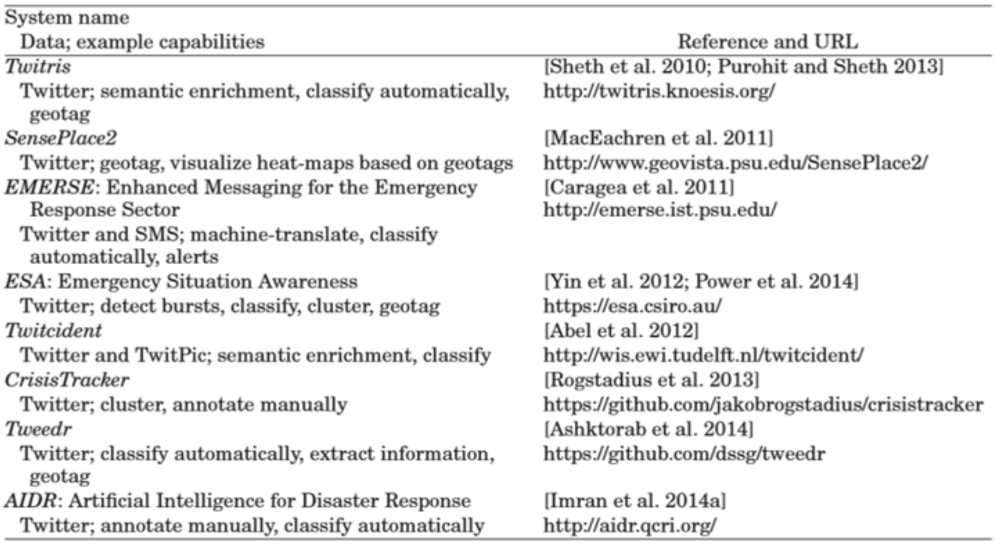
- Ability to extract information quickly.
- Identify trends.
- Ability to visualize data.
- Improved response time.
- Open source API for others to develop their own GUIs.

1.6 Background

During the Background research similar systems were found which were using twitter as their main data source. (Figure 1.6-1). [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.

Proposed tool /system has

- Extended number of categories
- Prioritization of entries in each category.



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

Figure 1.1 Existing Systems

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.

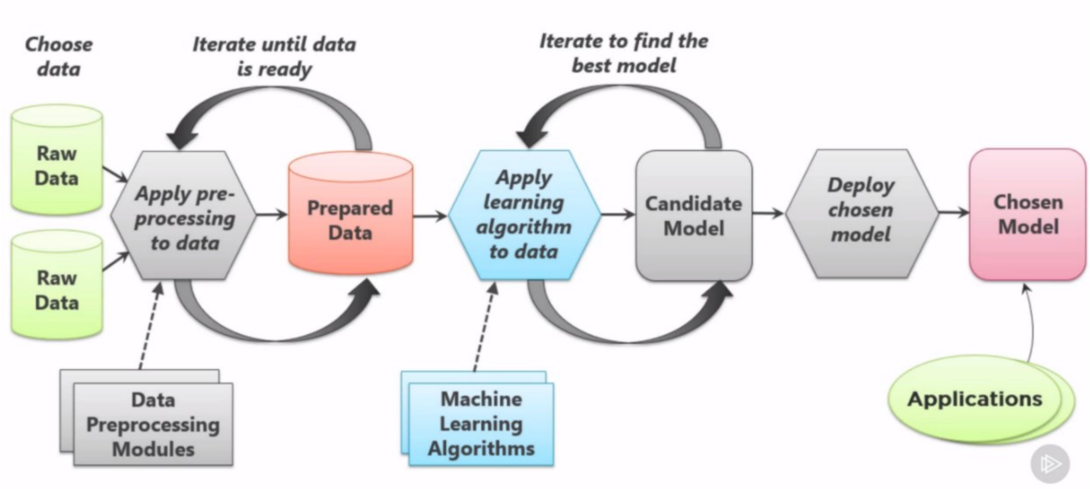

Choose data
Iterate until data is ready
Iterate to find the best model
Raw Data
Raw Data
Apply pre-processing to data
Prepared Data
Apply learning algorithm to data
Candidate Model
Deploy chosen model
Chosen Model
Data Preprocessing Modules
Machine Learning Algorithms
Applications

Figure 3.2 Process of developing a Machine Learning Algorithm

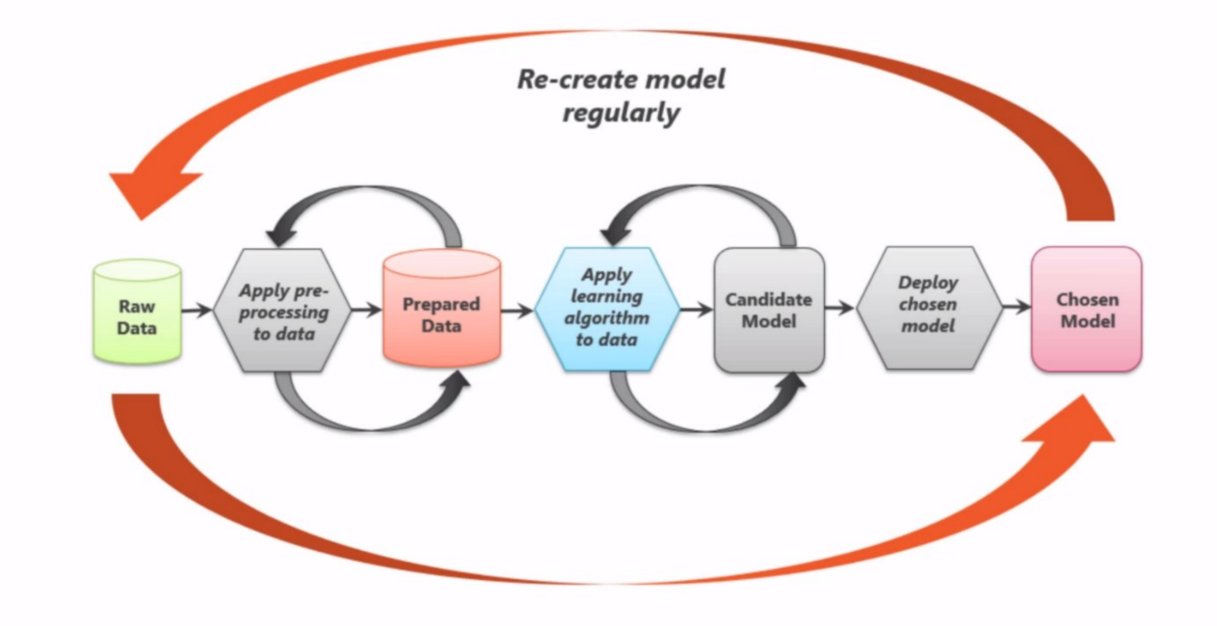

Re-create model regularly
Raw Data
Apply pre-processing to data
Prepared Data
Apply learning algorithm to data
Candidate Model
Deploy chosen model
Chosen Model

Figure 3.3 Repetitive Process of generating Machine Learning Model

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 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.

3.1.4 Evaluating the model

Cross Validation

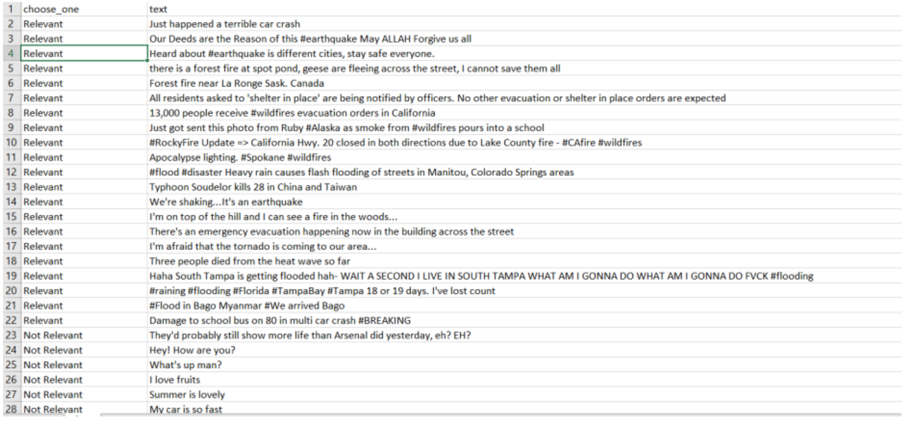
Cross validation attempts to avoid overfitting (training on and predicting the same datapoint) while still producing a prediction for each observation dataset. This is accomplished by systematically hiding different subsets of the data while training a set of models. After training, each model predicts 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

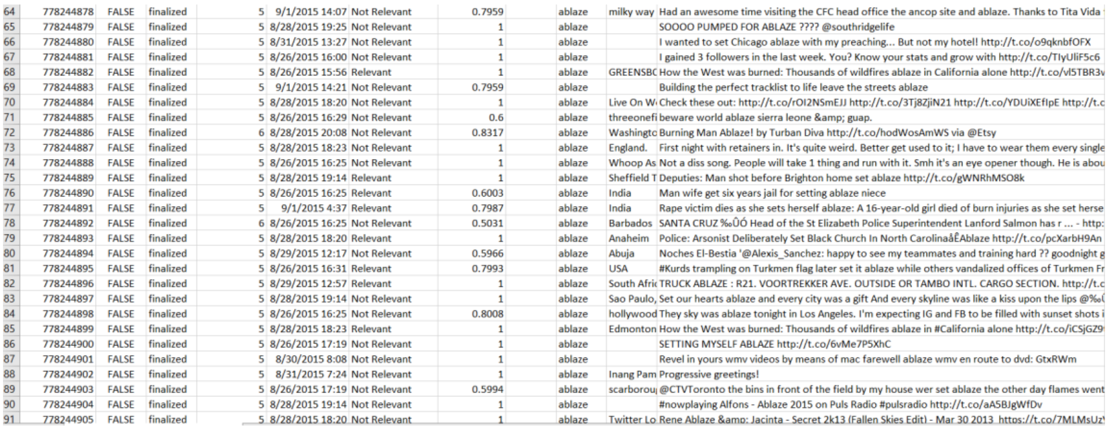
Supervised learning algorithms requires data with labels to train a model. The dataset that was used to develop the classification algorithm contains over 10,000 labeled tweets Relevant or Not Relevant. Original dataset contained more than one column which were meta data for the data. The cleaned dataset contains only two columns namely "Class" which can either be Relevant or Not Relevant and "Text" which contains the actual tweet.



1	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

Figure 4.1 Cleaned Dataset

Original Dataset contained tweets with a lot of noise like URL, Similes. Data Preprocessing techniques were used to reduce such impurities.



64	778244878	FALSE	finalized	5	9/1/2015 14:07	Not Relevant	0.7959	ablaze	milky way	Had an awesome time visiting the CFC head office the ancop site and ablaze. Thanks to Tita Vida
65	778244879	FALSE	finalized	5	8/28/2015 19:25	Not Relevant	1	ablaze		SOOOO PUMPED FOR ABLAZE ???? @southridgelife
66	778244880	FALSE	finalized	5	8/31/2015 13:27	Not Relevant	1	ablaze		I wanted to set Chicago ablaze with my preaching... But not my hotel! http://t.co/o9qknbfOFX
67	778244881	FALSE	finalized	5	8/26/2015 16:00	Not Relevant	1	ablaze		I gained 3 followers in the last week. You? Know your stats and grow with http://t.co/TIyUliF5c6
68	778244882	FALSE	finalized	5	8/26/2015 15:56	Relevant	1	ablaze	GREENSBC	How the West was burned: Thousands of wildfires ablaze in California alone http://t.co/vl5TBR3v
69	778244883	FALSE	finalized	5	9/1/2015 14:21	Not Relevant	0.7959	ablaze		Building the perfect tracklist to life leave the streets ablaze
70	778244884	FALSE	finalized	5	8/28/2015 18:20	Not Relevant	1	ablaze	Live On W	Check these out: http://t.co/rOI2NSmEJJ http://t.co/3Tj8ZjiN21 http://t.co/YDUiXEfIpE http://t.c
71	778244885	FALSE	finalized	5	8/26/2015 16:29	Not Relevant	0.6	ablaze	threeonefi	beware world ablaze sierra leone &amp; guap.
72	778244886	FALSE	finalized	6	8/28/2015 20:08	Not Relevant	0.8317	ablaze	Washingto	Burning Man Ablaze! by Turban Diva http://t.co/hodWosAmWS via @Etsy
73	778244887	FALSE	finalized	5	8/28/2015 18:23	Not Relevant	1	ablaze	England.	First night with retainers in. It's quite weird. Better get used to it; I have to wear them every single
74	778244888	FALSE	finalized	5	8/26/2015 16:25	Not Relevant	1	ablaze	Whoop As	Not a diss song. People will take 1 thing and run with it. Smh it's an eye opener though. He is abou
75	778244889	FALSE	finalized	5	8/28/2015 19:14	Relevant	1	ablaze	Sheffield T	Deputies: Man shot before Brighton home set ablaze http://t.co/gWNRhMSO8k
76	778244890	FALSE	finalized	5	8/26/2015 16:25	Relevant	0.6003	ablaze	India	Man wife get six years jail for setting ablaze niece
77	778244891	FALSE	finalized	5	9/1/2015 4:37	Relevant	0.7987	ablaze	India	Rape victim dies as she sets herself ablaze: A 16-year-old girl died of burn injuries as she set herse
78	778244892	FALSE	finalized	6	8/26/2015 16:25	Not Relevant	0.5031	ablaze	Barbados	SANTA CRUZ ‰ÛÓ Head of the St Elizabeth Police Superintendent Lanford Salmon has r ... - http:
79	778244893	FALSE	finalized	5	8/28/2015 18:20	Relevant	1	ablaze	Anaheim	Police: Arsonist Deliberately Set Black Church In North CarolinaåÊAblaze http://t.co/pcXarbH9An
80	778244894	FALSE	finalized	5	8/29/2015 12:17	Not Relevant	0.5966	ablaze	Abuja	Noches El-Bestia '@Alexis_Sanchez: happy to see my teammates and training hard ?? goodnight g
81	778244895	FALSE	finalized	5	8/26/2015 16:31	Relevant	0.7993	ablaze	USA	#Kurds trampling on Turkmen flag later set it ablaze while others vandalized offices of Turkmen Fr
82	778244896	FALSE	finalized	5	8/29/2015 12:57	Relevant	1	ablaze	South Afric	TRUCK ABLAZE : R21. VOORTREKKER AVE. OUTSIDE OR TAMBO INTL. CARGO SECTION. http://t.c
83	778244897	FALSE	finalized	5	8/28/2015 19:14	Not Relevant	1	ablaze	Sao Paulo,	Set our hearts ablaze and every city was a gift And every skyline was like a kiss upon the lips @‰Û
84	778244898	FALSE	finalized	5	8/26/2015 16:25	Not Relevant	0.8008	ablaze	hollywood	They sky was ablaze tonight in Los Angeles. I'm expecting IG and FB to be filled with sunset shots i
85	778244899	FALSE	finalized	5	8/28/2015 18:23	Relevant	1	ablaze	Edmonton	How the West was burned: Thousands of wildfires ablaze in #California alone http://t.co/iCSjGZ9
86	778244900	FALSE	finalized	5	8/26/2015 17:19	Not Relevant	1	ablaze		SETTING MYSELF ABLAZE http://t.co/6vMe7P5XhC
87	778244901	FALSE	finalized	5	8/30/2015 8:08	Not Relevant	1	ablaze		Revel in yours wmv videos by means of mac farewell ablaze wmv en route to dvd: GtxRWm
88	778244902	FALSE	finalized	5	8/31/2015 7:24	Not Relevant	1	ablaze	Inang Pam	Progressive greetings!
89	778244903	FALSE	finalized	5	8/26/2015 17:19	Not Relevant	0.5994	ablaze	scarborou	@CTVToronto the bins in front of the field by my house wer set ablaze the other day flames went
90	778244904	FALSE	finalized	5	8/28/2015 19:14	Not Relevant	1	ablaze		#nowplaying Alfons - Ablaze 2015 on Puls Radio #pulsradio http://t.co/aA5BJgWfDv
91	778244905	FALSE	finalized	5	8/28/2015 18:20	Not Relevant	1	ablaze	Twitter Lo	Rene Ablaze &amp; Jacinta - Secret 2k13 (Fallen Skies Edit) - Mar 30 2013 https://t.co/7MLMsUz

Figure 4.2 Raw Dataset

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 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
- Soket.io

Figure 4.1 Tools and Technologies

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.

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

- **Finding Trends**

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.

5.2 Architecture

The entire system is hosted on AWS (Amazon Web Service) which is a cloud service. It provides auto scaling for the services and pay as you go (pay for only what you use) scheme. AWS provides free limited access to its services (free tier) for 12 months. The component architecture is as follows.

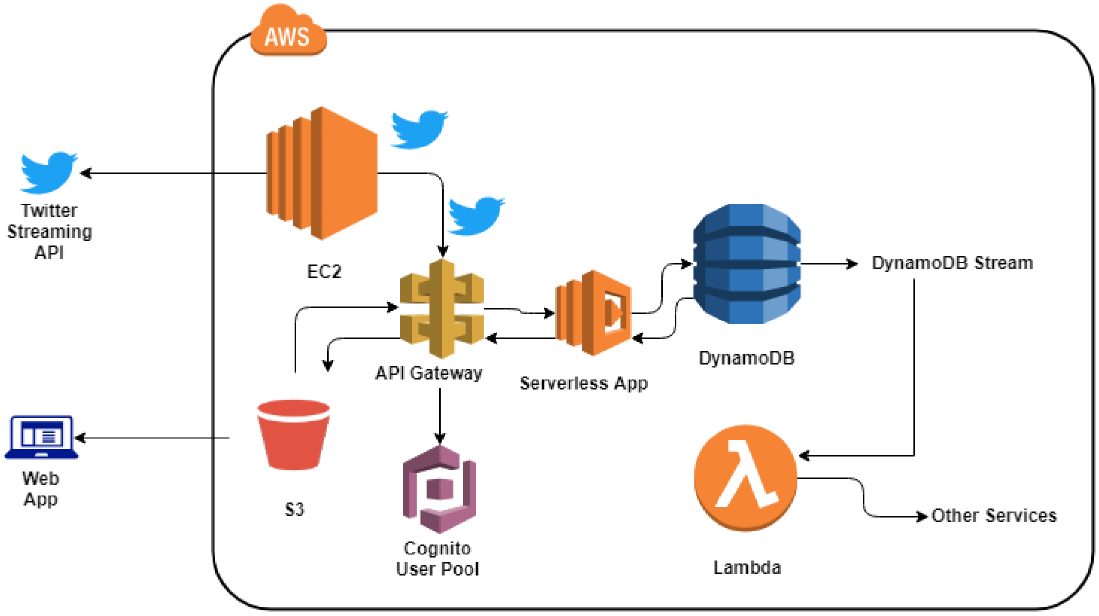
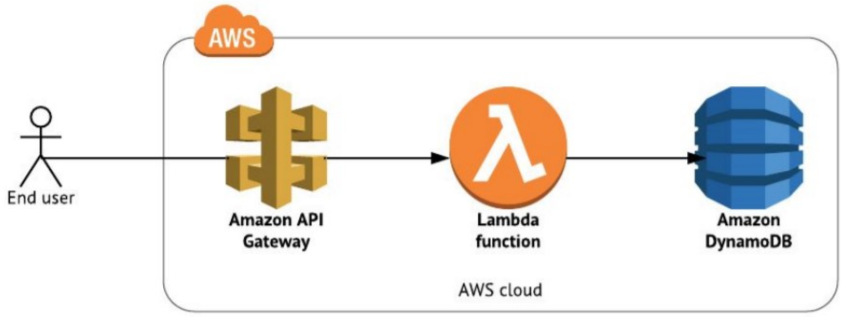
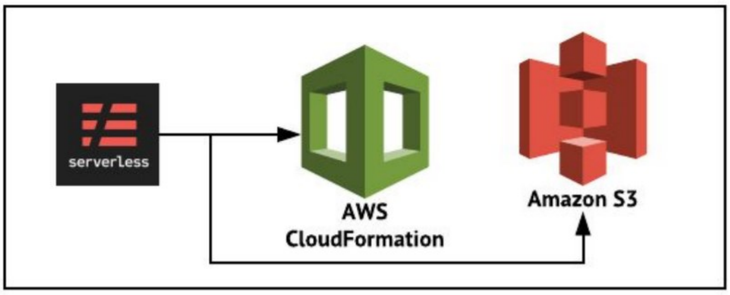


Figure 5.2 Architecture of the component

EC2 (Elastic Cloud Compute) is an Ubuntu Server running on AWS cloud. The algorithm for filtering/ identifying written as a Python script runs on this server. All the tweets identified as related from the algorithm are saved in the DynamoDB.



To access DynamoDB a RESTful service is developed using serverless. Serverless applications are helpful since they are managed by AWS.



Process starts when there are tweets available from Twitter Streaming API (is a freely available but limited service which provides tweets near real time). For more information about DynamoDB refer to APPENDIX B.

Global Secondary Index (GSI) has been created to reduce latency and to query using different partition key and sort key.

DynamoDB Streams allow us to listen to DB operations such as INSERT, DELETE and UPDATE as it happens. Connected to stream is a lambda function which triggers for prior mentioned events. This lambda function calls other service components depending on the event type.

Cognito user pool provides authentication services for the API. Web application is hosted on S3 (Simple Storage Service) and Served through Cloud Front Distribution.

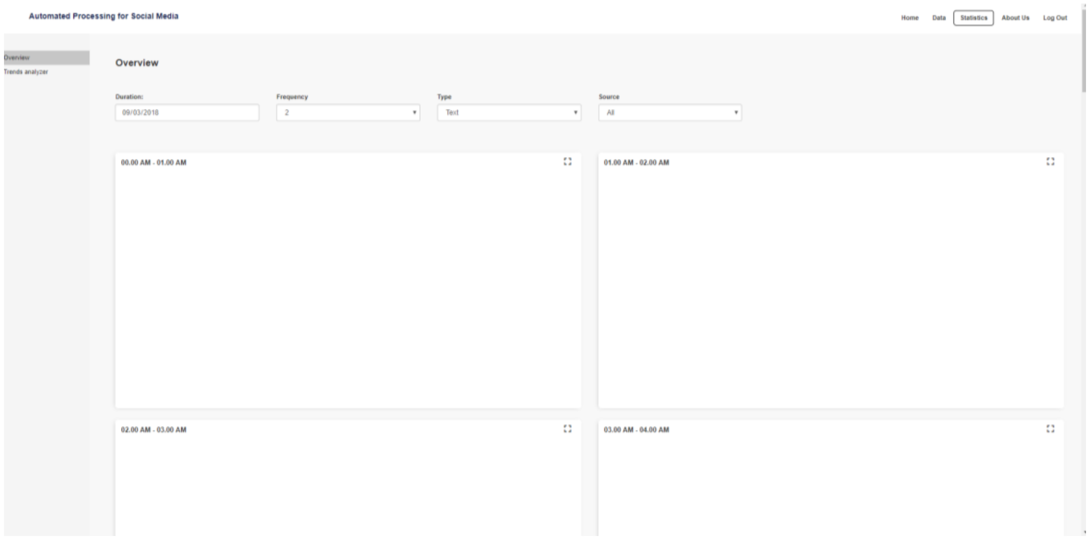
In DynamoDB attribute for Time to Live is created in order to remove data automatically after expiration.

5.3 User interfaces

Provided user interfaces are only for 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.

The following view shows the overview of a specific day hourly. User can also view the trending hashtags, select a threshold for the trends (word count)



Duration: 09/03/2018

Frequency: 2

Type: Text

Source: All

Figure 5.3. User Interface for trending word clouds (hourly)

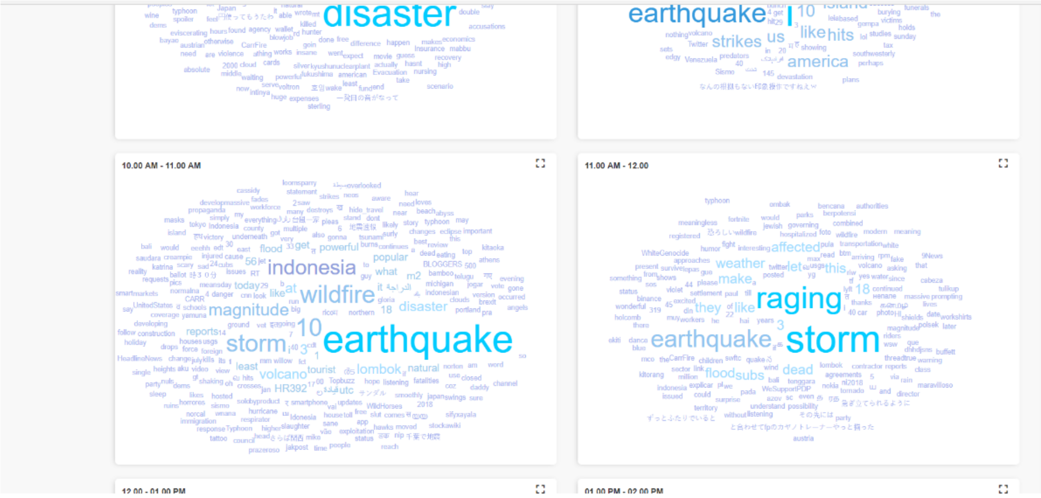
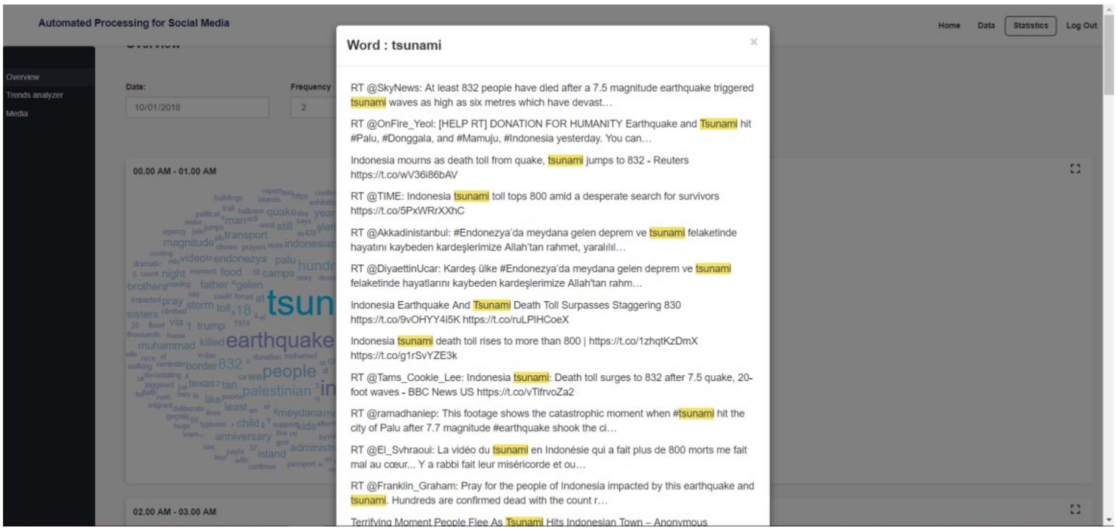


Figure 5.3 Example of hourly generated word cloud during earthquake



Word : tsunami

RT @SkyNews: At least 832 people have died after a 7.5 magnitude earthquake triggered tsunami waves as high as six metres which have devast…

RT @OnFire_Yeol: [HELP RT] DONATION FOR HUMANITY Earthquake and Tsunami hit #Palu, #Donggala, and #Mamuju, #Indonesia yesterday. You can…

Indonesia mourns as death toll from quake, tsunami jumps to 832 - Reuters https://t.co/wV36i86bAV

RT @TIME: Indonesia tsunami toll tops 800 amid a desperate search for survivors https://t.co/5PxWRrXXhC

RT @Akkadinistanbul: #Endonezya'da meydana gelen deprem ve tsunami felaketinde hayatını kaybeden kardeşlerimize Allah'tan rahmet, yaralılıl…

RT @DiyaettinUcar: Kardeş ülke #Endonezya'da meydana gelen deprem ve tsunami felaketinde hayatlarını kaybeden kardeşlerimize Allah'tan rahm…

Indonesia Earthquake And Tsunami Death Toll Surpasses Staggering 830 https://t.co/9vOHYY4i5K https://t.co/ruLPIHCoeX

Indonesia tsunami death toll rises to more than 800 | https://t.co/1zhqtKzDmX https://t.co/g1rSvYZE3k

RT @Tams_Cookie_Lee: Indonesia tsunami: Death toll surges to 832 after 7.5 quake, 20-foot waves - BBC News US https://t.co/vTifrvoZa2

RT @ramadhaniep: This footage shows the catastrophic moment when #tsunami hit the city of Palu after 7.7 magnitude #earthquake shook the ci…

RT @El_Svhraoui: La vidéo du tsunami en Indonésie qui a fait plus de 800 morts me fait mal au cœur... Y a rabbi fait leur miséricorde et ou…

RT @Franklin_Graham: Pray for the people of Indonesia impacted by this earthquake and tsunami. Hundreds are confirmed dead with the count r…

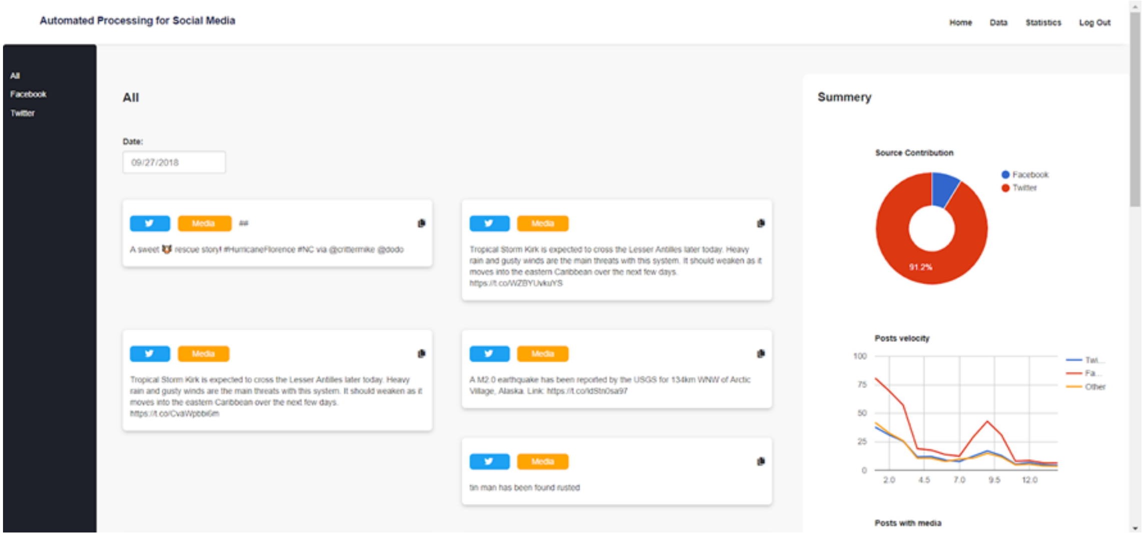
Terrifying Moment People Flee As Tsunami Hits Indonesian Town – Anonymous

Figure 5.4 Detailed view of a selected word from a word cloud.

Once the user clicks on a word system a window with all the posts would popup. The selected word would be highlighted in each of those posts.

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.



A sweet 🐱 rescue story! #HurricaneFlorence #NC via @crittermike @dodo

Tropical Storm Kirk is expected to cross the Lesser Antilles later today. Heavy rain and gusty winds are the main threats with this system. It should weaken as it moves into the eastern Caribbean over the next few days. https://t.co/WZBYUvkuYS

Tropical Storm Kirk is expected to cross the Lesser Antilles later today. Heavy rain and gusty winds are the main threats with this system. It should weaken as it moves into the eastern Caribbean over the next few days. https://t.co/CvaWpbbi6m

A M2.0 earthquake has been reported by the USGS for 134km WNW of Arctic Village, Alaska. Link: https://t.co/ldStn0sa97

tin man has been found rusted

Figure 5.5 Filtered Data

5.4 Testing

On 28 September 2018, a shallow earthquake struck in the neck of the Minahasa Peninsula, Indonesia, with its epicenter located in the mountainous Donggala Regency, Central Sulawesi. The quake was located 77 km (48 mi) away from the provincial capital Palu and was felt as far away as Samarinda on East Kalimantan and in Tawau, Malaysia. This event was preceded by a sequence of foreshocks, the largest of which was a magnitude 6.1 tremor that occurred earlier that day.

Following the mainshock, a tsunami alert was issued for the nearby Makassar Strait, but was called off half an hour later. A localized tsunami struck Palu, sweeping shore-lying houses and buildings on its way. The combined effects of the earthquake and tsunami led to the deaths of 2,000 people.

This was a real-world scenario which we were able to capture. Figure shows the most trending words cloud which was generated during the disaster.

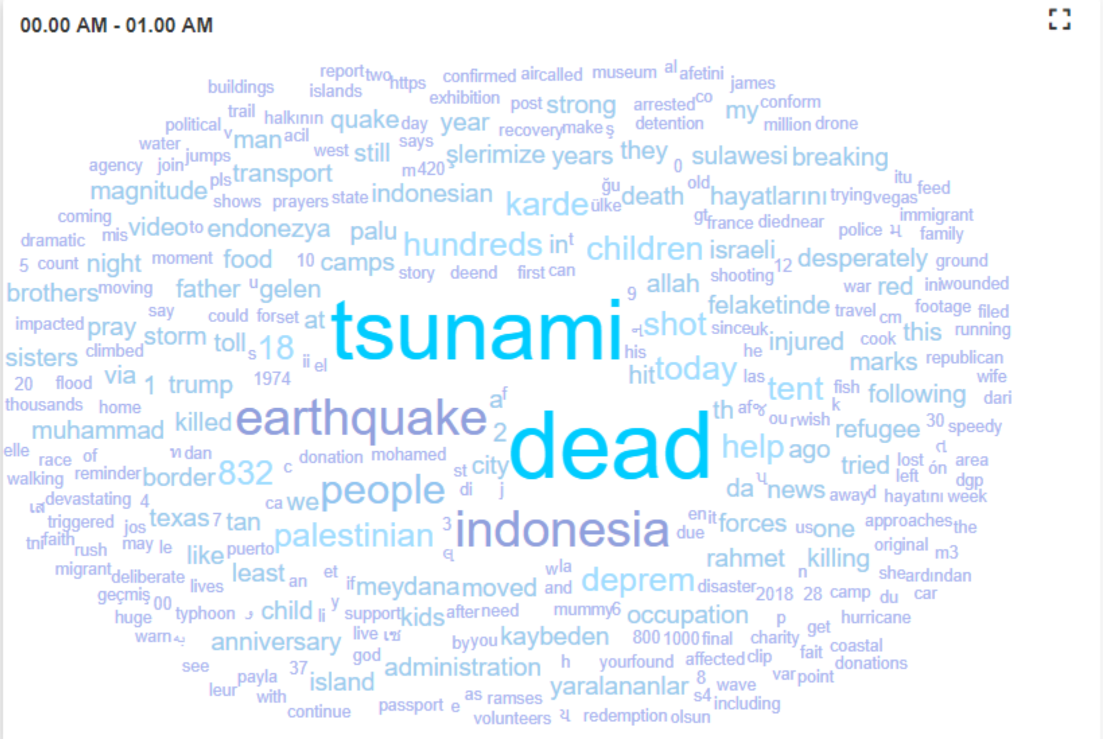


Figure 5.6 The word cloud after Tsunami occurred in Indonesia

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 comes to free usage. Part of the API called Streaming API which will be used in the project as a source of data.

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

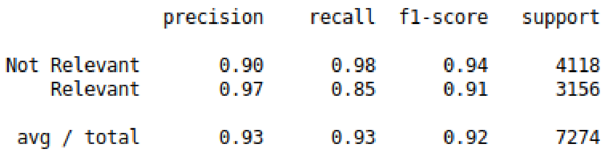
For filtering relevant posts from the twitter streaming API three algorithms were tested from those Naïve Bayes algorithm outperformed other algorithms namely random forest and Support Vector Machine. In machine learning, Naive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naive) independence assumptions between the features. With a multinomial event model, samples (feature vectors) represent the frequencies with which certain events have been generated by a multinomial with counting the number of times events observed in an instance. This is the event model typically used for document classification, with events representing the occurrence of a word in a single document. I have used TF*IDF with Multinomial Naïve Bayes to develop the classification alogirithm.

TF*IDF is an information retrieval technique that weighs a term's frequency (TF) and its inverse document frequency (IDF). Each word or term has its respective TF and IDF score. The product of the TF and IDF scores of a term is called the TF*IDF weight of that term. In simple terms the higher the TF*IDF score (weight), the rarer the term and vice versa. The TF*IDF algorithm is used to weigh a keyword in any content and assign the importance to that keyword based on the number of times it appears in the document. More importantly, it checks how relevant the keyword is throughout the web, which is referred to as corpus. For a term **t** in a sentence **s**, the weight $\mathbf{W_{t,s}}$ of term t in sentence s is given by:

$$\mathbf{W_{t,s} = TF_{t,s} \log (N/DF_{t})}$$

Where:

- $TF_{t,s}$ is the number of occurrences of t in sentence s.
- DFt is the number of sentences containing the term t.
- N is the total number of sentences in the corpus.



```
              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
```

Figure 6.1 Algorithm Test Accuracy Results

7 Conclusion

In this thesis, we addressed the problem of identifying related information and trends. One of the main contributions of work is to express the usability of such a system in a real-life situation. A discussion on different algorithms been provided. We proposed novel approach for the domain. This thesis only focuses on the data filtering and trending component.

Social media receives overwhelming number of posts during an emergency. Finding meaningful insightful information from this massive about of information with noise is hard. A lot of posts regarding emergency for example earthquake are about prayers and news which are not useful. Posts from victims and eyewitnesses are rare but very useful. However, for this information to be useful they must be process quickly since they get obsolete quickly. Existing solutions only provides the basic operation to process that information. Such as filtering related information. Organizations who responds to mass emergencies such and Disaster Management Centers (DMC) have employed teams called Intelligent Teams who go through these data as they come and make reports on the incident to improve the Situational Awareness. Doing this manually is inefficient and time-consuming. To make it more efficient in the using the same traditional way may require increase of the number of people in a team.

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. The proposed system 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.

8 Future Works

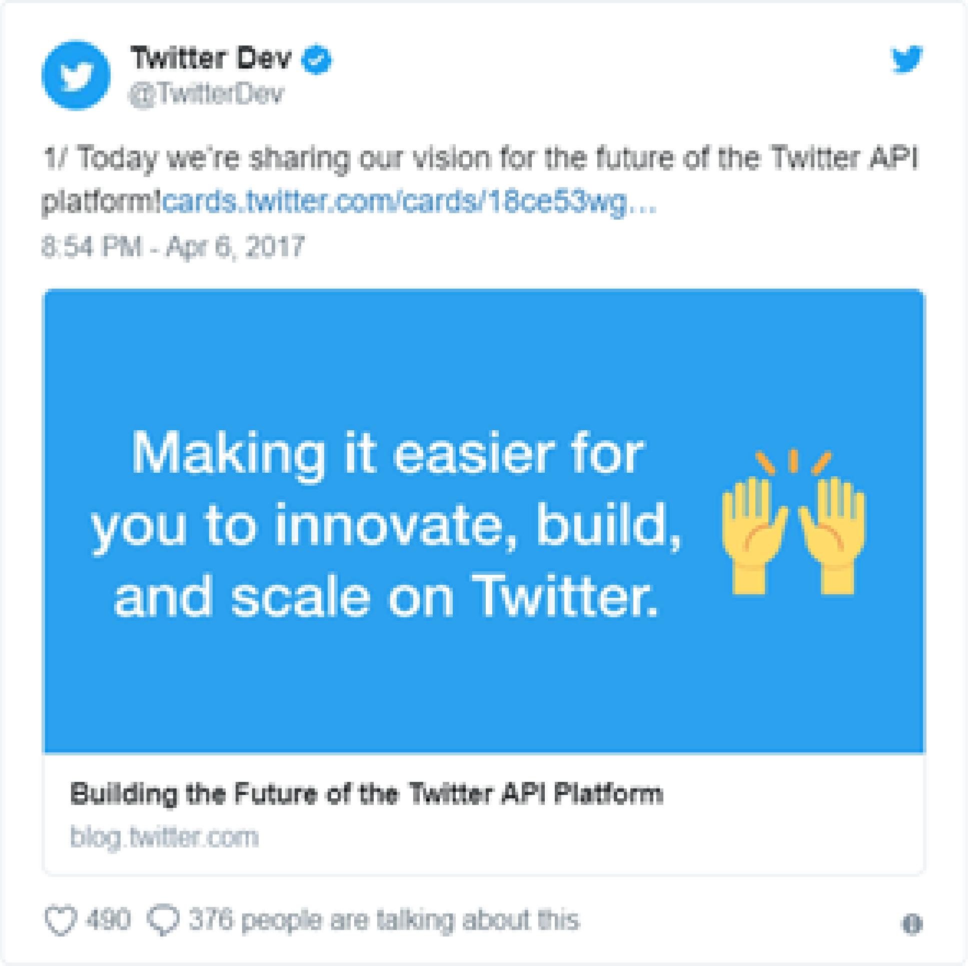
Many different adaptations, tests, and experiments have been left for the future due to lack of time (i.e. the experiments with more real situation). Future work concerns deeper analysis of tweets, solving issues like identifying language of a tweet and the location of a tweet. This thesis has been mainly focused on the data filtering and finding trends out of them Text Classification for different requirements is left out due to the lack of time and data which we did not have until the end of the project. Following are some interesting ideas which we can develop in the future.

1. It could be interesting to consider the analysis of regions of the tweets.
2. It was evident that a lot of media (images/ videos) are passed through in times of emergencies. Analysis of these content using image processing would be interesting.

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.



Twitter Dev
@TwitterDev

1/ Today we're sharing our vision for the future of the Twitter API platform!cards.twitter.com/cards/18ce53wg…

8:54 PM - Apr 6, 2017

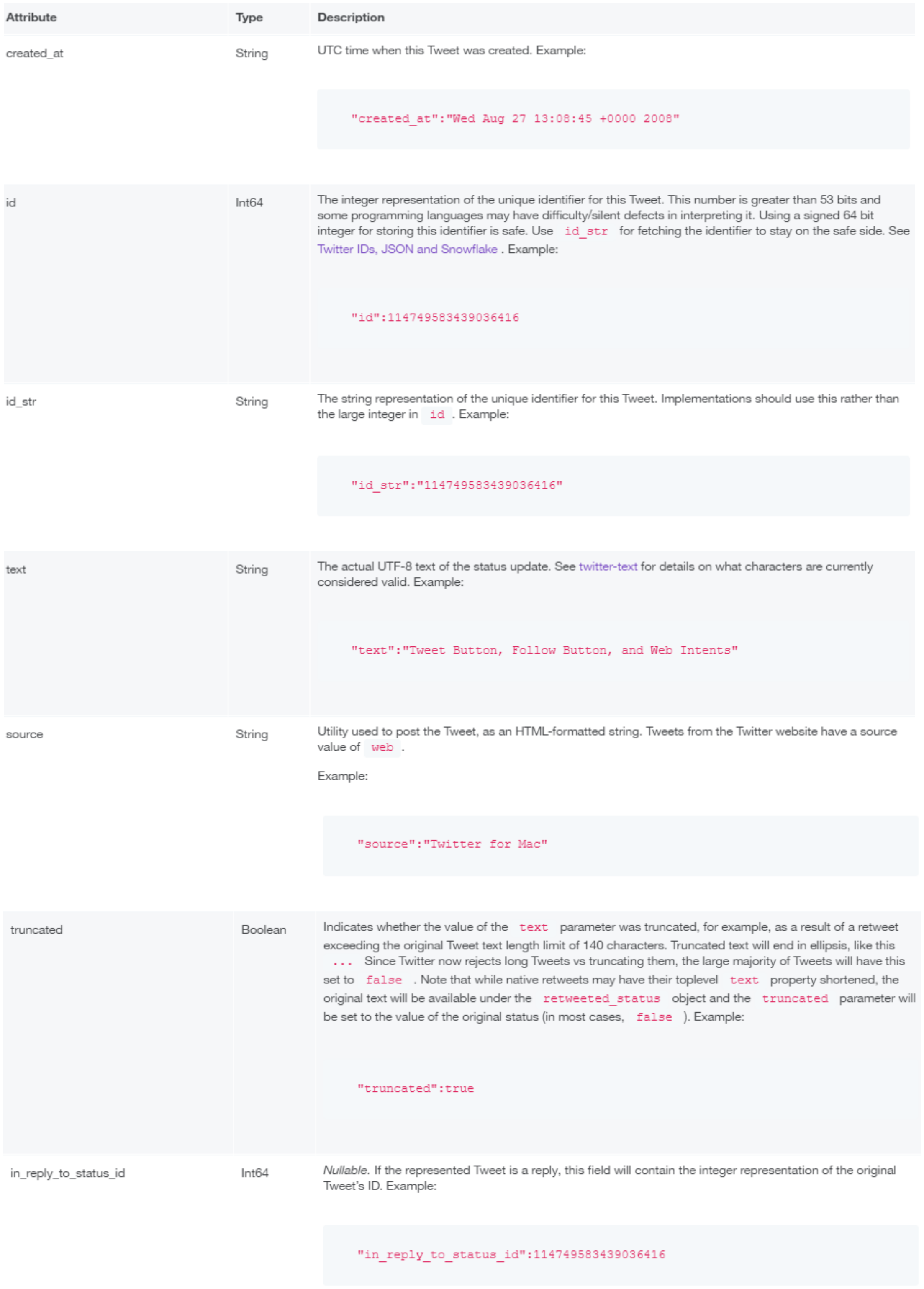
Making it easier for you to innovate, build, and scale on Twitter.

Building the Future of the Twitter API Platform
blog.twitter.com

490 376 people are talking about this

Figure A Sample Tweet

Table 2 JSON Object key description [11]



Attribute	Type	Description
created_at	String	UTC time when this Tweet was created. Example: `"created_at":"Wed Aug 27 13:08:45 +0000 2008"`
id	Int64	The integer representation of the unique identifier for this Tweet. This number is greater than 53 bits and some programming languages may have difficulty/silent defects in interpreting it. Using a signed 64 bit integer for storing this identifier is safe. Use `id_str` for fetching the identifier to stay on the safe side. See Twitter IDs, JSON and Snowflake . Example: `"id":114749583439036416`
id_str	String	The string representation of the unique identifier for this Tweet. Implementations should use this rather than the large integer in `id` . Example: `"id_str":"114749583439036416"`
text	String	The actual UTF-8 text of the status update. See twitter-text for details on what characters are currently considered valid. Example: `"text":"Tweet Button, Follow Button, and Web Intents"`
source	String	Utility used to post the Tweet, as an HTML-formatted string. Tweets from the Twitter website have a source value of `web` . Example: `"source":"Twitter for Mac"`
truncated	Boolean	Indicates whether the value of the `text` parameter was truncated, for example, as a result of a retweet exceeding the original Tweet text length limit of 140 characters. Truncated text will end in ellipsis, like this `...` Since Twitter now rejects long Tweets vs truncating them, the large majority of Tweets will have this set to `false` . Note that while native retweets may have their toplevel `text` property shortened, the original text will be available under the `retweeted_status` object and the `truncated` parameter will be set to the value of the original status (in most cases, `false`). Example: `"truncated":true`
in_reply_to_status_id	Int64	*Nullable.* If the represented Tweet is a reply, this field will contain the integer representation of the original Tweet's ID. Example: `"in_reply_to_status_id":114749583439036416`

in_reply_to_status_id_str	String	*Nullable.* If the represented Tweet is a reply, this field will contain the string representation of the original Tweet's ID. Example: `"in_reply_to_status_id_str":"114749583439036416"`
in_reply_to_user_id	Int64	*Nullable.* If the represented Tweet is a reply, this field will contain the integer representation of the original Tweet's author ID. This will not necessarily always be the user directly mentioned in the Tweet. Example: `"in_reply_to_user_id":819797`
in_reply_to_user_id_str	String	*Nullable.* If the represented Tweet is a reply, this field will contain the string representation of the original Tweet's author ID. This will not necessarily always be the user directly mentioned in the Tweet. Example: `"in_reply_to_user_id_str":"819797"`
in_reply_to_screen_name	String	*Nullable.* If the represented Tweet is a reply, this field will contain the screen name of the original Tweet's author. Example: `"in_reply_to_screen_name":"twitterapi"`

user	User object	The user who posted this Tweet. See User data dictionary for complete list of attributes.
		Example highlighting select attributes:
		`{`
		`  "user": {`
		`    "id": 2244994945,`
		`    "id_str": "2244994945",`
		`    "name": "TwitterDev",`
		`    "screen_name": "TwitterDev",`
		`    "location": "Internet",`
		`    "url": "https://dev.twitter.com/",`
		`    "description": "Your source for Twitter news",`
		`    "verified": true,`
		`    "followers_count": 477684,`
		`    "friends_count": 1524,`
		`    "listed_count": 1184,`
		`    "favourites_count": 2151,`
		`    "statuses_count": 3121,`
		`    "created_at": "Sat Dec 14 04:35:55 +0000 2013",`
		`    "utc_offset": null,`
		`    "time_zone": null,`
		`    "geo_enabled": true,`
		`    "lang": "en",`
		`    "profile_image_url_https": "https://pbs.twimg.com/"`
		`  }`
		`}`

coordinates	Coordinates	*Nullable.* Represents the geographic location of this Tweet as reported by the user or client application. The inner coordinates array is formatted as geoJSON (longitude first, then latitude). Example:
		`"coordinates":`
		`{`
		`"coordinates":`
		`[`
		`-75.14310264,`
		`40.05701649`
		`],`
		`"type":"Point"`
		`}`

place	Places	*Nullable* When present, indicates that the tweet is associated (but not necessarily originating from) a Place . Example:
		`"place":`
		`{`
		`"attributes":{},`
		`"bounding_box":`
		`{`
		`"coordinates":`
		`[[`
		`[-77.119759,38.791645],`
		`[-76.909393,38.791645],`
		`[-76.909393,38.995548],`
		`[-77.119759,38.995548]`
		`]],`
		`"type":"Polygon"`
		`},`
		`"country":"United States",`
		`"country_code":"US",`
		`"full_name":"Washington, DC",`
		`"id":"01fbe706f872cb32",`
		`"name":"Washington",`
		`"place_type":"city",`
		`"url":"http://api.twitter.com/1/geo/id/0172cb32.json"`
		`}`

quoted_status_id	Int64	This field only surfaces when the Tweet is a quote Tweet. This field contains the integer value Tweet ID of the quoted Tweet. Example: `"quoted_status_id":1147495834390364l6`
quoted_status_id_str	String	This field only surfaces when the Tweet is a quote Tweet. This is the string representation Tweet ID of the quoted Tweet. Example: `"quoted_status_id_str":"1147495834390364l6"`
is_quote_status	Boolean	Indicates whether this is a Quoted Tweet. Example: `"is_quote_status":false`
quoted_status	Tweet	This field only surfaces when the Tweet is a quote Tweet. This attribute contains the Tweet object of the original Tweet that was quoted.

retweeted_status	Tweet	Users can amplify the broadcast of Tweets authored by other users by retweeting . Retweets can be distinguished from typical Tweets by the existence of a `retweeted_status` attribute. This attribute contains a representation of the *original* Tweet that was retweeted. Note that retweets of retweets do not show representations of the intermediary retweet, but only the original Tweet. (Users can also unretweet a retweet they created by deleting their retweet.)
quote_count	Integer	*Nullable.* Indicates approximately how many times this Tweet has been quoted by Twitter users. Example: `"quote_count":1138` Note: This object is only available with the Premium and Enterprise tier products.
reply_count	Int	Number of times this Tweet has been replied to. Example: `"reply_count":1585` Note: This object is only available with the Premium and Enterprise tier products.
retweet_count	Int	Number of times this Tweet has been retweeted. Example: `"retweet_count":1585`
favorite_count	Integer	*Nullable.* Indicates approximately how many times this Tweet has been liked by Twitter users. Example: `"favorite_count":1138`

entities	Entities	Entities which have been parsed out of the text of the Tweet. Additionally see Entities in Twitter Objects . Example:
		`"entities":`
		`{`
		`"hashtags":[],`
		`"urls":[],`
		`"user_mentions":[],`
		`"media":[],`
		`"symbols":[]`
		`"polls":[]`
		`}`

extended_entities	Extended Entities	When between one and four native photos or one video or one animated GIF are in Tweet, contains an array 'media' metadata. Additionally see Entities in Twitter Objects . Example:
		`"entities":`
		`{`
		`"media":[]`
		`}`

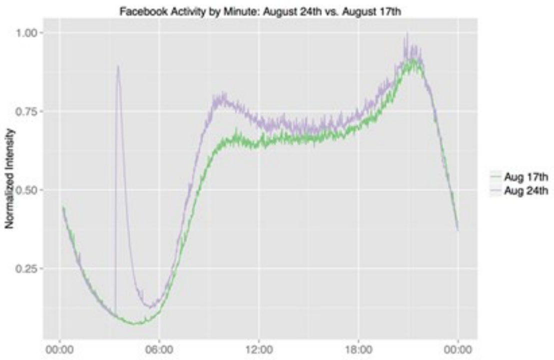
favorited	Boolean	*Nullable.* Indicates whether this Tweet has been liked by the authenticating user. Example: `"favorited":true`
retweeted	Boolean	Indicates whether this Tweet has been Retweeted by the authenticating user. Example: `"retweeted":false`

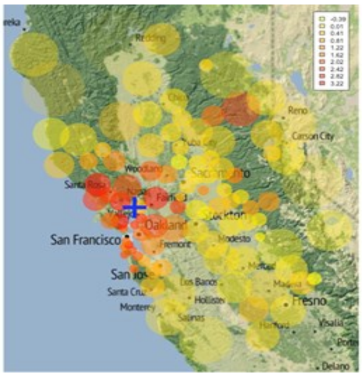
APPENDIX B – Literature

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.

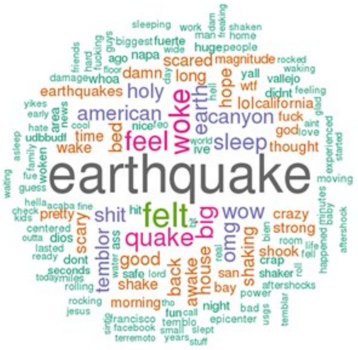
During a crisis, people turn to Facebook to stay connected to their friends and family. They use it to receive social support and keep the people they care about informed on how they are doing.



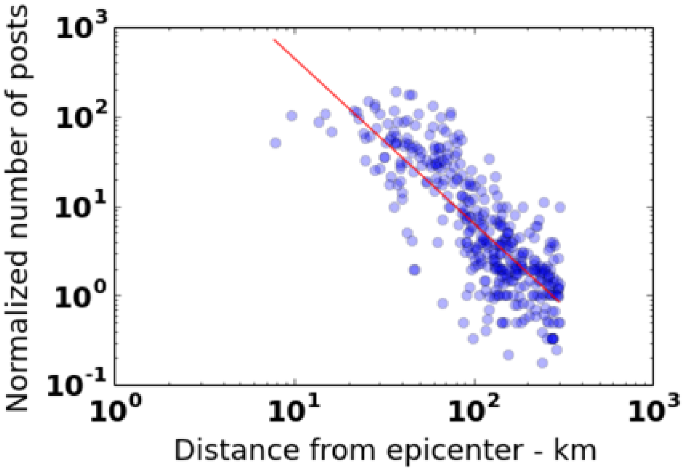


The map above shows the relative difference in activity on Facebook between the 24th of August, 3:21 a.m. and 3:26 a.m., and the same time period one week earlier. For visualization, we cluster together nearby cities which showed similar changes in activity. The color represents the percent variation 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.


sleeping work dam freaking shaken
friends hard fucking guys biggest fuerte man home
wide huge people
ago napa scared magnitude rocked
floor waking
damage whoa damn day long hope yall vallejo
earthquakes holy earth hell wtf didnt feeling
yikes woke lol california glad
early area news american canyon fuck
hate cool nice feo aint
asleep udbbudf time bed feel world god love experienced
family wake ive sleep thought started
fue woken
waiting guess earthquake moving
hella acaba fine
check kids pretty shit hit felt big wow crazy minutes baby
centered bien strong aftershock
outta dios scary temblor ass quake omg room happened
lasted shaking life
ready real shook fell
dont water good crap
seconds safe lord back awake house san shaker roll
today miles
rolling shake bay aftershocks
power
rocking morning tho night bad usgs temblar
jesus fun call
single francisco temblo epicenter
facebook small slept
terremoto years stuff

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).

APPENDIX C – DynamoDB

DynamoDB is a fast and flexible NOSQL database which provides reliable performance. It is fully managed by Amazon Web Services (AWS) which means we don’t have to manage servers, scaling, backup, security. AWS assures to provide consistent single-digit millisecond latency at any scale.

There are three core components to DynamoDB

- **Tables**

  Like other DBs DynamoDB stores data in tables. There is a limitation to the number of tables that can be created region wise. (as of 2018 the initial limit is 256 table although you can request AWS for more). There is no limitation to the size of a table in DynamoDB.

- **Items**

  An item is a group attributes that is uniquely identifiable among all other items. Each DynamoDB table consists of zero or more items. The maximum item size in DynamoDB is 400 KB, which includes both attribute name binary length (UTF-8 length) and attribute value lengths (again binary length). The attribute name counts towards the size limit.

- **Attributes**

  Each item is composed of one or more attributes. An attribute is a fundamental data element, something that does not need to be broken down any further. Attributes in DynamoDB are similar in many ways to fields or columns in other database systems.

Partition Key / Sort Key

Partition key act as the primary key for a table. Sort Key is an optional field which is used with the partition key. When both Partition key and the sort key is used it becomes a Composite key for the table. If we are using Partition key alone then it must be unique. When we are using Partition key along with a Sort key the partition key can duplicate. But two Items with the same Partition key value cannot have the same Sort key value.

When decide to use DynamoDB we must first think about how we are going to retrieve data, and how we handle the provisioned capacity units.

- **Partition Key Only** (Primary key, must be unique)
- **Partition Key + Sort Key** (composite primary key)

DynamoDB Data Retrieval

There are two ways

- **Scan**

  When used it goes through all the items in a table. Scan operation is less efficient, and it consumes more RCUs.
- **Query**

  The query operation finds items based on primary key values. We can query a table of a secondary **index** that has a composite primary key (a partition key and a sort key)

We can optionally provide a “**FilterExpression**”. A “FilterExpression” determines which items within the results should be returned. All other results are discarded. Each of these operations returns only 1mb of data in a single request. A single Query operation will read up to the maximum number of items set (if using the Limit parameter) or a maximum of 1 MB of data and then apply any filtering to the results using FilterExpression. If “**LastEvaluatedKey**” is present in the response, we will need to paginate the result set.

Capacity Units

Capacity Units are the measurement of how much resources are used when using DynamoDB services. (Prices may vary depending on the region that is used to create the table)

- **Read Capacity Units (RCU)**

One RCU = one strongly consistent read per second, or two eventually consistent reads per second, for items up to 4 KB in size.

- **Write Capacity Units (WCU)**

One WCU = one write per second, for items up to 1 KB in size.

DynamoDB Eventual Consistency.

The eventual consistency option maximizes your read throughput. However, an eventually consistent read might not reflect the results of a recently completed write. All copies of data usually reach consistency within a second. Repeating a read after a short time should return the updated data.

DynamoDB Strong Consistency.

In addition to eventual consistency, DynamoDB also gives us the flexibility and control to request a strongly consistent read if your application, or an element of your application, requires it. A strongly consistent read returns a result that reflects all writes that received a successful response before the read.

DynamoDB Indexes

There are two types of indexes we can create in DynamoDB.

- **Global Secondary Index (GSI)**
  The primary key of the index can be any two attributes from its table.
- **Local Secondary Index (LSI)**
  The partition key of the index must be the same as the partition key of its table. However, the sort key can be any other attribute.

Time to Live

Time to Live (TTL) for DynamoDB allows you to define when items in a table expire so that they can be automatically deleted from the database. TTL is provided at no extra cost to reduce storage usage and reduce the cost of storing irrelevant data without using provisioned throughput. With TTL enabled on a table, you can set a timestamp for deletion on a per-item basis, allowing you to limit storage usage to only those records that are relevant.

TTL is useful if you have continuously accumulating data that loses relevance after a specific time period. For example: session data, event logs, usage patterns, and other temporary data. If you have sensitive data that must be retained only for a certain amount of time according to contractual or regulatory obligations, TTL helps us ensure that it is removed promptly and as scheduled.

DynamoDB Streams

DynamoDB Streams is an optional feature that captures data modification events in DynamoDB tables. The data about these events appear in the stream in near real time, and in the order that the events occurred. Each event is represented by a stream record. If you enable a stream on a table, DynamoDB Streams writes a stream record whenever one of the following events occurs:

- A new item is added to the table: The stream captures an image of the entire item, including all its attributes.

- An item is updated: The stream captures the "before" and "after" image of any attributes that were modified in the item.

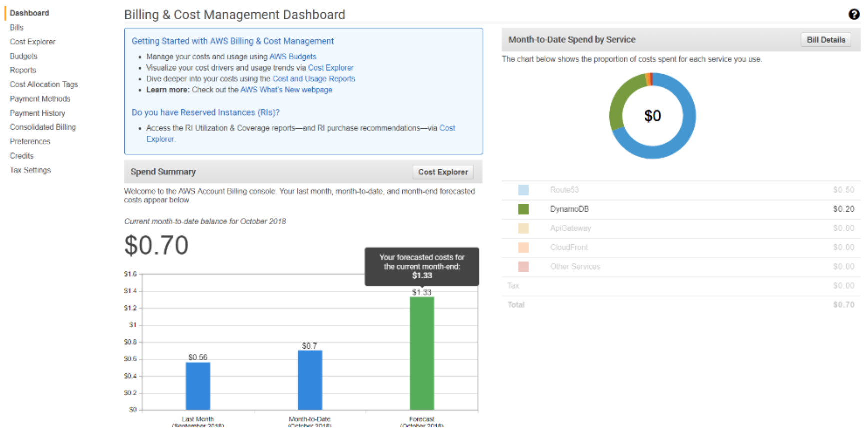
- An item is deleted from the table: The stream captures an image of the entire item before it was deleted.

Each stream record also contains the name of the table, the event timestamp, and other metadata. Stream records have a lifetime of 24 hours; after that, they are automatically removed from the stream.

We can use Streams together with Lambda to create a trigger—code that executes automatically whenever an event of interest appears in a stream.

APPENDIX D – Pricing for AWS (As of 2018)

Amazon Web Services (AWS) is a secure cloud **services** platform, offering compute power, database storage, content delivery and other functionality to help businesses scale and grow. The AWS Cloud provides a broad set of infrastructure services, such as computing power, storage options, networking and databases, delivered as a utility: on-demand, available in seconds, with pay-as-you-go pricing. (means you only have to pay for what you use and scalable as you want).



Getting Started with AWS Billing & Cost Management

- Manage your costs and usage using AWS Budgets
- Visualize your cost drivers and usage trends via Cost Explorer
- Dive deeper into your costs using the Cost and Usage Reports
- **Learn more:** Check out the AWS What's New webpage

Do you have Reserved Instances (RIs)?

- Access the RI Utilization & Coverage reports—and RI purchase recommendations—via Cost Explorer.

Spend Summary

Welcome to the AWS Account Billing console. Your last month, month-to-date, and month-end forecasted costs appear below

Current month-to-date balance for October 2018

$0.70

Your forecasted costs for the current month-end: **$1.33**

Month-to-Date Spend by Service

The chart below shows the proportion of costs spent for each service you use.

■	Route53	$0.50
■	DynamoDB	$0.20
■	ApiGateway	$0.00
■	CloudFront	$0.00
■	Other Services	$0.00
Tax		$0.00
Total		$0.70

Figure D AWS Billing Dashboard

AWS provides a free tier which helps developers to try out their services. It is valid for 12 months. There are few services that provides free resources to some extent which does not expire. For example, 25 RCU and WCU are free for every AWS account (As of 2018 Sept)

Every AWS resource/service used for the project are located in the Sydney Region

For the research project we used

- 4 EC2 instances to run the services.
- 1 S3 bucket
- 2 DynamoDB Tables (25 RCU and 25 WCU)

For running these services, it did cost less than 10$

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

[10] Amazon Web Services Documentation
https://docs.aws.amazon.com/index.html#lang/en_us

[11] Twitter Documentation for Tweet Object
https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/tweet-object.html