

AUTOMATED PROCESSING FOR SOCIAL MEDIA DATA IN A MASS EMERGENCY: VALIDATING ACCURACY OF POST

Sepala Dahanayake Saumya Madushani

IT15028310

Degree of Bachelor of Science

Department of Information Technology

Sri Lanka Institute of Information Technology

Sri Lanka

October 2018

AUTOMATED PROCESSING FOR SOCIAL MEDIA DATA IN A MASS EMERGENCY: VALIDATING ACCURACY OF POST

Sepala Dahanayake Saumya Madushani

IT15028310

The dissertation was submitted in partial fulfilment of the requirements for the
B.Sc. Honors degree in Information Technology

Department of Information Technology

Sri Lanka Institute of Information Technology

Sri Lanka

October 2018

DECLARATION

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

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

Signature: Sepala Dahanayake Saumya Madushani

Date: 10/05/2018

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

Signature of the supervisor:

Date:

ABSTRACT

Knowledge and information give people the power to make decisions and take action, they are the key to the success of real-time decision-making. The world is full of emergencies caused by natural disasters. In such situations, Social media can be valuable where information can be shared to save lives and minimize human and social impact. During such disaster situation it is impossible to stop spreading false information and rumors through those social media. Fostering the quality of information is important to validate the information collected for decision-making. With the empowerment of the general public and the abundance of information on social media, the quality of information becomes essential to achieving an effective and efficient emergency response. This research work aims to introduce an accuracy measuring feature which is completely missing or limited characteristic of existing system since it is hard to achieve, using modern concepts such as, Semantic Analysis, Natural Language Processing (NLP), Machine Learning (ML).

ACKNOWLEDGEMENT

The work described in this document was carried out as my 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 my duty to express my gratitude to all who gave me the support to complete this major task.

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

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

TABLE OF CONTENTS

DECLARATION	i
ABSTRACT.....	ii
ACKNOWLEDGEMENT	iii
TABLE OF CONTENTS.....	iv
LIST OF TABLES	v
LIST OF FIGURES	vi
LIST OF ABBREVIATIONS.....	vii
1.INTRODUCTION	1
1.1 Background.....	1
1.2 Research Gap	2
1.3 Research Problem	3
1.4 Research Objectives.....	4
1.4.1 Main Objective.....	4
1.4.2 Specific Objectives	4
2.METHODOLOGY	5
2.1 Methodology	5
2.1.1 System Architecture and Implementation	5
2.1.2 Tools	10
2.1.3 Technologies	11
2.2 Testing.....	12
3.RESULTS & DISCUSSIONS	14
3.1 Research Finding and Discussion	14
4.CONCLUSION.....	15
5.REFERENCE.....	16
6.GLOSSARY	17
7.APPENDICES	19

LIST OF TABLES

Table 1:Information Quality Criteria	2
Table 2:Comparison between existing systems	3

LIST OF FIGURES

Figure 1: System Architecture diagram	5
Figure 2: disaster.com logo.....	6
Figure 3: Dataset.csv file	6
Figure 4: Python code for load csv file.....	6
Figure 5: Implementation of Naive Bayes Model.....	7
Figure 6: Save classified dataset to sav file	8
Figure 7: Implementation to accuracy check	8
Figure 8: Results of accuracy check	8
Figure 9: Implementation of API Part 01	9
Figure 10: Implementation of API Part 02	9
Figure 11: Anaconda Navigator logo.....	10
Figure 12: Jupyter Notebook Logo	10
Figure 13: Install NLTK to Anaconda Navigator	11
Figure 14: Python logo.....	12
Figure 15: Scikit-Learn logo.....	12
Figure 16: Test 01 results.....	12
Figure 17: Test 02 results.....	13
Figure 18: Test 03 results.....	13
Figure 19: Database connection.....	19

LIST OF ABBREVIATIONS

NLP	Natural Language Processing
API	Application Programming Interface
ML	Machine Learning
AIDR	Artificial Intelligence for Disaster Response
APDM	Automated Process for Disaster Management
NLTK	Natural Language Tool Kit
CSV	Comma-Separated Values

1. INTRODUCTION

1.1 Background

There is a growing interest in social media. One of the reasons is because of its rapid success in attracting participants. There have been claims that social media is valuable for businesses, mainly in marketing. Another area for investigating the value of social media is the need to share information on social media to save lives, such as in an emergency. Emergencies pose a challenge to managing information on social media. Emergency response can present a completely unique situation that traditional systems are not optimally configured to support. However, there is a need to ensure quality and reliability, in order to verify information collected for decision making. With the empowerment of the general public and the abundance of information on social media, the quality of information has become the key to achieving an effective and efficient results in emergency response and saving lives.

Communication of quality information is critical in emergency management. When a disaster strikes, the rapid gathering and sharing of crucial information among public safety agencies, emergency response units, and the public can save lives and reduce the scope of the problem. Emergency management audiences can include general public, disaster victims, all levels of government, business community, media, elected officials, community officials, first responders, NGOs, law enforcement, medical communities, scientific communities, volunteer groups and others. Communicating in the midst of an emergency response and recovery effort can be difficult because of conflicting reports, confusion and expectations of the public. The cost of poor communication and coordination can be high and increase human loss.

With tools such as online social media there are new communication and information channels that have the potential to ensure timeliness of information delivered from multiple sources, avoidance of information overload and establishment of accountability. The public can even develop their own methods of informing others during disasters, when it is unable to get what it deems as reliable, timely information from the central responding agencies. Online forums have allowed people to transcend geographical distances that normally constrain the reach of helping efforts, to share information and coordinate citizen-led efforts. [1]

There are a number of examples of social media being used during emergencies. A couple of examples are the Canterbury University's response to inform educational design with Facebook after the Canterbury Earthquake, where Facebook enabled on-going dialogue and

information sharing between the institution's staff and the wider educational community [2]. Another example is Sam Johnson's famous UC Student Volunteer Army where up to 10,000 students helped Christchurch people deal with the aftermath of the Canterbury Earthquake [3]. There are also practical examples of low quality information communicated during disaster [4]. The 2010 Haiti Earthquake was the first time the US government agencies employed social media and collaborative workspaces as the main knowledge sharing mechanisms [5].

1.2 Research Gap

Data quality criteria in emergency management are identified as accuracy, timeliness, consistency, completeness, relevancy and fitness for use [6]. However, data is not information, but structure and context make information out of data. There is a gap in research of what criteria leads to improved information quality in emergency management and how to achieve accuracy of information that generated for decision making. However, this research suggests that general criteria for information quality may be applicable to social media used for emergency management and methodology for calculate accuracy of the information that generated. No research suggests how information quality is achieved on social media used for emergency management and how to achieve accuracy of information However, some research discuss how information quality is achieved in general on social media: use a series of textual and social characteristics in order to separate high quality content from the rest on Yahoo! Answers [7].

Here are the information quality criteria that I use to validate accuracy of entries (posts).

Criteria	Type	Definition
Accuracy	Intrinsic	Information is corrected and verified
Consistency	Intrinsic	Information is brief and to the point
Relevancy	Contextual	Information is useful
Reliability	Intrinsic	Information is reliable and trustworthy
Timeliness	Contextual	Information is current for the task

Table 1:Information Quality Criteria

Social media clearly has strengths and presents opportunities in emergency management. Following table represent most significant tools of disaster response which is based on Twitter, a popular social media platform.

Features	Twitris	Senseplace 2	AIDR	EMERSE	APDM
Dedicate for disasters	✗	✓	✓	✗	✓
Analyze twitter posts	✓	✓	✓	✓	✓
Analyze retweets	✗	✗	✗	✗	✓
Validate accuracy of posts	✗	✗	✗	✗	✓

Table 2: Comparison between existing systems

From the above table, we can clearly figure it out that all existing systems are focus on finding information about disaster situation, not to validate the accuracy of the information.

1.3 Research Problem

Decision makers face a number of challenges during an emergency. Followings are some of them

- information overload
- insufficient streams of information
- low information quality
- unreliability of information
- incorrect and false information

All of above mention challenges are around accuracy of incoming information. Managing information flows is critical in an emergency because decision making is bounded by time urgency or information may become outdated as the conditions change.

Social media used for emergency management and emergency response must satisfy requirements for high standards of reliability and accuracy. But the most lacking feature of current systems available is measuring accuracy and 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. The unstructured data needs to be cleaned in order to be used in other stages. With the empowerment of the general public and the abundance of information on social media, identifying ways of calculating accuracy levels for entries is central for decision makers to achieve an effective and efficient outcome in the emergency response. This part of research project covered the accuracy measuring of incoming information during emergency situation.

1.4 Research Objectives

1.4.1 Main Objective

Social media is a key source of people's help and information in disaster situations. We cannot trust everything on the internet. There will be the same false information as the correct information. It is important to communicate accurate and up-to-date information in crisis situations. Sharing false information can have catastrophic consequences if you switch resources from where you actually need them. Therefore, the accuracy of validation should be high despite the delay in response time. The main objective of this research part is validating accuracy of entry information (entry post).

1.4.2 Specific Objectives

- To stop spreading false information
- To give accurate and up-to-date information
- To avoid hear-say rumors

2. METHODOLOGY

2.1 Methodology

This section includes detailed descriptions about the techniques and mechanism used to develop the propose research part. The descriptions include how software implementation of the project is carried out, what are the materials and data needed, and how they will be collected.

2.1.1 System Architecture and Implementation

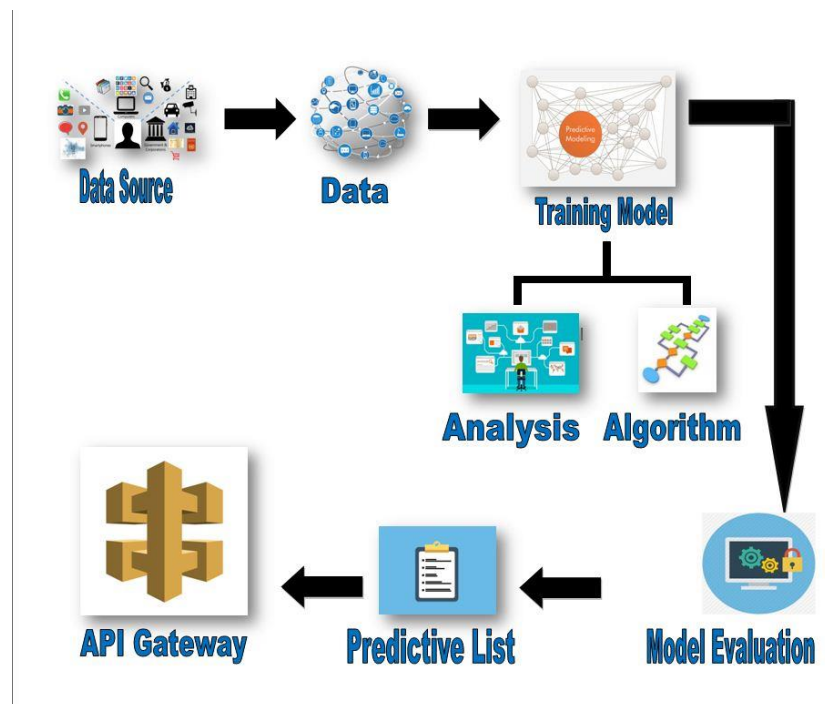


Figure 1: System Architecture diagram

Step 1: Data Integration

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. Therefore, integration of data is much important in this project. Data integration can be done by three steps.

- First step is data extraction – this means choosing right data to carry-out the project and extract data from the data source. To build my model I user disaster.com website as my main data source.



Figure 2: disaster.com logo

- Second Step is transforming data in to useful format which means a format that can access by the model. In machine learning we usually collect data into csv file because it's easy to access csv file from python code.

	A	B	C	D
1		post	comment	lable
2	1	I feel for these people deeply. They can not drink their water	You are absolutely right about that! I read that there are thousands and thousands of children who have been exposed	High
3	2	Tropical cyclone Vardah headed for Chennai	Tropical cyclone Vardah packing winds of 100 kilometers per hour (62 miles per hour) and is headed for the Bay of Bengal	Medium
4	3	Naivasha Tragedy: Burst tanker leaves 44 dead, 14 vehicles buried	Thirty nine people have died in a horrific accident after a tanker carrying highly flammable chemical lost control and ran into a wall	High
5	4	A shooter is at Washington DC, at the Capitol.	They are just reporting this shooting; so there is not a lot of information at this time. Apparently, at least one person was killed	Medium
6	5	A five storey building has collapsed in Istanbul killing 5 people	Like the building in Taiwan, this is definitely a combination of poor structural integrity and mild earthquake tremors. If it was a stronger earthquake, it would have been a disaster	Low
7	6	A splinter group of the Taliban is claiming "credit" for the murder of a US soldier	I heard similar type of news few months ago. I don't really understand what the terror group wants. They are just killing innocent people	High
8	7	It is being reported that there is a terrorist attack, or multiple attacks	What a horrendous situation. But sadly, I am not surprised. The UK is also on alert for up to '10 simultaneous attacks' in London	Medium
9	8	62 killed in Russian plane crash	Damn! That's so depressing that in 2016 travelling by plane is still considered kind of a risk.	Low
10	9	Tropical cyclone Vardah packing winds of 100 kilometers per hour	On Dec 11 all is pleasant about sudden an cyclone Vardah created a huge loss and sad to Chennai. Cyclone Vardah headed for Chennai	High

Figure 3: Dataset.csv file

- Third Step is load data in to model which means load the csv file in to machine learning model using python code to train the model.

```

8
9
10 df = pd.read_csv('dataset.csv')
11 df.head()
12
13

```

Figure 4: Python code for load csv file

Step 2: Training Model

In model training we build algorithm to analyze input training data and generate classified dataset to use future prediction purpose. Those classified data will store in sav file and use when actual project runs.

Algorithm Used – Naïve Bayes Classifier

It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. For example, a

fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as ‘Naive’. Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods [11].

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

Likelihood
Class Prior Probability

Posterior Probability
Predictor Prior Probability

$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \dots \times P(x_n|c) \times P(c)$$

```
from sklearn.feature_extraction.text import TfidfVectorizer

tfidf = TfidfVectorizer(sublinear_tf=True, min_df=1, norm='l2', encoding='latin-1', ngram_range=(1, 2), stop_words='english')
features = tfidf.fit_transform(df.comment).toarray()
labels = df.value
features.shape

from sklearn.feature_selection import chi2
import numpy as np

N = 2
for label, value in sorted(value_to_id.items()):
    features_chi2 = chi2(features, labels == value)
    indices = np.argsort(features_chi2[0])
    feature_names = np.array(tfidf.get_feature_names())[indices]
    unigrams = [v for v in feature_names if len(v.split(' ')) == 1]
    bigrams = [v for v in feature_names if len(v.split(' ')) == 2]
    # print("# {}:".format(label))
    # print(" . Most correlated unigrams:\n. {}".format('\n. '.join(unigrams[-N:])))
    # print(" . Most correlated bigrams:\n. {}".format('\n. '.join(bigrams[-N:])))

X_train, X_test, y_train, y_test = train_test_split(df['comment'], df['label'], random_state = 0)
count_vect = CountVectorizer()
X_train_counts = count_vect.fit_transform(X_train)
tfidf_transformer = TfidfTransformer()
X_train_tfidf = tfidf_transformer.fit_transform(X_train_counts)

classifier = MultinomialNB().fit(X_train_tfidf, y_train)
```

Figure 5: Implementation of Naive Bayes Model


```

53
54 pickle.dump(classifier, open('ValidityAlgo.sav', 'wb'))
55

```

Figure 6: Save classified dataset to sav file

Step 3: Model Evaluation

Evaluate models to find out the best model fit with Naïve Bayes model to increase the accuracy of the training model. Here I compare another three modules with Naïve Bayes model to find the most accurate model.

```

86
87 from sklearn.linear_model import LogisticRegression
88 from sklearn.ensemble import RandomForestClassifier
89 from sklearn.svm import LinearSVC
90 from sklearn.model_selection import cross_val_score
91 models = [
92     RandomForestClassifier(n_estimators=200, max_depth=3, random_state=0),
93     LinearSVC(),
94     MultinomialNB(),
95     LogisticRegression(random_state=0),
96 ]
97 CV = 5
98 cv_df = pd.DataFrame(index=range(CV * len(models)))
99 entries = []
100 for model in models:
101     model_name = model.__class__.__name__
102     accuracies = cross_val_score(model, features, labels, scoring='accuracy', cv=CV)
103     for fold_idx, accuracy in enumerate(accuracies):
104         entries.append((model_name, fold_idx, accuracy))
105 cv_df = pd.DataFrame(entries, columns=['model_name', 'fold_idx', 'accuracy'])
106 import seaborn as sns
107 sns.boxplot(x='model_name', y='accuracy', data=cv_df)
108 sns.stripplot(x='model_name', y='accuracy', data=cv_df,
109              size=8, jitter=True, edgecolor="gray", linewidth=2)
110 plt.show()
111
112

```

Figure 7: Implementation to accuracy check

model_name

LinearSVC: 0.822890

LogisticRegression: 0.792927

MultinomialNB: 0.688519

RandomForestClassifier: 0.443826

Figure 8: Results of accuracy check

According to the results Linear SVM (Support Vector Machine) has more accuracy with Naïve Bayes model.

Step 4: Create API

Python API is created to predict incoming new Twitter posts' accuracy and update database with the accuracy of new post



```
jupyter ValidityController.py 30/09/2018
File Edit View Language
46
47 @app.route('/validation/partition/<string:partition_key>/sort/<string:sort_key>', methods=['GET'])
48 def get_post(partition_key, sort_key):
49     try:
50         response = table.query(
51             TableName='Posts',
52             KeyConditionExpression=Key('partition').eq(partition_key) & Key('hour-id').eq(sort_key)
53         )
54
55         print('The response is ')
56         print(response)
57         item = response['Items']
58         print('The item ')
59         print(item)
60         data = item[0]['text']
61         tokens = nltk.word_tokenize(data)
62         validity = "PENDING"
63
64         arr = []
65         arr.append(data)
66
67         if tokens[0] == 'RT':
68             result = loadModal.predict(count_vect.transform(arr))
69             validity = str(result[0])
70
71
```

Figure 9:Implementation of API Part 01



```
jupyter ValidityController.py 30/09/2018
File Edit View Language
63
64     arr = []
65     arr.append(data)
66
67     if tokens[0] == 'RT':
68         result = loadModal.predict(count_vect.transform(arr))
69         validity = str(result[0])
70
71
72     if data is not None:
73         response = table.update_item(
74             Key={
75                 'partition': item[0]['partition'],
76                 'hour-id': item[0]['hour-id'],
77             },
78             UpdateExpression="set validity = :val",
79             ExpressionAttributeValues={
80                 ':val': validity
81             },
82             ReturnValues="UPDATED_NEW"
83         )
84     return jsonify({'partition': item[0]['partition'], 'hour-id': item[0]['hour-id'], 'validity': validity})
85
86 except ClientError as e:
87     print(e.response['Error']['Message'])
88
```

Figure 10:Implementation of API Part 02

2.1.2 Tools

- Anaconda Navigator

Anaconda Navigator is a desktop graphical user interface (GUI) included in Anaconda distribution that allows you to launch applications and easily manage conda packages, environments and channels without using command-line commands [8]. I used Anaconda Navigator to access Jupyter Notebook for implementation purpose.



Figure 11:Anaconda Navigator logo

- Jupyter Notebook

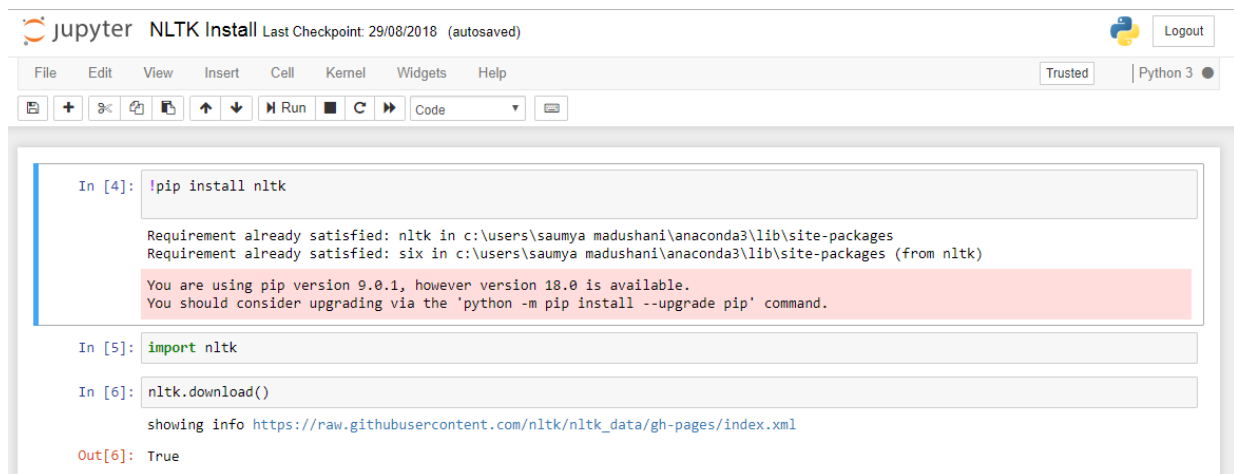
The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning [9]. I used Jupyter Notebook to implement my python code for model building.



Figure 12:Jupyter Notebook Logo

- NLTK (Natural Language Tool Kit)

NLTK (Natural language tool kit) is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum.



```

In [4]: !pip install nltk

Requirement already satisfied: nltk in c:\users\saumya madushani\anaconda3\lib\site-packages
Requirement already satisfied: six in c:\users\saumya madushani\anaconda3\lib\site-packages (from nltk)
You are using pip version 9.0.1, however version 18.0 is available.
You should consider upgrading via the 'python -m pip install --upgrade pip' command.

In [5]: import nltk

In [6]: nltk.download()

showing info https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/index.xml

Out[6]: True

```

Figure 13: Install NLTK to Anaconda Navigator

2.1.3 Technologies

- Python

Python is an interpreter, object-oriented, high-level programming language with dynamic semantics. Its high-level built in data structures, combined with dynamic typing and dynamic binding, make it very attractive for Rapid Application Development, as well as for use as a scripting or glue language to connect existing components together. Python's simple, easy to learn syntax emphasizes readability and therefore reduces the cost of program maintenance. Python supports modules and packages, which encourages program modularity and code reuse. The Python interpreter and the extensive standard library are available in source or binary form without charge for all major platforms and can be freely distributed



Figure 14:Python logo

- Scikit-Learn

Scikit-learn is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-mean and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.[10]



Figure 15:Scikit-Learn logo

2.2 Testing

Testing is done in stages while building algorithm

- Test to find the training dataset is loading

Out[5]:

Unnamed: 0		post	comment	lable
0	1	I feel for these people deeply. They can not d...	You are absolutely right about that! I read th...	High
1	2	Tropical cyclone Vardah headed for Chennai	Tropical cyclone Vardah packing winds of 100 k...	Mediam
2	3	Naivasha Tragedy: Burst tanker leaves 44 dead,...	Thirty nine people have died in a horrific acc...	High
3	4	A shooter is at Washington DC, at the Capitol.	They are just reporting this shooting; so ther...	Mediam
4	5	A five storey building has collapsed in Istanb...	Like the building in Taiwan, this is definitel...	Low

Figure 16:Test 01 results

- Test the model to find the most frequent terms in dataset

```
# 'High':
. Most correlated unigrams:
. state
. people
. Most correlated bigrams:
. damn depressing
. terror group
# 'Low':
. Most correlated unigrams:
. travelling
. kind
. Most correlated bigrams:
. travelling plane
. kind risk
# 'Mediam':
. Most correlated unigrams:
. hour
. uk
. Most correlated bigrams:
. hit awful
. horrendous situation
```

Figure 17:Test 02 results

- Manually test if the model is working correctly

```
In [7]: print(clf.predict(count_vect.transform(["China has really been getting it lately, especially the Asian continent. Good thing it w
['Mediam']
```

Figure 18:Test 03 results

3. RESULTS & DISCUSSIONS

3.1 Research Finding and Discussion

The root cause behind the idea for developing an API for analyze social media data in disaster management was there is not many social media analysis tools around this area. But there is vast amount of information generate during disaster situation. Therefore, we did some literature survey to find the current usage of social media in disaster management. From those findings, we realized that social media can play vital role in emergency management. One of biggest requirement was to identify accuracy level of social media posts during emergency situation as literature survey clearly shown that false information is spreading while disasters.

This requirement can achieve by analyzing all the comments related to a particular post on a social media page. There is a possibility that the same term may reoccur. Therefore, it was proposed to identify the recurring count of a term as well.

During implementation, it was discovered using Python for implementations of algorithms would be efficient due to the fact that machine learning algorithm can easily develop using python code. Further, another finding was that the Naïve Bayes algorithm would be the most efficient in this scenario to extract keywords. This conclusion was made comparing accuracy and efficiency with bag-of-words algorithm and tf-idf implementation as Naïve Bayes Classifiers had the upper hand when both accuracy and efficiency combined. After 90% of development, it is fortunate to state that the validating accuracy feature was developed successfully and works as expected.

4. CONCLUSION

Previous studies have agreed that social media has the potential to save lives in an emergency, but it is not without its challenges. With the empowerment of the general public and the abundance of information on social media, calculating accuracy of information is central for decision makers in achieving an effective and efficient outcome in the emergency response. However, there is a gap in previous literature because there is no recognized way to achieve accuracy of information in the use of online social media for emergency management.

Using verified, validated and timely information can decrease human loss in emergency situation. Throughout this research work I tried to fill the gap in previous literature using current solution.

5. REFERENCE

- [1] Turoff, M., Van de Walle, B. & Hiltz, S. (2009). Emergency Response Information Systems: Past, Present, and Future in B. Van de Walle., M, Turoff and S, Hiltz (Eds.) Information Systems for Emergency Management, New York, M. E. Sharpe, 369-387.
- [2] Dabner, N. (2012). ‘Breaking Ground’ in the use of social media: A case study of a university earthquake response to inform educational design with Facebook. Internet and Higher Education, 15, 69 – 78
- [3] MacManus, R. (2011, February 27). #EQNZ: Social Media response to Christchurch Earthquake. Accessed 5 April 2018
- [4] Sam Sachdeva, (Nov 15 2016) Earthquake: Why do people share fake pictures and videos on social media after a disaster? Accessed 20 March 2018
- [5] Yates, D. & Paquette, S. (2011). Emergency knowledge management and social media technologies: A case study of the 2010 Haitian earthquake. *International Journal of Information Management*, 31, 6 – 13.
- [6] Figueiredo, F., et al. (2012). Assessing the quality of textual features in social media. Information Processing and Management,
- [7] Agichtein, E. et al. (2008). Finding high-quality content in social media. WSDM Conference 2008, 183-193.
- [8] Anaconda Navigator Documentation [Online]. Available: <https://docs.anaconda.com/anaconda/navigator/> [Accessed: Apr.04,2018]
- [9] [Online]. Available: <http://jupyter.org/> [Accessed: Apr.04,2018]
- [10] Wikipedia [Online]. Available: <https://en.wikipedia.org/wiki/Scikit-learn> [Accessed: Apr.20,2018]
- [11] Sunil Ray, (September 11, 2017) 6 Easy Steps to learn Naïve Bayes Algorithm (with codes in Python and R) [Online]. Available: <https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/> [Accessed: June 22, 2018]

6. GLOSSARY

Term	Definition
Natural Language Processing	Natural language processing (NLP) is a field of computer science, artificial intelligence and computational linguistics concerned with the interactions between computers and human (natural) languages.
Machine Learning	Machine learning is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of computer programs that can change when exposed to new data.
API – Application Programming Interface	Application program interface (API) is a set of routines, protocols, and tools for building software applications. An API specifies how software components should interact.
Twitter	A popular social network where people share information such as news, entertainment details, sports details and also everyday activities.
DynamoDB	Amazon DynamoDB is a fully managed proprietary NoSQL database service that supports key-value and document data structures and is offered by Amazon.com as part of the Amazon Web Services portfolio

SAV file extension	SAV is a file extension used for the saved data of SPSS (Statistical Package for the Social Sciences)
--------------------	--

7. APPENDICES

```
36
37 dynamodb = boto3.resource('dynamodb',
38                             aws_access_key_id="AKIAIRBULYW7MZY2WTCQ",
39                             aws_secret_access_key="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",
40                             region_name="ap-southeast-2",
41                             endpoint_url="https://dynamodb.ap-southeast-2.amazonaws.com"
42                             )
43
44 table = dynamodb.Table('Posts')
45
```

8. Figure 19: Database connection