VISUALIZATION OF SOCIAL MEDIA DATA IN TRACKING CRIME

A Project

Presented to the faculty of the Department of Computer Science

California State University, Sacramento

Submitted in partial satisfaction of
the requirements for the degree of

MASTER OF SCIENCE

in

Computer Science

by

Jitender Singh

SPRING
2022

VISUALIZATION OF SOCIAL MEDIA

DATA IN TRACKING CRIME

A Project

by

Jitender Singh

Approved by:

_________________________________, Committee Chair
Dr. Anna Baynes

_________________________________, Committee Chair
Dr. Xiaoyan Sun

____
Date

Student: Jitender Singh

I certify that this student has met the requirements for format contained in the University format manual, and this project is suitable for electronic submission to the library and credit is to be awarded for the project.

____________________________, Graduate Coordinator ____________________
Dr. Jinsong Ouyang Date

Department of Computer Science

Abstract

of

VISUALIZATION OF SOCIAL MEDIA DATA IN TRACKING CRIME

by

Jitender Singh

Predicting crime has become very important for the safety of civilians as there has been an unprecedented growth in crime related to social media. Police departments monitor social media platforms in a standard way to get information on crime. This project aims to facilitate law enforcement bodies in tracking crime data. We have proposed a system to classify the Twitter data into positive and negative sentiment tweets using deep learning. In the first part, we have explored several deep learning models to compare the result of sentiment analysis and implemented the model on the project. The Kaggle dataset has been used to train and test the model, which is subsequently larger than IEEE dataset. In the second part, we have used the computer visualization technique to shortlist the event from the challenge. We have developed several interactive graphs and dashboard systems to analyze the dataset in the project. For our research, we have taken IEEE's vast challenge dataset to validate the system, and the results we have achieved are encouraging. The IEEE challenge is a fictitious scenario where the challenge encourages finding the whereabouts of the displaced employees using Twitter data.

___________________________________, Committee Chair

Dr. Anna Baynes

____

Date

ACKNOWLEDGEMENTS

I would like to thank Professor Anna Baynes for introducing me to computer visualization and giving me a chance to work on such an innovative project. I appreciate her continuous involvement in the project and support throughout the journey. I am grateful to Professor Xiaoyan Sun for reviewing my project and readiness to take interest in my project. Lastly, I would like to thank my parents, family, and friends for their support and involvement in the journey toward completing a master’s degree.

TABLE OF CONTENTS

Page

LIST OF TABLES

Tables Pages

LIST OF FIGURES

Figures Pages

CHAPTER 1. INTRODUCTION:

1.1 Overview:

Social media has evolved into an essential driver for broadcasting information in diversified domains such as entertainment, business, crisis management, politics, and science. The low cost and ubiquitously are the main factors behind making social media a popular platform. The unprecedented growth of social media has resulted in the accumulation of data, termed social media Big Data [1].

Criminals use the Web for mobilization, online training, recruiting new members, financing, and mitigation. The Web and Social networks are clustering individuals with common interests closer. Sometimes such association leads to illicit events. Web content is becoming a source of criminal acts. At the same time, Web can also help devise tools to investigate and prevent criminal acts.

Data analysis can be performed based on two categories, the first is the crime against people, and the second is the crime against property. The crime of property is 72% of occurrences, while crimes against people has a preponderance of 61%. Automated mechanisms face the challenge as the language used in social networks is short and informal, making effective computational automated solutions complex. Secondly, Web content is primarily unstructured data, so complexity is even higher when Criminal Slang Expression (CSE) is used[2].

1.2 Motivation:

Technology's advancement and growth rate provide low price electronic devices to people, making them more dependent on smartphones and internet-connected devices. Social media prediction models help to establish a new business or new ideas. Twitter is popular social media platforms for machine learning projects. It allows sending messages and images, and users of social media comment about a particular crime resulting in more violence in public. On average, millions of tweets are shared on Twitter in one second. 37 billion tweets collected for analysis in 7 years have shown the trend of Twitter usage to spread worldwide [3].

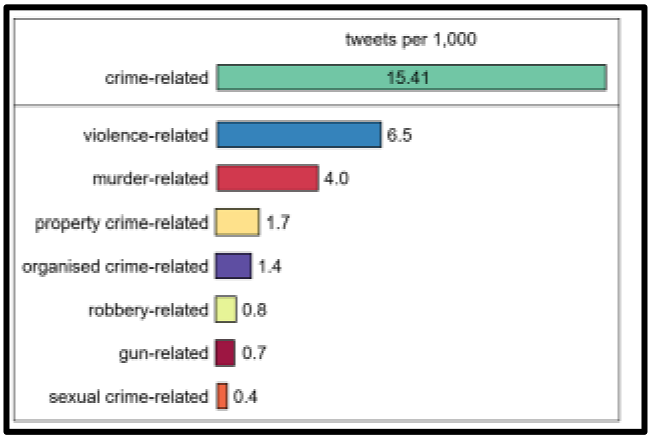
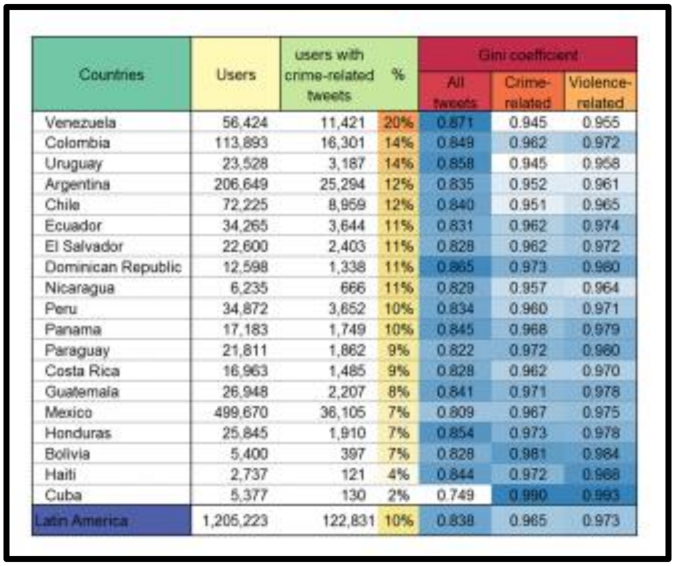


Figure 1: Crime tweets in Latin America per 1,000 [4]

Figure 1 gives an insight into the relationship between crime and tweets. Most of the crime related tweets are violence related, with more than 40% of all tweets under consideration.



Countries	Users	users with crime-related tweets	%	Gini coefficient		
				All tweets	Crime-related	Violence-related
Venezuela	56,424	11,421	20%	0.871	0.945	0.955
Colombia	113,893	16,301	14%	0.849	0.962	0.972
Uruguay	23,528	3,187	14%	0.858	0.945	0.958
Argentina	206,649	25,294	12%	0.835	0.952	0.961
Chile	72,225	8,959	12%	0.840	0.951	0.965
Ecuador	34,265	3,644	11%	0.831	0.962	0.974
El Salvador	22,600	2,403	11%	0.828	0.962	0.972
Dominican Republic	12,598	1,338	11%	0.865	0.973	0.980
Nicaragua	6,235	666	11%	0.829	0.957	0.964
Peru	34,872	3,652	10%	0.834	0.960	0.971
Panama	17,183	1,749	10%	0.845	0.968	0.979
Paraguay	21,811	1,862	9%	0.822	0.972	0.980
Costa Rica	16,963	1,485	9%	0.828	0.962	0.970
Guatemala	26,948	2,207	8%	0.841	0.971	0.978
Mexico	499,670	36,105	7%	0.809	0.967	0.975
Honduras	25,845	1,910	7%	0.854	0.973	0.978
Bolivia	5,400	397	7%	0.828	0.981	0.984
Haiti	2,737	121	4%	0.844	0.972	0.988
Cuba	5,377	130	2%	0.749	0.990	0.993
Latin America	1,205,223	122,831	10%	0.838	0.965	0.973

Figure 2: Frequency of people who posted a tweet about crime[4]

Figure 2 explains the ratio of crime related and violence related tweets divided by all tweets among users. Figure 2 shows that in Latin America only, there are 1.2 million users of Twitter, out of which 122831 have posted crime-related tweets. This is a strong motivation behind Visualization of Social Media Data in Solving Crime, mainly based on analyzing crime-related tweets.

1.3 Objective:

The main goal of Visualization of Social Media Data in solving crime and to analyze the data related to criminal activities. Although we have taken a fictitious scenario, our purpose is to generalize it to any real-world Twitter data. Thus, this research will focus on the effectiveness of Twitter in analyzing and visualizing crime using a set of words that can estimate the probability of a crime.

This research utilizes the 2014 IEEE VAST Challenge to help track the crime. This research can help law enforcement forces visualize the crime data and allocate resources for crime investigation. The annual IEEE VAST Challenge goal is to advance visualization in practice, visualization in Data Science, and data analysis challenge competitions.

1.4 IEEE VAST Challenge:

The 2014 IEEE VAST challenge presented a fictitious scenario: the scenario has mentioned about disappearance of GASTech employees, GASTech is an oil and gas company in Abila. The Tethys-based GAStech had been operational for almost 20 years. It was known for natural gas production in Kronos. It was a profitable organization for the Kronos government, but it was not an environment-friendly organization. The Gastech company was celebrating the success of its 1st public offering in January 2014 in the city of Abila. Few employees went missing during this celebration. The Protectors of Kronos (POK) are suspected of employees' disappearance. Our research has focused on the mini-challenge 3, which requires the identification of events for further investigation. The data provided are extracted from social media and emergency service data from real-time data. We have used the above fictitious scenario to conduct our research. The challenge provides access to a CSV data stream from 2 primary sources:

1. Automated filtered Microblog records potentially relevant to the incident occurring in real-time.
2. The Kronos and Abila police and other emergency responder departments' emergency Texts

Challenge focuses on identifying risks and mitigating them more effectively.

1.5 Methodology:

The relevant data will be mined and observed, focusing mainly on Tweets, responses, and various other cases of people who have posted relevant to committing a crime or have shown attention in a crime. The information collected needs to be processed and grouped. The 1st step in this approach is to preprocess the Twitter data. The preprocessing will involve Data Pre-processing, Tokenization of Data, Removing the Stop Word, and combining words to form sentences. The machine learning algorithm will be applied to classify disaster and non-disaster tweets and then analyze them onto the graph. We have used the Natural Language processing challenge datasets from Kaggle (Open-Source Platform) to provide accuracy to our machine learning algorithm. Valence Aware Dictionary and Sentiment Reasoner (VADER) is used to perform sentiment analysis based on lexicon and rule-base before applying deep learning models like (LSTM+DNN).

1.6 Dataset:

Twitter has become an important communication source in challenging times. The availability of smartphones has helped its users to announce emergencies instantly. That is why news agencies and disaster relief organizations are keen to monitor Twitter. We have used two datasets in this project. The first dataset is the **Disaster Tweets** used by Kaggle competitions which determines whether a tweet is related to crime. The second dataset is related to a fictitious situation where the developed models help law enforcement bodies from Kronos and Tethys analyze the situation and try to trace the employees who are marked as missing to come back home. The dataset Mini-Challenge 3 is based on feeds of microblogs (Twitter Data) and emergency calls (Message Broadcasted by Emergency Responder).

1.7 Report Organization:

The rest of the report is ordered as below. Chapter 2, Literature Study, contains essential information about the literature review. Chapter 3 talks about the Design and Analysis of Data. It talks about various sub-topics like an overview of data, data preprocessing methods, and data sources. Chapter 4 - Design and Implementation- talks about design choices and implementation procedures of deceptive and non-deceptive interactive visualizations used in this project. Next, we have Chapter 5 - Results, where we have widely discussed the results implemented in this proposed project. Chapter 6 is the Conclusion.

CHAPTER 2: LITERATURE REVIEW:

Sentiment analysis is the art of classifying text about a particular product as positive sentiment sentences, negative sentiment sentences and then third type of neither positive nor negative sentences. The analysis is conducted using lexicon-based algorithms or machine learning models. The lexicon-based approach counts the negative and positive words related to the data. The machine learning approach is based on algorithms to infer and classify sentiment from data. Many practical and accurate models have been developed to conduct sentiment analysis for different use cases [5].

2.1 Overview of the Sentiment Analysis Methods Used:

The sentiment analysis can be based on either the Lexicon method which is calculating sentiment based on semantic orientation, ML methods, or a fusion of above mention methods. The Lexicon method of semantic orientation works like an unsupervised learning method as it doesn’t need any training of data and is based on the dictionary. Most of the research in this domain uses Sentiwordnet and TF-IDR methods. This approach considers the occurrences of the positive or negative words in the sentence text with other lexicons like Sentiwordnet [6]. The research by Ashir is based on a rule and semantic-based procedure. It’s combined with unsupervised ML methods for generalization on different social media platforms. The approach uses a ruled-based technique for word contraction, expansion, emoticon detection, semantics-based text preprocessing, and noise removal. The research proves that lexical features in unsupervised learning perform better than word embeddings. Twitter samples' datasets give an accuracy of 91.1% [7]. The research detects a person's mental state by extracting emotions from the text and applying NLP techniques, using data posted on social media platforms [8]. However, lexicon-based solutions are easy to handle as these are based on only counting positive and negative words. It works for a different language and provides a reasonable speed to complete analysis.

The research by Wagh et al. uses a 4-layer system for Twitter data Sentiment Analysis. It uses a Bidirectional Encoder Representations from Transformers (BERT) model for encoding to do sentence generation depictions. The research has shown reasonable results on all datasets. The model has shown an accuracy of 71.82% on the SemEval 2017 dataset. [9]. The research works on the polarity of words from twitter data by performing feature extraction and using dictionary-based methods. These results are compared with methods like Word2Vec, TF-IDF, and CountVectorizer. The word scoring is done using VADER dictionaries and SentiWordNet. The proposed method works well in detecting the polarity of words [10].

2.2 Social Media Platforms for Sentiment Analysis:

Social media analytics have established that tweet posts can help strengthen post-disaster management [11] [12] [13], predicting elections [14], national revolutions [15], and predicting infectious disease outbreaks [16]. Text mining can help in predicting crimes. Tracking crime can also be conducted through electronic networks, virtualized community surveillance, and official criminal history [17]. Figure 3 shows a choropleth map for the predicted crime which is derived from geo-location.

Social media can be classified into the following four groups: Micro-blogs (Twitter, Tumblr), Blogs (Reddit, Quora), Content communities like Youtube, Tiktok, Instagram, and Social networking sites like (FB, Linkedin, discord) [18]. The site which allow users to post small blogs, specifically Twitter, are the top-rated social media platforms for collecting user opinions. It is estimated that approximately 85% of research is based Twitter sentiment analysis. It is one of the top-visited websites and helps users to share their views on any issue. APIs in Twitter help people to retrieve data on any topic using hashtags keywords. It is estimated that about 500 million tweets are sent on Twitter daily, publicly accessible through API [19]. Having the most prominent social media users, Facebook is not the first preference for sentiment analysis. Its data is messy, unstructured, and full of short forms/abbreviations and spelling errors. Analysis is done to understand the criminal activity and design procedures for prevention and effective crime control [20]. Another study collected data from various social media sources, including blogs, forums, Expedia, YouTube, aggregators, Facebook, Twitter, mainstream media, and WordPress. The research confirmed that 88% of the data for natural language processing comes from Twitter [21].

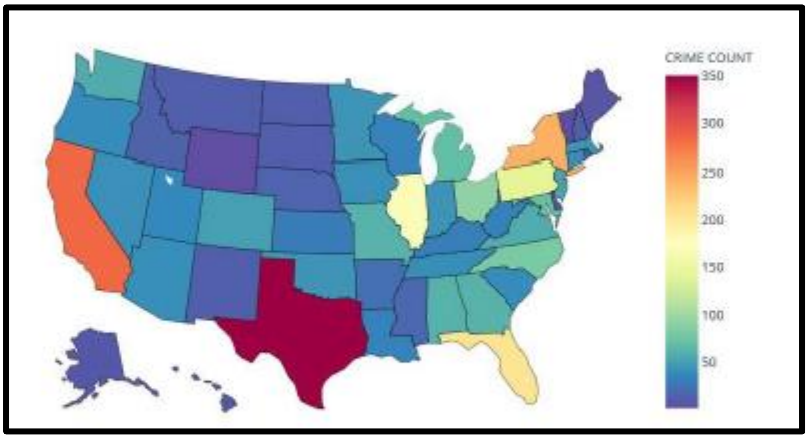


Figure 3: USA Choropleth map for the geo-location Tweets Crime[21]

2.3 Automatic Crime Prediction using Twitter:

The research at [22] has used social media, particularly Twitter, to predict future criminal incidents. Their research used the tweets from a news agency covering Virginia. A model was developed to predict threat levels and potential hit-and-run crimes. 3600 plus tweets regarding 290 hits and run incidents were collected to test the predictive model. Dimensionality reduction and prediction via linear modeling were used to train the model.

The research by Govindasamy has analyzed famous people's tweets and hashtags learn the pattern of their thinking and response. The conducted research extracts tweets from Twitter through a Twitter API to collect data. These tweets assess the importance of those events and the polarity of the tweets [23].

2.4 IEEE VAST Challenge 2014:

VAST 2014 Mini-Challenge 3 investigates the cause behind the disappearing GAStech employees. Clues to the investigation are crimps inside social media data. Much research has been carried out using interactive and visual analysis. The first author in [24] used a tableau to generate the timeline of events.

The generated timeline helped discover dense areas and visualize them using a stream graph, mood analysis graph, and word cloud. The visual analysis has helped predict the outcome of different events by using time stamps and the study of mood graph of what is going on and extract essential information from Twitter data. The analysis contains the information about the place of events mentioned in figure 4: TAG-Trouble At Gelato, Black Van, Police Standoff, APD-Abila Police Department, Terrorist, (Midsize event events gives an idea of a situation), i.e., Shooting Fire, Terror, Standoff. The Figure 4. Shows the word cloud for research discussed in [24].

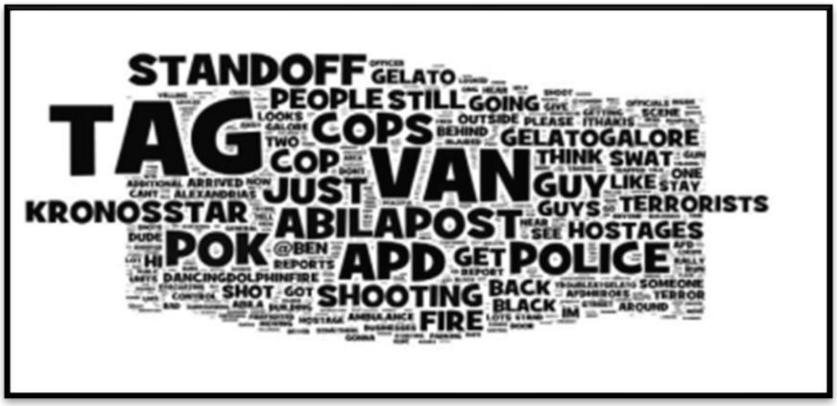


Figure 4: Word cloud [24]

Minjing Mao [25] has proposed the work using JavaScript, Java, and Postgres. The author [25] has drawn three synchronized charts of subjects, action, and location on the same dashboard. The drawn chart shows the peak of an event, then the author uses a cursor to pick an event and view the message. The Twitter message can be seen on the second screen, shown in figure 5.

The abstracted three critical events from the research are the fire at the Dancing Dolphin apartment, the accident between a black van and a bike, and the gunfire. The events are later visualized on a bubble chart to find the relationship between words and their importance based on the size of the bubble chart. Figure 5 illustrates the Dashboard for the author's work.

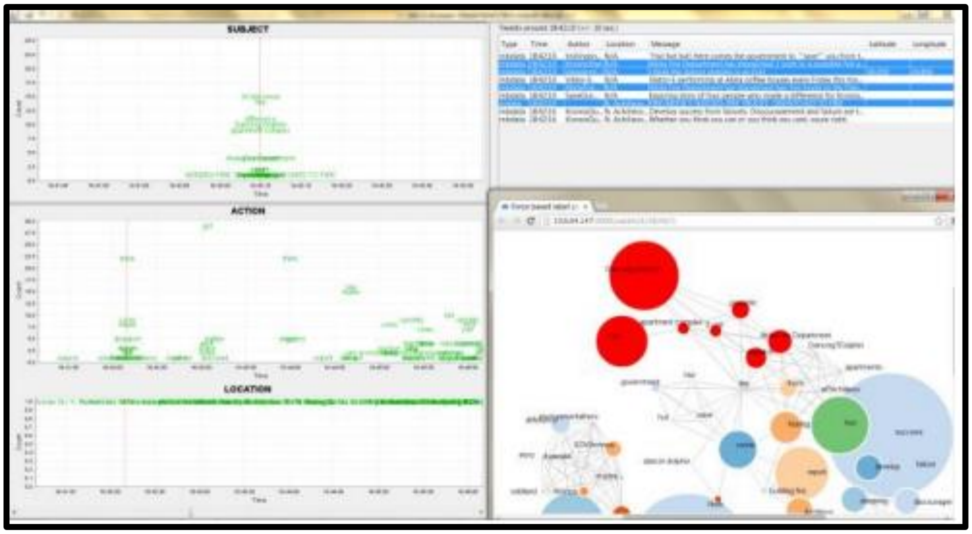


Figure 5: Dashboard [25]

Fabian Fischer [26] has focused on the most critical time segment to analyze and extract into the sliding slices. The essential findings of this research are Dolphins Apartment Fire, Black Van, Hostages, and Shooting near Gelato Galore. The authors have used REST Service, Spark Service, and NVisAware (real-time visual analytic tool to enhance situational awareness). The REST service is an interface to connect data stream and preprocess; Events stored in DB and real-time summaries created for these events. Later these real-time summaries are fed into the NVisAware web application [27].

NvisAware web application provides visual analytics and clustering to condense the data stream to meaningful sentences, but with certain limitations of performance and scalability issues due to HTML and JavaScript. The backend can cap collection and provide data circulation and keeping strategies to multiple slides, but the graphical user interface cannot. Figure 6 explains VAST 2014 Mini-Challenge 3 Recognition architecture, while Figure 7 illustrates research work conducted by authors [28].

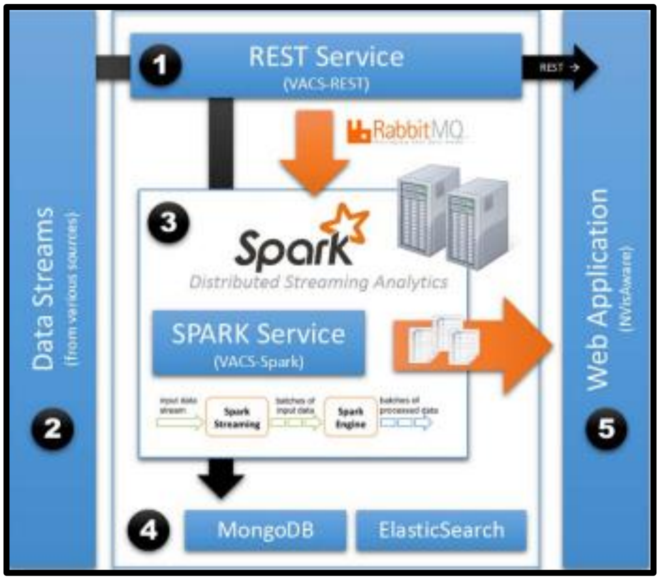


Figure 6: NStreamAware Architecture [29]

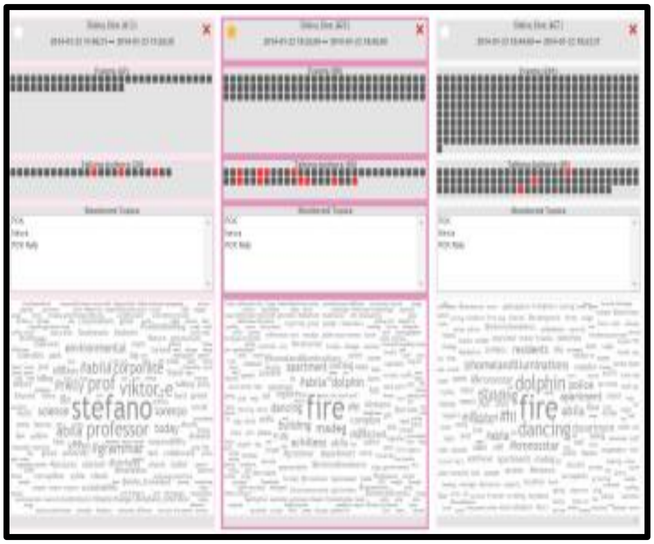


Figure 7: Sliding Slices [26]

Furthermore, the authors [26] have proposed a 2-monitor setup. The first monitor display contains a few linked components, including

A. Table to show the user selected message

B. Overall volumes of the sentimental value of messages

C. List of words selected previously

D. Geo-location of words listed in the second component

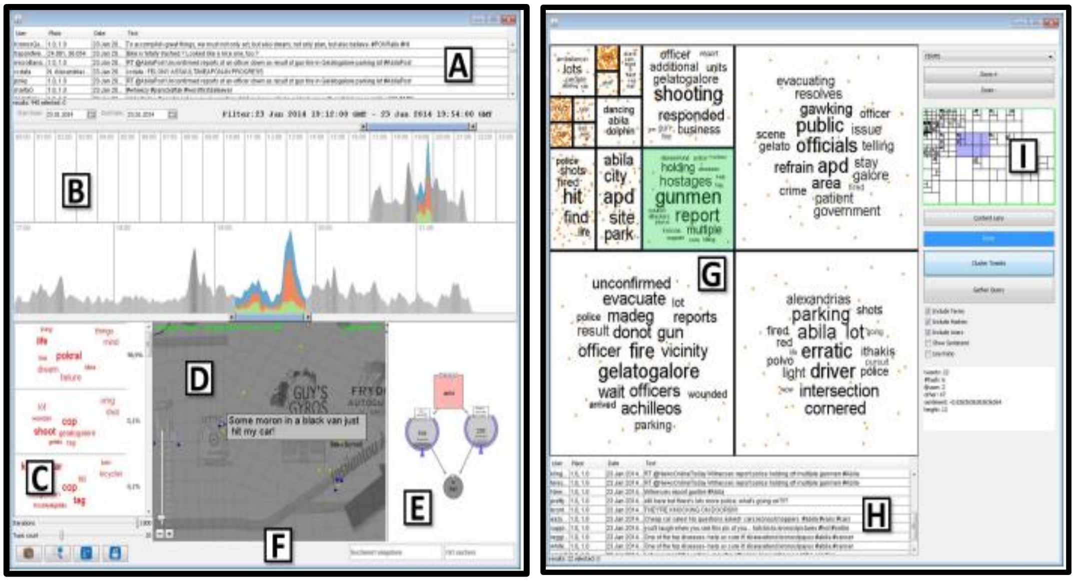
E. Filter-flow graph to store and combine the configurations

F. Movable lens.

While the second monitor setup shows the keyword shortlisted in the first screen setup. The main finding of the research work is stated below:

- 18:40 at the Dolphin Apartment building of Abila
- the second-largest peak was at 19:30 when the shooting took place between the police and the driver of a black van.
- The hostages are kept prisoner in the vehicle.

This work modified Scatterblogs [26], which has setup situational awareness based on tweets located on map and employed an SVM classifier trained on extensive historical data to find relevant information. Figure 8 elaborates the 2--Monitor Screen Setup.



Filter:23 Jan 2014 19:12:00 GMT - 23 Jan 2014 19:54:00 GMT

Some moron in a black van just hit my car!

Figure 8: 2-Monitor Screen Setup [29]

CHAPTER 3: DESIGN AND ANALYSIS OF DATA:

This chapter discusses the methodologies used to preprocess and overview the datasets. The dataset for this research is obtained from the open-source platform Kaggle. The dataset name is" Natural Language Processing with Disaster Tweets". Another dataset we have used is "Mini-Challenge 3" from IEEE Vast Challenge 2014.

3.1 Data Pre-processing:

Data preprocessing is the crucial steps in text analysis. We consider several factors while performing the data preprocessing for any visualization work. It is one of the most critical steps in generating the desired result, especially while handling the raw data from social media data. The text from social media platforms is full of noise and irrelevant characters such as HTML tags. So, the algorithms which are used to clean data from Twitter messages include eliminating commas, and illustration, performing tokenization, stemming, and stop word. A clean tweet must not have weblinks, hashtags, or tag (i.e., tag to some person or place). In addition to this, line breaks [30] and tabs should be replaced.

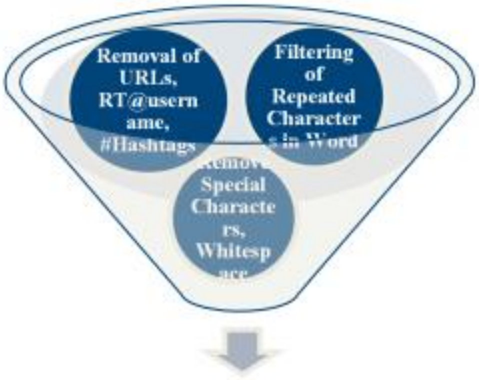

Removal of URLs, RT@username, #Hashtags
Filtering of Repeated Characters in Word
Remove Special Characters, Whitespace

Figure 9: Structure for preprocessing user tweets on Twitter [31]

The text might contain vowels like “hellooooo”, so these vowels should be removed repeatedly in sequence. Another replacement can be on laughs, where a sequence of \a” and \h” are to be analyzed. These can be replaced by smiling” Tag [32]

Removing emoji and emoticon from a dataset is another step of preprocessing. Emoji are small digital images or icons used to express emotion. These emojis symbol are small enough to be considered in the text. Emoticons represent facial expressions formed using a combination of keyboards [33].

3.2 Tokenization of Data and Removing Stop Word:

3.2.1 Tokenization:

After preprocessing and cleaning the noise from that data, we are left with raw words in sentences [34]. The word in sentences has significance and may represent feeling expressed by the person. Tokenizing is the first step in the NLP pipeline. Tokenization is splitting a sentence into tokens or words and returning a list of words. Every sentence gets its meaning by the word itself, not by the spaces, punctuation, and commas. There are different ways of tokenizing, i.e., using the inbuilt python method, regular expression, NLTK, Spacey, Keras, and Genism to analyze the sentences correctly and interpret the meaning of a text. For the project, I have performed tokenization using NLTK [35].

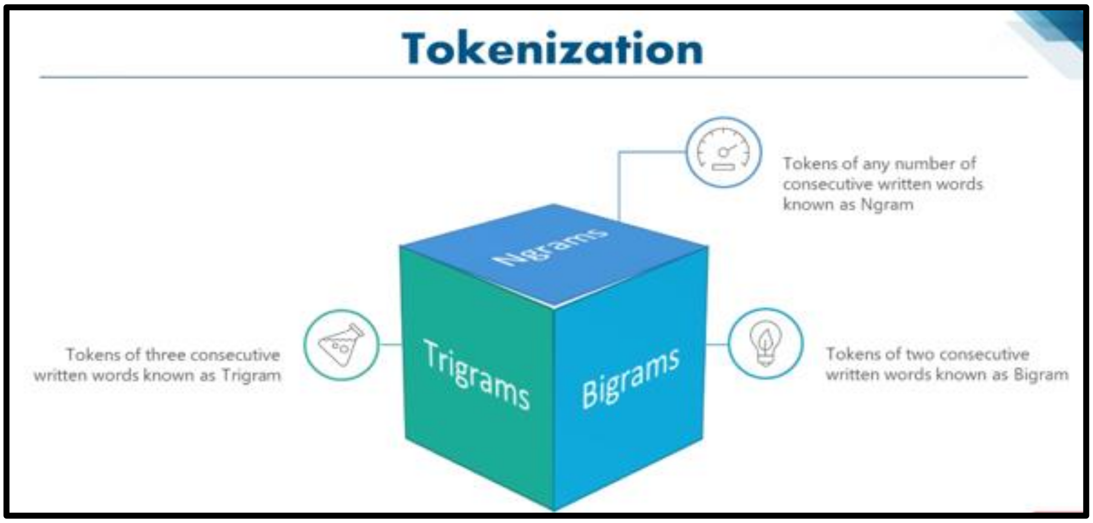

Tokenization
Tokens of any number of consecutive written words known as Ngram
Ngrams
Tokens of three consecutive written words known as Trigram
Trigrams
Bigrams
Tokens of two consecutive written words known as Bigram

Figure 10: Dividing text into tokens

3.2.1 Removing Stop Words:

The "stop words" or "stop word list" or "stop list", are used in sentences for grammar and tenses. They are very commonly used words in each language, so deleting them will not affect the data for analysis. If we eliminate them, we can focus on the actual issue. Similarly, the stop word takes space in DB and valuable processing time in analyzing the text. Removing them can save tons of computation power for large datasets. NLTK has inbuilt python library which has a list of stop words stored in Sixteen different languages. We have removed the stop words from tokenized tweets for our research. Some examples of stop words are "a", "an", "the", "our", and "hers".

Figure 11 demonstrates the process of tokenization and removing stop words. The input is the sentence: “our Deeds are the reason for this earthquake god may forgive us”. The sentence is tokenized into several words: In Figure 11 we can see the words.

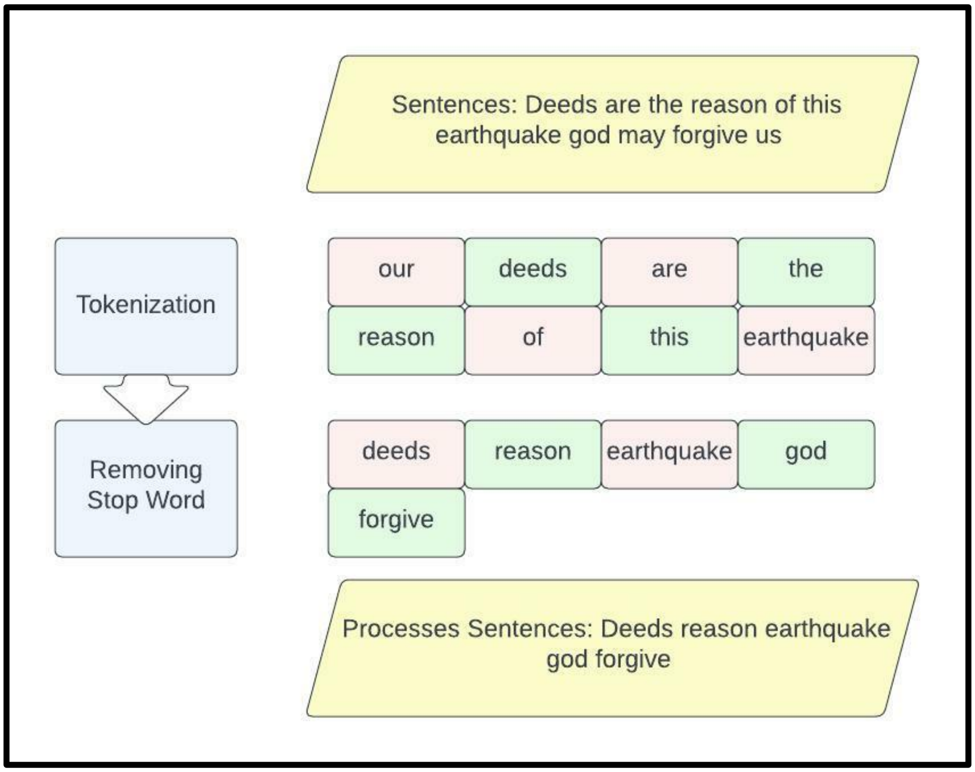

Sentences: Deeds are the reason of this earthquake god may forgive us
Tokenization
our
deeds
are
the
reason
of
this
earthquake
Removing Stop Word
deeds
reason
earthquake
god
forgive
Processes Sentences: Deeds reason earthquake god forgive

Figure 11: Tokenization and Removing Stop Word

After completing the tokenization process, the algorithm to remove stop words is applied. The output from the tokenization module is Deeds, reason, earthquake, God, forgive. It is evident that removing stop words significantly reduces computations as the input size to the algorithm is decreased.

3.3 Overview of Datasets:

Our research is based on two different datasets from open-source platforms i.e., the IEEE Vast Challenge dataset and Kaggle Natural Language Processing with Disaster tweets dataset. Both datasets are in an Excel/CSV format.

3.3.1 IEEE VAST Challenge Dataset:

The dataset is related to a fictitious situation on January 23rd, 2014. Few workers at GAStech have been missing for three days from their workplace. It is the focus of law enforcement agencies to trace the missing employees so that they can go back home safely. A visual analytic expert from law enforcement focuses on maintaining situation awareness as the crime situation becomes clear. The analyst can use one of the streams coming from these two major sources:

The automated filter separate microblog records that are possibly relevant to the incident

1. Text mined extracts of data broadcasted by law enforcement bodies of Kronos, fire departments, and police.
2. This data helps to find the actual happening behind the incident to trace missing employees. Maps of Abila and other relevant documents are available to the visual analytic expert.

There are three segments of data in this challenge:

Segment 1 This data is related to the period 1700 - 1830 Abila time on January 23. The data is available in comma-separated values (CSV) format.

Segment 2 This data is related to tweets in the period 1830 - 2000 Abila time on January 23.

Segment 3 This data covers the period 2000 – 2130 or a few extra moments after 2130 Abila time on January 23. The server allows access to it multiple times.

The proposed system has processed 2 types of data.

Microblog messages:

Microblog message is represented by mbdata. These blogs contain conventions from Twitter e.g., @, symbol (to attach a person name with the body of text), Hashtags (#) (to relate a message to a topic), "RT" (showing the retweets tweets). Few messages are either spam or junk. As messages can be sent in multiple short batches, more than one message may hold the same timestamp. They may also be delivered out of order. Their size may be up to 200 characters. Their format is as under:

1. Date: in YYYYMMDDHHMMSS
2. Author: author name, a text string
3. Message: maybe up to 200 characters
4. Latitude: latitude of the locations (optional)
5. Longitude: The locations (optional)

Call center data

Call center data is represented by ccdata. The data from call center is also part of the data stream. The data is in the format written below:

Date: in YYYYMMDDHHMMSS format.

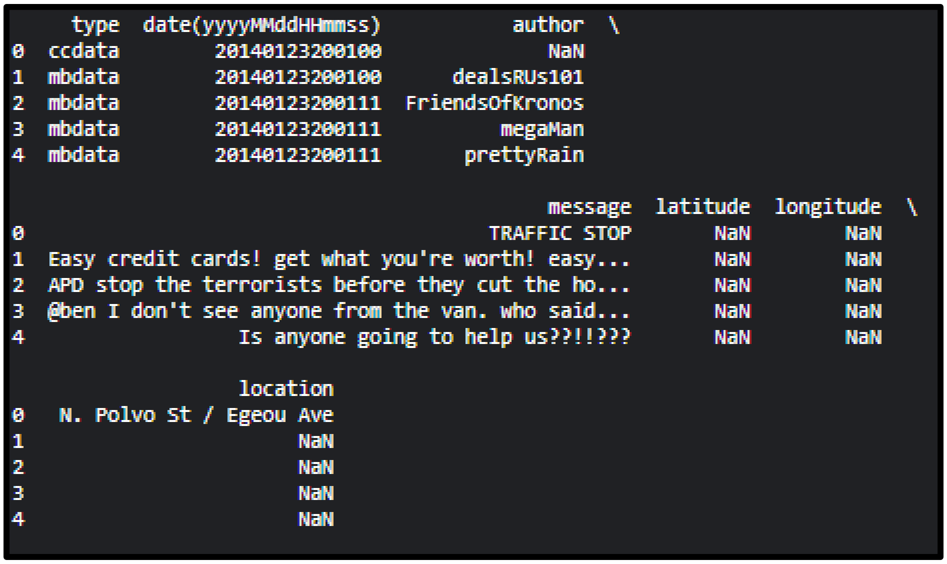
Message: this may be up to 200 characters

location: the address at which the incident happened (optional)

Each excel table in the VAST Challenge dataset contains the following attributes, i.e., message, date, location, latitude, and longitude. The data consists of three-time segments, i.e., the first segment is from 1700 to 1830, the second segment is from 1830 to 1900, and the last segment is from 1900 to 2000. The three CSV files used in coding are MC3a, MC3b, and MC3c as illustrated in Table 3.1.

Table 3.1: IEEE VAST dataset

Dataset Name	Rows	Columns
MC3a	1033	7
MC3b	1815	7
MC3c	1215	7



```
     type  date(yyyyMMddHHmmss)           author  \
0  ccdata        20140123200100              NaN
1  mbdata        20140123200100      dealsRUs101
2  mbdata        20140123200111  FriendsOfKronos
3  mbdata        20140123200111          megaMan
4  mbdata        20140123200111       prettyRain

                                          message  latitude  longitude  \
0                                    TRAFFIC STOP       NaN        NaN
1  Easy credit cards! get what you're worth! easy...       NaN        NaN
2  APD stop the terrorists before they cut the ho...       NaN        NaN
3  @ben I don't see anyone from the van. who said...       NaN        NaN
4                Is anyone going to help us??!!???       NaN        NaN

                location
0   N. Polvo St / Egeou Ave
1                      NaN
2                      NaN
3                      NaN
4                      NaN
```

Figure 12:IEEE VAST Challenge Dataset

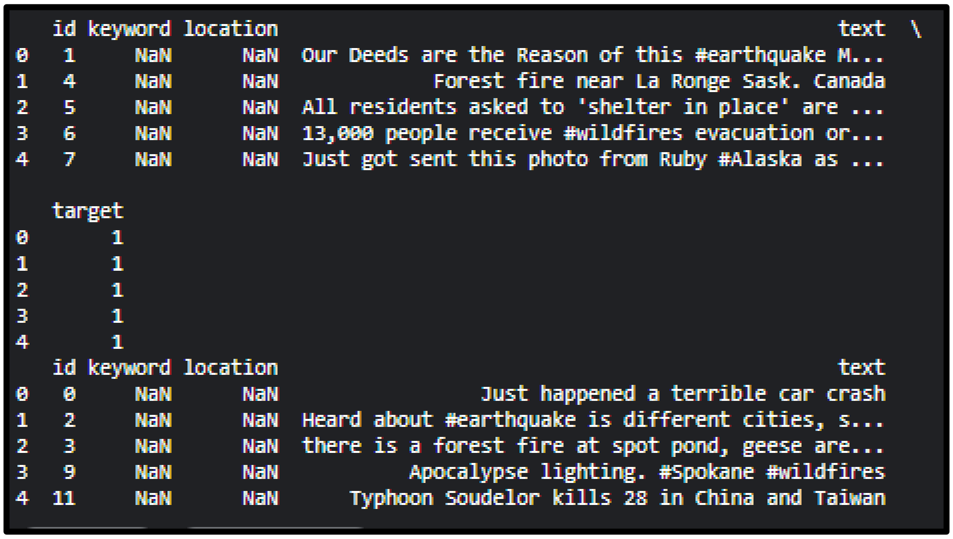
3.3.2 Kaggle Natural Language Processing Dataset:

The initial objective was to use the mini-challenge 3 dataset. But the data size is not enough to train the Machine learning model. To achieve more accuracy and better results for the sentiment analysis model, we have included the Natural Language processing Disaster tweets data from the open-source platform Kaggle.

The NLP twitter dataset has 10,000 tweets and five columns that hold information about the text of a tweet, the unique id for each tweet, the location of user, the keyword from that tweet, and the target (in training csv file only, it specifies whether a tweet is about disaster tweets). Moreover, the tweets dataset into two CSV datasets for training and testing. The number of rows and columns in the Kaggle dataset is substantially higher than in the Mini challenge dataset. The reason for using the Kaggle dataset for training and testing is that it contains a large number of tweets.

Table 3.2: IEEE VAST dataset

Dataset	Rows	Columns
Training	7615	5
Testing	3263	4



```
   id keyword location                                               text  \
0   1     NaN      NaN  Our Deeds are the Reason of this #earthquake M...
1   4     NaN      NaN             Forest fire near La Ronge Sask. Canada
2   5     NaN      NaN  All residents asked to 'shelter in place' are ...
3   6     NaN      NaN  13,000 people receive #wildfires evacuation or...
4   7     NaN      NaN  Just got sent this photo from Ruby #Alaska as ...

   target
0       1
1       1
2       1
3       1
4       1
   id keyword location                                               text
0   0     NaN      NaN                 Just happened a terrible car crash
1   2     NaN      NaN  Heard about #earthquake is different cities, s...
2   3     NaN      NaN  there is a forest fire at spot pond, geese are...
3   9     NaN      NaN           Apocalypse lighting. #Spokane #wildfires
4  11     NaN      NaN      Typhoon Soudelor kills 28 in China and Taiwan
```

Figure 13: Kaggle Natural Language Dataset

CHAPTER 4: DEVELOPMENT AND IMPLEMENTATION:

4.1 Introduction:

This chapter talks about different tools and techniques used in developing and implementing the project. It will highlight all-important programming languages, frameworks, and technologies. We are performing sentiment analysis on the tweet dataset discussed in the previous section to categorize them into disaster and non-disaster tweets. To train and test our model, we have used LSTM and VADER.

4.2 Glove Model:

This model uses the concept of one-word embedding applied to the whole corpus. GloVe model trains itself on existing counts of word in data, globally. It minimizes least-squares error and creates a word vector space with useful structure. The GloVe is an unsupervised ML algorithm it generates the vector representations for text.

The project was first tested on the Glove model with LSTM, but the results were not very favorable. So, the not-very-promising results led to exploring other models. The encouraging results of the project on VADER+LSTM+DNN led to shifting the architecture to VADER from Glove Model. When this model was compared with VADER, VADER seems to work better with things like slang, emojis. It is sensitive to punctuation and capitalization. Hence, it performs well, especially for social media sentiment analysis.

4.3 VADER Sentiment Analysis:

Sentiment analysis categorizes text into a positive or negative opinion. Based on specific rules. This analysis is used to measure sentiments, attitudes, and emotions. Text

Sentiment Analysis is done based on the lexical and machine learning approaches. VADER sentimental analysis is based on a dictionary that combines semantic features with sentiment scores. To calculate the sentiment score, we can accumulate the intensity of words in the text.

VADER is based on a lexical approach. Lexical approaches relate words with sentiment by developing a sentiment dictionary by creating a lexicon. This dictionary classifies the sentiment of phrases and sentences without training data. Sentiment is categorized into happy sentences as positive, sad sentences as negative, and third category of neither negative nor positive. It can have integer values like scores or intensities. This approach does not need any training or labeled data. Only the dictionary is used as it contains everything required to assess the sentiment of sentences in the dictionary of emotions. Vader sentiment analysis based on the predefined dictionary that maps semantic features to excitement intensities of the sentences which is called as sentiment scores [33]. Vader uses the advantages of parsimonious rule-based modelling to create a sentiment analysis engine that:

1) performs well for all platforms of social networks and generalizes well.

2) it needs no training data for its working

3) gives reasonable accuracy to be used with live streaming data

4) It is not affected by the speed-performance tradeoff.

As mentioned earlier, the sentiment calculated value of data can be calculated by adding up the powers of each word in the text. E.g., “the person doesn't love his dog”, in this sentence, the “love” has positive sentiment, and “doesn't” make it a negative statement.

The word “love” has a positive score of around 0.42, while for the word “doesn't,” the negative score is higher, resulting in a negative sentence overall. Vader can understand the positive, negative, capitalization, punctuation, and neutral statements, but can't determine the difference between the objective and subjective statements and facts. It aggregates whole sentence sentiment to find the opinion.

Compound Scores metric used for sentences:

1. Positive: more than or equal to 0.05
2. Negative: score of -0.05 or less
3. Neutral: score more than -0.05
4. Compound: less than 0.05

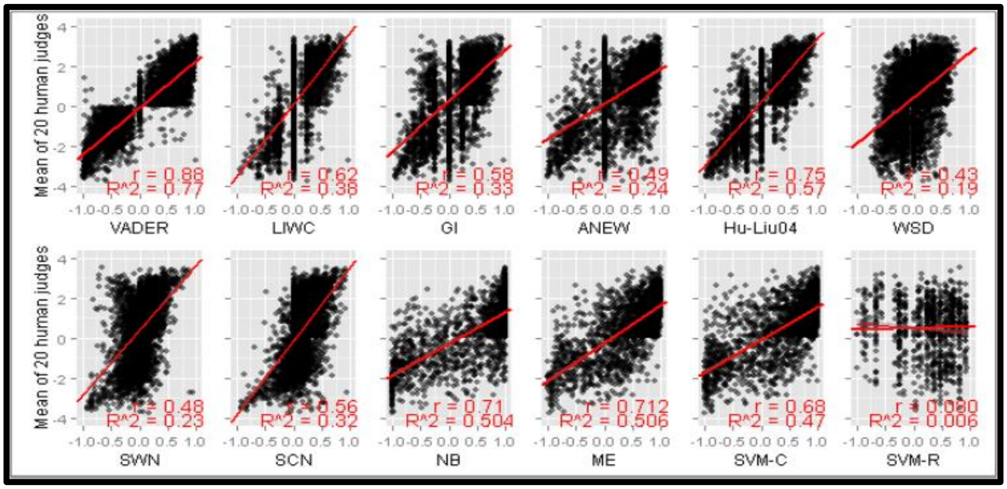


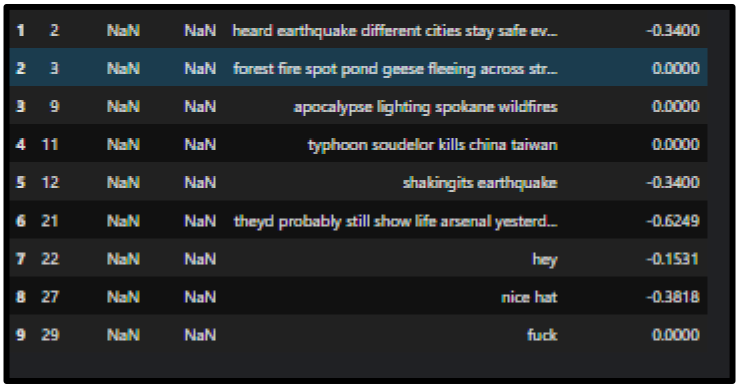
Figure 14: Comparison of Vader and others’ Sentiment Analysis Technique [33]

As Figure 14 shows, the VADER lexicon performs perfectly in all social media domains, even better than an individual human. Surprisingly, VADER (*F1* = 0.96), while the individual human raters classified the sentiment of tweets at (*F1* = 0.84). Figure 14 also compares different techniques applied to 4 different kinds of datasets.

The Figure 14 establishes that VADER is computationally efficient. Although ME and NB have outperformed VADER for the movie dataset, all other datasets in Table 4.1 have shown reasonable performance. Like some SVM models, the machine learning-based models could not fully process the data, especially for movie reviews and NYT editorials. Therefore, we have chosen VADER as part of our sentiment analysis model.

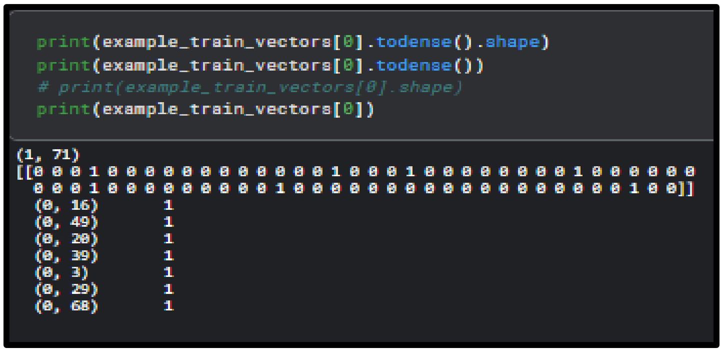
4.4. Count Vectorization:

Count Vectorization is the process of tokenization of the text with very little preprocessing. It involves eliminating punctuation marks and converting all text to lowercase. To encode unseen texts, a vocabulary of the known words is formed. An encoded vector is constructed for all the words from the vocabulary and the frequency of each word that appeared in the document. We have performed the count vectorization to convert the text document into the matrix of the token count. It is used for generating the vector representation which helps it to represent the feature [34]. Figure 15 shows the Vader prediction, while Figures 16 and 17 show Count Vectorization on Kaggle and VAST.



1	2	NaN	NaN	heard earthquake different cities stay safe ev...	-0.3400
2	3	NaN	NaN	forest fire spot pond geese fleeing across str...	0.0000
3	9	NaN	NaN	apocalypse lighting spokane wildfires	0.0000
4	11	NaN	NaN	typhoon soudelor kills china taiwan	0.0000
5	12	NaN	NaN	shakingits earthquake	-0.3400
6	21	NaN	NaN	theyd probably still show life arsenal yesterd...	-0.6249
7	22	NaN	NaN	hey	-0.1531
8	27	NaN	NaN	nice hat	-0.3818
9	29	NaN	NaN	fuck	0.0000

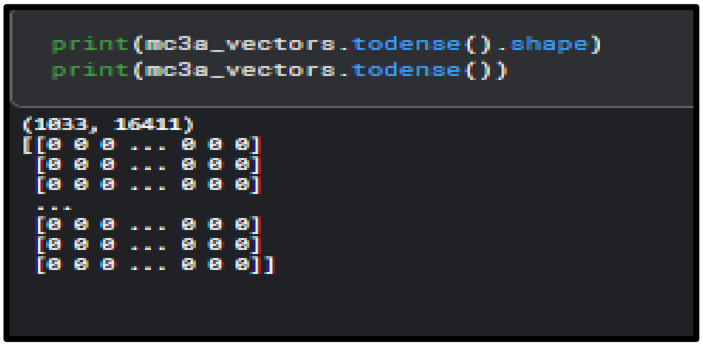
Figure 15: VADER Prediction



```
print(example_train_vectors[0].todense().shape)
print(example_train_vectors[0].todense())
# print(example_train_vectors[0].shape)
print(example_train_vectors[0])
```

```
(1, 71)
[[0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0
  0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0]]
  (0, 16)	1
  (0, 49)	1
  (0, 20)	1
  (0, 39)	1
  (0, 3)	1
  (0, 29)	1
  (0, 68)	1
```

Figure 16: Count Vectorization on Kaggle

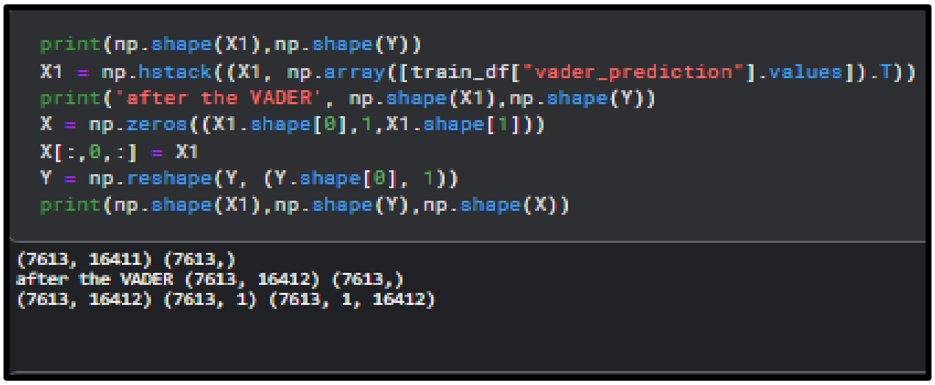


```
print(mc3a_vectors.todense().shape)
print(mc3a_vectors.todense())
```

```
(1033, 16411)
[[0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 ...
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]]
```

Figure 17: Count Vectorization on VAST

We have performed the count vectorization and Vader sentiment analysis for the NLP data. The sentiment score column of tweets sentence adds to the table of 16412 columns and 7613 rows shown in Figure 18, resulting in one additional column. Figure 18 elaborates the output at this stage after performing Vader sentiment analysis. The output vectors are of the shape 7613 x 16413, while the label has the dimension 7613.



```
print(np.shape(X1),np.shape(Y))
X1 = np.hstack((X1, np.array([train_df["vader_prediction"].values]).T))
print('after the VADER', np.shape(X1),np.shape(Y))
X = np.zeros((X1.shape[0],1,X1.shape[1]))
X[:,0,:] = X1
Y = np.reshape(Y, (Y.shape[0], 1))
print(np.shape(X1),np.shape(Y),np.shape(X))
```

```
(7613, 16411) (7613,)
after the VADER (7613, 16412) (7613,)
(7613, 16412) (7613, 1) (7613, 1, 16412)
```

Figure 18: Code Snippet of Vader sentiment analysis

4.5 LSTM Sequential Model:

LSTM is a recurrent neural network that uses deep learning. It is a unique model that will store separate memory cells that will only update and display their contents if necessary. It is a special type of Recurrent Neural Network, used for learning in long run and it will update cells accordingly. Figure 19 shows the architecture of an LSTM deep learning model.

This neural network has two types of memory: long-term and short-term. LSTM uses four gates for its working. Forget gate (decide whether to keep info or not), remember gate (combines long term memory and short-term memory to get new long-term memory), learn gate (combines short term memory and the current event), use gate (combines long term memory and short-term memory to get new short-term memory, it discards what is not needed). Figure 20 elaborates the relationship between these gates.

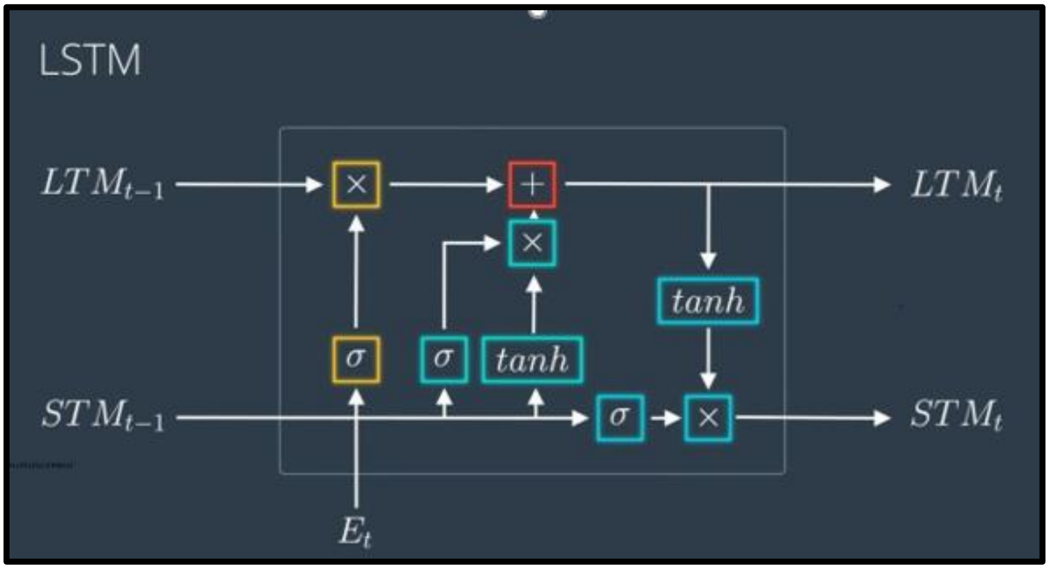


Figure 19: LSTM Network Block Diagram

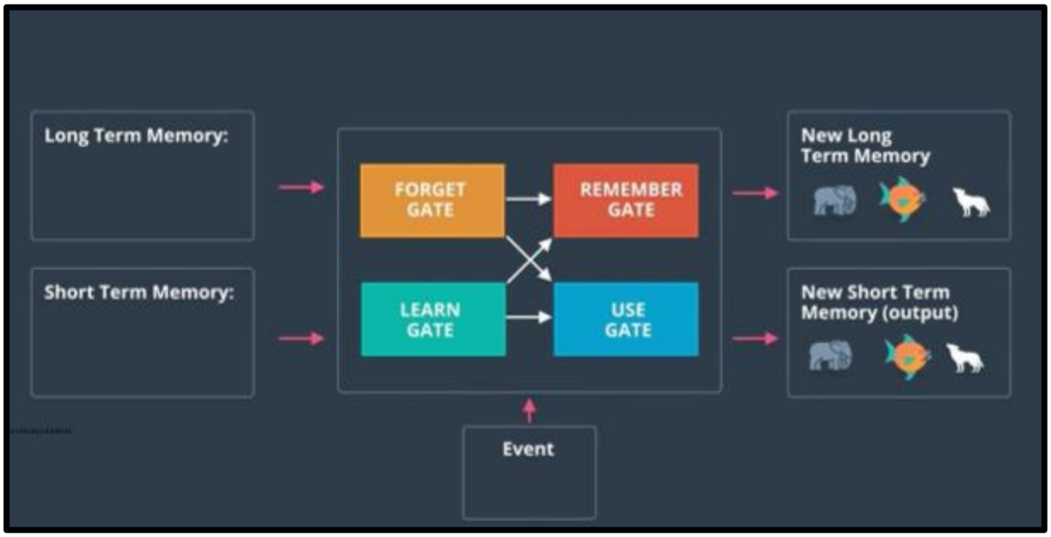


Figure 20: The Four Gates in LSTM Network

The proposed research has used 64 layer- LSTM network; its output is fed into the DNN component of the proposed model.

4.6 Deep Neural Network:

These networks are interconnected mazes of neurons. They are the most efficient models for text classification. These types of models view the texts as a bag of words. Every word they learn is in the form of vector representation produced by count vectorization. They take the vector sum or average of those values as a form of text and pass it through more than one feed-forward layer known as a multi-layer perceptron. At the output, the model performs classification.

In this project, we have used a combination of multiple dense layers:

1. 64-unit dense layer
2. 8-unit dense layer
3. 1-unit dense layer

Figure 21 elaborates on the architecture of the DNN used in the proposed solution. It is fed 64-neuron input, which is dropped down to 8 dense neuron layers. Its output is fed to the next layer with one neuron only.

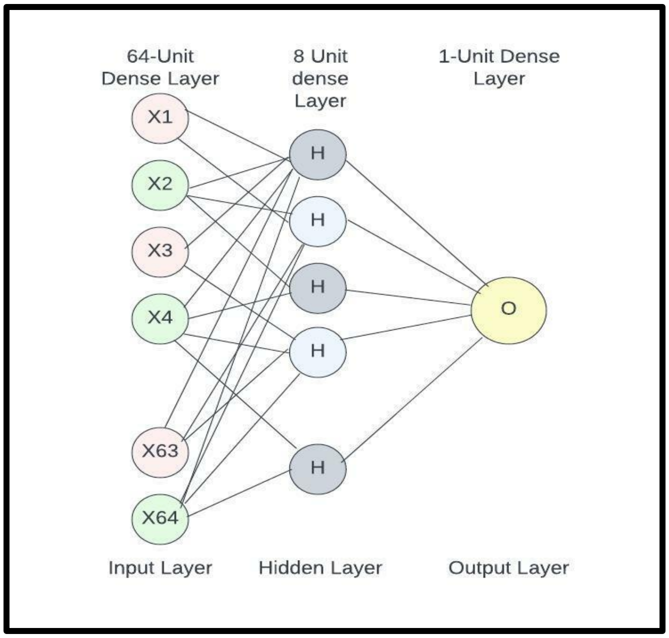


Figure 21: Maze of Neurons

Figure 22 shows the system's overall architecture starting from input to final prediction. The process starts with input from the Vast Challenge dataset and the data from the Kaggle dataset; this data is preprocessed. After processing by VADER, it is transferred to the LSTM model, and its output is passed through the visualization process, which gives the processed tweets. In the end, VADER processing and LSTM processing output are combined and fed to a 4-neuron layer that combines the output from the two. The last layer is a 1 neuron layer, which gives the final output.

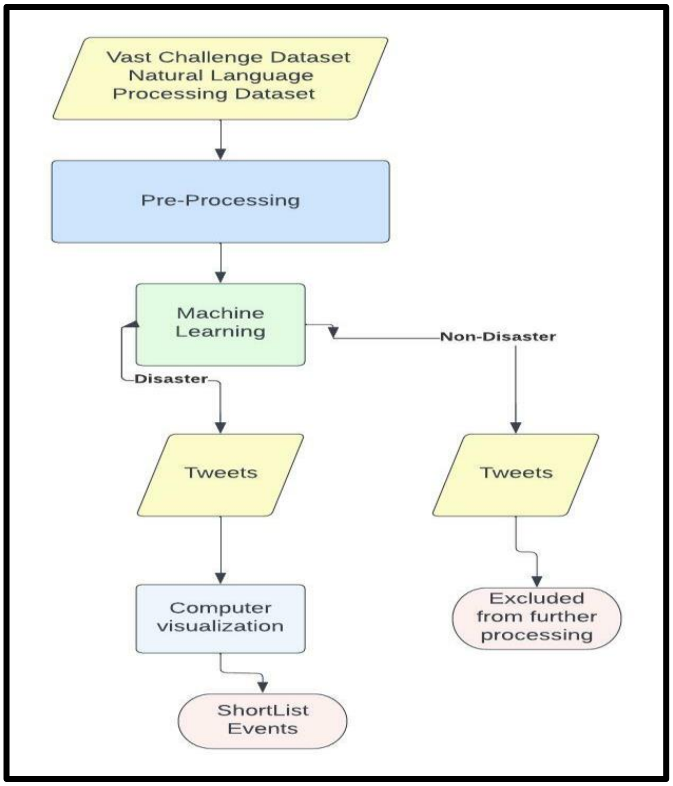


Figure 22: System Architecture

4.7 Model Architecture:

1. The input data with sentiment score added column fed into the input layer of the model.
2. The vectorized data will be sent into the LSTM layer, while the last column data will get fed into the Vader sentiment analyzer.
3. Then data from the LSTM layer is transferred to a 64-unit dense layer, 8-unit dense layer, and finally to a 1-unit dense layer simultaneously. LSTM layer uses a dropout of 20%.

4. The Dense layers after LSTM use “Relu” as an activation function. The last output dense layer uses “Sigmoid” as activation for last year.
5. The model uses “Adam” as the optimizer, the loss function is “BinaryCrossEntropy”, while the matrix of performance measure is “Accuracy”.
6. The previous layer's output concatenated with Vader output is passed to a 4-unit dense layer, it takes as input 2 neurons and after processing generates 4 neurons.
7. Then finally, into the 1-unit dense layer to get the output. Its input layer is fed 4 neurons.

Figure 23 elaborates on this architecture in detail, specifying the architecture used at each layer.

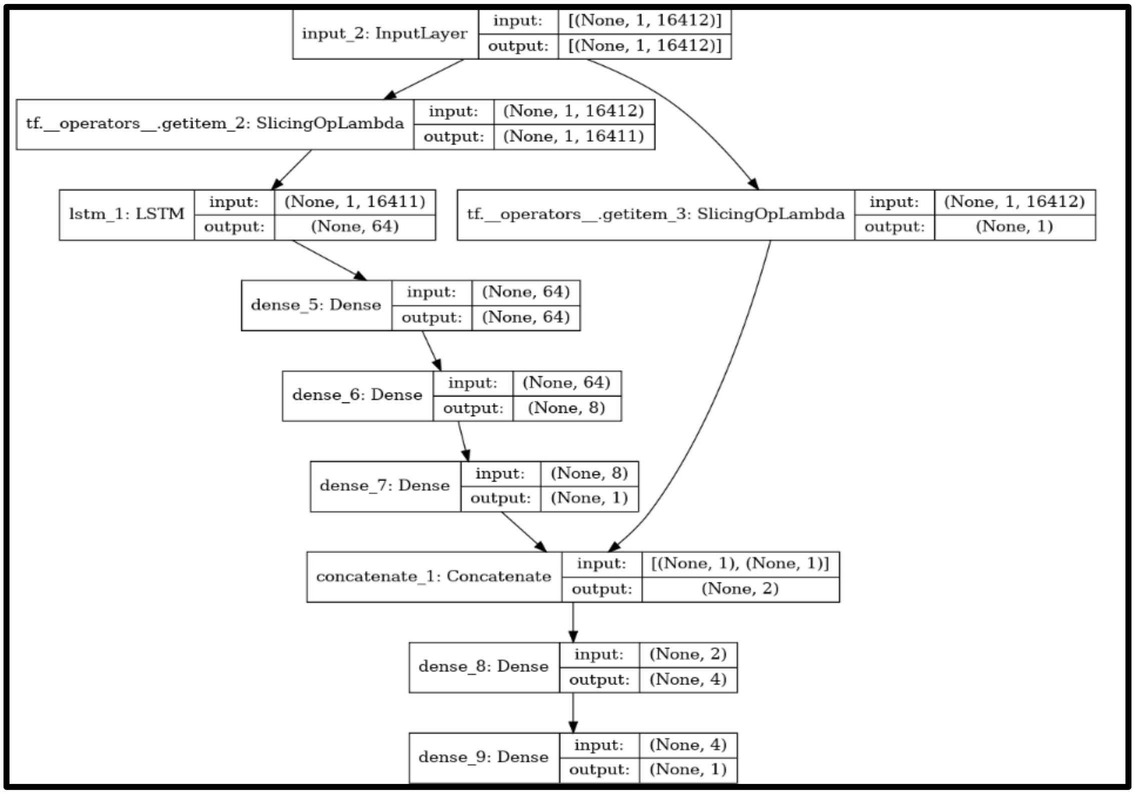
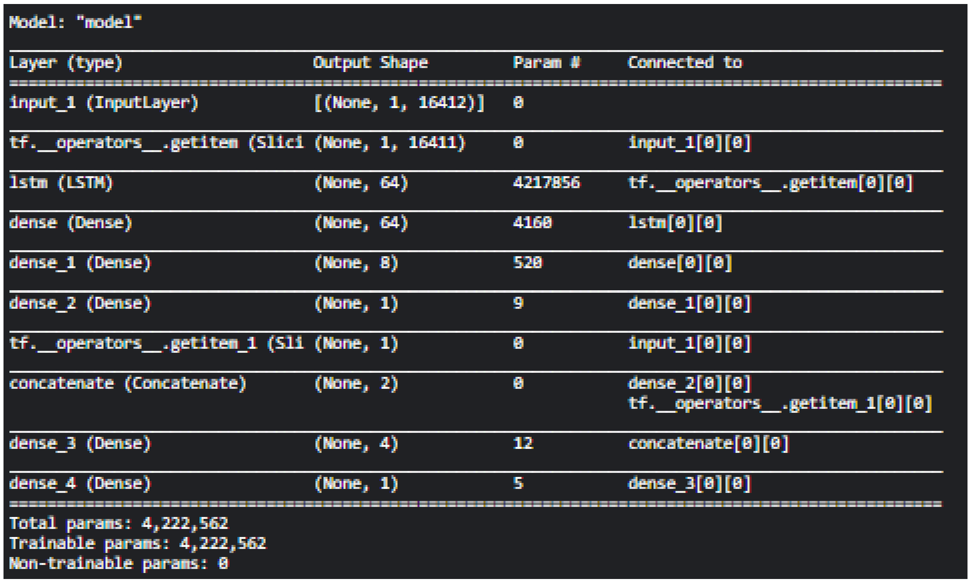


Figure 23: Model architecture

Figure 24 elaborates the model architecture with hyperparameter details at each layer. The system has 4,222,562 trainable parameters.



```
Model: "model"
__________________________________________________________________________________________________
Layer (type)                    Output Shape         Param #     Connected to
==================================================================================================
input_1 (InputLayer)            [(None, 1, 16412)]   0
__________________________________________________________________________________________________
tf.__operators__.getitem (Slici (None, 1, 16411)     0           input_1[0][0]
__________________________________________________________________________________________________
lstm (LSTM)                     (None, 64)           4217856     tf.__operators__.getitem[0][0]
__________________________________________________________________________________________________
dense (Dense)                   (None, 64)           4160        lstm[0][0]
__________________________________________________________________________________________________
dense_1 (Dense)                 (None, 8)            520         dense[0][0]
__________________________________________________________________________________________________
dense_2 (Dense)                 (None, 1)            9           dense_1[0][0]
__________________________________________________________________________________________________
tf.__operators__.getitem_1 (Sli (None, 1)            0           input_1[0][0]
__________________________________________________________________________________________________
concatenate (Concatenate)       (None, 2)            0           dense_2[0][0]
                                                                 tf.__operators__.getitem_1[0][0]
__________________________________________________________________________________________________
dense_3 (Dense)                 (None, 4)            12          concatenate[0][0]
__________________________________________________________________________________________________
dense_4 (Dense)                 (None, 1)            5           dense_3[0][0]
==================================================================================================
Total params: 4,222,562
Trainable params: 4,222,562
Non-trainable params: 0
```

Figure 24: Model architecture with hyperparameters details

CHAPTER 5: RESULT:

5.1 Introduction:

This section mainly focuses on the graphical representation of information and data by using visual elements like a bar graph, words cloud, area graph, and pie chart. After performing the preprocessing of data and implementing the model we have described in chapter 4 on the VAST challenge dataset, we have covered the contextual inquiry and the implementation of a dashboard to classify the disaster and non-disaster tweets. We have plotted an interactive graph and created a dashboard to visualize processed data.

5.2 Graph:

Bar graph:

In our research, we have developed bar graphs for two types of data, and we are analyzing the peak of the bar chart to track the occurrence of non-disaster tweets from a particular time frame. We used a 5-time interval for the bar graph, i.e., 5-min, 10 min, 15 min, 30 min, and 1 hour in our implementation. These time intervals mentioned above show the peak of the bar graph at

The peak of a smaller time interval bar graph can help analyze the particular time frame; for example, average tweets in 5 min timestamp are 60-100 tweets. However, the smaller timeframe can miss actual events covered in the larger timeframe of 1 hour and 30 minutes. On the other hand, we have a prominent timeframe peak of the bar graph which can cover up to 700-1000 tweets data and can be used to find essential information.

Examining the CSV file data over an extensive timeframe is not pragmatic, so we have used the word cloud for that finding.

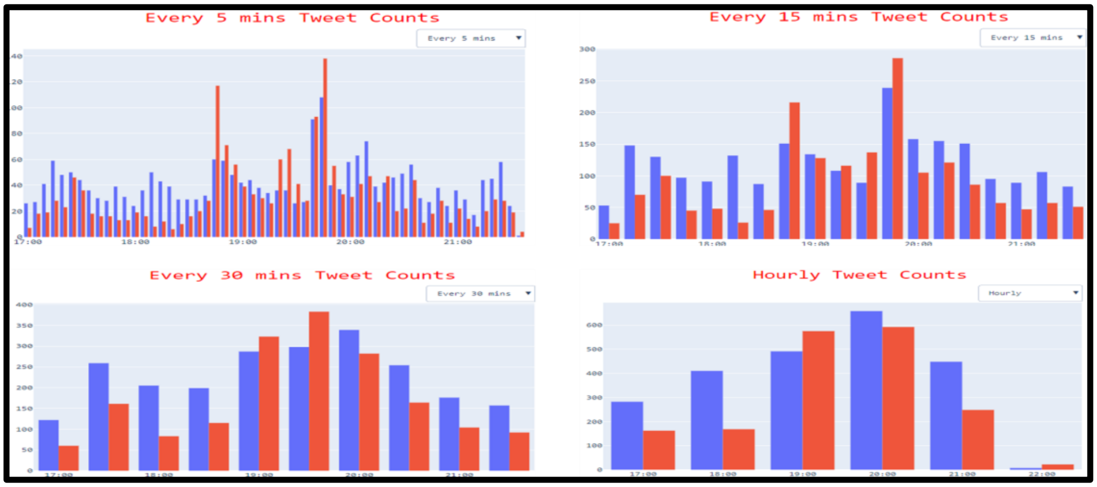


Figure 25: Bar Graph

Word cloud:

A Word cloud is a visual portrayal of words, displaying the most prominent words or frequent words in the body of text. Word cloud is beneficial in representing the keywords and shortlisting them. Nevertheless, to process with a word cloud, we must remove the stop word, weblinks, emoji, and emoticons as discussed in the Design and analysis of data chapter (Third Chapter).

Like the bar graph, we have developed a word cloud for five different timeframes, i.e., 5-min, 10-min, 15-min, 30-minutes, and hourly. Each word cloud represents the peak of its respective bar graph; a 5-min word cloud is created from a 5-min bar graph peak. For example, the peak in the 5-min bar graph was achieved at 19:42, and the word cloud is developed from 19:42 to 19:47 timeframe tweets data.

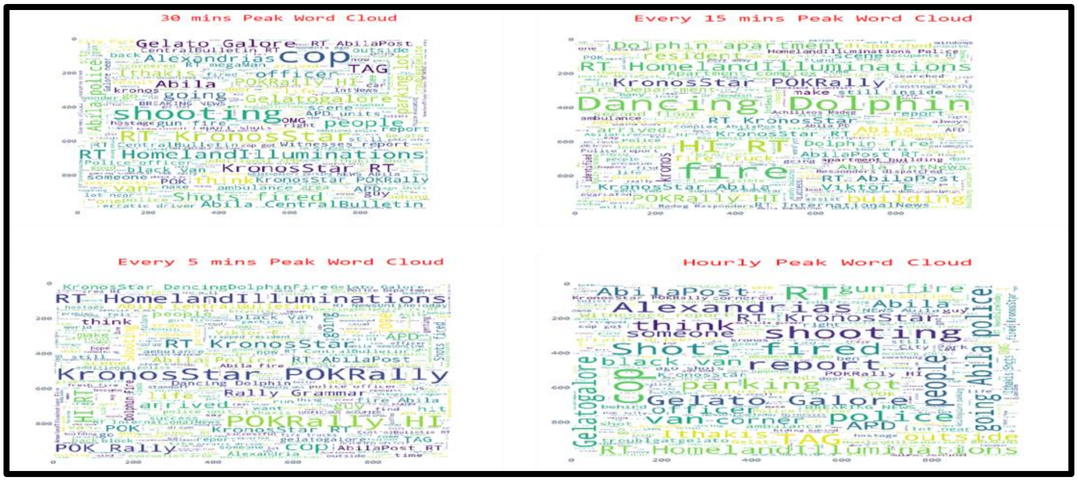


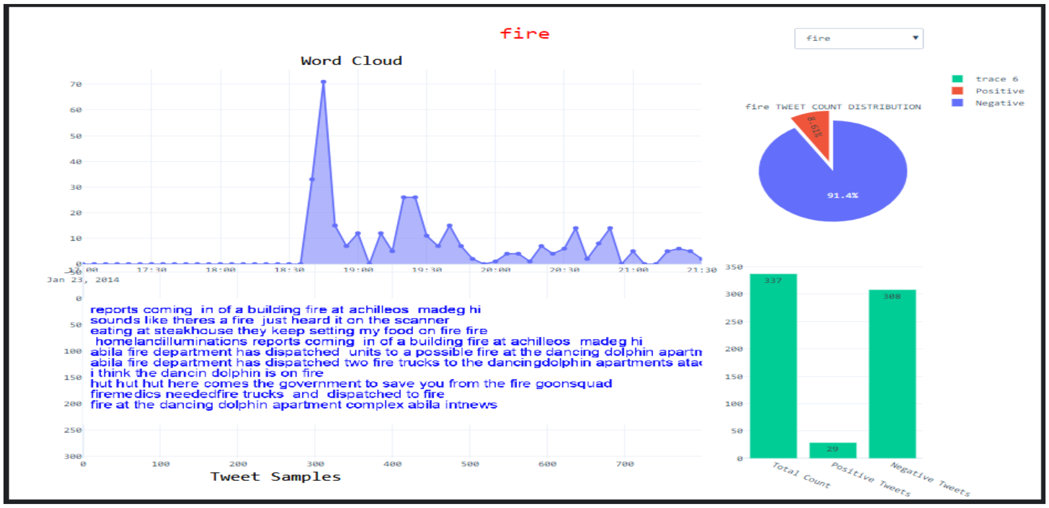
Figure 26: Word cloud

5.3 Dashboard:

Dashboard aims to enhance visibility, better decision making, real-time data analytics, and many more. It is a data management tool that uses data visualization technology to analyze and visualize the information practically.

The dashboard has a total of five components:

1. Dropdown List
2. Area Chart
3. Pie chart
4. Bar graph
5. Window to see the tweets

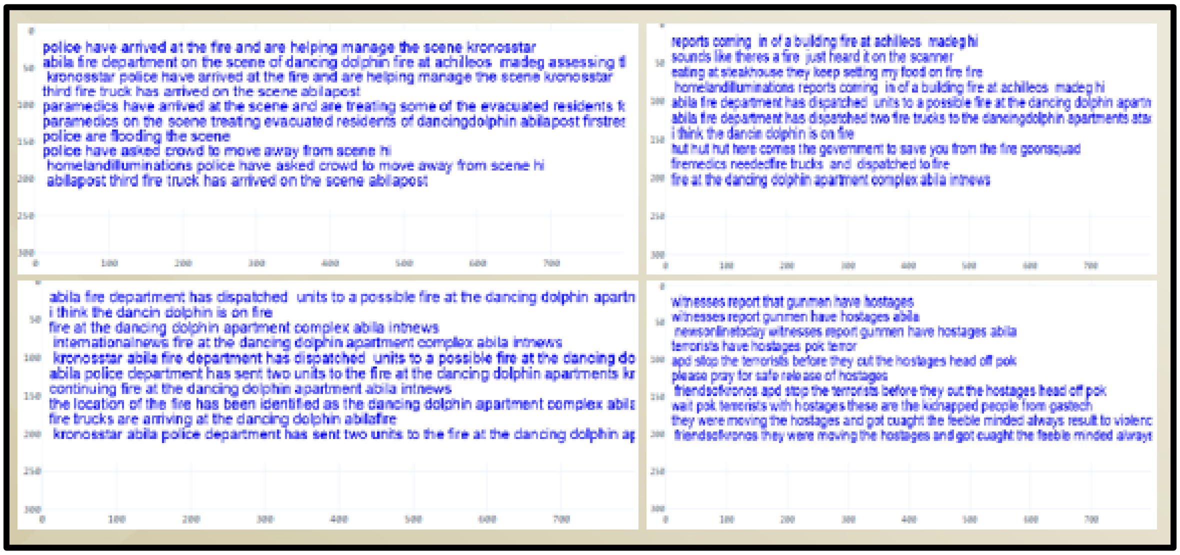


reports coming in of a building fire at achilleos madeg hi
sounds like theres a fire just heard it on the scanner
eating at steakhouse they keep setting my food on fire fire
 homelandilluminations reports coming in of a building fire at achilleos madeg hi
abila fire department has dispatched units to a possible fire at the dancing dolphin apartn
abila fire department has dispatched two fire trucks to the dancingdolphin apartments atac
i think the dancin dolphin is on fire
hut hut hut here comes the government to save you from the fire goonsquad
firemedics neededfire trucks and dispatched to fire
fire at the dancing dolphin apartment complex abila intnews

Figure 27: Dashboard

The dropdown list contains all-important keywords shortlisted from the word cloud, the area chart shows us the keyword usage throughout the time, the pie chart & the bar graph combinedly gives us the knowledge of keyword presence in negative or positive tweets, and the additional window will provide the top 10 tweets related to the keyword.

It is essential to find the relevant events in the IEEE Vast challenge, which will help resolve the case against POK in the disappearance of GasTech Employees. Some keywords we have procured from the challenge, i.e., kidnapping of employee -> hostages, scenes, police, etc., While some keywords are acquired from the word cloud images, i.e., black van, dolphin, fire, etc., Combining the list of keywords from a challenge and word cloud can give the result list. The result list is the dropdown component in the dashboard, which will contain the selected keywords. Selecting from the dropdown component will change the dashboard's other component. The area chart, bar graph, and pie chart will show the usage of a particular keyword in whole and positive/ negative tweets data. In contrast, the additional window will show the top ten related tweets to the keyword, which can help analyze the actual tweets data.



police have arrived at the fire and are helping manage the scene kronosstar
abila fire department on the scene of dancing dolphin fire at achilleos madeg assessing tl
kronosstar police have arrived at the fire and are helping manage the scene kronosstar
third fire truck has arrived on the scene abilapost
paramedics have arrived at the scene and are treating some of the evacuated residents fc
paramedics on the scene treating evacuated residents of dancingdolphin abilapost firstres
police are flooding the scene
police have asked crowd to move away from scene hi
homelandilluminations police have asked crowd to move away from scene hi
abilapost third fire truck has arrived on the scene abilapost

reports coming in of a building fire at achilleos madeg hi
sounds like theres a fire just heard it on the scanner
eating at steakhouse they keep setting my food on fire fire
homelandilluminations reports coming in of a building fire at achilleos madeg hi
abila fire department has dispatched units to a possible fire at the dancing dolphin apartn
abila fire department has dispatched two fire trucks to the dancingdolphin apartments atal
i think the dancin dolphin is on fire
hut hut hut here comes the government to save you from the fire goonsquad
firemedics neededfire trucks and dispatched to fire
fire at the dancing dolphin apartment complex abila intnews

abila fire department has dispatched units to a possible fire at the dancing dolphin apartn
i think the dancin dolphin is on fire
fire at the dancing dolphin apartment complex abila intnews
internationalnews fire at the dancing dolphin apartment complex abila intnews
kronosstar abila fire department has dispatched units to a possible fire at the dancing do
abila police department has sent two units to the fire at the dancing dolphin apartments kr
continuing fire at the dancing dolphin apartment abila intnews
the location of the fire has been identified as the dancing dolphin apartment complex abila
fire trucks are arriving at the dancing dolphin abilafire
kronosstar abila police department has sent two units to the fire at the dancing dolphin ap

witnesses report that gunmen have hostages
witnesses report gunmen have hostages abila
newsonlinetoday witnesses report gunmen have hostages abila
terrorists have hostages pok terror
apd stop the terrorists before they cut the hostages head off pok
please pray for safe release of hostages
friendsofkronos apd stop the terrorists before they cut the hostages head off pok
wait pok terrorists with hostages these are the kidnapped people from gastech
they were moving the hostages and got cuaght the feeble minded always result to violenc
friendsofkronos they were moving the hostages and got cuaght the feeble minded always

Figure 28: Window of tweets sample scene, fire, dolphin, and hostage

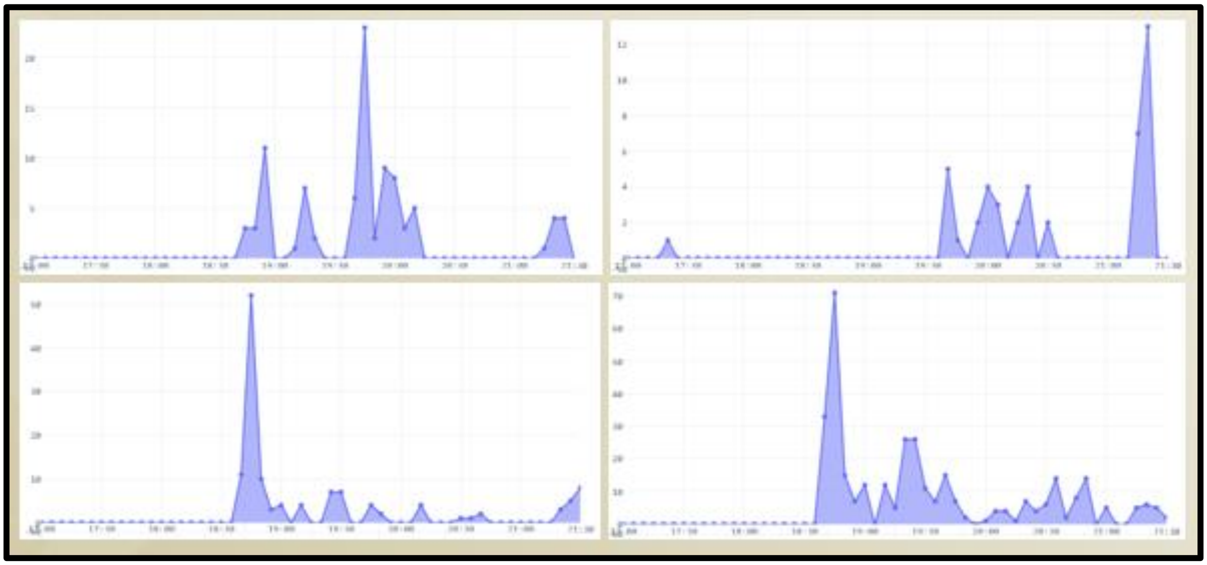
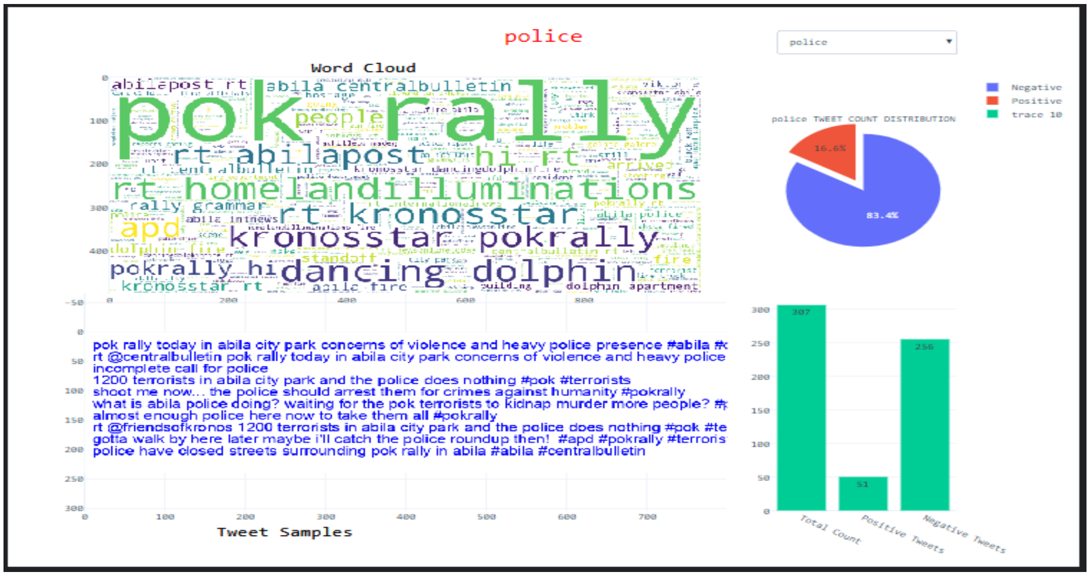


Figure 29: Area chart of Scene, fire, dolphin, and hostage

We have developed a second screen dashboard, as shown in figure 5.6, that includes the word cloud image, replacing the area chart with a word cloud. The word cloud incorporates all disaster tweets of data. The python framework dash provides a feature of magnifying into the dash component. The feature can help verify whether the keyword is present inside the word cloud.



pok rally today in abila city park concerns of violence and heavy police presence #abila #c
rt @centralbulletin pok rally today in abila city park concerns of violence and heavy police
incomplete call for police
1200 terrorists in abila city park and the police does nothing #pok #terrorists
shoot me now… the police should arrest them for crimes against humanity #pokrally
what is abila police doing? waiting for the pok terrorists to kidnap murder more people? #
almost enough police here now to take them all #pokrally
rt @friendsofkronos 1200 terrorists in abila city park and the police does nothing #pok #te
gotta walk by here later maybe i'll catch the police roundup then! #apd #pokrally #terroris
police have closed streets surrounding pok rally in abila #abila #centralbulletin

Figure 30: Dashboard with word cloud

5.4 Short Listed Events:

Creating a dashboard from an open-source framework library for building a data visualization interface. Getting relevant results from the dashboard is a crucial task. Monitoring the results from the dashboard can give a clue about the events that are taking place in city of Abila and the disappearances of GasTech employees. The shortlisted events are listed below:

1. POK Rally, Abila, Terrorist, Police, Kronosstar
2. Dancing Dolphin, Kronosstar, Dolphin Fire, Shots Fired, Shooting, Gelato Galore
3. Fire, Black Van, Abilapost, Dancing Dolphin, Gelato galore, Standoff, Terrorist

CHAPTER 6: CONCLUSION:

We have presented the infrastructure to analyze social media data in this project. We have used a deep learning model to scale the performance and increase the scalability and computer visualization to visualize the data. We have tried three deep learning models and, out of those three models, selected one LSTM model for implementation based on its computation power, efficiency, and accuracy. Created an interactive graph to find out essential keywords from the data; they were able to shortlist them and use them in the dashboard. We reduced the data size by selecting the disaster tweets data only from the data. Improve the performance of the model by using RNN.

References

[1] R. R. de Mendonça, D. F. de Brito, F. de Franco Rosa, J. C. dos Reis, and R. Bonacin, “A Framework for Detecting Intentions of Criminal Acts in Social Media: A Case Study on Twitter,” *Information 2020, Vol. 11, Page 154*, vol. 11, no. 3, p. 154, Mar. 2020, doi: 10.3390/INFO11030154.

[2] M. Magalhães *et al.*, “Predicting Frequent and Feared Crime Typologies: Individual and Social/Environmental Variables, and Incivilities,” *Social Sciences 2022, Vol. 11, Page 126*, vol. 11, no. 3, p. 126, Mar. 2022, doi: 10.3390/SOCSCI11030126.

[3] Y. Liu, C. Kliman-Silver, and A. Mislove, “The Tweets They Are a-Changin’: Evolution of Twitter Users and Behavior”, Accessed: Apr. 29, 2022. [Online]. Available: https://stream.twitter.com/1.1/statuses/sample.json,

[4] R. Prieto Curiel, S. Cresci, C. I. Muntean, and S. R. Bishop, “Crime and its fear in social media,” *Palgrave Communications 2020 6:1*, vol. 6, no. 1, pp. 1–12, Apr. 2020, doi: 10.1057/s41599-020-0430-7.

[5] S. Sitaram, K. R. Chandu, S. K. Rallabandi, and A. W. Black, “A Survey of Code-switched Speech and Language Processing,” Mar. 2019, doi: 10.48550/arxiv.1904.00784.

[6] B. das Sarit Chakraborty Student Member and I. Member, “An Improved Text Sentiment Classification Model Using TF-IDF and Next Word Negation,” Jun. 2018, doi: 10.48550/arxiv.1806.06407.

[7] A. M. Ashir, “A Generalized Method for Sentiment Analysis across Different Sources,” *Applied Computational Intelligence and Soft Computing*, vol. 2021, 2021, doi: 10.1155/2021/2529984.

[8] A. U. Hassan, J. Hussain, M. Hussain, M. Sadiq, and S. Lee, “Sentiment analysis of social networking sites (SNS) data using machine learning approach for the measurement of depression,” *International Conference on Information and Communication Technology Convergence: ICT Convergence Technologies Leading the Fourth Industrial Revolution, ICTC 2017*, vol. 2017-December, pp. 138–140, Dec. 2017, doi: 10.1109/ICTC.2017.8190959.

[9] N. Azzouza, K. Akli-Astouati, and R. Ibrahim, “TwitterBERT: Framework for Twitter Sentiment Analysis Based on Pre-trained Language Model Representations,” *Advances in Intelligent Systems and Computing*, vol. 1073, pp. 428–437, 2020, doi: 10.1007/978-3-030-33582-3_41.

[10] C. Dhaoui, C. M. Webster, and L. P. Tan, “Social media sentiment analysis: lexicon versus machine learning,” *Journal of Consumer Marketing*, vol. 34, no. 6, pp. 480–488, 2017, doi: 10.1108/JCM-03-2017-2141/FULL/XML.

[11] K. Banujan, B. T. G. S. Kumara, and I. Paik, “Twitter and Online News analytics for Enhancing Post-Natural Disaster Management Activities,” *2018 9th International Conference on Awareness Science and Technology, iCAST 2018*, pp. 302–307, Oct. 2018, doi: 10.1109/ICAWST.2018.8517195.

[12] B. Kuhaneswaran, B. T. G. S. Kumara, and I. Paik, “Strengthening Post-Disaster Management Activities by Rating Social Media Corpus,” *https://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/IJSSOE.2020010103*, vol. 10, no. 1, pp. 34–50, Jan. 1AD, doi: 10.4018/IJSSOE.2020010103.

[13] N. Said *et al.*, “Natural disasters detection in social media and satellite imagery: a survey,” *Multimedia Tools and Applications 2019 78:22*, vol. 78, no. 22, pp. 31267–31302, Jul. 2019, doi: 10.1007/S11042-019-07942-1.

[14] N. Anstead and B. O’Loughlin, “Social Media Analysis and Public Opinion: The 2010 Uk General Election,” *Journal of Computer-Mediated Communication*, vol. 20, no. 2, pp. 204–220, Mar. 2015, doi: 10.1111/JCC4.12102.

[15] P. N. Howard, A. Duffy, D. Freelon, M. M. Hussain, W. Mari, and M. Mazaid, “Opening Closed Regimes: What Was the Role of Social Media During the Arab Spring?,” *SSRN Electronic Journal*, Apr. 2011, doi: 10.2139/SSRN.2595096.

[16] C. Research Paper Bendler, “Crime Mapping through Geo-Spatial Social Media Activity,” 2014.

[17] S. P. C. W. Sandagiri, B. T. G. S. Kumara, and B. Kuhaneswaran, “ANN Based Crime Detection and Prediction using Twitter Posts and Weather Data,” *2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy, ICDABI 2020*, Oct. 2020, doi: 10.1109/ICDABI51230.2020.9325660.

[18] K. Ali, H. Dong, A. Bouguettaya, A. Erradi, and R. Hadjidj, “Sentiment Analysis as a Service: A Social Media Based Sentiment Analysis Framework,” *Proceedings - 2017 IEEE 24th International Conference on Web Services, ICWS 2017*, pp. 660–667, Sep. 2017, doi: 10.1109/ICWS.2017.79.

[19] J. Hao and H. Dai, “Social media content and sentiment analysis on consumer security breaches,” *Journal of Financial Crime*, vol. 23, no. 4, pp. 855–869, 2016, doi: 10.1108/JFC-01-2016-0001/FULL/XML.

[20] H. Isah, P. Trundle, and D. Neagu, “Social media analysis for product safety using text mining and sentiment analysis,” *2014 14th UK Workshop on Computational Intelligence, UKCI 2014 - Proceedings*, Oct. 2014, doi: 10.1109/UKCI.2014.6930158.

[21] S. W. Phoong, “Social Media Sentiment Analysis on Employment in Malaysia Investigate the Effect of Structural Change of the Oil Price on Consumer Price Index View project The Effectiveness of Social Media on Lifelong Learning View project,” 2017, Accessed: Apr. 29, 2022. [Online]. Available: https://www.researchgate.net/publication/322222593

[22] M. Wang and M. S. Gerber, “Using twitter for next-place prediction, with an application to crime prediction,” *Proceedings - 2015 IEEE Symposium Series on Computational Intelligence, SSCI 2015*, pp. 941–948, 2015, doi: 10.1109/SSCI.2015.138.

[23] C. Shobana, B. Rangasamy, R. K. Poopal, S. Renuka, and M. Ramesh, “Green synthesis of silver nanoparticles using Piper nigrum: tissue-specific bioaccumulation, histopathology, and oxidative stress responses in Indian major carp Labeo rohita,” *Environmental Science and Pollution Research 2018 25:12*, vol. 25, no. 12, pp. 11812–11832, Feb. 2018, doi: 10.1007/S11356-018-1454-Z.

[24] M. C. Landoni, A. Rukavina, M. L. Traverso, and P. Verasay, “VAST challenge 2014: The kronos incident-mini-challenge 3,” *Proceedings - 2015 International Workshop on Data Mining with Industrial Applications, DMIA 2015: Part of the ETyC 2015*, pp. 79–81, Aug. 2016, doi: 10.1109/DMIA.2015.16.

[25] M. Mao, “VAST 2014: Summary on Mini Challenge 3 work,” *2014 IEEE Conference on Visual Analytics Science and Technology, VAST 2014 - Proceedings*, pp. 377–378, Feb. 2015, doi: 10.1109/VAST.2014.7042574.

[26] F. Fischer and F. Stoffel, “NStreamAware: Real-Time visual analytics for data streams (VAST Challenge 2014 MC3),” *2014 IEEE Conference on Visual Analytics Science and Technology, VAST 2014 - Proceedings*, pp. 373–374, Feb. 2015, doi: 10.1109/VAST.2014.7042572.

[27] D. Thorn, M. Worner, and S. Koch, “ScatterScopes: Understanding events in real-time through spatiotemporal indication and hierarchical drilldown: VAST 2014 Mini Challenge 3 Recognition: ‘Honorable mention for good support for situation awareness,’” *2014 IEEE Conference on Visual Analytics Science and Technology, VAST 2014 - Proceedings*, pp. 387–388, Feb. 2015, doi: 10.1109/VAST.2014.7042579.

[28] F. Fischer and F. Stoffel, “NStreamAware: Real-Time visual analytics for data streams (VAST Challenge 2014 MC3),” *2014 IEEE Conference on Visual Analytics Science and Technology, VAST 2014 - Proceedings*, pp. 373–374, Feb. 2015, doi: 10.1109/VAST.2014.7042572.

[29] H. Bosch *et al.*, “ScatterBlogs2: Real-time monitoring of microblog messages through user-guided filtering,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2022–2031, 2013, doi: 10.1109/TVCG.2013.186.

[30] R. Patel and K. Passi, “Sentiment Analysis on Twitter Data of World Cup Soccer Tournament Using Machine Learning,” *IoT 2020, Vol. 1, Pages 218-239*, vol. 1, no. 2, pp. 218–239, Oct. 2020, doi: 10.3390/IOT1020014.

[31] P. Tyagi and R. C. Tripathi, “A Review towards the Sentiment Analysis Techniques for the Analysis of Twitter Data,” *SSRN Electronic Journal*, Feb. 2019, doi: 10.2139/SSRN.3349569.

[32] L. Zhang, S. Wang, and B. Liu, “Deep learning for sentiment analysis: A survey,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, p. e1253, Jul. 2018, doi: 10.1002/WIDM.1253.

[33] “VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text | Proceedings of the International AAAI Conference on Web and Social Media.” https://ojs.aaai.org/index.php/ICWSM/article/view/14550 (accessed Apr. 29, 2022).

[34] “Tokenization in NLP | Kaggle.” https://www.kaggle.com/code/satishgunjal/tokenization-in-nlp/notebook (accessed Apr. 29, 2022).

[35] D. Deepa, Raaji, and A. Tamilarasi, “Sentiment Analysis using Feature Extraction and Dictionary-Based Approaches,” *Proceedings of the 3rd International Conference on I-SMAC IoT in Social, Mobile, Analytics and Cloud, I-SMAC 2019*, pp. 786–790, Dec. 2019, doi: 10.1109/I-SMAC47947.2019.9032456.