# MASTER THESIS

**In Order to Obtain the**

**PROFESSIONAL MASTER**

**in**

**Geographic Information Systems**

**(GIS)**

**Presented and defended by:**

**Ali Mohammad Harakeh**

**On Wednesday October 28, 2020**

**Title**

**Improving Governments' Public Decisions During Covid-19 Using Geospatial Intelligence on Social Media Data**

**Supervisors**

**Ms. Gretta Antoine Kelzi**

**Mr. Ahmad Mohammad Khater**

**Reviewers**

**Dr. May Abbas Dhayni**

**Ms. Manal Zuheir Sayed**

# TABLE OF CONTENTS

# ACKNOWLEDGEMENTS

I would first like to thank Dr. Kifah Tout of the Science & IT Department at Lebanese University's Science Faculty for providing me the opportunity to experience the GIS master and learn a lot of things. He also helped provide the license needed to use the ArcGIS platform that was needed in this research.

I would also like to thank my thesis advisor Ahmad Khater who works at Khatib & Alami Egypt's Branch for all his help and supervision during this research period as he was always ready to provide any advice or help that he could provide.

Lastly, I thank Khatib & Alami for providing the research's general topic, Esri Lebanon for helping with the ArcGIS license, and everyone who encouraged me during the research period.

# ABSTRACT

Covid-19, also known as the coronavirus, is an ongoing pandemic that spreads most often when people are physically close, and it has caused a global social and economic disruption that requires the efforts of both governments and individuals to overcome it. But as governments applied their planned measures, they lacked the understanding of how individuals were moving and interacting with these measures. For that purpose, the goal of this research was to develop a geographically visual and context-aware mechanism that used location-based analysis of social media data, such as Twitter's tweets, to improve governments' public decision making by taking into consideration the social reactions and interactions in this pandemic during the research's period. This mechanism depended on geographically tracking individuals' sentiments and emotions to help the government improve their interactions with them while also considering their discussions as a way to understand their current thoughts and opinions. It also helped extract the most crowded geographic locations at some periods where there were interactions between individuals that might have been a cause for the suddenly increased cases of Covid-19 cases. Working on and analyzing a sample data of 50,000+ Twitter tweets, this research could find the approximate source location of tweets, the geographic distribution of people's sentiment and emotions spanning throughout this research period, and the geographic gathering spots which contributed to the disease spreading at some periods. It also provided a visual output of the analysis in a geographical dashboard using the ArcGIS platform that could help the decision-maker better understand the geographic perspective of the situation during the pandemic. For future work, we can work on extracting more data from other social media sources and enhance the extraction methods to better understand the text context and provide more concrete insights.

**Keywords**: Covid-19, Corona, Social Media, Twitter, Tweets, ArcGIS, Sentiment Analysis, Emotion Analysis, Topics Extraction, Public Activities.

# CHAPTER 0: INTRODUCTION

COVID-19, first being identified in December 2019 in Wuhan, China, was declared a Public Health Emergency of International Concern in January 2020 and a pandemic in March 2020. It is an ongoing pandemic caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), that spreads most often via physical contact of contaminated surfaces or through the air when people are at close range of an infected person as he breathes, coughs, sneezes, talks, or sings. As of October 9, 2020, there have been 36.5 million cases confirmed globally as well as more than 1.06 million deaths attributed to this pandemic. Both the pandemic and response measures have contributed to social and economic disruption, including the largest global recession since the Great Depression [1]. According to recent data, around 100 million people are expected to fall into extreme poverty and global famines for 130 million people. It has also led to the postponement or cancellation of events, widespread supply shortages, partial or fully closing of Educational Institutes, with various alternatives used, and many incidents of xenophobia and racism against Chinese people and against those perceived as being Chinese or as being from areas with high infection rates.

As governments try to ease these pandemic effects on society and apply their planned measures, they need a way to track how these changes affected society and social stability as this pandemic requires the combined efforts of both government and society to prevent further spread of this virus. People usually care about decisions that affect their current situation more than logical or scientific measures that they can't see the result of directly. For that purpose, governments should have more insights into people's movements, thoughts, and opinions that they can reflect upon when making decisions.

Nowadays, as more people enter the online world, social media platforms became the new virtual space where people can interact with each other without the obstacle of distance. Social media provide straightforward, cost-free, and familiar ways to communicate with all people categories all over the country, and by nature reflects individuals' opinions and statuses, at the moment of posting, that refers to their current feelings and sentiment as well as what topics they are discussing as a whole society. They are now a powerful communication mechanism where billions of data are generated through them every day. And this made them a great source of data that directly relates to the public and provide an information bank that can be used by scientists and developers to research this data and make use of it. This is especially true in our case as they can provide valuable geographic insights that help governments track all of these social activities and separate them by region for a better understanding of where and when each activity happened.

For that reason, analyzing this content by location and visualizing it geographically on a map, while still respecting people's privacy and not sharing their identities, can improve governments' public decisions as it helps maintain public sentiment, deliver more accurate public measures, and understand people's thoughts and opinions. It can also extract the locations and places where most people gathered during the pandemic as having a geographic understanding of how people moved is very important in controlling and monitoring the spreading of the disease.

# CHAPTER I: PROBLEM DEFINITION

By nature, social media content is unpredictable and relates to individuals who each have their unique way of expressing themselves and sharing their thoughts and opinions. These posts can be in different forms and formats with distinct languages or language accents unique to each country or country part. That's why it's difficult to consider every aspect of this content in our analysis as there is no specific standard that we can base on. As people mostly tweeted using three languages: Lebanese, Arabic, and English, we found the Lebanese language to be especially tricky as it was very volatile that neither its form nor the characters used when writing the same word remained the same throughout all the data. Furthermore, using both Arabic and English characters, it created two ways of writing: Lebanese-Arabic and Lebanese-English (also known as the Internet Language).

*Table 1 - Tweets Languages*

| Language | Phrase |
|---|---|
| English | How are you? |
| Arabic | كيف حالك؟ |
| Lebanese (English) | Kifak? or Kefak? or Keefak? |
| Lebanese (Arabic) | كيفك؟ |

Another problem we faced was how to get the geographic location of these posts' owners as we wanted to differentiate between different country areas in our analysis. As this research uses Twitter tweets, there were two public ways for us to get users' locations. Twitter's "User Location Access" option, which was disabled by default, allowed Twitter to know the user's location when he posts his tweets while also permitting other people to search for these geo-tweets via the geo-search function. Another way was the "User-Provided Location" in the User's Twitter page that users customized as they liked. Twitter then linked this given location to an actual real-world location and included these users' tweets in the geo-search results. However, the user location access option was OFF by default and not known to many users, while most others ignored it. As for the user-provided location, many users either didn't fill their profile location correctly or left it empty. These limitations lead to fewer geo-tweets resulting from our geo-search, which required us to find another way to get users' locations or at least approximate that location.

# CHAPTER II: BACKGROUND & RELATED WORK

## 1. Geographic Information Systems (GIS)

A geographic information system (GIS) is a conceptualized framework that enables spatial and geographic data to be collected and analyzed. GIS applications are computer-based tools that allow the user, by presenting them as maps, to create interactive queries (user-created searches), store and edit spatial and non-spatial data, evaluate the performance of spatial information and visually share the results of these activities. Multiple tools, procedures, strategies and approaches make use of geographic information systems. It is related to various activities and multiple applications relating to: engineering, planning, management, transport / logistics, insurance, telecommunications, and business. For this purpose, GIS and location intelligence applications are the basis of geographical analysis and visualization-based location-enabled services. Geospatial intelligence, which is part of GIS, is the intelligence of human activity on Earth derived from the exploitation and study of imagery and geospatial knowledge that identifies, analyses, and visually portrays physical characteristics and geographically referenced activities on Earth. It helps understand real-world data and incorporate the everyday activities and operations of people into valuable knowledge that reflects geographical activity. With the help of tracking tools such as GPS devices, environmental sensors, and many others, we can specify where things happen and provide data with contextual information.

## 2. ArcGIS

ArcGIS [2] is an implementation of a GIS create by ESRI [3]. It offers unique capabilities and flexible licensing for applying location-based analytics to businesses practices, provide greater insights using contextual tools to visualize and analyze data, and allows collaboration and sharing via maps, apps, dashboards and reports. Its features include spatial analytics and data science, field operations, mapping, real-time visualizations and analytics, 3D GIS, Imagery and remote sensing, and data collection and management.

## 3. Text Mining

The advancement of computational technology over time has enabled the dramatic development of text mining methods and tools. Text mining is a computational process to understand the meaning and context of text documents as it helps categorize the text and determining the category according to the content. Over the past decades, many text mining methods incorporated machine learning algorithms and deep learning to achieve their goal.

## 4. Machine Learning and Deep Learning

Machine learning is an artificial intelligence (AI) technology that gives systems the ability to learn and develop from experience automatically without being programmed specifically. Machine learning focuses on the development of computer programs that can access and use data to learn on their own. Deep learning is a function of artificial intelligence (AI) that mimics

the functioning of the human brain in data processing and the development of patterns for decision-making use. Deep learning is a subset of artificial intelligence machine learning that has networks capable of learning unsupervised from unstructured or unlabeled knowledge.

# 5. Fuzzy Searching

Fuzzy Searching is the process of locating terms that are likely to be relevant to a searched argument even when this argument doesn't exactly match these terms. This process is done through employing fuzzy matching algorithms and return a list of the most relevant terms it could find. For example, if a user types "Misissippi" into Yahoo or Google (both of which use fuzzy matching), a list of hits is returned along with the question, "Did you mean Mississippi?" where alternative spellings and words that sound the same but are spelled differently may are given. A fuzzy matching program can operate like a spell checker, compensate for common input typing errors, and correct errors introduced by optical character recognition (OCR) scanning of printed documents. The program can also return hits with content that contains a specified base word along with prefixes and suffixes. For example, if "planet" is entered as a search word, hits occur for words containing words such as "protoplanet" or "planetary". Fuzzy Searching algorithms are categorized into many categories where Distance and Phonetic Algorithms are two of the main ones. Distance Algorithms aims to check the similarity of two words by how many edits are required to change word b into word a. For example, Levenshtein Distance (Edit Distance) [4] is one of the most popular Distance algorithms. As for phonetic algorithms, they aim to find how words are similar by their pronunciation sound, where Soundex [5] is one of the popular phonetic algorithms for English words.

# 6. Related Work

The increasing popularity of social media services and smartphones has enabled the public to share their daily activities online and to leave their digital footprint in public areas. Collecting social media messages and their coordinates within public areas could help researchers understand dynamic spatial-oriented human activities on Earth. For example, Tsou et al. (2013) [6] demonstrated a research framework for tracking and analyzing the spatial content of social media that can facilitate the tracking of social events (2012 U.S. presidential election) from a spatial-temporal perspective. Liu et al. (2014) [7] used location-based social media data to analyze the underlying patterns of trips and spatial interactions in cities and revisited spatial interaction and distance decay in spatially-embedded networks. Several scholars used social media, as crowdsourced spatiotemporal data content, to understand emergency events, enhance emergency situation awareness, and improve the efficiency of emergency response [8–11]. These researchers worked with different social media platforms and depended mainly geo-tagged social media posts which, as mentioned in the problem definition section, are scarce in Twitter. Thus, one of the main points that this research focuses on is how to extract locations from users' data and use them in our analysis and visualization.

# CHAPITRE III: CONTRIBUTION

The purpose of this research was to find a way to analyze and visualize social media content through a mechanism that could provide useful insights about public reactions and interactions during the Covid-19 pandemic period of this research. This mechanism aimed to help decision-makers to have an overview of the social situation so that they could react to and manage any public disruption that might have been the result of some planned measures they applied or due to some public disturbance. Furthermore, we wanted it to be the link that could fill a little bit of the connection gap between the government and the people.

We achieved this mechanism using text mining techniques such as Sentiment Analysis, Emotion Analysis, Keywords Extraction, and Rule-Based Extraction. We also extracted the most crowded locations that might have been a cause for the sudden increase in Covid-19 cases. We then separated all analysis results by their geo-location and grouped them by their Kadaa and Mohafaza that was visualized using the ArcGIS platform.

## 1. Getting Data

The data source used for this research was the Twitter Social Media Platform as it provided the most direct public and social content without the need to follow or join communities like in other platforms. Search keywords related to Covid-19 and its pandemic were selected to get the sample tweets for this research.

**Search Keywords**: Covid-19, Corona, Healthcare, Medical, Pandemic, Virus, كورونا, and كورونا_لبنان

## 2. Location Data

Due to the problems discussed in the problem definition section above, we needed a new way to get users' locations when tweeting for our analysis and visualization. Our proposed solution was to use each user's tweets as a data bank that reflected the user's interests and concerns so that we can then extract any location reference in these tweets and consider the most frequently mentioned location as a place of great importance to this user, and so, it might represent his address. This method requires two things: a list of known locations details and a way to extract these location references.

### 2.1. Locations Names and Details

We managed to get a list of 3000+ locations in Lebanon with their different names (English and Arabic), latitude, longitude, mohafaza, and kadaa details.

These location details were scraped from the Lebanon section of the Global Gazetteer [12]

### 2.2. Extracting Locations

We discovered after many tests that we couldn't combine the Arabic location references extraction method with other languages like English and Lebanese-English method as the characters of each were too apart from each other. Thus, we separated both Arabic and

English by checking the characters used and applied the appropriate fuzzy search methods for each case.

### *2.2.1.Arabic and Lebanese-Arabic Location Names*

We discovered that:

- Both languages had the same location names.
- A prefix was often added before the location name.
- Some Lebanese-Arabic location names often had slight characters differences.

To handle these problems, we used two fuzzy search methods that took into consideration the many aspects we discovered before. The first method helped solve the prefix problem as we checked for any partial sub-string inside the word provided to check for any location reference. As for the other characters' mismatch situation, we used something called the Levenshtein Distance Formula, also called the Edit Distance, to check how different the provided word was from our location names and find any similar location name to this word.

*Table 2 - Fuzzy Search Methods*

| Word | Problem | Location Result | Method |
|---|---|---|---|
| ببيروت | Prefix "ب" | بيروت | Partial sub-string search |
| تل اخضر | Missing "ال" in "اخضر" | تل الاخضر | Levenshtein Distance Formula (Edit Distance) |

### *2.2.2.English and Lebanese-English Location Names*

We discovered that:

- Both languages used the same letters but differed greatly in usage.
- Both languages often had mismatched characters for the same location name.
- There were many ways to write the Lebanese-English location names in.
- Both languages location names often had the same sound.

Due to these discoveries, we were able to use the Soundex phonetic fuzzy search method to compare the sounds of the words while using the same Levenshtein Distance Formula as before to help us handle characters mismatching and finding similar words.

*Table 3 - Words with similar sounds*

| Word | 4-digit Soundex Codex |
|---|---|
| Beirut | B630 |
| Beyrut | B630 |
| Beyrouth | B630 |
| بيروت | ب000 (Doesn't work for Arabic) |

*Table 4 - Levenshtein Distance (Edit Distance)*

| Word 1 | Word 2 | Edit Distance |
|---|---|---|
| Beirut | Beyrut | 1 edit(s) required |
| Beirut | Beyrouth | 3 edit(s) required |
| Beyrut | Beyrouth | 2 edit(s) required |

# 3. Handling Different Languages

As mentioned before, social media content was full of different types of text forms and formats. This research approached this problem from its outer perspective as it proposed a solution to unify all of these forms and formats while focusing on the Lebanese language that was specific to Lebanon's Social Media Platforms.

## 3.1. Unifying Languages

This research proposed to unify these languages into one language that was easy to handle and work on. For that purpose, we chose the English language as our output language as it was the most wildly used language in the world and one of the easiest and straightforward languages to work on. As the English language was very popular, many tools converted other languages to it, and one of these tools was Google's "Google Translate" which we used to unify all other languages found in our tweets. However, even though it could translate almost everything, it was still lacking on the Lebanese language side as it couldn't translate all of the Lebanese text in our tweets.

*Table 5 - Wrong Google Translations*

| Lebanese | Google Translate |
|---|---|
| Kn mnsab bas sa7 halla2 | Be the position, but correct Hala |
| Eh walla nsab bl marad | Uh, not lineage, but murad |
| Wasfi men edoctoor | My description is from Adster |

## 3.2. Translating Lebanese Language

As Google Translate didn't work well on all the Lebanese text we had, we decided that we needed to create a tool to help translate the rest of these texts that couldn't be translated.

### *3.2.1.Arabic Lebanese*

Although Arabic Lebanese is different than standard Arabic, it still had some similarities to it and Google Translate did a good job translating it to English.

*Table 6 - Lebanese-Arabic Translation*

| Lebanese | Translation |
|---|---|
| مرحبا | Hello |
| كيفك | How are you |
| تمام | Ok |

### *3.2.2.English Lebanese (The Lebanese Internet Language)*

As English Lebanese used English letters & numbers without having a sentence structure or fixed vocabulary, it was hard to translate its text with Google Translate. For that purpose, an external tool was needed to help handle these texts as much as we could. This tool depended on two methods:

#### 3.2.2.1. Lebanese-To-English Dictionary Of ~2500 Words

The first thing we did was gathering the most used Lebanese words and translate them manually to create a dictionary of Lebanese-To-English translation mapping. Some of these words were gathered from Google's Lebanese-To-English dictionary [13] containing around 1500 Lebanese words translated into English. The other words were gathered and filtered from a data bank of 10,000 random tweets after removing any words related to any other language and manually translating the rest of the Lebanese words. This Dictionary also included many variations of the same word where we tried our best to keep the definition as global and general as possible to match as many use-cases as we could. We should also note that some words' meanings could differ according to the context.

*Table 7 - Lebanese-To-English Dictionary Sample*

| Lebanese | English |
|---|---|
| 2ahwe, 2hwe, 2ahwi | Coffee |
| Adiim, Adiime, Adeem, Adeeme | Old |
| Dahab, Dahabi, Dahabiiyi | Gold |
| 2sm, Esm | Name |
| Eta2es | The weather |
| Jeser | Bridge |
| 7arara | Temperature |

As our dictionary didn't cover everything and still lacked a lot of words, we tried to check for similar words by using Levenshtein Distance Formula (Edit Distance) Fuzzy Searching method as some words differed by 1-2 characters and still retained the same meaning.

*Table 8 - Similarity between words*

| LB | EN | Word | Similarity |
|---|---|---|---|
| Hone | Here | Hon | 86% |
| 2hwi | Coffee | 2ahwi | 89% |
| Kifak | How are you | Kefak | 80% |

#### 3.2.2.2. English Lebanese → Arabic Lebanese → English

Another way was to convert these Lebanese-English words into Lebanese-Arabic by mapping English characters to their respective Arabic version so that we could

approximate the Arabic word. After that, we passed this Arabic version of the word to Google Translate as it had a function that could help us correct the word or find the pure Arabic version which was then translated into English.

This Mapping depended on Single and Double Mapping tables that were created after many tests and observations.

*Table 9 - Single Characters Mapping*

| EN | AR |
|---|---|
| A | ا |
| B | ب |
| D | د |
| E | ي |
| F | ف |
| G | ج |
| H | ه |
| I | ي |
| J | ج |
| K | ك |
| L | ل |
| M | م |
| N | ن |
| O | و |
| Q | ق |
| R | ر |
| S | س |
| T | ت |
| W | و |
| Y | ي |
| Z | ز |
| 2 | ا |
| 3 | ع |
| 5 | خ |
| 7 | ح |
| 8 | غ |

*Table 10 - Double Characters Mapping*

| EN | AR |
|---|---|
| aa | ع |
| th | ث |
| sh | ش |
| sa | ص |
| da | ض |
| ta | ط |
| fa | ف |
| 2e | ء |
| eh | اي |
| en | ين |
| ll | لا |

*Table 11 - Lebanese-English Translation Flow*

| LB | LB-AR | AR | EN |
|---|---|---|---|
| mr7aba | مرحابا | مرحبا | Hello |
| 2hla | اهلا | اهلا | Welcome |

## 4. Data Study & Analysis

Our main aim was to study the social reactions and interactions between people during the Covid-19 pandemic period of this research, which needed to be preceded by location data extraction and unifying of the content to be in the English language for easier handling. We chose to split our work into main categories and subcategories. (Figure 1)

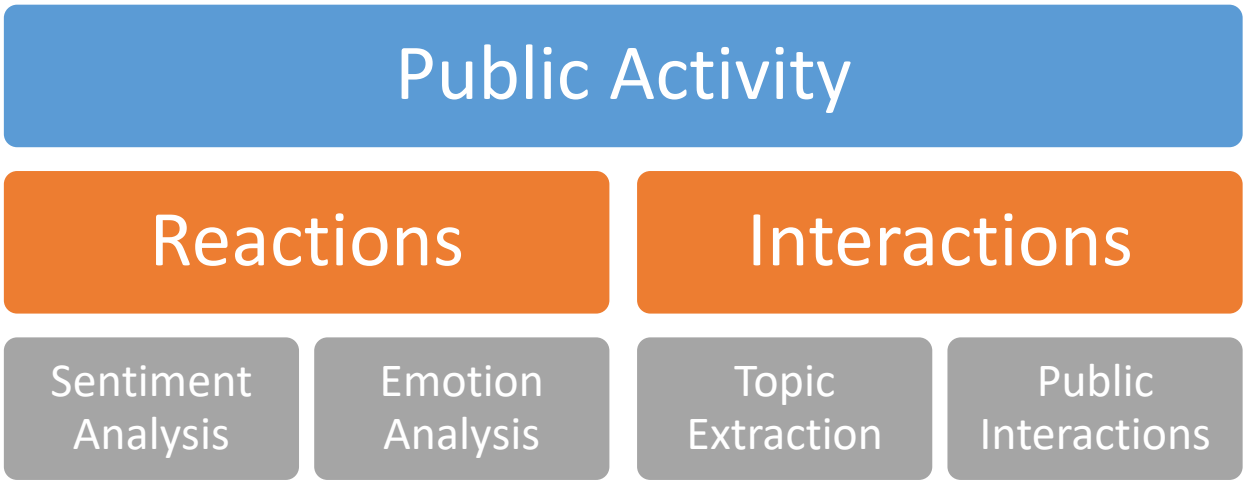


*Figure 1 - Analysis Workflow*

## 4.1. Sentiment Analysis

Our starting point was the study of public reactions through sentiment analysis to discover how people were viewing and handling the situation during the pandemic. The main goal was to observe whether people were accepting or complaining about their situation as taking any new measures for this pandemic without knowing how it would affect the public negatively or positively should never happen.

Sentiment analysis was an already known Natural Language Processing (NLP) technique that extracted text sentiment and categorized it. It was a tool that didn't need much work to be useable. However, it was essential to note that the sentiment analysis model differed based on what it had been trained on as there were many variations of trained models. For this research, we used the TextBlob [14] python tool as it provided all of the three sentiment categories (positive, negative, and neutral) that we based our study on during this research period.

## 4.2. Emotion Analysis

Another way to study public reactions was to extract the emotions that people expressed in their tweets as it allowed us to observe how people felt during this pandemic. As the Covid-19 pandemic needed the efforts of both governments and people to overcome it, it was essential to take the appropriate measures to keep people as hopeful and positive as much as possible since the overflow of negative emotions would cause public disruption and social upset.

The study consisted of examining the tweets to probe for any reference to the eight basic emotions (anger, disgust, fear, sadness, surprise, anticipation, joy, and trust). Using an emotion lexicon word list from the National Research Council Canada (NRC) [15] containing ~14,000 words related to emotions, we were able to find how people felt through this research pandemic period.

## 4.3. Topic Extraction

After observing how the public reacted to the pandemic, we decided to dig deep into the data and observe how people interacted with each other to get more insights. The goal was to create a clearer picture of social activity by observing the discussed topics between people.

We based our Topic Extraction Model on the popular TF-IDF model [16], a numerical statistic model intended to reflect how important a word is to a document in a collection or corpus. What we did here was to optimize this model's settings to best suit our use case.

### *4.3.1. Cleaning Data*

Removing any non-alphabet characters from our text, correcting spelling mistakes, and then lemmatizing [17] words to return them to their original form so that there wouldn't be any redundancy.

#### *4.3.2. Creating Stopwords*

Stopwords were a list of words that we tell our TF-IDF model to neglect and never consider when extracting topics. These words contained frequently used English stopwords, Arabic stopwords, and words related to Twitter (twitter, http, pic, com, ...) that didn't provide any useful insights.

#### *4.3.3. N-grams*

As extracting single keywords was lacking, we set our model to return the most relative bi-grams (2 words phrases) and tri-grams (3 words phrases) so that we could have some context and not just lonely keywords.

#### *4.3.4. Manual Filtering*

We chose to set our model to return the best 25 bi-grams and 25 tri-grams combinations which we then filtered manually for better viewing and clarity.

## 4.4. Public Interactions

Covid-19 mostly spread via physical contact or being in the range of an infected individual who wasn't taking the needed measures to protect himself or others. So, we decided to use our tweets data to find any public activity that would lead people to gather and have close contact with others. For our case, we chose to search for events that already happened to take note of them and never let them repeat as the pandemic was not going to disappear anytime soon.

To achieve this, we referred to the Covid-19 record sheet that we gathered from external sources, mainly from the Ministry of Health's daily reports as it was the most accurate, to get some hints about when there was a sudden increase in Covid-19 cases. That was done by studying the change in Covid-19 cases over time by comparing the difference between the previous day and next day cases with the cumulative average change in cases since the pandemic started. These "Hot Days" were then used as a reference to search for what happened at that time and whether there was any public activity that helped in spreading the virus.

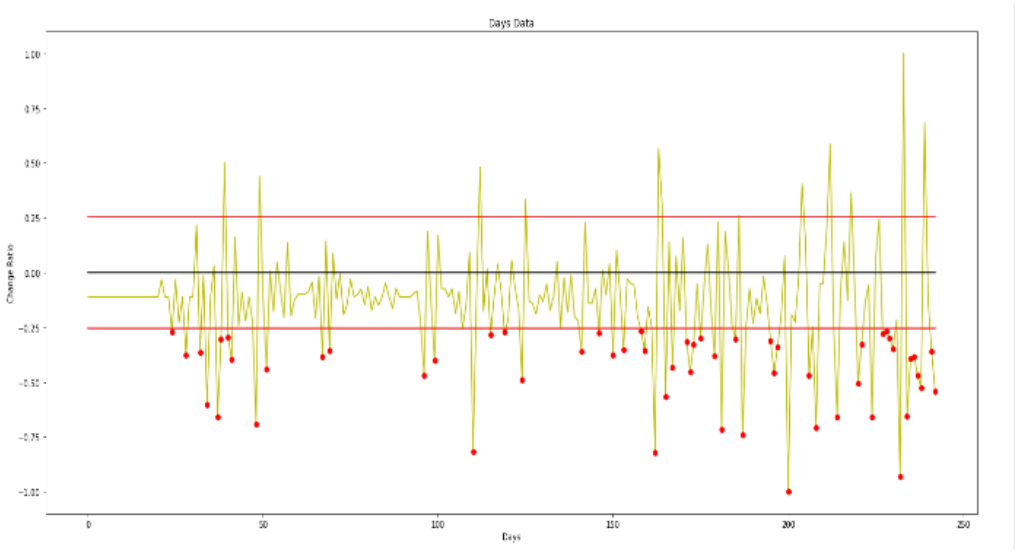


*Figure 2 - Days where Covid-19 cases suddenly increased (55 days)*

#### *4.4.1. Getting and Cleaning Data*

Our method was to check what was trending on twitter in the range of 4-5 days before each of these days as the sudden increase might not be directly due to that specific day's events. But before that, we needed to filter those 55 days (hints) we got and prioritize those with the greatest increase in cases and group the others that were in range of each other. After that, we got the tweets related to those trends where we cleaned and filtered them to remove tweets without a location reference in its content. Our filtering method was to check for non-dictionary words as location-related words are usually nouns and aren't included in a dictionary (ex: Beirut).

#### *4.4.2. Building Geo-Database*

After cleaning and filtering the data, we then got the needed geo data for any location reference in the remaining tweets and created a record of these activities in each day.

# CHAPTER IV: RESULTS AND DISCUSSION

In this research, we combined the location data with the time factor to give more useful location/time-aware insights. The results have been analyzed and grouped in daily and monthly basis where other time basis can be also considered. Daily analysis included the analysis results of daily tweets distributed by their locations, while monthly analysis grouped the results by month at the level of the Kadaas and Mohafazat, and overall situation. This research included data for eight months (Feb to Sep), so as to incorporate everything done, we will discuss the overall situation results in what follows.

## 1. Tweets Sample Data

This was the distribution of 50,000+ tweets related to Covid-19 we got according to each month (Feb to Sep).

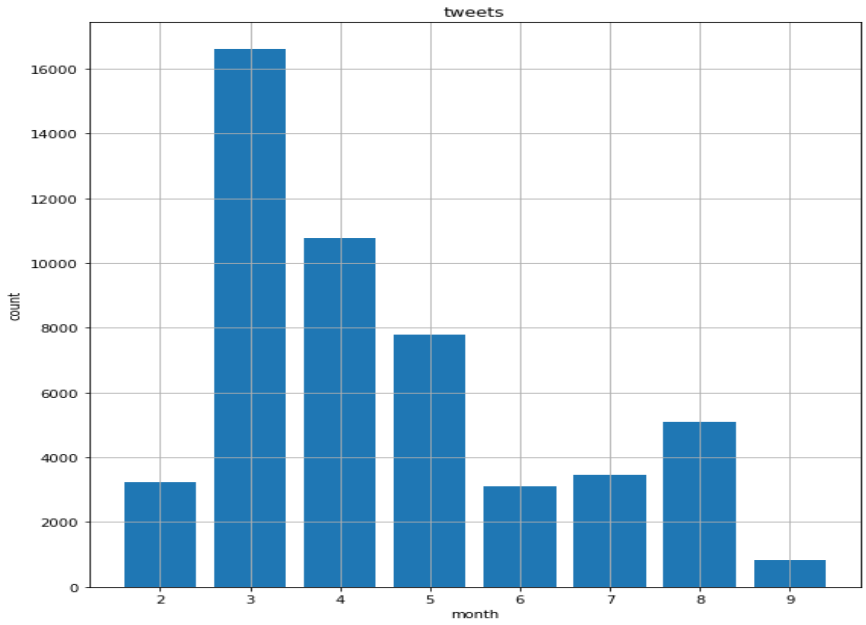


*Figure 3 - Gathered Tweets for each month*

## 2. Location Data

The final tool combined everything we talked about before and followed a specific workflow to determine the source location of the tweets we got. In this workflow, we first check the profile page of the user who posted the tweet to check if he provided a location reference there or not. Depending on that, we either used that location if it was similar to any of the location names we had or we got the user's last 500 tweets to check for any location reference. Finally, we choose the most frequent location from these references as this user's location address.

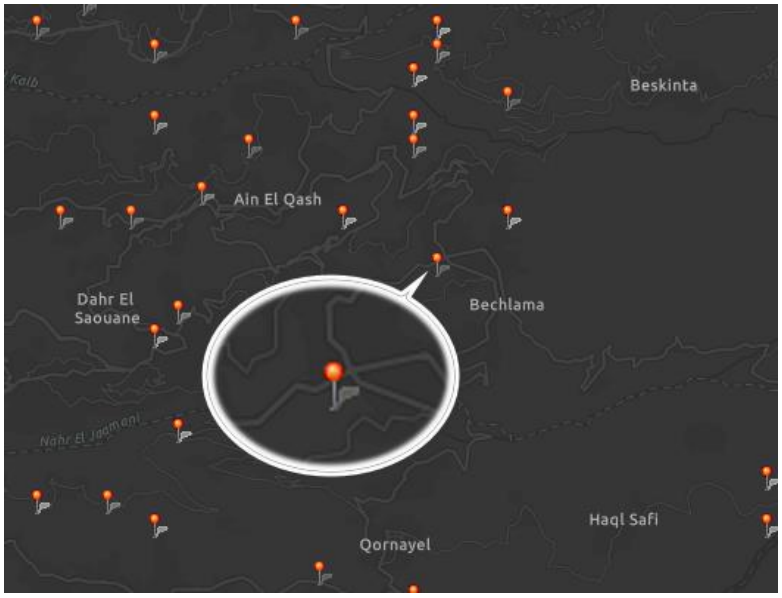


*Figure 4 - Mapping Tweets according to their location on a map*

# 3. Unifying Languages

We unified all languages by translating what we could using google translate and the rest with our tool. In this tool, we first checked for similar words in our dictionary and translated the rest to English after converting to Arabic.

After some observations, we found that these two methods combined gave very good results for individual word translation but lacked in keeping the text connected as our tool translated each word individually. Although the resulted text was not connected, we still managed to keep the mood and general intent of the text as it still contained some of its original meaning.

*Table 12 - Lebanese to English Translation*

| Lebanese | Google Translate | Tool |
|---|---|---|
| kn mnsab bas sa7 halla2 | Be the position, but correct Hala | was Position enough correct now |
| eh walla nsab bl marad | Uh, not lineage, but murad | yes indeed set up at disease |
| wasfi men edoctoor | My description is from Adster | prescription from doctor |

# 4. Sentiment Analysis

The sentiment analysis was done to observe whether people were accepting or complaining about their situation as taking any new measures for this pandemic without knowing how it would affect the public negatively or positively should never happen.

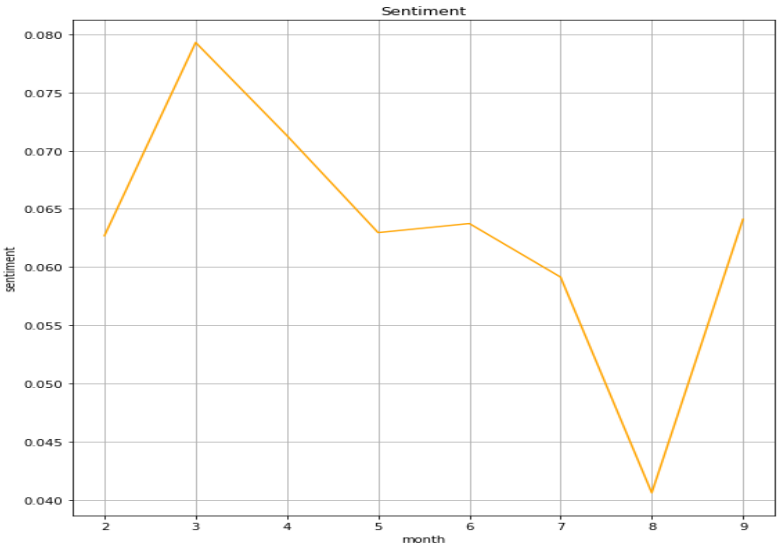


*Figure 5 -Monthly Sentiment Distribution*

**Observations**: Through this data (Figure 4), we found that people were the most positive at first (months 2 and 3) as Lebanon shut its airport, and the situation was pretty good in comparison to other countries that the pandemic started to increase in. However, that didn't continue as we started seeing a less positive response in later months (months 4 and 5) where the airport reopened and caused an increase in Covid-19 cases as infected travelers came into the country. The situation was a little better the next month (month 6) as cases reached a specific level that wasn't that bad in comparison to the total population where many measures taken by the government worked in quarantining most infected people and anyone who interacted with them. Alas, things didn't stay good as people started getting bored from staying in their homes and became careless about protecting themselves after seeing that there weren't many cases happening, which made people angry at these ignorant individuals after seeing the sudden increase in Covid-19 cases (month 7). As that was not enough, Lebanon reached a record-breaking of Covid-19 cases per day as people went out and gathered, especially after the great explosion that shook all of Lebanon at Beirut's Port that lead to a large number of negative responses from the public (month 8). Although the government applied measures in those two months (months 7 and 8), it was neither sufficient nor decisive enough in preventing crowds of people to gather especially near Beirut's Port. Finally, we saw a positive shift in people's sentiment in the last month data during the period of this research changes but as we checked the number of tweets we got that month and the average cases happening per day, we found that it was due to people not caring much anymore and Lebanese being the Lebanese they are as it's known that a Lebanese person can manage to get used to anything after some time.

**Conclusion**: This analysis can help the government in two ways. The first is to track public sentiment and see how people reacted to the measures they took so that they know if there is a need for more explanation or improvement. While the second is that they can also observe the current public sentiment to see how people might react to a new measure they want to apply especially when this measure includes actions that have complaints about them.

# 5. Emotion Analysis

As the Covid-19 pandemic needed the efforts of both governments and people to overcome it, it was essential for governments to take the appropriate measures to keep people hopeful and positive as much as possible since an overflow of negative emotions would cause public disruption and social upset.

We based this analysis on the eight basic known emotions: Anger (Red), Fear (Green), Disgust (Purple), Sadness (Blue), Surprise (cyan), Anticipation (orange), Joy (yellow), and Trust (light green). These emotions were extracted from the tweets and colored according to the emotion they represent to show how people felt during this research period. Also, by observing this data (Figure 5), we could further improve our previous observations from the sentiment analysis and dig deep into how people reacted to the Covid-19 pandemic during this research period.

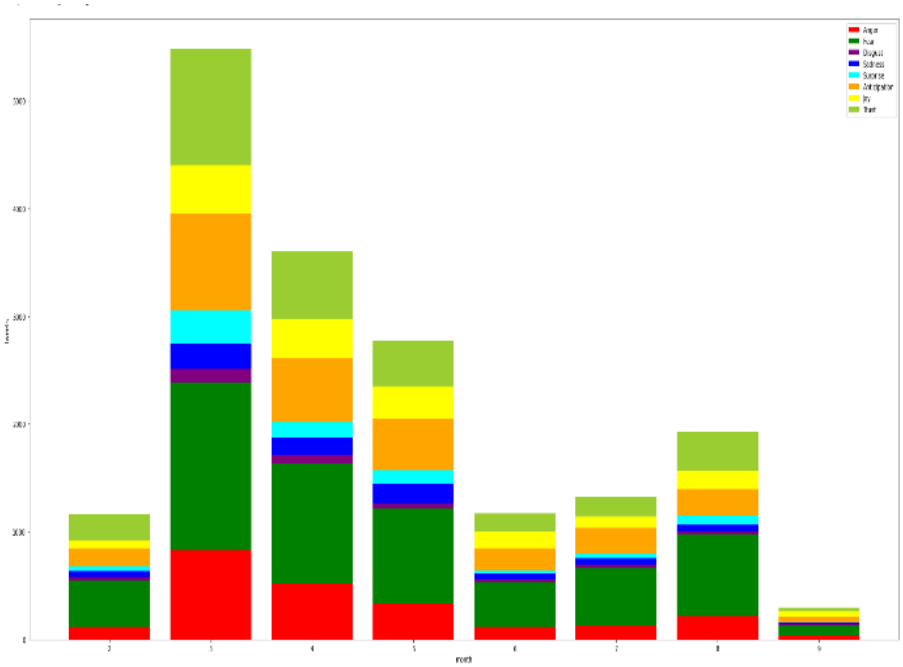

*Figure 6 - Monthly Emotions Distribution*

**Note: This analysis is word-based and not context-based, so there might be some misinterpretation of context in the results.**

**Observations**: We first observed how people felt fearful when hearing about this pandemic the first time and knowing about its fast spread in China (month 2). And as more time went by, people felt the most fear and anger when this pandemic started showing in Lebanon (month 3) but also remained positive after the airport was managed and shut down to contain the virus spread. This situation continued but with less impact as time went by and as more cases were being reported (months 4 and 5) after opening the airport, as people started to get used to living with this pandemic. That was especially visible in later days where people's interest was getting weaker (month 6). However, that carelessness brought with it a sudden increase in Covid-19 cases that ignited that fear again (month 7), which increased even

further, momentarily, when Beirut's Port explosion happened (month 8), just to die out the next month (month 9) as people got used to it and started being even more careless.

**Conclusion**: Using this analysis, governments can know when to take the appropriate measures to help people stabilize their psychological situation as this pandemic not only affected the infected people but also others who were afraid and anxious about what would happen to them and their loved ones.

# 6. Topic Extraction

The goal was to create a clearer picture of social interaction by observing the discussed topics between people.

*Table 13 - Mohafaza Sample*

| Month | Mohafaza | Topics |
|---|---|---|
| 6 | الجنوب | medical \| كورونا #لبنان \| #كورونا #كورونا_لبنان# \| Syrian refugees خليك_بالبيت #كورونا_لبنان \| حسان دياب \| # \| marijuana مدينة_صور # \| marijuana dispensary \| #كورونا_فايروس #مدينة_صور #جنوب_لبنان \| وزير الداخلية |

*Table 14 - Kadaa Sample*

| Month | Kadaa | Topics |
|---|---|---|
| 4 | بيروت | كورونا #لبنان \| #كورونا_لبنان #خليك_بالبيت \| فيروس كورونا \| جديدة # كورونا #لبنان_ينتفض# \| medical cannabis \| بفيروس \| إصابة جديدة |

**Observations**: Through these topics, we observed how people interacted with what was going during the pandemic period of this research and dug deep into what thoughts and opinions they had. We saw what was trending in every month in separate regions of the country as every region had its own mentality and interaction with this pandemic and the measure taken by the government.

**Conclusion**: These Topics can enhance our previous public reaction analysis as it digs deep into the tweets to provide us with the most important context that was being discussed during a specific time. This allows the government to further observe how their measures were perceived by the public and what they can do to improve their decisions.

# 7. Public Interactions

In this research, we searched for events that happened during the pandemic period of this research that might have resulted in the sudden increase of Covid-19 cases at specific periods of time.

*Table 15 - Crowded Spots Sample*

| Date | Source | Location | Event |
|---|---|---|---|
| 4/8/2020 | travelers | Rafic Hariri International Airport - مطار بيروت الدولي | دخول وافدين من الخارج الى لبنان |
| 8/19/2020 | locals | مار مخايل - بيروت | اعمال ازالة الركام في محيط مرفا بيروت |

**Observations**: The search we did provided us with data about possible events that happened at a specific date that might have increased Covid-19 cases. This data included the source of this increase by stating whether it was from locals or non-locals (Travelers) and both the location and description of the events that happened in the range of 4-5 days of this date that lead to this increase.

**Conclusion**: This search can help the government to not allow the repetition of events like this and take the appropriate measures to either prevent these events in the future or at least take precaution measures when they happen.

# 8. Dashboard

The dashboard contains all of the analysis mentioned projected on maps. These maps are time enabled and distributed into regions according to Kadaas and Mohafazat. **The dashboard contains four maps**:

**Tweets Map**: contains all of our data, including the sentiment, emotions, posting location, kadaa, and mohafaza of every tweet.


*Figure 7 - Tweets Map for month 2*

**Kadaas & Mohafazat Maps**: contains monthly Kadaa/Mohafaza data, including monthly sentiment, tweets count, emotions distribution, and topics. Each Kadaa/Mohafaza is highlighted according to its sentiment value where sentiment values are categorized into four sections (ranges), lowest to high, from light-green to deep-green to represent how positive the sentiment is.

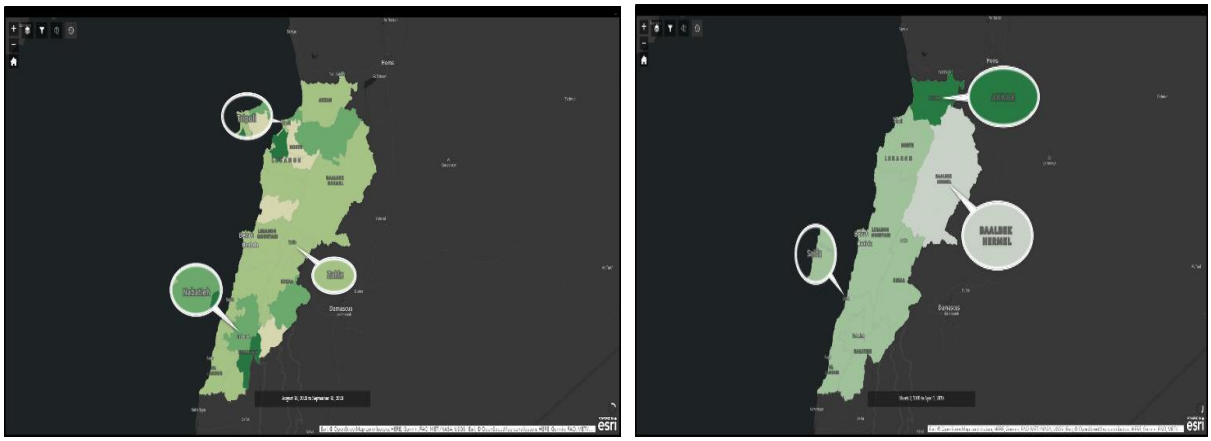


*Figure 8 - Kadaa Map for month 9, showing 3 Kadaas (Nabatieh, Tripoli, and Zahle) / Mohafazat Map for month 3, showing 3 Mohafazat (Akkar, Baalbak-Hermel, and Saida)*

**Public Interactions Map**: includes all crowded spots relating to the sudden increase of Covid-19 cases where people gathered and were in close range of each other, which might help the spread of the virus in case an infected person was there.

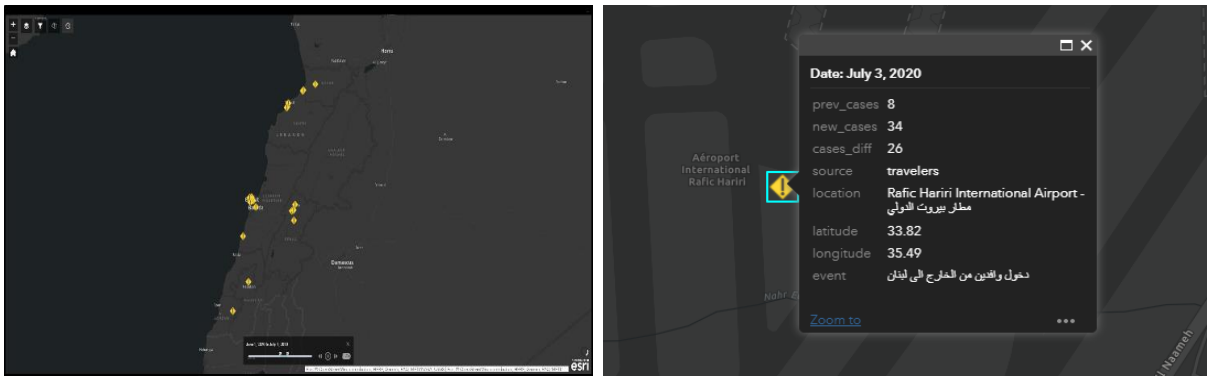


*Figure 9 – Public Interactions Map & details*


*Figure 10 - Dashboard*

**Charts**:

- **Overall Sentiment Distribution:** The overall distribution of sentiment of all tweets
- **Overall Emotions Distribution**: The overall distribution of emotions of all tweets
- **Monthly Sentiment Distribution**: The monthly mean sentiment for each month (Feb to Sep)
- **Monthly Emotions Distribution**: The monthly emotions distributions for each month (Feb to Sep)
- **Monthly Tweets Distribution**: The monthly count of tweets (Feb to Sep)
- **Daily Covid-19 Cases Change**: The daily change in Covid-19 cases where the change is the difference between the day cases and its previous day cases compared to the cumulative average of case change, i.e (PREVIOUS – NEXT) / AVERAGE

**Queries**:

- **Mohafazat & Emotions Distribution**: A query to select a specific mohafaza and view its monthly topic, sentiment, tweets count, and emotions distribution
- **Kadaas & Emotions Distribution**: A query to select a specific kadaa and view its monthly topic, sentiment, tweets count, and emotions distribution
- **Crowded Spots**: A query to view the crowded locations where people gathered at a specific chosen time

**Extra tools to dig deep into data**:

- **Layer Management**: Shows/Hides layers on the map
- **Filter**: filters the visible layer by choosing what data to show
- **Situation Inspect**: Inspects what is nearby a point, line, polygon you draw

**Time Slider**: Enables Time slideshow to view each map situation at every month

# CHAPITRE V: CONCLUSION AND FUTURE WORK

Public opinion incorporates all the desires, needs, and thoughts of people in society. It can be said to be the collective opinion of the country, which portrays the importance of this opinion in making decisions that affect the country.

As seen in our research, there are many ways to improve governments' public decision making through location-based analysis of social media content by providing a situational analysis according to the geographical locations of said content over a period of time. It showcases that we can base on people's reactions to extract key information about the situation and how we can further enhance that observation by digging deep into social interactions and actions.

As for future work, we hope that this research can be used as a real-time monitoring system where it can help governments or related organization in observing the public opinion in the hope they take it into consideration in times of need, like during this pandemic, to improve their decisions and lessen the weight on their fellow citizens as much as they can.

# BIBLIOGRAPHY


[1] "Great Depression," [Online]. Available: https://en.wikipedia.org/wiki/Great_Depression.

[2] "ArcGIS," [Online]. Available: https://en.wikipedia.org/wiki/ArcGIS.

[3] "Esri," [Online]. Available: https://en.wikipedia.org/wiki/Esri.

[4] "Levenshtein Distance," [Online]. Available: https://en.wikipedia.org/wiki/Levenshtein_distance.

[5] "Soundex," [Online]. Available: https://en.wikipedia.org/wiki/Soundex.

[6] M.-H. Tsou and L. Madsda, "Visualization of social media: Seeing a mirage or a message?," pp. 55-60, 2013.

[7] Liu, Y.; Sui, Z.; Kang, C.; Gao, Y. Uncovering patterns of inter-urban trip and spatial interaction from social media check-in data. PLoS ONE 2014, 9, e86026

[8] Yates, D.; Paquette, S. Emergency knowledge management and social media technologies: A case study of the 2010 Haitian earthquake. Int. J. Inf. Manag. 2011, 31, 6–13

[9] Yin, J.; Lampert, A.; Cameron, M.; Robinson, B.; Power, R. Using social media to enhance emergency situation awareness. IEEE Intell. Syst. 2012, 27, 52–59

[10] Sakaki, T.; Okazaki, M.; Matsuo, Y. Tweet analysis for real-time event detection and earthquake reporting system development. IEEE Trans. Knowl. Data Eng. 2013, 25, 919–931

[11] Bakillah, M.; Li, R.-Y.; Liang, S.H.L. Geo-located community detection in Twitter with enhanced fast-greedy optimization of modularity: The case study of typhoon Haiyan. Int. J. Geogr. Inf. Sci. 2015, 29, 258–279

[12] "Locations in Lebanon," [Online]. Available: http://www.fallingrain.com/world/LE/.

[13] "Google's Lebanon-To-English," [Online]. Available: https://sites.google.com/site/lebaneselanguage1/dictionary.

[14] "TextBlob," [Online]. Available: https://textblob.readthedocs.io/en/dev/.

[15] "Emotions Lexicon," [Online]. Available: http://sentiment.nrc.ca/lexicons-for-research/.

[16] "TF-IDF," [Online]. Available: https://en.wikipedia.org/wiki/Tf%E2%80%93idf.

[17] "Lemmatisation," [Online]. Available: https://en.wikipedia.org/wiki/Lemmatisation.

# LIST OF ABBREVIATIONS

**GIS**: Geographic Information Systems

**AI**: Artificial Intelligence

**NLP**: Natural Language Processing

**NRC**: National Research Council Canada

**TF**: Term Frequency

**IDF**: Inverse Data Frequency

# LIST OF FIGURES

# LIST OF TABLES