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 both the efforts of governments and individuals to overcome it. But as governments apply their planned measures, they lack the understanding of how individuals are moving and interacting with these measures. For that purpose, the goal of this research is to develop a geographically visual and context-aware mechanism that uses social media data, such as Twitter's tweets, to improve governments' public decision making by taking into consideration social reactions and interactions. Social media provide straightforward and familiar ways to communicate with all people categories all over the country, and that can be used by governments to communicate with the people and understand their needs and satisfaction level with the provided services. By nature, social media content reflects individuals' opinion or status at the moment of posting that refers to their current feelings and sentiment as well as what topics they are discussing between them as a whole society, while also containing data on the activities they announce to the public. The challenge here is how to analyze and understand the social media content with different languages, accents, and contexts, while also extracting the locations (in words) and convert them to a digital format (geocoding) to be projected visually on a map. This research uses NLP (Natural Language Processing) methods such as Sentiment Analysis, Emotion Analysis, Keyword Extraction, and Rule-Based Extraction to help achieve this intended mechanism of this research. This mechanism depends on geographically tracking individuals' sentiments and emotions to help the government improve on their interactions with them, while also considering their discussions as a way to understand their current thoughts and opinions. It also helps extract the most-visited 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 at these periods. Working on and analyzing a sample data of 50,000+ Twitter tweets, the 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 more to the disease spreading at some periods. It also provides a visual output of the analysis in a geographical dashboard using the ArcGIS platform that helps the decision-maker to more understand the geographic perspective of the situation of the pandemic .For future work, we can work on extracting more data from other social media sources and enhance the extraction methods that better understand text context and provide more concrete insights.

1. INTRODUCTION

COVID-19, also known as the coronavirus, was declared a Public Health Emergency of International Concern in January 2020, and a pandemic in March 2020 after first being identified in December 2019 in Wuhan, China. 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 as an infected person 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 often forget that what they are doing might not be accepted by the people as they wanted it to be. This pandemic requires both the government and society combined effort to prevent further spread of this virus. There is always this communication gap between governments and people as most of the decisions are taken without considering what people are currently feeling and experiencing. 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 must have more insight into people's thoughts and opinions while also having a better way to express their decisions in a way that allows people to relate to it and achieve the intended results. So, there must be a way to better understand people's thoughts and opinions and consider them when making decisions. Social media data can be one of the best sources of insights on how people currently feel and think, as they provide a way for people to express their feelings, interests, complaints, and thoughts.

People all over the world use Social Media as a platform to express themselves as it allows them to have direct access to each other to share their thoughts and feelings about what's most important to them. It is a content that expresses the different emotions, thoughts, topics, and sentiment of society and relates to their needs and priorities as a whole community. What makes this content important is that it is coming from the people who are experiencing this pandemic directly in their everyday life and have their problems and hardships that they want to overcome. Analyzing this content and visualizing it geographically on a map can help improve governments' public decisions in many fields such as maintaining public sentiment, better decision delivery, prioritizing people's needs, and getting some opinions that might be of help to the current situation.

1. PROBLEM DEFINITION

Posts are the building blocks for any Social Media Platform as they are the content that provides this platform with context and meaning. By nature, this content is unpredictable and relates to individuals who each have their unique way of expressing themselves and sharing their thoughts and opinions. These posts come in different forms and formats as they can be sent in the form of text, images, or videos while including distinct languages or language accents unique to each country or country part. That's why it's difficult to take every aspect of these forms and formats into consideration in our analysis as there is no specific standard that we can base on. Such a problem requires us to think of a way to convert and unify these different forms into one in a way that keeps the original context intact while also allowing easier text handling for analysis. Then, we face another problem of how to get the geographic location of these posts owners as we want to differentiate between different country areas in our analysis, and that will require us to think of ways to get this data taking into consideration the source we are getting the data from.

1. BACKGROUND AND RELATED WORK

intended

1. CONTRIBUTION

The purpose of this research was to find a way to visualize and analyze 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.

This research proposed many solutions to reach this intended mechanism where we gathered social media posts, discovered their origins, handled many different text forms and formats, and g finally got the geographical distribution overview of public reactions and interactions.

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, Corona, Healthcare, Medical, and كورونا

1. Location Data

Getting the source locations of tweets was one of the main problems that this research wanted to solve, and by default, Twitter had its own mechanism for locating tweets as they provide two public ways to do that.

1. Method A: User Location Access (Settings Enabled/Disabled Option)

This option, which was disabled by default, allowed Twitter to know the user's location when he posted his tweets while also permitting other people to search for these geo-tweets via the geocode search function in the Twitter search field.

**Search Parameters**: <search- keyword> geocode:<latitude>,<longitude>,<radius>km

1. Method B: User's Profile Location

Another way was the user-provided location in the User's Twitter page that users customized as they like. Twitter then tried to link this given location to an actual real-world location and included these users' tweets in the geo-search results.



1. Limitations

The problem with method A was that it only showed the geo-tweets of people who enabled Twitter location access from their account settings, which was OFF by default. Moreover, this option was not known to many users and mostly ignored by the rest. As for method B, many users either didn't fill their profile location correctly or just left it empty. These limitations lead to fewer geo-tweets resulting from our geo-search, as our benchmarks showed that searching for non-geo-tweets with the same search parameters gave us ??x times the tweets than when we used the geo-search function to get the geo-tweets.

1. Proposed Solution

Our proposed solution was to use each user's tweets as a data bank that reflects 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.

* 1. Locations Data

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

* 1. 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 used appropriate fuzzy search methods for each case.

* + 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.

|  |  |  |  |
| --- | --- | --- | --- |
| 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) |

* + 1. 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 same the 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 - Soundex Fuzzy Search Method Example

|  |  |
| --- | --- |
| Word | 4-digit Soundex Codex |
| Beirut | B630 |
| Beyrut | B630 |
| Beyrouth | B630 |
| بيروت | 000ب (Doesn’t work for Arabic) |

Table - Edit Distance Fuzzy Search Method Example

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

1. Final Tool

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.

1. 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.

1. Tweets Languages

People mostly tweeted using four languages: Lebanese, Arabic, English, and French. These languages had their own form and format except the Lebanese language. This language 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 - Lebanese Language Example

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

1. Unifying Languages

This research proposed to unify these languages into one language that is 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 - Lebanese Language Translation by Google Translate

|  |  |
| --- | --- |
| 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 |

1. 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.

* 1. Arabic Lebanese

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

Table - Google Translation of Lebanese Arabic

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

* 1. 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 depends on two methods:

* + 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 [?] 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 – Lebanese to English Dictionary Example

|  |  |
| --- | --- |
| 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 |

* + 1. 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 of it which was translated into English.

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

Table - Mapping of Single letters

|  |  |
| --- | --- |
| EN | AR |
| A | ا |
| B | ب |
| D | د |
| E | ي |
| F | ف |
| G | ج |
| H |  |
| 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 | ء |
| 3 | ع |
| 5 | خ |
| 7 | ح |
| 8 | غ |

Table - Mapping of Double Letters

|  |  |
| --- | --- |
| 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 - Translation Example

|  |  |  |  |
| --- | --- | --- | --- |
| LB | LB-AR | AR | EN |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

* + 1. Final Tool

As our dictionary didn't cover everything and still lacked a lot of words, we tried to check for similar words by using different methods (Insertion, Deletion, and Substitution) as some words differ by 1-2 characters and still retained the same meaning. These methods were Fuzzy Searching [1] methods and could be observed in the following example (Table 10).

Table - Fuzzy Search methods

|  |  |  |  |
| --- | --- | --- | --- |
| LB 1 | LB 2 | EN | Actions |
| Hone | Hon | Here | Insert “e” into “Hon” |
| 2hwi | 2ahwi | Coffee | Delete “a” from “2ahwi” |
| Kifak | Kefak | How are you | Substitute “e” with “i” |

After this check, similar words were translated according to the dictionary while the other words were translated from Lebanese to English by mapping them to Arabic then translating to English.

1. Observations

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 - Lebanese to English Translation Samples

|  |  |  |
| --- | --- | --- |
| 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 |

1. 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 - Analysis Workflow

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 is an already known Natural Language Processing (NLP) method that extracts text sentiment and categorizes it. It was a tool that didn't need much work to be useable. But, it was essential to note that sentiment analysis can differ based on what it has trained on as there were many variations of trained sentiment analysis models. For this research, we used the TextBlob 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.

|  |  |
| --- | --- |
| Text | Sentiment |
|  | Positive |
|  | Neutral |
|  | Negative |

1. 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 they could 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) [??] containing ~14,000 words related to emotions, we were able to find how people felt through this research pandemic period.

1. 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 the social activity by observing the discussed topics between people.

We based our Topic Extraction Model on the popular TF-IDF model [??]. It is 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.

* 1. Cleaning Data

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

* 1. Creating Stopwords

Stopwords are 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, ...).

* 1. 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.

* 1. 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 results and clarity.

1. Public Actions

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. These "Hot Days" were 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.

* 1. Getting Tweets 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. 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).

* 1. Manual Filtering

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 gathering spots in each day.