Media Engineering and Technology Faculty
German University in Cairo

Spatio-Temporal Urban Sentimental Analysis

Bachelor Thesis

Author: Amr Khaled Elsafy
Supervisors: Dr. Amr ElMougy
Dr. Mervat Abuelkheir
Submission Date: 22 June, 2020

Media Engineering and Technology Faculty
German University in Cairo

Spatio-Temporal Urban Sentimental Analysis

Bachelor Thesis

Author: Amr Khaled Elsafy
Supervisors: Dr. Amr ElMougy
Dr. Mervat Abuelkheir
Submission Date: 22 June, 2020

This is to certify that:

(i) the thesis comprises only my original work toward the Bachelor Degree

(ii) due acknowlegement has been made in the text to all other material used

---

Amr Khaled Elsafy
22 June, 2020

Acknowledgments

I would like to show gratitude to Dr. Amr ElMougy who continued to supervise and guide this thesis to completion during the times of pandemic. I also would like to thank Jose Portilla on his classes that helped go from zero to hero in the field of Data Science. I would like to thank my friends whom have given me an advice, lended me a hand of help or an ear to listen during this tough times. Finally, I would like to thank Allah for giving me the strength and health during this times and the wisdom to contribute to science for the benefit of the whole world.

Abstract

In the era of digitalization, each day people leave behind their primitive lifestyles and fous their attention on ugrading digitally however not all region are affected the same way, considering the difference between rural and urban areas. We decided to put that thought into question, given a specific location and timeframe bound, we investifgate and quantify the variation in both the interaction and sentiment on the internet regarding the urban features of the region. Moreover, we made a case study on the society's mood and behavior during the times of the COVID-19 virus, showing how much people are affected and the way they are affected. Previous works have investigated sentiment analysis on city events and exposure to urban areas with several methods being introduced in analyzing them. None of them however, has been in the Middle East area nor applying their investigations in the Arabic language. This study analyzes the activity and mood of the Egyptian governorates on social media in terms of their urban features during the two weeks COVID-19 was declared a pandemic and quarantine measures were taken. First we collect all spatial and urban features regarding the Egyptian governorates to build our own dataset along with building a dictionary of terms associated with the coronavirus for detection. Afterwards we introduced four approaches in tweets collection regarding locations and collected a governorate labelled corpus that would undergo our machine learning based model for sentiment annotation. Through training and testing, our model achieves best accuracy through normalizing and stemming the tweets, extracting both unigrams and bigrams features and classifying using a Naive-Bayes classifier. Our results show that Cairo, the capital, is the most interactive on social media with a total urban land use and most population, however it also showed it was the least happy and a significant difference in the results to the rest of the governorate which could be justified given the highest rates in infection[1]. It also shows that the impact of the stay at home procedures negatively on people, halting the education on campus and remote work, but people have been adapting given the show of a better mood through the end adding more positivity when mentioning the virus

[1]https://www.care.gov.eg/EgyptCare/Index.aspx

Contents

Chapter 1

Introduction

1.1 Motivation

The rapid growth in the number of users in social media contributing to the public expression in publicly available data is splendid for data scientists, especially with a lot of developed tools nowadays for scrapping and analyzing the continuous data stream. Several works have been made in the field of sentiment extraction from text content, and eventually these studies came to shed light on the field regarding Arabic language along with various preprocessing techniques that inspired us to perform sentiment analysis on Arabic real life data.

Furthermore, various attempts have been done regarding the analysis of spatio-temporal data to present patterns and investigate the behaviour of the people in their cities yielded outstanding outcomes in the analysis and gave explanations to unresolved behaviours which have been lacking studies over the Middle East, leaving a lot of room for unanswered questions. This study is being done during the time of the COVID-19 virus pandemic, a major event that has been affecting humanity globally, leading the people to stay in quarantine for months so far. We are encouraged to look into having a spatio-temporal study, especially in the Arab world, to analyze the sentiment and behaviour of the people during a specific period in the pandemic.

1.2 Aim of the Project

The purpose of having this study is to first, overcome the challenges faced when dealing with untagged location tweets through introducing several approaches in data collection in terms of locations. We also discuss and include several techniques in Arabic text preprocessing and testing them with different combinations in order to get the best possible accuracy for Arabic sentiment classification. Using our annotated corpus, we aim to analyze and draw patterns regarding two factors: Urban Analysis and the COVID-19 impact

all over the regions of Egypt. We examine the activity and sentiment over the governorates of Egypt to report and inform about the social media interactivity and mood over each governorate. Finally we choose to specify a 14 day period over the 2nd till the 4th week in March to highlight the impacts of the rapid increase in the virus spreading and declarations of the virus as a pandemic and the quarantine procedures on people. All in all, this study puts its labelled corpora for investigating the variation of activity and sentiment in terms of the urban features and the virus' effect on Egypt's governorates during the beginning days of the pandemic.

1.3 Outline

The second chapter represents a background on the fields of text data mining and preprocessing in Arabic along with the works done on sentiment classification and a literature review on the applications done on spatio-temporal data. The third chapter introduces the methodology through various data collection and preprocessing methods as well as building a model for Arabic sentiment annotation. It also includes building a dataset on the Egyptian governorates spatial and urban features and a dictionary for coronavirus detection. In the fourth chapter we analyze and draw insights out of our analysis in terms of Governorates' activity with respect to urban features and the coronavirus related situations. Finally, in the fifth and last chapter, we conclude all our work and results in our study and propose further work in the fields.

Chapter 2

Related Work

2.1 Internet text data mining and preprocessing

It is without a doubt the influence of the internet on people has been clear , with the evolution coming with Web 2.0 allowing users to have access to numerous rich online content and resources with fast accessibility, it has grabbed the attention of a lot of researchers to invent data mining techniques and analyze a study of the behaviour of people on the internet through the various forms of expression on the internet with different linguistic styles to get insights better than usage of public polls.[1]

2.1.1 Data collection

Arabic Language has shown difficulties when it comes to data collection from the internet, especially when compared to a similar task of the English language. These difficulties occur due to the limited and small number of Arabic activity over the internet in reviewing and e-commerce websites for example. Another problem which has surfaced over the internet in the middle east is that most Arabic internet users tend to use the English language in communication or Arabic transliterated in Roman characters to write in their posts and reviews. Fortunately, there has been an emergence of pure Arabic websites that made more Arabic content available for reviewing and analyzing.[2]

ElSahar and El Beltagy built their own customized web crawler to meet their requirements to scrape over the Arabic (or English-based with Arabic content) review websites using the open source Scrapy framework, to automate the generation of the annotated dataset. Unfortunately, review websites will have a more formal content and focus on a certain domain which would yield in a lower accuracy model than what social media websites can offer of a more day to day general discussions and moods that can just be domain-less. [2]

Twitter's API is used for collecting numerous amounts of text content from tweets through

Twitter's own search API where it has critical information to our research like the geocode and the timestamp as well as the text mainly[3]. Data and sentiment can also be derived from images and videos through image recognition, but is more expensive to analyze than text and would require a much higher computational power and access is more often restricted due to user privacy.[4].

Shoukry and Rafea scraped a corpus from Twitter of 20k tweets from different news topics, taking into consideration having enough large size to get more accurate classifier results. Afterwards, comes the annotation process where the dataset had two raters labelling the sentiment polarity of tweets with an 80% agreement on the labels on tweets and leaving up the remaining tweets for a third rater to be the deciding vote.[5].

An 84,000 tweets dataset was collected in two phases by Nabil et al. . The first phase was defining the top active accounts in Egypt on Twitter using SocialBakers website that yielded in the top 30 usernames and got the latest tweets for them in their last two years at the time ending up with 36,000 tweets total. In addition, they crawled EgyptTrends, a twitter page that specifically tweets about top hashtag trends in Egypt, getting 2,500 different hashtags that were used to later get 48,000 tweets.[6]

Gong et al. have used sentimentally annotated datasets that have already been collected to the size that is both sufficient and small enough for the ML methods to stabilize performance with the available resources which they found best at 8,000 posts. They have classified them into two types: Common-based dataset and Event-based dataset. The common-based datasets have no context or knowledge about city events allowing a wide variety of data and variety of overall sentiment. However, the event based datasets are data collected around and about events which were found to be more expressive and polarized.[4].

Using the Twitter API and the search API provided by twitter, Shoukry et al. have focused on the searched tweets to be of Arabic language by setting the lang=ar. They made sure the data is of sufficient quantity for the classifier to be trained, having over 4000 tweets.[7]

Abdulla et al. scraped 2000 tweets equally balanced in labelling, using a tweet crawler including various topics and different type of documents written in both Modern Standard Arabic (MSA) and the Jordanian dialect. The aim of the scraped tweets is to be captured based on emotional value for the system to use this information to determine the polarity of the text.[1]

2.1.2 Data processing and filtration

The classification of a text differs from having the variety to be a sentence, phrase or a document to standardize the form of all sentences. The choice of the type of documents in

the corpus impacts the analysis depending on the length, formality, cultural and sarcasm features present.[1] In addition, social media texts are often noisy, more often in the negative texts [5] with a lot of mess due to lack of having a standard form from misspellings, missing punctuations to redundant repetitions and thus requires a lot of preprocessing. Going into the text preprocessing process, with its three main stages: Normalization, Stemming and Removing stopwords.[7].

Text Normalization

Normalization is the stage to transform the text to a common standard form for all the data, there have been various approaches to normalizing the text content taking into consideration the background, the domain and the formality of such text.

When dealing with reviews from various domains such as hotels and restaurants, El-Sahar and El Beltagy made sure to eliminate the irrelevant reviews along with multiple duplicates of spam reviews from their datasets. For example the dataset collected from product reviews were so redundant that only 5k reviews were unique out of the 14k collected reviews, due to the fact most of these reviews contain one word. Also while they were looking into movie reviews, they discovered that people tend to thoroughly express their opinion and reviews on movies. They annotated this dataset by extracting the ratings each review had, then normalizing the rating to be of three categories: positive, negative and neutral.[2].

Schwartz et al.[3] and Mitchell et al.[8] are analyzing twitter data to observe individuals behaviour by analyzing word frequencies and time series to derive patterns in mental changes. This also calls to the techniques used to filter out the automated bots, businesses and private accounts from the data, but it may have been more illustrative to not filter out non English data as well[3]. Whereas Mitchell et al. saw that get the results needed they focused on the 50,000 most frequent words in each region[8].

Shoukry et al.[7] filtered out tweets that only held one opinion and from different topics, and for a more understandable format to the classifier, they have removed usernames, URLs, hashtags and all non Arabic words to construct feature vectors using term frequency. They have also expressed the problem of the spam and misleading tweets which could affect the classification when the classifier is built on such tweets.[6]

In addition to collecting data from Twitter API, Gong et al. used SocialGlass which is an integrated system specialized in collecting and processing of social media data. The filtration process was to remove spam accounts along with short posts (shorter than 30 characters). For non English posts, Google Translate API was used to translate all posts to English to be the best translate service compared to others, but is still very limited to its accuracy of its translation. Also, Figure Eight, a crowd-sourcing platform, was used to perform crowd-sourcing operations.[4]

Shoukry and Rafea have used an already made normalizer claiming that it's very efficient, on the contrary the normalizer appears to be unavailable at this time, but fortunately they have mentioned some of the normalization rules they followed to normalize Arabic text such as removing 'tashkeel'(diacritics) and 'tatweel'(elongation) and standardizing Arabic characters to the simplest normal form like the letters 'alef', 'yeh' and 'ta'a', due to the similarity of the letters in use the Arabic internet users often don't settle for one character.[5]

Abdulla et al. separates the positive and negative tweets into separate files, they use MS Word's dictionary in their tool as a reference to correct the misspellings in the text. For the repeated characters, they are removed through an algorithm that detects every word with length over 5 and automatically removes any repeated characters and checks up the word in the MS Word dictionary.[1]

Text Stemming

Stemming is the stage of reduction of the words to either its root or another type of base form, so usually the commonly related words would map to the same stem, making the ideal objective of a stemmer is to reduce the word to the shortest form possible without having a different meaning. Stemmers can also reduce the number of features extracted significantly due to features' convergence in similarity, especially when using unigrams which makes for better accuracy when classification and reduction of the overfitting problem.[5]

The stemming process has been highlighted over research studies for making a difference in the field of text mining yet still facing problems with the Arabic language. Not only Arabic researches into stemming have expressed the complexity of the process, but also the studies were made over modern standard Arabic (MSA) which can't handle the variety of existing Arabic dialects like the Egyptian dialects, stemming words that shouldn't have been stemmed and leaving out words that it didn't recognize, it needed stemming. There are two types of stemmers existing for Arabic text: an aggressive stemmer that takes the word and directly returns back its corresponding root and a light stemmer which looks for certain shape of prefixes, affixes and suffixes to determine the removal of unnecessary character and achieve a more simpler form.[5]

An approach was then made to create a custom light stemmer specified to the domain of the Egyptian dialect for the simplicity in implementation and high effectiveness. A light stemmer was also chosen over the aggressive due to the weakness of aggressive stemmers in making decisions that would map to too many common terms with different meanings to the same root. Also in the works of Abdulla et al., both aggressive and light stemmers went various experiments and each time the light stemmer would outperform the aggressive one [1]. The Arabic light stemmers have some deficiencies dealing with the broken plurals, nevertheless Shoukry and Rafea came up with a set of rules to deal with such problem by adding and removing certain infixes and affixes given the length and

the shape and order of characters. They have also presented two lists both normalized and stemmed for the Egyptian dialect: one for the irregular terms where the input word first passes by to check if it matches otherwise stemming rules will apply, the other is for the irregular plurals. As a result, the customized stemmer for the Egyptian dialect outperformed the default Arabic light stemmer.[5]

The implementation of the stemmer passes through three phases: prefix removal, suffix removal and infix removal (mostly for broken plurals) in the same corresponding order. They check the stemmed word in the list to see if it is there, if not, the stemming continues till all three phases are done.[5]

Stopwords Removal

Lastly is the stopwords removal stage which is a reference to words that add little to no meaning to the content of the text such as prepositions. Different stop words are used in classifying the Arabic sentence for the variety of dialects making it harder for a general classifier to perform in each different dialect, causing very small improvements after removing the stop words. This causes important words to be removed when it shouldn't have been and/or leave other stop words that need to be removed.[5]

Determining the Egyptian dialect stopwords to build a dictionary had to come from scratch due to the lack of Egyptian stopword dictionary. In the beginning, they identify words in the dataset, calculate their frequency value and classify them into frequency ranges in which they proved Zipf's law stating that there exists an inverse relation between frequency range and word counts. Moreover, Shoukry and Rafea began collecting top words after removal of sentiment words, named recognition entities and verbs, despite the classifier's accuracy was dropping, they have claimed its due to removal of important words and maintaining unimportant ones and thus the process had to be done over manual speculation ending with a 128 stopwords that would increase the classifier's accuracy by a significant 1.5%.[5] When dealing with stopwords, Abdulla et al. got their stopwords list from the Khoja stemmer tool and further added different stopwords from different Arabic dialects.[1]

2.2 Sentimental Analysis

Sentiment analysis (also referred as opinion mining)[1] has been the focus of many scientific researches due to its potential to be involved in many various of applications for its relation to fields like text mining and natural language processing (NLP). It is used for the purpose of identifying positive and negative opinions and emotions automatically, in order to determine the attitude of the writer or the tone of its content.[7]

A study by Korayem et al. has classified sentimental analysis and subjectivity to be

considered through these four categories: Predicted class, either the text is subjective or objective. Predicted polarity, text is of positive, negative or neutral nature. Level of classification for sentiment analysis, whether the text is of word, sentence or document level. Finally the Applied approach, either a supervised or an unsupervised one.[9]

When sentiment is taken into consideration, there are two schemes that are best followed when categorizing a sentiment, simplified and detailed. A simplified sentiment scheme is one with two classes: a polarized class that only features two categories: positive and negative and the "3 class" class which involves same two categories with the addition on the neutral classification. On the other hand, the detailed scheme includes more categories ranging with extreme, very and slightly negative and positive values along with neutral.[4, 2] Also, sarcasm content has shown to have ambiguity in the misinterpretation of whether the intention is a positive or negative one leading to a consequently wrong polarity classification.[1]

The sentimental polarity is defined as a metric from word to document level's opinion and subjectivity, whether such word or document is of a positive opinion or a negative one. Moreover, the annotation process is a process of giving each piece of text a sentiment value, but it is a human made one, thus it is prone to have a large human error especially when done with a detailed scheme which gives more the reason why the existing datasets are based on simplified scheme annotation missing out on the strength and detail of the detailed scheme, in addition to the difficulty of labeling cultural and sarcasm content. Also for supervised ML approaches, a huge dataset requires building and manual annotation that would be costly and very time inefficient.[4, 1]

The negation mechanism (or switch negation) is simply inverting the polarity of a sentence when it proceeds by one of the negation words in the sentence. It is found hard in Arabic to implement such a mechanism easily due to the language having around 20 negation words.[7] Abdualla et al. handle negation through adding a negation list with the main negation words in the MSA in addition to negation words from different dialects. Furthermore, intensification mechanism or the booster words are used to strengthen the word's polar intensity, in which the main Arabic booster words can also come not only after it, but also before the adjective.[1]

2.2.1 Feature Extraction

Feature vectors are built to be used then as input parameters for machine learning models to train the classifier upon and have similar parameters to be used when testing and generating sentiment with several approaches in extracting features from a document. Dealing with word-document level, features chosen to work with are unigrams, bigrams and trigrams. Unigrams, where each word in the text is considered as one token, are the easiest in extraction and provide good insights for the data, however working with bigrams and trigrams (two words and three words as one token respectively) would prove more

context in cases of negation and patterns of sentiment expression. Significance in accuracy can be noted when using bigrams along with unigrams as features, however not much significance can be shown when adding trigrams to the features' list, but rather adding more cost for computation than accuracy according to Shoukry and Rafea [5] whom on the contrary had problems with a relatively small dataset to show actual significance.

ElSahar and El Beltagy have applied several famously used methods in sentiment classification such as word frequency, word existence and TF-IDF (Term Frequency - Inverse Document Frequency Matrix). Also they have a used Delta TF-IDF, a derivative from TF-IDF assigning n-grams a weight corresponding to the difference of the normal TF-IDF values which is promising with efficiency due to better performance in documents with common recurring subjective words more likely to appear in a large number of documents resulting to a small number of IDF values. When conducting experiments, it has shown that there isn't a significance difference in the performance between a traditional TF-IDF and a Delta one as some datasets tend to work better with the Delta and others with the traditional, however they will both work much better when using a 2 class sentiment classification rather than the 3 class and with lexicon based feature vectors.[2]

Another feature representation ElSahar and El Beltagy tackled is a feature vector made out of entries from previously generated lexicons, where they referenced both lexicon domain specific and a lexicon with all domains for comparison. The document is represented by matching its content to its corresponding entry in the feature representation and their counts. Their results have shown that the performance when referencing a lexicon of all domains outperformed a lexicon domain specific and not referencing any lexicons over all experiments with different feature representations and datasets.[2]

In the exploration of creating a feature vector for each tweet, Shoukry and Rafea had taken out unigrams, bigrams and trigrams from a labelled corpus of 1000 tweets as features. For each feature, they calculate their frequency in the main corpus of 20k tweets and add them to a dictionary such that each feature has its frequency opposite to it. Then for each tweet, we check if any of the candidate features exists, so that it's matching frequency can be fetched and placed into this tweet's feature vector. By the end the tweet would end up having a feature vector of the form (word1:frequency1, word2:frequency2 . . . , "polarity").[5]

Along with ngrams, Cargea et al. have identified sentiment features such as polarity cues, emoticons, internet acronyms and punctuation and SentiStrength. Polarity cues consist of having three different features: PosDensity, NegDensity and PosVsNegDensity. PosDensity is the number of positive words normalized by the tweet's word count, same goes with NegDensity for negative words and PosVsNegDensity is the number of positive per negative polarity cues ((PosDensity+1)/(NegDensity+1)). Emoticons are widely used through social media to express emotion, so tweets are checked for emoticons by searching through Wikipedia's emoticon dictionary. Internet Acronyms commonly used were collected (ex. lol) and constructed a positive and negative acronym dictionary and calculated the counts. Punctuations were taken as features for their show of intensity

(ex. I hate this!!!!!!) most frequently the exclamation and question marks counts. Lastly, SentiStrength calculate sentiment strength score using the toll's own algorithm for short informa online texts.[10]

2.2.2 Sentimental Analysis Methods

Sentimental analysis methods can be classified to the most commonly used methods of analysis into lexicon-based methods, machine learning-based methods and hybrid methods and it is still can't be determined which method proves to be most effective in each case or scenario. The performance of each method is determined through three accuracy measures: Precision, Recall and Accuracy for supervised and unsupervised methods. Each metric is an equation formed of TP, TN, FP, FN values, which are true positive, true negative, false positive, and false negative.[1] Python's Scikit Learn library offers the usage of classification reports to calculate the weighted accuracy of a given classifier's performances.[6]

Lexicon based method

Also known as the semantic orientation method is an unsupervised dictionary based method where it assigns a consecutive combination of words a sentimental score (or the polarity strength if available in the detailed scheme to show variety of scores) according to its sentiment dictionary and then calculates the weighted average sentiment score of the whole text (filtered of stop words), but its weakness shows when taking context into consideration [4, 7], as well as adapting to different domains, but many researchers prefer the lexicon approach for to avoid the costly effort of manually labelling a corpus.[1]

Machine Learning based method

A method that trains a model with datasets that are annotated and verified of their sentiment in their contextual content, so it will have a combination of different features that will result in a specific class. The annotated datasets are splitted into two parts: a training and validation part and testing part in order to test the model on unseen data to make sure the model isn't over fitted, other than the data selected to train on, where most research studies preferred the 80:20 split ratio. Some researchers tend to a three partition split for having a validation set acting like a mini test[6]. The testing can also be done using different n-folds cross validation taking into consideration the memory limitations as well[1], having a balanced and an unbalanced version of the annotated dataset and if available testing a three class (the neutral class) version of the dataset compared to the polarized dataset (positive and negative only)[2]. The model is then used calculate sentiment giving a better accuracy due its semantic orientation for better generality.[4] For instance, when using a balanced dataset version over an unbalanced version, the balanced version will always have fewer training examples leading to data sparsity for many

ngrams and resulting in a less accurate classification, so it is always better to have the difference to be close in number.[6]

The most popular machine learning classifiers in sentiment analysis are Naive Bayes (NB) and support vector machines (SVM) where multiple works have shown both methods to be the best performance between them, nevertheless are supervised classifiers and would need a large annotated corpus. Naive Bayes (NB) is a supervised linear machine learning algorithm that classifies its data according to Bayes' Theorem, popular to use when classifying texts. Support vector machines (SVM) are a group of learning models used linear and nonlinear classification analysis having an edge over NB that is based on probabilities. The performance of SVM has an inverse relationship with the user parameter C into it, whereas the value of C increases, the accuracy of SVM classifiers decreases[10]. Weka Suite Software has been used to implement SVM and NB classifications and testing.[4, 7, 5]. The RapidMiner software tool has also been used for data mining and building ML models containing text preprocessing methods including tokenization, stemming and stopword removal.[1]

In the efforts of Gong et al. LinearSVC (Linear support vector classification) was found as the best method with the least estimation error performance. Results also showed from ElSahar and El Beltagy's work that using classifiers as Linear Regression, K-Nearest Neighbors and Stochastic Gradient Descent have performed worse than SVMs and NB classifiers, with the worst classifier to be the K Nearest Neighbors[2]. ML methods have outperformed the lexicon based methods when estimating sentiment especially when training on an event-based dataset being more expressive, but might also influence the data with the construction bias of chosen events. Moreover, sentiment estimation error was lower when Neutral Polarity was removed allowing only positive and negative values to be set in the nature of crowds reacting to events expressively.[4]

Hybrid method

Hybrid method is a combination of both the lexicon and the machine learning based methods that requires more future work in development. It is inspired by the qualities and the performance of both the lexicon and machine learning based have[4]. Read and Carroll have added a predefined lexicon to their machine learning supervised approach, and with implementing similarity methods, they were able to increase the lexicon size and outperform an unsupervised lexicon.[11]

2.2.3 Sentimental Analysis Tools

The Hedonometer

The Hedonometer, is an analysis tool introduced by Dodds et al.[12] for microblogging services like Twitter to analyze happiness, with a dictionary of sentiment for over 10,000

word that analyzes each word in the sentence and gives it a happiness score from 1 to 9 from least to most happy with scores 4 to 6 being neutral and calculates average happiness score of each sentences to determine whether this sentence is has positive or negative sentiment. The Hedonometer was used by Frank et al.[13] to classify sentiment of 180,000 people with respect to their movements. Unfortunately, the Hedonometer is limited in word context and word order to strengthen its reliability.[3]

To address this problem Schwarz et al.[3] have calculated sentiment as the weighted average score to each word relative to its frequency and subtract the differences between scores to show the sentimental effect between before and after park exposure.

Language Assessment by Mechanical Turk (LabMT)

Another tool used to analyze sentiment was the Language Assessment by Mechanical Turk (LabMT) word list by Mitchell et al. which has a similar scale to the Hedonometer but Amazon's Mechanical Turk's users scored 10,000 words of them. Mitchell et al. went further filtering out the words that were of neutral sentiment (from 4 to 6) to not affect the total score. A weakness, however, is that the LabMT tool only supports the English Language and may face the problem too of identifying other languages words as English with a very different meaning.[8]

The tool was also used by Nabil et al. to manually label the corpus using the Boto AP, where they applied four tags: subjective positive, subjective negative, subjective mixed, objective. Tweets labelled with the same rating by minimum of two raters are classified as conflict free, otherwise they are labelled as conflicted and ignored. The four tag approach would yield a lot of objective tweets, along with the redundant of having mixed labelled tweets that would limit the insights followed from the same approach.[6]

SentiStrength and SentiWordNet

Two lexicon based tools were used by Gong et al. SentiStrength and SentiWordNet. SentiStrength[1] was created to estimate the strength of a sentiment on MySpace website on short English texts having either negative score from -5 to -1, a neutral score of 0 or a positive score of 1 to 5. SentiWordNet was similar to the dictionary approach but rather assigns sentimental values to WordNet sets than social media data.[4]

Caragea et al. had used the SentiStrength algorithm to convert the three classification problem into two binary classification problems: Polar vs neutral Positive vs Negative. The neutral class will take into consideration the texts given the score of -1 and +1 as part of it, and for the other scores lower and higher in range respectively will be classified as polar. SVM and NB classifiers will then be used to classify from the polarized texts the positive and the negative, where a text is positive if its positive sentiment score is greater than its negative sentiment score and vice versa.[10]

[1] http://sentistrength.wlv.ac.uk/

2.2.4 Building Sentiment Lexicons

ElSahar and El Beltagy introduced a semi-supervised method for building lexicons from annotated multi-domain datasets through selecting the phrases most significant to the accuracy of their sentiment, and making use of SVM's feature selection abilities. This lexicon's aim was to target only the polarity of a positive and negative sentiment classifications and therefore only the positive and negative annotated data were taken out to use.[2]

After an 80:20 split ratio to the datasets to train the model and further testing it, a bag of words model is built upon unigram and bigram features of the training examples with frequency values assigned to each one of them. Furthermore, they have used cross validation to tune the soft parameter margin to choose the best performing classifier with the highest accuracy and lowest selected features. Finally, they get the model with best parameters, rank and map the non-zero coefficients to their corresponding n-grams features. The highest discriminative features are the ones with value farthest away from zero with their sign determining their sentimental polarity, this is used as the basis to classify the label to the ngrams. Resulting unigrams and bigrams are then reviewed and filtered by Native Arabic graduate students for incorrect and irrelevant terms.[2]

With a 1000 annotated tweet corpus, Shoukry and Rafea have decided on a 60:40 split to use for training and testing respectively. 600 tweets balanced equally between the two polarities are preprocessed to be used to build two sentiment word lists for most frequent positive and negative words respectively, with each word a frequency weight is given. The class of a tweet is determined by the sum of weight scores of all sentiment words present in the tweet. If dealing with a 3 class sentiment classification, the neutral class is either determined through a cumulative score of exact 0, or the absence of all sentiment words in the lists in the tweet, but the authors have claimed that this way is unacceptable of determining the neutral class, and settled for this conditions to be only labelled as "other" with a binary sentimental classification. The authors have also mentioned that building a larger more comprehensive sentiment words list shall be considered as an improvement to the performance, given that their corpus is relatively small.[5]

Removal of stopwords from sentiment words' list shouldn't alter the accuracy of the classification, rather it enhances the speed of the performance to get the results. In lexicon based approaches, preprocessing is the only form of improvement to the performance for the sentiment words, and thus the outcome of the accuracy has a great higher significance after preprocessing both the sentiment words and the corpus to the same form. While stemming is a great improvement, it still can't capture all the inconsistency in the text which affects the dictionary based approach more. Given that a word is in the dictionary with a sentiment value, the same word of a different form (extra prefix or suffix) even after preprocessing wouldn't be able to map to its value in the dictionary and would be declared as neutral or void.[5]

An automatic method introduced by Abdulla et al. begins with having 300 seed words from the SentiStrength dictionary, that are additionally translated to Arabic using an English-Arabic dictionary. Using several extensions, synonyms of each word is added to the dictionary and given the same polarity, as well as giving emoticons a sentiment and adding them, jumping the dictionary length from 300 to 3479 annotated words. Any missing word of such dictionary was declared as neutral or zero score. They also use the switch negation mechanism to change the polarity if any negation words present, as well as the intensification mechanism to add more score to the positive and negative words in both cases, the word followed by or after an intensification word. The lexicon is still limited because it fell short when used against annotated corpus ML method, and also showed that the bigger the lexicon the better the results, but still the accuracy gap became shorter as we get to the larger lexicons and that smart preprocessing is the better way to get better accuracies.[1]

2.2.5 Arabic Sentiment Analysis

With strides and development done in Sentiment Analysis,it is still clear that the studies done are insufficient and in early stages compared to the English language whether document or sentence based, it was needed that studies shall be made developing this field in the Arabic region facing the challenges of the Arabic language. When it comes to resources for Arabic sentimentally annotated datasets and dictionaries, there were insufficient resources due to lack of availability of such resources, being labelled to a certain domain or of small sizes. ElSahar and El Beltagy went on to gather, create and annotate datasets coming from a variety of backgrounds. Datasets were collected from annotated reviews based on ratings from movies and restaurants to hotels and products. They also added a dictionary built from the same datasets, as well as the code used for the purpose of publicly releasing them for further use.[2]

Non-English words and phrases are still in the early stages of developing sentimental analysis with the accuracy of the English language at the moment. The Arabic language is of importance due its representation of the large scale audience of the Middle East region and being a top 10 most used languages on the Internet being in the 4th place in the languages' world ranking,[2] has lacked sufficient research in the region. Difficulties are facing the Arabic language from reaching a better sentiment analysis due to the complexity of the language's structure. Machine learning based approach is the one used so far in Arabic sentiment analysis due to the lack of an Arabic semantic dictionary.[7]

For sentiment of Arabic words, an approach is to get the Arabic word roots after classifying and create and store their sentiment into a dictionary, which will be later checked if the word root exists for reference to extract its polarity, or the addition of the unknown words to the dictionary by the user to learn. For a more ML based approach, the SVM

[2]http://www.internetworldstats.com/stats7.html

method has been suited for the Arabic language edging over the NB method for having the better use of the syntactic and stylistic features to cover over the Arabic language.[7]

2.3 Applications on spatio-temporal data

This section is about how the data obtained can be sorted in geographical areas with common interests that is seen other than the fact that they are grouped within regional proximity and how to make sense in doing so. Several efforts and approaches have been made into classifying the data geographically to get insights out of analyzing them. By 2050, it has been estimated to have 66% of the global population to be living in urban areas.[14]

The usage of physical sources of information such as questionnaires and surveys have been good references in urban related studies, however they fall short in determining the land cover due to the difficulty in extracting land utilization from physical infrastructure.[15]. With the rise in the digital era and microblogging services within urban areas, several efforts were made using location enabled tweets to help draw a pattern of behavior in response with geographical bounds, and how people interact with urban spaces. However, with the exponential growth in the big data and the lack of structuring the spatial information has caused it a challenge that researchers have been active about.[15]

2.3.1 Characterizing spatio-temporal patterns

Three mechanisms have been identified by Frank et al. on how the Twitter user can report his/her location when tweeting. First of all when the user signs up for Twitter, they are given the option to give their location info that will show up in their profile. Second, the user while posting the tweet has the option to tag their tweets with a place (first option is the city where the device's IP address is found). Last but not least is the tweets located using GPS, where the user chooses to add the geo-location with the latitude and longitude with accuracy of the device's GPS and the tweet could be placed down to a 10 meter radius circle. Using the third mechanism, Frank et al. have found about only 1% of the tweets collected are geolocated.[13]

Soliman et al. were able to find temporal patterns at key locations of most users actively tweeting at that time to be reasonable and significantly related to the urban types of these locations. They have identified four critical times where the tweets activity exhibited significant change: 7am, 12pm, 3pm, 8pm. The 7-3 pm period was identified for schools, whereas for the activity that remains till 6pm has been identified as people in workplaces. Moreover, shoppings were identified as the activity for tweets from 6-8 and residentials peak activity till midnight. They are also taking a look into geotagged tweets of the city of Chicago for its frequently updated datasets to characterize the types of urban land use. The data has been collected using Twitter API using a bounding

box of coordinates to eliminate all tweets outside Chicago. They also have expressed the unreliability of geo-located info on social media being inconsistent and require further research.In their study they introduced four assumptions to classify the Twitter user patterns: Random walker scenario, preferential return scenario, semantic coherence scenario and temporal coherence scenario.[15]

A random walker is more like tourist behaviour, tweeting only from new places, moving randomly across the city, causing a set of locations randomly distributed across the city. The preferential return is based on the fact that people will spend 90% of their time in key locations to expect that number of tweets is to increase with time spent in these key locations, where the results showed that people follow a preferential return pattern in few key locations rather than the random walker pattern. Semantic coherence is an assumption that the most frequent places for any person is to be home and work, so it will assign the top location tweeted from is the user's home and later show that there is absence of semantic coherence in social media users. Temporal coherence relates each key location to a period of time during the day based on the period with most tweets, and how the aggregation of tweets over a long period would show the time of the day related to such key location. If timing of tweets is dependent on land use type, classification algorithms could deduce types of land use of key locations based on the timing at each location.[15]

2.3.2 Urban analysis

The data could show in time periods when analyzing the differences of the results and the anomalies and spikes let us know about what that time had occurred and how the people behave around it. The data could also show difference in results prior, during or post a certain activity or event by defining a baseline data prior the event for comparison[3]. While the region of data could go to compare in hourly basis per day to monitor the impact of a short term event[3] or a yearly basis to monitor the impact of large usually catastrophic events to receive insights about each time period.[16] Temporal events can also show the impacts and how it varies through temporal distances from such events and the lifetime of the event's impact.

In the study of Ferriera et al.[16], they proposed a model to deal with millions of data which was limited by commonly used tools, where they focused on New York City as their region. These data contain spatial components to query it through visual manipulated queries on a map, where the user requires no background around the science of it .This is important because the user can easily navigate through the model's map, specifying the origin-destination points desired or by brushing through the area needed to query upon allowing the results to come in different forms visualized on the map. Also they have added widgets in their model. With a regular and a recurrent widget, the user can specify temporal constraints in data selection whether through an atomic or an easily made complex constraints.

Schwartz et al.[3] had there approach tackling green spaces in the San Francisco area identified through the NDVI proxy used for quantifying the green intensity in the area. In addition to Mitchell et al.[8] focusing on the states of the US and urban areas that have been identified by the United States' Census Bureau's MAF/TIGER database. These studies expressed the limitation of acquiring exact locations as it depends on the accuracy of the mobile GPS and bounded by the US offices and borders.

Gong et al. focused on urban data that involved large crowds that being common based on daily lives or event based spreading through the city to help crowd managers have a clearer insight on crowd behavior in both the cities of Amsterdam and Rotterdam for their population variety and sufficient quantity.[4]

Frank et al. have collected 37 million geolocated tweets from 180,000 people to be able to describe a person's movement patterns for its importance in urban planning models due to the better accuracy of locations of tweets, rather than mobile phones and antenna distance methods that were restricted to cell towers locations. The tweets location is quantified using the radius of gyration (or gyradius) and visualized using Matlab. In their findings they have shown that individuals that move farther away from their expected locations express more happiness with more positive words and that a depressive emotional state can be detected by the decline in movement and social interactions.[13]

A field that acquired the attention of data science researchers to drawing out spatio-temporal patterns in presidential elections. A study was made to infer sentiment of the population among both the republican and the democratic parties in the USA towards the 2016 elections using geolocated tweets and creating the Compass framework to analyze and visualize the data, but the results only showed that the tweets can only show popularity and time periods, but not a definitive winner to the elections considering different last minute factors that had changed the 2016 elections vote.[17] Another study analyzed the presidential election in Brazil in 2014 where the authors have shown good results in having a model that predicts how the vote will end up along with sentiment analysis to the tweets involved around the period of presidential campaigns.[18]

2.3.3 Disease and disaster events data analysis

Healthcare information has had a triumph since the digitalization evolution from both healthcare industries to people sharing their experiences with a certain medical condition. With restrictions applied to the access of healthcare databases, social media is now offering a generous amount of information about people's interactions with diseases publicly. Paul et al. have adopted the SNOMED-CT terminology being the most inclusive in the medical field globally to search for tweets and geographically bound them to see most relevant diseases with respect to a certain region. In order to study the lifetime of a disease, they also surveillanced rise and fall period of a disease by monitoring the tweets

mentioning the disease over such period for better understanding of such an event as well as the most relevant diseases at a certain time period.[19]

There have been very little studies made into classifying a user's sentiment during disaster events, that could help understand the user's concerns and panics, where people tend to turn to social media to describe their experiences to others affected and even for responders as a cry for help for their needs. Caragea et al. have mentioned the usage of micro-blogging during disasters to help bystanders reach responders and made their scope on the Hurricane Sandy event to show how people's sentiment can vary on how close in distance to the hurricane impacts, how far can the impact go, and responders to assess how affected the emotional status of the population. To represent the hurricane spatially, the National Oceanic and Atmospheric Association (NOAA) wind speed approximation was used, and as the strength of the hurricane decreased with time, the search diameter also decreased. Their study has shown Twitter users seem to be most active about the disaster during and at the impact's location, and the difficulty connecting a location to a certain event, because not all people who mention that certain event, are actually affected by it. The spatial mapping of sentiment to the hurricane was visualized using ArcGIS[3].[10]

In Brazil, a study was made to monitor the spread and behaviour of an epidemic disease called Dengue. The authors argue that surveys that collect the spatial and temporal information for the public health service, but the process is so slow for the info to be made public and too late to take countermeasures, thus they have used Twitter continuous data to represent the spatial and temporal spread and surveillance of the disease. The surveillance how people on social media react to epidemics according to four factors: Volume, location, time and public perception. Volume is the tweet count with the word "Dengue" in them, is predicted using regression, while location and time of 'Dengue' tweets are predicted using clustering and public perception is analyzed using sentiment analysis, in addition to focusing on tweets for users who had personal experience with the disease. Furthermore, the authors used Google Geocoding API to filter out users that have location info in their profiles by converting the text data in coordinates of latitude and longitude, however many users lacked their location info in their profiles leading a lot of tweets discarded.[20]

[3]http://www.esri.com/software/arcgis

Chapter 3

Sentiment and COVID-19 Annotation on Spatio-Temporal Data

Studies in the US [8, 17] that took their states into account used the United States' Census Bureau's MAF/TIGER database for states geographical and urban information. Since our scope is domestically as well in the Egyptian scope, we wanted to create a similar database for urban and geographical information needed on the Egyptian governorates, but unfortunately there didn't exist a single updated database that involved all needed features.

3.1 Creating the governorates database

To create the database we needed, we collected three datasets from three different sources and combined them to their respective values. The first dataset was obtained from the City Population website[1] for statistics on major cities in countries, containing information about the area and population of the governorate. The second dataset was obtained from the Egyptian Central Agency for Public Mobilization and Statistics[2] that contained information about the urbanity of each governorate. Finally, the third dataset was obtained from SimpleMaps website[3] for interactive maps that featured the geographical coordinates for each dataset.

After combining each dataset, we are having a governorates database that has all the features needed. The features in this dataset included:

- **Name**: The English formal name for the governorate.

[1]https://www.citypopulation.de/en/egypt/cities/
[2]http://www.msrintranet.capmas.gov.eg/
[3]https://simplemaps.com/static/data/country-cities/eg

- **Native**: The Arabic formal name for the governorate.
- **Area**: The Area in km squared for the governorate.
- **Capital**: The capital city of the governorate.
- **Population**: The population residing in the governorate.
- **% Urban**: The percentage of urban land of the governorate.
- **Urban Population**: The calculated percentage of the population exposed to urban land.
- **Latitude**: The latitude coordinate of the governorate.
- **Longitude**: The longitude coordinate of the governorate.

This database, however, shows weakness in reliability given the fact that it is not from one direct official source of the government, also some of the data collected such as the population and the %Urban are five years outdated as the most updated available data. The governorate database will be a huge fundamental later in this paper to be used in the data collection, and also show relations and insights for the performance of each governorate given a certain comparison.

3.2 Data Collection

The beauty of Twitter data is being expressive in text form on a continuous stream timeline with a huge number and variety of data to choose from. That being said, it causes a challenge to find the proper constraints and methods that would yield with the best kind of data acquired for a specific domain, given the nature of noise in all data around us.

Twitter offers its own API[4] for querying around its tweets, unfortunately, it limits its user to have a limited amount of requests up to 15 requests per window, as well as having a 15 minute interval between each window, allowing those features to only be of a premium use. For that matter, we have decided to alternate the method used for tweets collection to a Twitter web scraping tool called TWINT (Twitter Intelligence Tool)[5] which doesn't use the Twitter API with the same effectiveness avoiding the API's limitations.

We set the spatio-temporal constraints in the TWINT tool by setting the duration of collection from 08-03-2020 till 22-03-2020 spanning a 14 day representation in the Egyptian governorates. Also, we have collected tweets both in English and Arabic language separately defining the language constraints in different windows of collection.

[4]https://developer.twitter.com/en/docs
[5]https://github.com/twintproject/twint

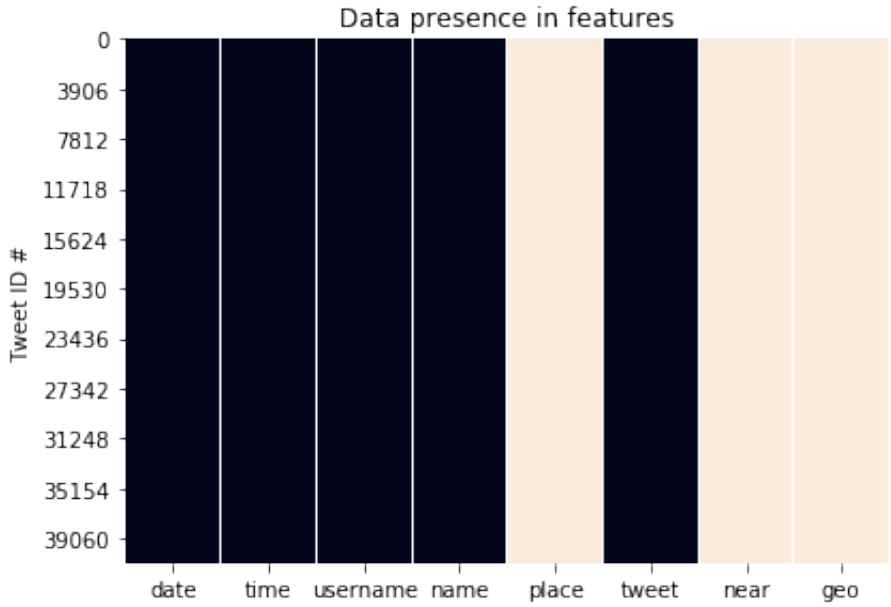

Figure 3.1: The presence of each feature for each tweet.

A sample of 40,000 tweets that have been collected using a traditional search query about Egypt, and we have extracted the needed features that have returned back to have the form in.

As observed in figure 3.1, the missing data for the tweet all data related to the spatial information of the tweet, which can be justified due to the fact that users mostly disable having their tweets geotagged and prefer not add a location of the tweet and shows how difficult it is determine the origin place of the tweets in Egypt.

In order to overcome this challenge, we propose four different approaches in collecting data to represent each governorate's own tweets: The Geocode Approach, The Nearby Approach, The Keyword Search Approach and The Profile Info Approach.

3.2.1 Geocode Approach

Many studies have taken advantage of dealing with geotagged tweets to encapsulate a certain region, and since in Figure 3.1 we showed that most tweets in Egypt come with no geotag information, we decided to create our own encapsulation geocode for the region. Using the information in our governorates database, we were able to have the center coordinates of each governorate, in addition to the area feature we were capable of adding the radius feature by calculating it using the equation

$$r = \sqrt{\frac{A}{\pi}} \tag{3.1}$$

- **Radius**: The calculated radius of the circular area of the governorate.

And the geocode takes the form of (latitude,longitude,radius(in km)).

This approach already expresses some drawbacks despite dealing with geographical coordinates beginning with the fact that most governorates shape are not in circular form which makes it both leaving out areas uncovered and involving areas that are not within the given governorate, especially border governorates that would go beyond the scope of having non Egyptian tweets. Furthermore, when dealing with circle representation, there would appear intersections in these circles causing ambiguity for the tweet having more than one representative governorates and duplications of the same tweet.

3.2.2 Nearby Approach

In the direction of finding a geographical approach to represent the governorates, Twitter allows its user to search tweets nearby his/her location, so we have decided to try and feed the TWINT tool the names from the governorates database to get nearby tweets for each governorate. In an attempt to avoid the relocation of the nearby approach to somewhere else than the intended governorate by adding ", Egypt" to each governorate name.

3.2.3 Keyword Search Approach

This approach is quite simple and a straightforward way of how a Twitter user usually surfs for tweets regarding any topic. Once again we make use of both our Name and Native features of the governorate database for both the English and Arabic search queries respectively.

Moreover, we have noted that some of the Arabic names of the governorates come off as other meanings to commonly used words. Both Sharqia and Gharbia governorates have their Arabic names to also mean Eastern and Western respectively and would go off topic easily, but fortunately we had names of the capital cities for both governorates, so Az-Zagazig and Tanta were used as alternatives in the Arabic search query. Another governorate name that have shown controversy is the Qena governorate, due to its Arabic equivalent name being frequently present in the Holy Quran and would result in many tweets from the Quran, and sadly the capital city to the Qena governorate is Qena itself, so we compensated by searching for the whole term 'Qena governorate' to specify tweets relating to the governorate itself.

3.2.4 Profile Info Approach

In this approach, we decided not to rely on the limitations on the features of the tweets and rather shift our focus to the user him/herself where we can observe for two important

factors that determines the location of the user and thus assume his/her governorate from there and categorize their tweets accordingly.

The two factors are looking at the profile's hometown info and the profile's own bio. Twitter users when signing up, are given a chance to put in their hometown locations to be known where this user is actually from, but not necessarily tweeting from. For the sake of labelling our data we assume that their hometown is updated and tweeting from it. Some other users prefer not to share information about their hometown location to protect their privacy, but could add info in their own bio about their hometown such as hometown, school, university or work. We look into these two factors by checking whether the hometown location or the bio contains any words in our governorates' database's English and Arabic names, and assign each user a governorate where he/she is from. If it occurs that both factors have found a governorate in them, we prioritize the hometown location factor to be dominant, even in the case of different values.

To implement this approach we implemented the following steps that would guarantee us the most effective result

1. Obtained 11 of the most famous Egyptian accounts of various genres on Twitter identified by the SocialBakers[6] website on tracking social media info.
2. Scrapped 40,000 usernames following each account, having 440,000 users in total.
3. Eliminated all duplicate usernames, due to the fact that a user would be highly likely following more than one of these Egyptian famous accounts.
4. Began collecting the profile info for all the unique users.
5. Removed all private accounts, due to the inaccessibility to their tweets.
6. Removed all accounts with zero tweets.
7. Removed all accounts that joined Twitter after the end of the specified period.
8. Removed all accounts that had neither hometown information nor bio on their profiles.
9. Ran the search on governorates in their locations and bios.
10. Removed all accounts that had no Egyptian governorates identified.

Afterwards, we had obtained user accounts that we were able to label and verify their governorate and began collecting off English and Arabic tweets that were from the specified period and labelling where it came from according to its user.

Following each approach we add new features to the scrapped tweets along with the features in figure 3.1 and labelling them accordingly:

[6]https://www.socialbakers.com/statistics/twitter/profiles/egypt

- **method**: The method of approach used to collect this tweet
- **language**: The language of the tweet
- **governorate**: The governorate the tweet came from.

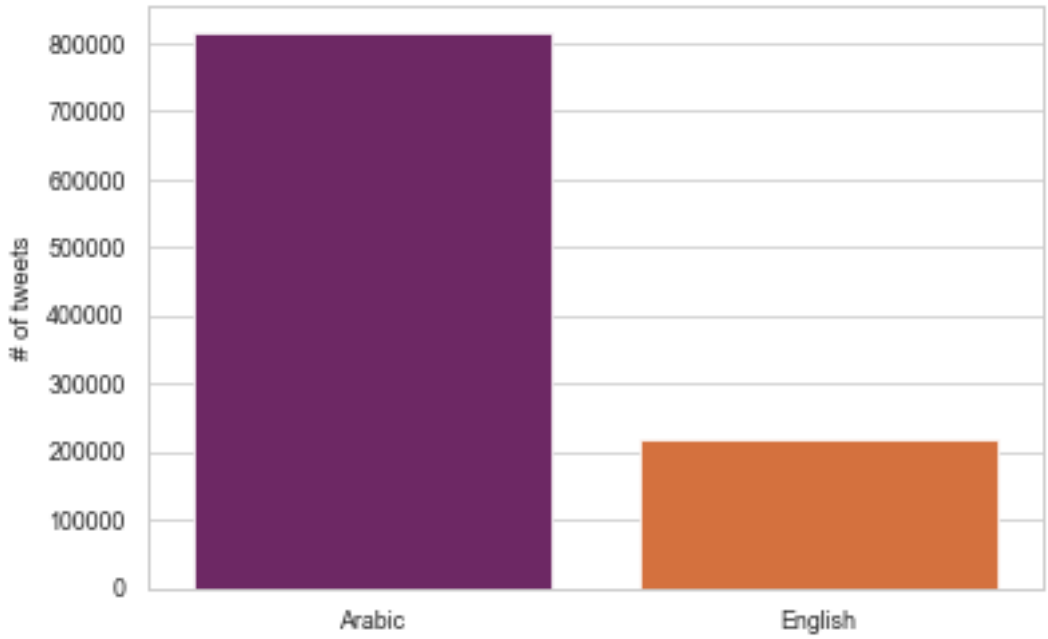

Figure 3.2: The number of tweets for each language

We had the fear that upon collection we wouldn't find a lot of Arabic tweets and so we began collection English tweets as well, but upon inspecting the number of tweets collected in figure 3.2 we were able to have much more Arabic tweets, which shows how much the influence of the Arabic language has become in Egypt in using in the microblogging services which lead us to consider only the Arabic tweets from this point.

3.3 Data Preprocessing

Social media as we know it contains numerous noisy features that come along the way, with Twitter being mostly known for text form of expression, the collected tweet corpus needed go through various stages of text preprocessing inorder to focus on the necessary and insightful content of these tweets. We have collected all Arabic tweets and binded them into one file, since the process is bound to be the same regardless of the collection method and prepare them to go through text preprocessing.

Most of the preprocessing stages are quite common with each language, but with some differences to take into account, we also need to take into account somethings like the

domain or dialect, nevertheless our tweets are going through the most common text preprocessing stages: Normalization, Stemming, Stopwords Removal.

3.3.1 Normalization

To begin with, our tweets have plenty of needless parts of its text that we have identified of no use to our study and so these parts had to be removed using various tools and algorithms that made sure that only the important parts of the text remains and as much of the redundant data identified is removed. Our text normalization considered the following to preprocess:

- Multiple line tweets are combined into a single line.
- Removal of any urls.
- Removal of any numbers.
- Limiting the number of consecutive repeated characters to only two.
- Breaking up hashtags and underscores into normal spaces.
- Removal of punctuations.
- Removal of emoticons.

The following steps are done especially to normalize the form of the Arabic language texts as they vary in forms and show of characters

- Removal of any English characters.
- Removal of Arabic punctuations, as they differ from the English ones.
- Removal of Arabic tanween. (’ مرحبًا ’)
- Normalizing the all alef characters into the form (’ ا ’)
- Normalizing the all yeh characters into the form (’ ي ’)
- Normalizing the all last heh characters into the form (’ ة ’)

3.3.2 Stemming

Following the normalization of tweets, we chose a light stemmer when it comes to choosing the types of stemmers rather than an aggressive stemmer to the roots of words. We decided to go with Tashaphyne, an Arabic light stemmer that was used by Shoukry and Rafea[5] to lightly stem their tweets. In addition, the Tashaphyne[7] stemmer offers the option to make a custom prefix and suffix list for stemming to reference, which would help in terms of trying out lists specifically for the Egyptian dialect.

[7]https://pypi.org/project/Tashaphyne/

3.3.3 Stopwords Removal

Stopwords are usually removed from texts given their likelihood of appearing to relevant documents as well as non relevant documents to the query in question.[21] To obtain Arabic stopwords, we relied mostly on the NLTK[8] (Natural Language ToolKit) corpus on Arabic stopwords. We also added in the governorates names as identified named entities that would most likely appear given our domain. Finally, we observed the most frequent words after stop removals, especially those of Egyptian dialect after each removal and manually added them to the stopwords list. It is important that the stopwords list would run through the same normalization process in order the change in characters to be done as well, so both the stopwords and the clean text are of the same form.

3.4 Sentiment Analysis

Now that we have a clean Arabic text, we can begin the application of labeling each tweet to its corresponding sentiment. We have chosen to go with a two class sentiment classification of positive (1) and negative (0) after multiple studies have found its performance than a 3 class classification as well as having a neutral feature would not provide much insight to tracking trends and patterns.

3.4.1 Feature Extraction

In order to prepare our tweets for sentiment classification, we have to extract features from our tweets that would make the process easier on the classifier to retrieve information from our text. We performed text features extraction using SciKit-Learn's[9] CountVectorizer to fit all the words in our tweets in a bag of words. We chose to experiment on having unigrams and bigrams and trigrams and choosing the most optimal for accuracy and performance. Afterwards, we used the Tf-Idf transformer to use the collected n-gram bag of words and transform it into TF-IDF(Term Frequency-Inverse Document Frequency) matrix form.

3.4.2 Sentiment Annotation

For sentiment annotation, we chose to go with the ML-based method regarding its prevalence in performance over the Lexicon-based along with having to extract multiple features and much work as well as having gathering an Arabic lexicon for reference.

[8]https://www.nltk.org/
[9]https://scikit-learn.org/stable/

Gathering the annotated dataset

First, we have to get ourselves an Arabic sentimentally annotated dataset for training. We used the ASTD dataset made by Nabil et al.[6] for training not only because it is collected from the same source of our data collection, but also being collected from Egypt which would make into account the dialect differences. The dataset contains 10,000 tweets annotated in a 4 class classification: positive, negative, neutral and objective. For the sake of our research we decided to discard all neutral and objective tweets, leaving us with around 2,500 unbalanced tweets between positive and negative classifications.

Model training and annotation

With our annotated dataset, we make sure it goes through the same stages of text preprocessing and feature extraction as our actual corpus and we should have a TF-IDF matrix of our annotated dataset preprocessed. To choose our classifiers, we decided on the two best performing classifiers for text features in multiple works[13, 6, 7], the Naive-Bayes (NB) classifier and the Support Vector Machines (SVM) classifier and in order to implement them we used SciKit-Learn's MultinomialNB and LinearSVC respectively to fit and train the model to our dataset. Once the model is trained and tested with the optimal accuracy and performances, we used the same settings to fit our corpus and used the model to predict the sentiment of each tweet. We made sure the sentiment feature for each tweet is normalized between 0 and 1 rather than the original 'pos' and 'neg' for later aggregation functions as averaging and summing.

- **sentiment**: The sentiment polarity of the tweet in 0 and 1

3.5 COVID-19 Tweet Classification

Our data collection took into consideration the spatio-temporal constraints without adding any filters to determine a specific domain, however there is a great influence in those tweets relating to the COVID-19 virus given the choice of time period to involve such topic to analyze. Unfortunately, there weren't any COVID-19 classified datasets for ML-based classification using ML classifiers. On the contrary, we went for a more lexicon based approach and developed a dictionary on the most important terms regarding the disease that have been collected initially from the WHO (World Health Organization) website[10] and translated to Arabic using Google Translate[11] as well as manually adding words by observing the most frequent words in the initial COVID-19 classification. We also kept the English terms in the dictionary and ran our search on the raw versions of the data, for the sake of catching trending hashtags and official scientific terms for the virus that were expressed in English. We added for the tweet a feature resembling the check for the tweet whether it related to the virus or not.

[10] https://www.who.int/emergencies/diseases/novel-coronavirus-2019/events-as-they-happen
[11] https://translate.google.com/

- **corona**: A check whether the tweet is Coronavirus related or not

Chapter 4

Results and Discussion

This section analyzes the performances of the methods used in the data collection as well as in the sentiment annotation. Furthermore, we use our annotated corpus song with the governorates database we made for the purpose of showing off relations and patterns regarding both the urban analysis and the COVID-19 case study.

4.1 Data Collection Methods

For the four approaches we proposed in section 3.2, we take a closer look on the performances and limitations of each approach.

4.1.1 Geocode Approach

We mentioned the limitation of circular areas going out of bounds and not encapsulating as much of the governorate area. In figure 4.1 we showed the top 100 users that are represented in the most governorates, where it confirms the circular areas intersecting mentioned in section 3.2.1. The top user is represented by 14 governorates which would mean that at a single point, there are 14 governorate circular areas intersecting.

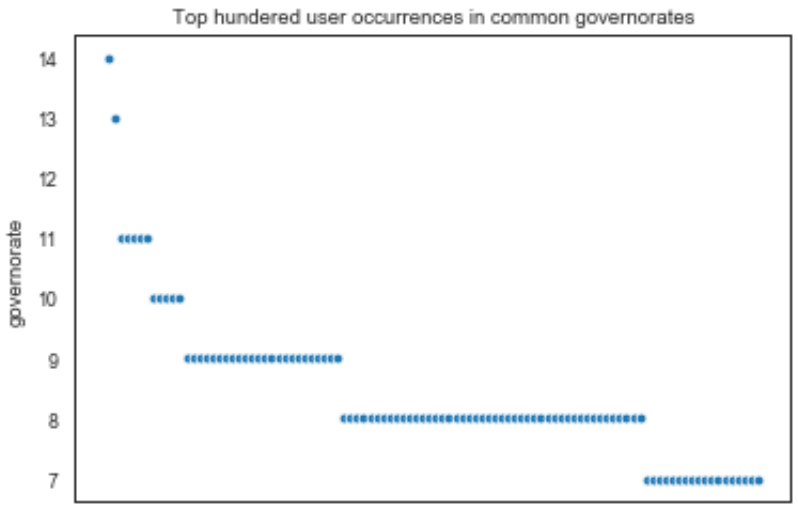

Figure 4.1: No. of governorates represented in the top 100 users

This deduction further made us ask how much of the users couldn't be identified as being part of a single governorate and in figure 4.2 it shows that the majority of the users are identified in more than one governorate. These graphs illustrate the weakness of the geocodes in determining the location of a tweet and would result in multiple representations of the same tweet in different regions.

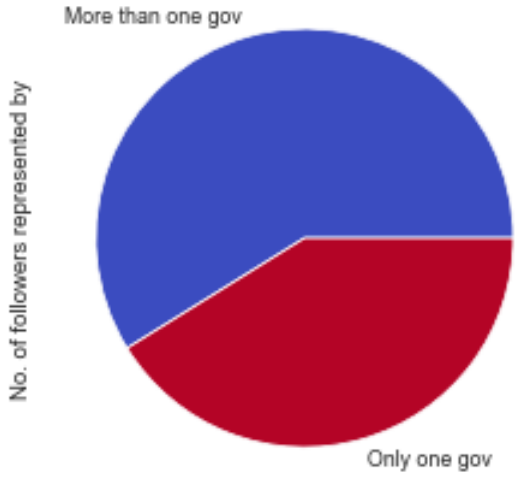

Figure 4.2: Pie chart of users by the no. of governorates represented

4.1.2 Nearby Approach

Another concern in the geographical approaches is going out of scope of the Egyptian domain, we attempted to add the ", Egypt" to each governorate to add more context to the location name, to prevent it from redirecting somewhere else. However, upon data collection we would find some problems collecting tweets for some of the governorates using the Nearby attribute, in which it would relocate somewhere else and open an endless stream of tweets that both didn't relate to the domain, nor stopped collection. We have identified which governorates in which languages that didn't complete their data collection process and as figure 4.3 shows almost half of the governorates didn't collect their tweets completely. For this matter we have decided to discard tweets collected from the Nearby approach completely from further analysis for the inconsistency in data and incomplete representations.

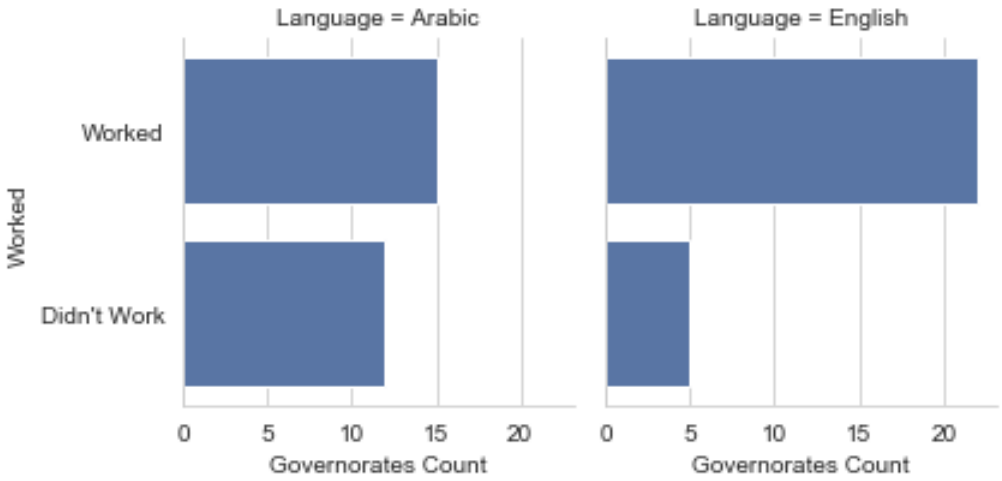

Figure 4.3: No. of governorates that complete and didn't complete data collection with respect to language

4.1.3 Keyword Search Approach

It is convenient for a twitter user that wants to know about any governorate is to search for its keyword, nevertheless the search query could redirect you to irrelevant tweets having the same names as the governorates with different meaning, especially in Arabic. We were able to highlight the cases that would really impact our collection in section 3.2.3 for the governorates "Sharqia", "Gharbia" and "Qena", despite that some governorates might be a little impacted given their double meanings. In the table below, we show the governorates whose names have a conflict in double meanings. The starred governorates are the cases we changed in our approach that we observed to be of significant influence on the domain of tweets

Native governorate name	Other meaning
القاهرة	The Conqueress
البحيرة	The Lake
البحر الأحمر	The Red Sea (The actual sea)
الغربية*	Western
الشرقية*	Eastern
قنا*	Protect us

Table 4.1: Governorates with double meanings

When we took a further look into the data collected by the keyword search, we discovered that most of the tweets are issued by news agencies rather than users. Upon looking at the most users that mention all about the governorates are the news accounts, we also noted that we could tell which news agencies makes the most coverage across the governorates.

Username	Mentioning governorates count
alahram	27
alahramgate	27
vetogate	26
aja_egypt	25
eldostoregypt	25

Table 4.2: Top users with mentioned governorates count in their tweets

4.1.4 Profile Info Approach

In our proposal to look into the user's profile info rather than the tweets, we began gathering about 440,000 usernames that were then filtered out using the methods discussed in section 3.2.4 until we were able to identify those users whom we could label out their governorate. In figure 4.4, we can observe the decay in the number of users left after each process of filtration till we reach about 4,500 users which would make about 1% of the initially collected usernames. This would show the lack of efficiency in the approach in reaching its goal with the little percentage of users that we were able to identify.

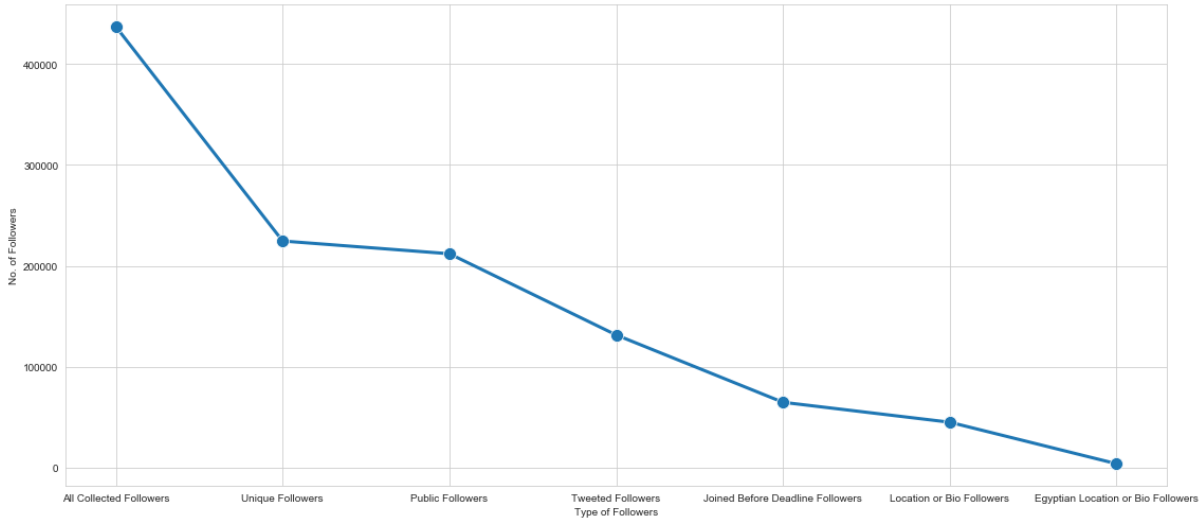

Figure 4.4: No. of users remaining after each filtration process

In the process of identifying the governorate using the profile info, we took matters into two key factors which are the hometown location and the profile's bio. In the figure below what factor is most likely to hold the information we need being the hometown location, nonetheless the significance of the bio approach is still significant and would add more users.

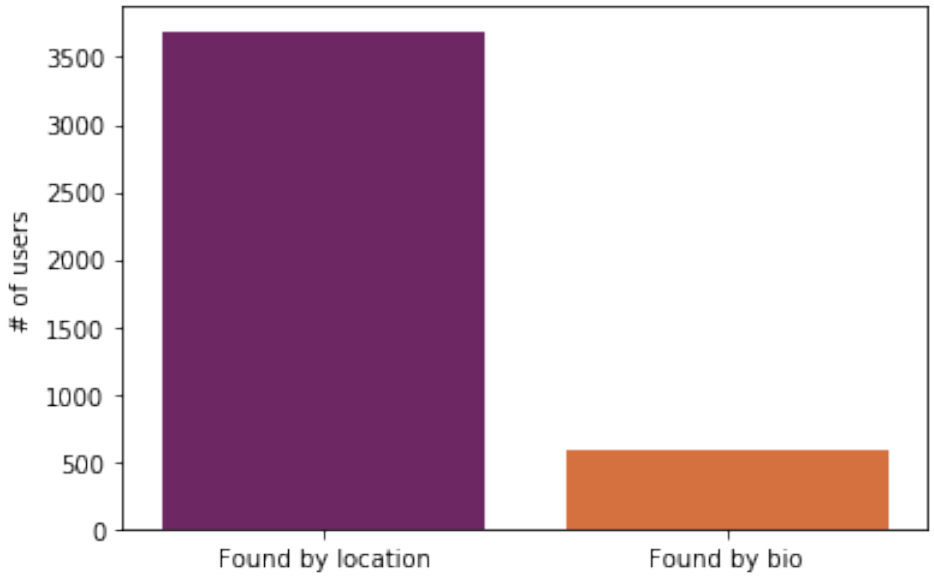

Figure 4.5: No. of users located using each factor in the profile info

After all the filtration and labelling were down, we've come to observe a limitation in the approach as shown in figure 4.6 where there were no representations for two governorates

Red Sea and New Valley in our users.

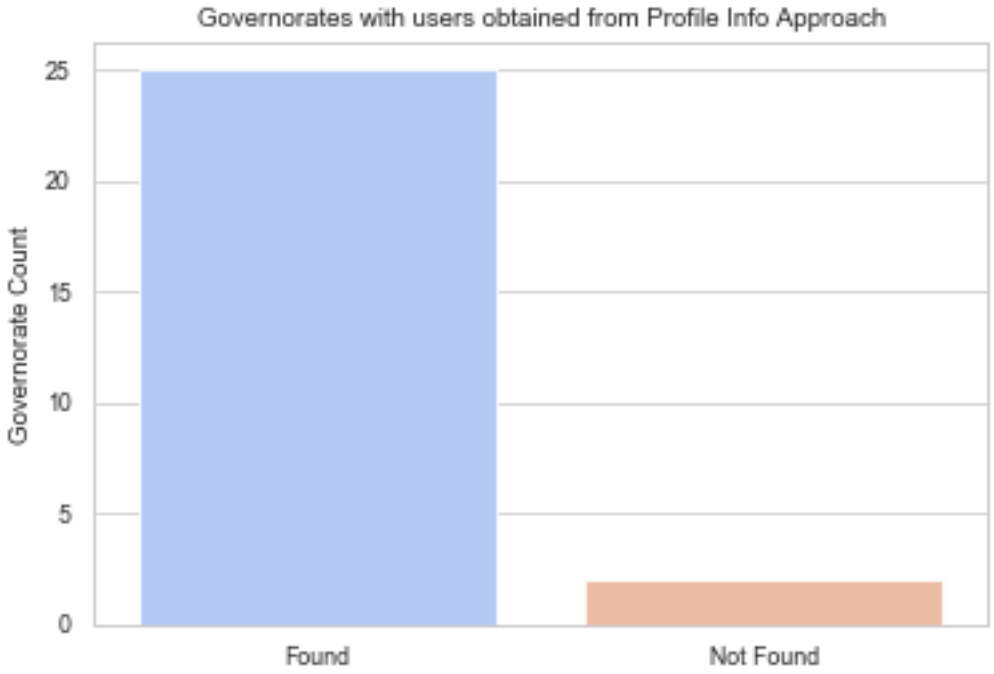

Figure 4.6: No. of governorates found and not found in the users using Profile Info approach

4.1.5 Comparison for all approaches

After being done with the data collection from each approach, we have collected a corpus with a total around 1 million tweets in both languages. Here we compare the performance of all approaches together regarding the data collection.

In the figure below, it reveals the performance of each approach regarding the number of tweets collected with regard to the language. It can be deduced that there is a pattern of having more Arabic tweets than English, however the errors and inconsistency in the Nearby Approach are further revealed here when observed being out of scope and getting more English tweets, which further influenced us to discard all the approach's tweets in future analysis.

Moreover, it can be seen the huge resemblence in the number of tweets gathered by the geocode approach, this could have been the influence of the limitation of representing a majority of the users in more than one governorate, and thus might contain alot of duplicated tweets that couldn't be labelled into a single governorate.

On the contrary, in figure 4.8 where we count the number of unique active users per

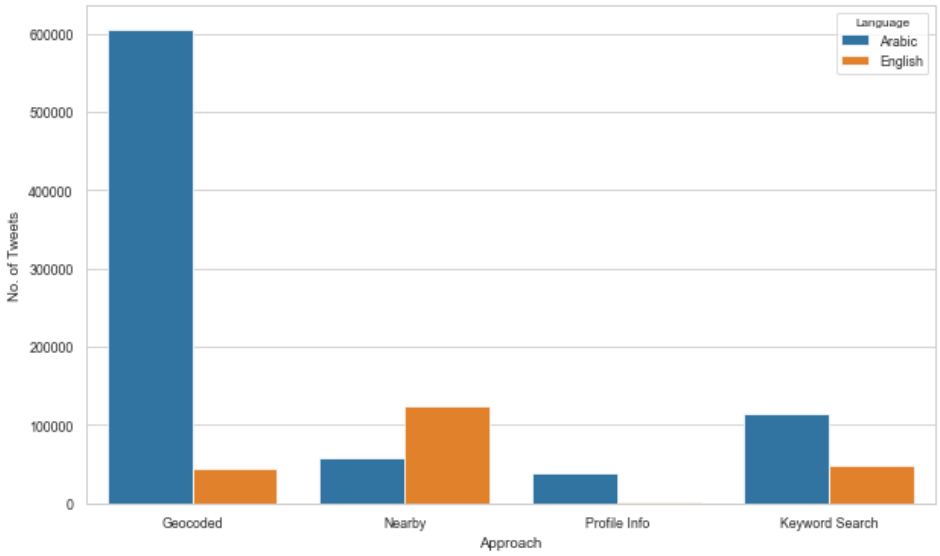

Figure 4.7: No. of tweets per language for each approach

approach, despite the Geocode approach having almost five times more tweets than the Keyword Search approach, the Keyword Search had shown to have more contributing users. This deduction made us take into consideration the weakness of the Geocode approach and its influence granted the huge number of its tweets in the corpus.

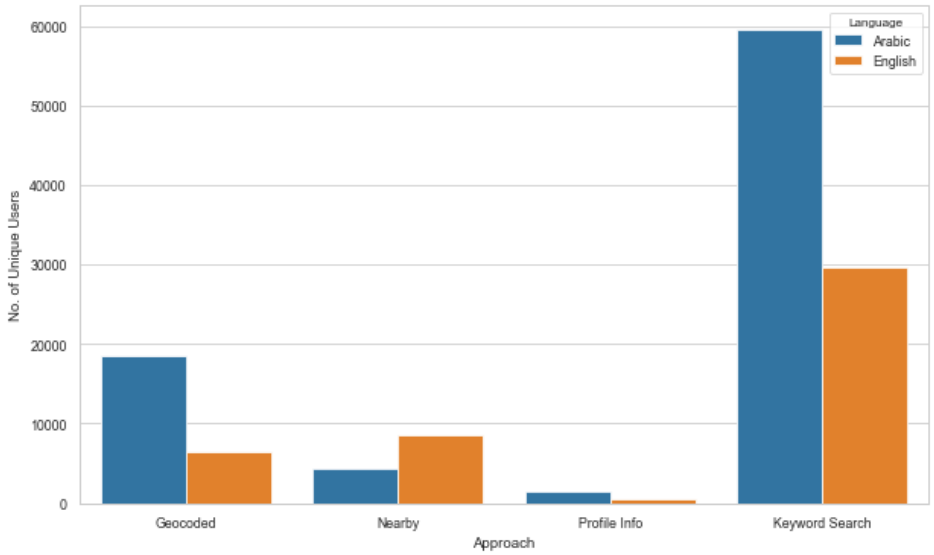

Figure 4.8: No. of tweets per language for each approach

In both graphs, the Profile Info approach have shown the correct patterns, however on a very small scale outperformed by both the Geocode and Keyword Search approaches. This is due to the small number of users collected with respect to the number of users in

the other two approaches and would need a much larger number of profiles to behave the same way, given that both approaches have completed their data collection to the max.

4.2 Preprocessing and Sentiment Annotation

For the purpose of finding the best settings for applying our sentiment classification to our corpus. We decided to experiment on a couple of variables, thereby trying to yield the best performance and accuracy. Using the ASTD[6] dataset we obtained for training, we decided on a 80:20 split to include as many examples for training without going overfitting in addition to the fact that the 80:20 split mostly produces the best result. The variables we take inconsideration are the preprocessing form of the tweet, the n-grams features extracted for TF-IDF matrix and lastly the classifier used for classification. The options for each variable are:

Variable	Option 1	Option 2	Option 3
Preprocessing Form	Raw (not processed)	Normalized	Stemmed
N-Gram Feautres	Unigrams	Unigrams + Bigrams	Unigrams + Bigrams + Trigrams
Classifier	Support Vector Machines (SVM)	Naive Bayes (NB)	-

Table 4.3: Experiment variables for testing sentiment classification accuracy

We have tested each combination and in the figure below, are the percentages of accuracy for each combination in testing the model's sentiment classification on unseen annotated data.

In terms of preprocessing, the findings suggest the influence of the normalization stages and further the stemming stages on the accuracy of the classification. Regarding the feature extractions, it is observed that there is a significance in performance when adding bigrams to the unigrams to the extraction, on the other hand adding trigrams didn't do much of an effect rather would make more computation time.

The NB classifier mostly has outperformed the SVM classifier in multiple scenarios, nevertheless both classifiers perform the same highest accuracy of 84 % using the same settings, however it is worth mentioning that the NB classifier has less range of accuracy (83-85) than SVM's range (82-87) making it more consistent.

Thus we have concluded that having stemming our texts, extracting unigrams and bigrams features and classifying using a NB classifier is the finest option to sentimentally classify our corpus.

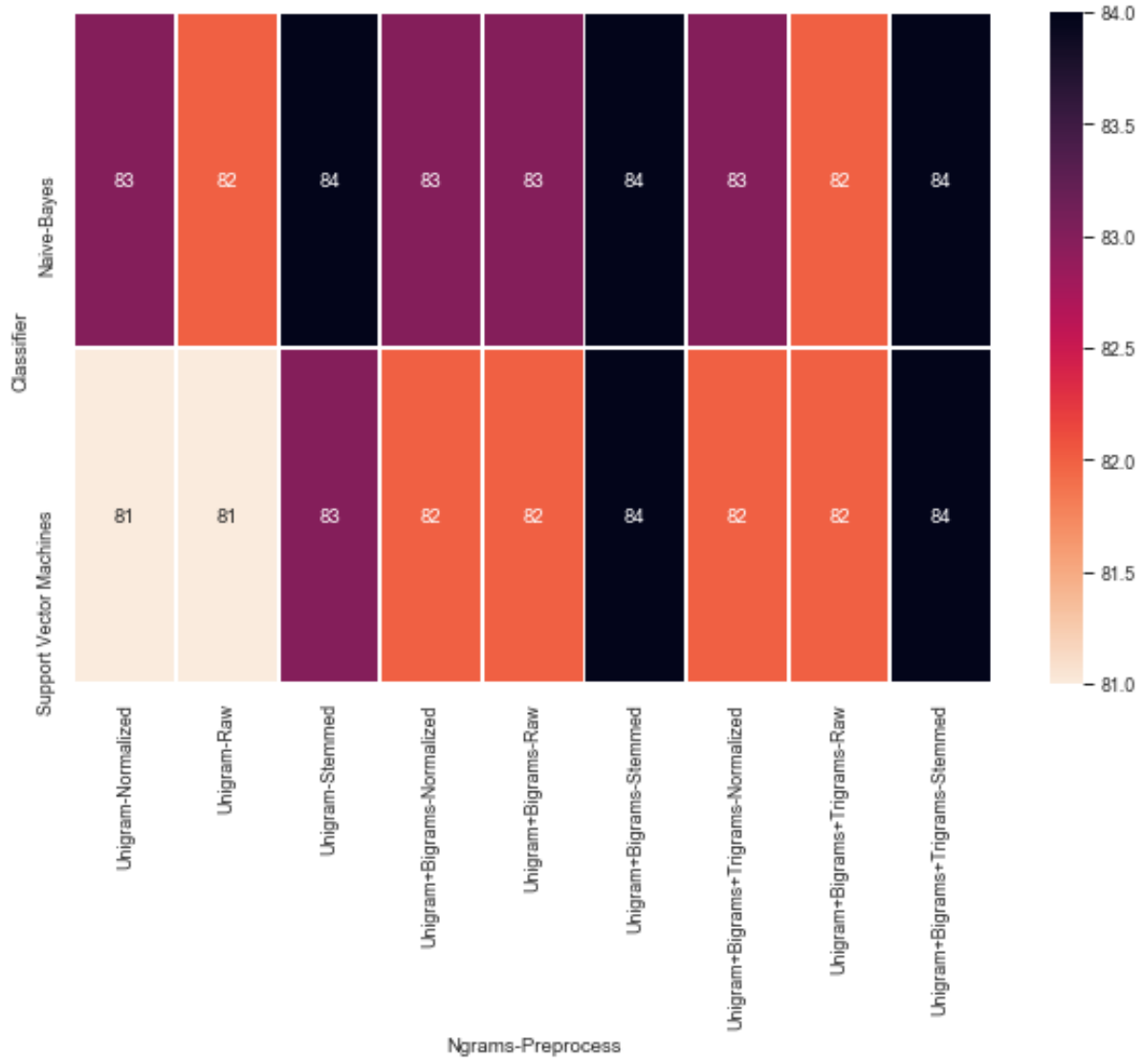

Figure 4.9: Heatmap of the classification accuracy for each combination of variables

4.3 Governorates Urban Sentiment Analysis

4.3.1 Governorates Dataset Analysis

In order to begin our analysis, we must refer back to our governorates database in section 3.1 that would serve as the fundamental for all info that is governorate related. We present visually the percentage of urban land use and the population count exposed for all governorates in the following two graphs respectively.

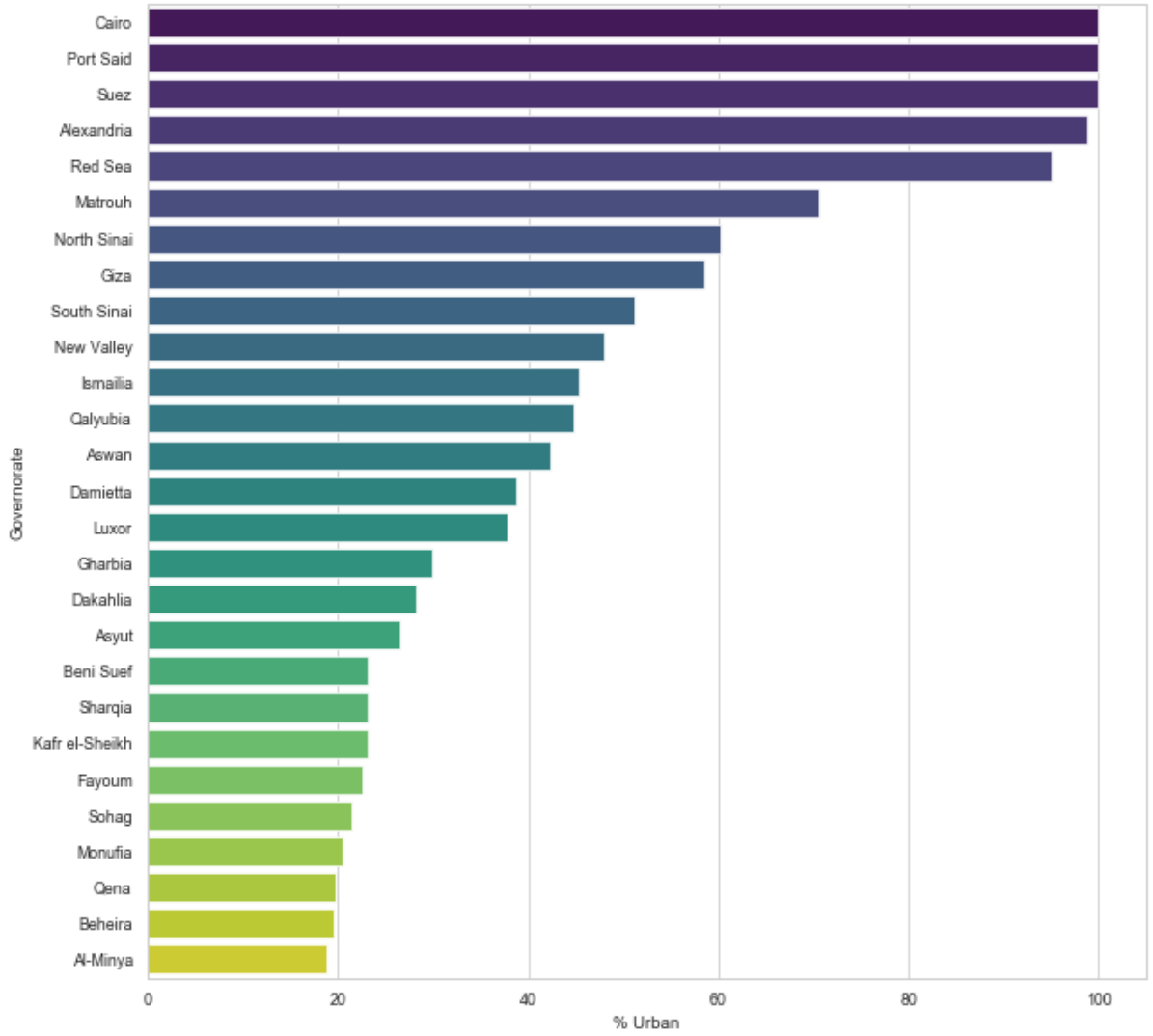

Figure 4.10: Representation of each governorate to its urban land use percentage

From figure 4.10, we can decide that there are only 9 governorates out of the 27 with over half their land use is urban. Cairo, the capital city, has all of its land use to be Urban alongside Port Said and Suez and almost Alexandria and Red Sea. Aside from

Cairo, the four governorates show their importance for having urban land use given they are considered majorly in their access to the offshore regions for naval transportations as well as leisure spots.

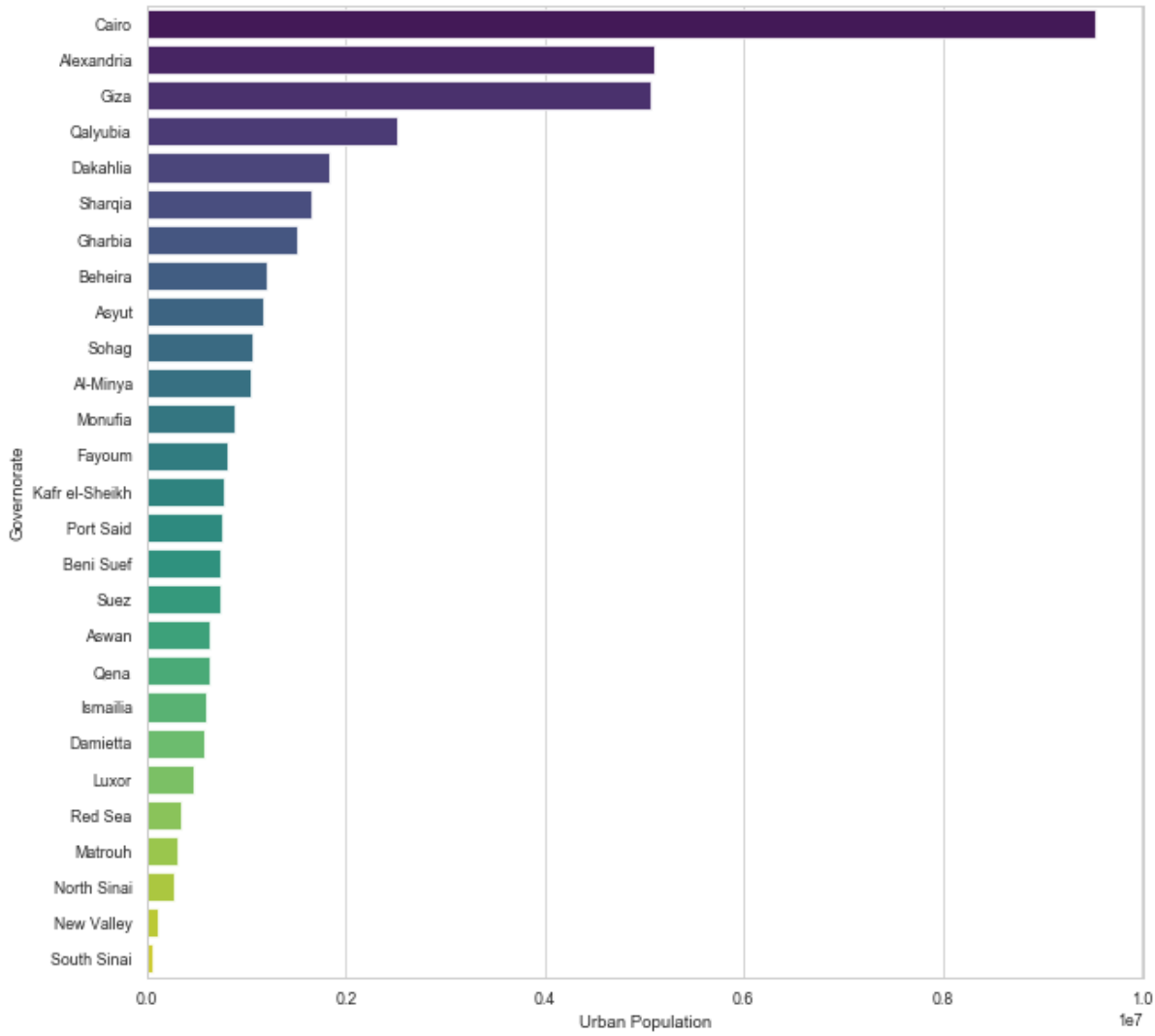

Figure 4.11: Representation of each governorate to its urban exposed population

For the purpose of visualizing how much of the governorates' population is living in urban areas, we assumed that the governorate population is distributed evenly along the governorate area. Once again, Cairo is showing significance in having the most population living in urban areas, along with being the capital despite its relatively small area, with Giza and Alexandria showing how much they vary from the rest of the governorates.

4.3.2 Tweets and Users Analysis

In this section, we take a closer look at the twitter corpus collected and labelled with the governorates and compare the findings with those we have deduced from section 4.3.1. This also presents the opportunity to evaluate the performances of each approach separately and see the relevance and consistency of the data.

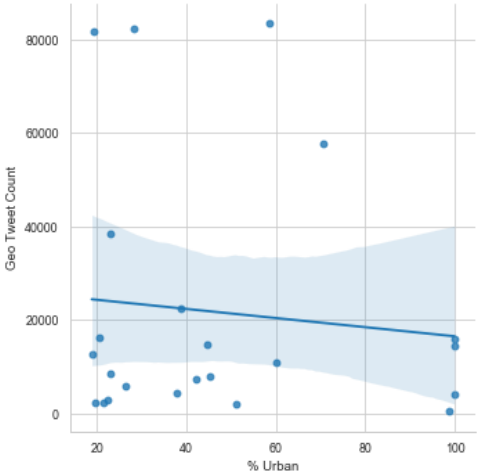

Figure 4.12: With Geocode tweets

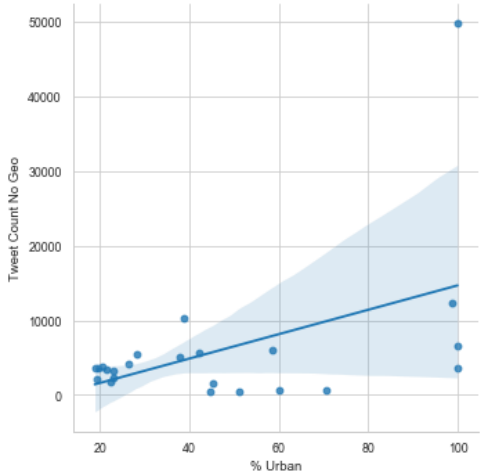

Figure 4.13: Without Geocode tweets

Figure 4.14: Relation between Tweet Count and Urban land %

As we describe the relation between the number of tweets and the percentage of urban area, we have seen the influence of the weakness of the Geocode approach on the relation to the point that the relation became inverse. Nevertheless we have illustrated from the graph without the Geocoded tweets and it shows we have a positive relationship.

When plotting the relation between the tweets count and the urban population, we can see a stronger relationship between the two than between urban land use, which shows the inspiration for us to estimate the population living in urban areas within the governorate to show a better relationship.

The governorate in each non-geo graph in figures 4.13 and 4.16 in the top right most corner is Cairo. Both graphs have shown the variance in estimation that is influenced on the vast difference Cairo has had in the performance on social media activity compared to the rest of the governorates.

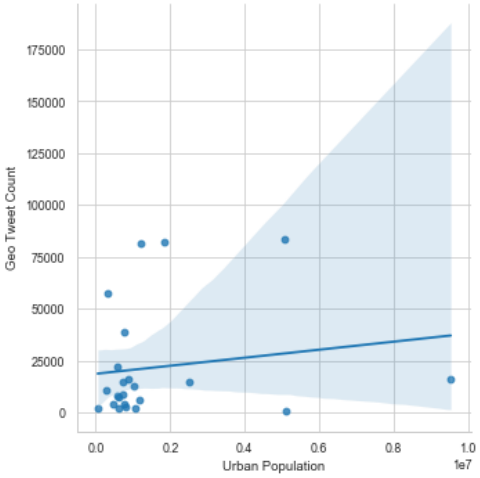

Figure 4.15: With Geocode tweets

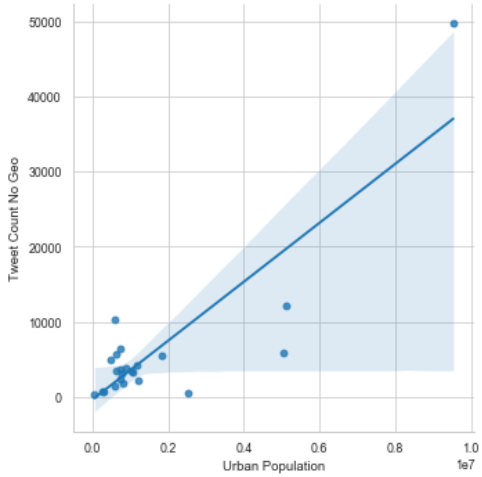

Figure 4.16: Without Geocode tweets

Figure 4.17: Relation between Tweet Count and Urban Population

Most Tweets and Active Users for each approach

From our deductions in the relation graphs regarding the effect of approaches on the outcome, we were motivated to cut down and have a closer inspection on the performance of each approach on its own as well as demonstrate the performance of the top governorates on each approach in terms of both the tweet counts and active users.

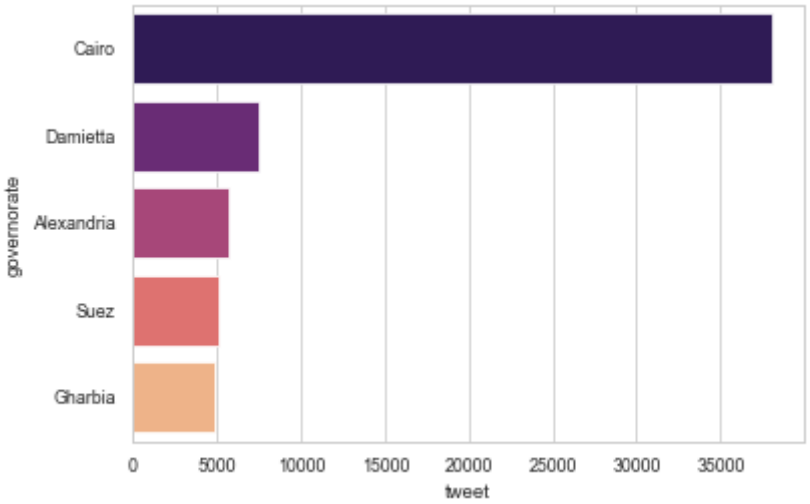

Figure 4.18: Top governorates with most tweets for Keyword Search Approach

The number of tweets that contain relevancy to the governorates in its content when searched upon. Cairo is the most mentioned governorates in tweets.

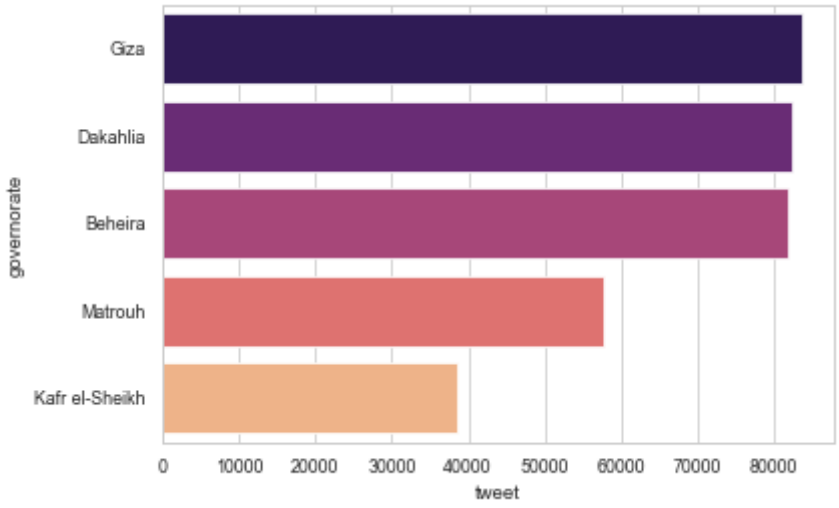

Figure 4.19: Top governorates with most tweets for Geocode Approach

As we can observe looking into the defect of the Geocode approach, we can see the unreliability of the results through the absence of the capital city Cairo from the top most active regions. Instead, as we look back in figure 4.9 and 4.10, we can see a governorate such as Matrouh which can be seen with in the five least population exposed to urban area making it into the top along with Kafr El-Sheikh which is just over 20% of urban land use.

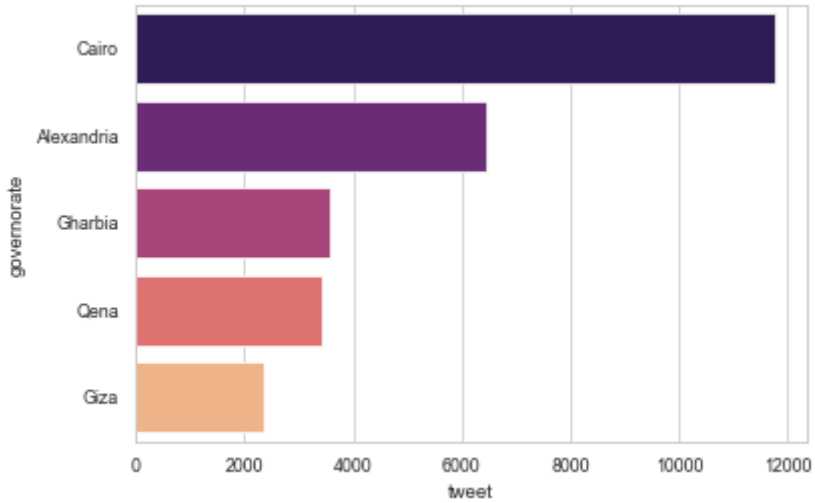

Figure 4.20: Top governorates with most tweets for Profile Info Approach

Once again Cairo is back at the top with most tweets, in addition to Alexandria coming in second place and Giza making it in the top most tweets makes the approach most relevant to the Urban Population graph in figure 4.10.

As we can notice how both the Profile Info approach and the Keyword approach are concurrent in behavior and thus revealing better relations together than with the Geocode approach which we have looked closer and seen its unreliability. It is worth noting that governorate Qena which appears in the top in the Profile Info approach had gone through changes in the Keyword Search approach that made the search results alot less.

Here we represent a similar representation as above, but for the active users, where we can see how many users are tweeting regarding each governorate.

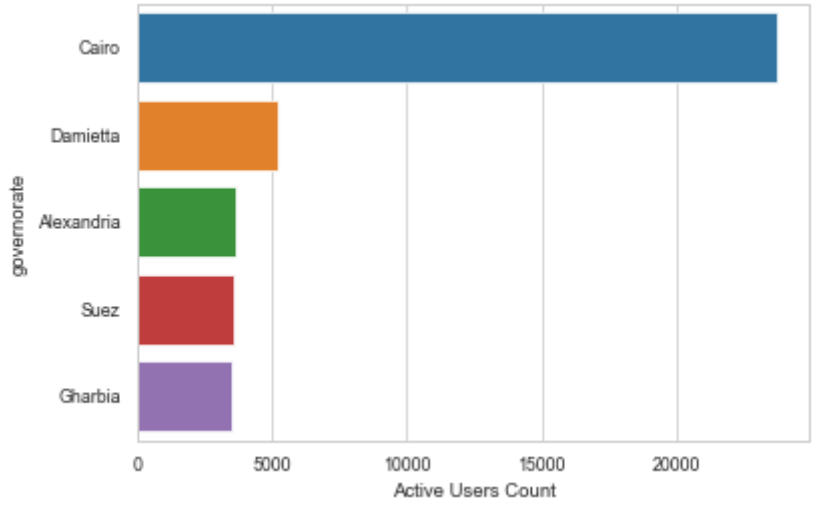

Figure 4.21: Top governorates with most active users for Keyword Search Approach

The graph above shows number of users mentioning these governorates in their tweets.

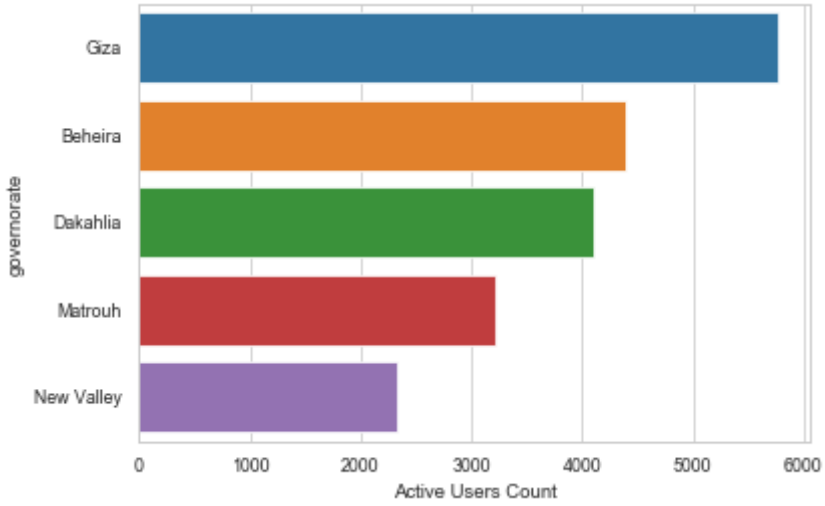

Figure 4.22: Top governorates with most active users for Geocode Approach

The graph above shows the number of users within the circular area of each of these tweets.

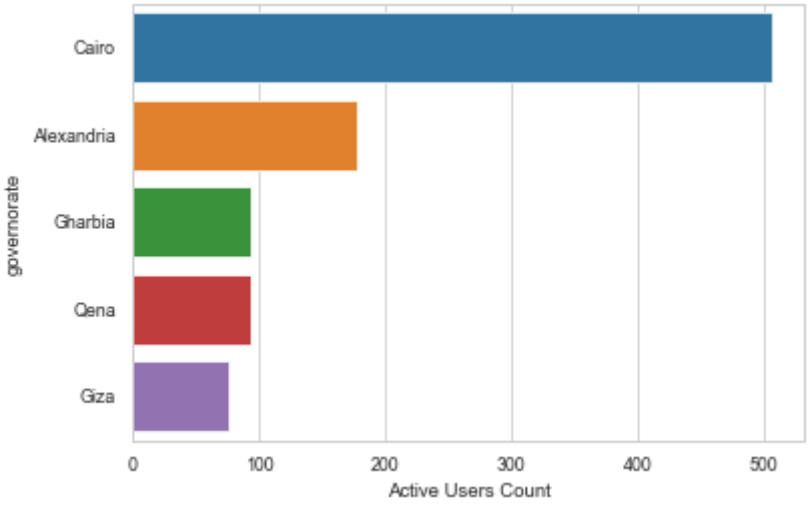

Figure 4.23: Top governorates with most active users for Profile Info Approach

The graph above shows the number of users that have been identified from their hometown locations or bio to be part of these governorates.
The active users graphs have shown that they are alot similar, if not the same, in terms of the top most governorates in each approach.

Sentiment Analysis

After analyzing the performances regarding the tweets and the most users, we focused our direction in reviewing the sentiment of the tweets itself. Using our model that we trained in section 3.4, we apply sentiment classification on the tweets after preprocessing them and aggregate the mean sentiment of each governorate during the period.

The findings in figure 4.24 suggests that despite the fact that Cairo has a 100% urban land use as well as the most population exposed to these urban areas, it still holds a significant negative sentiment amongst its people compared to the rest of the governorates.

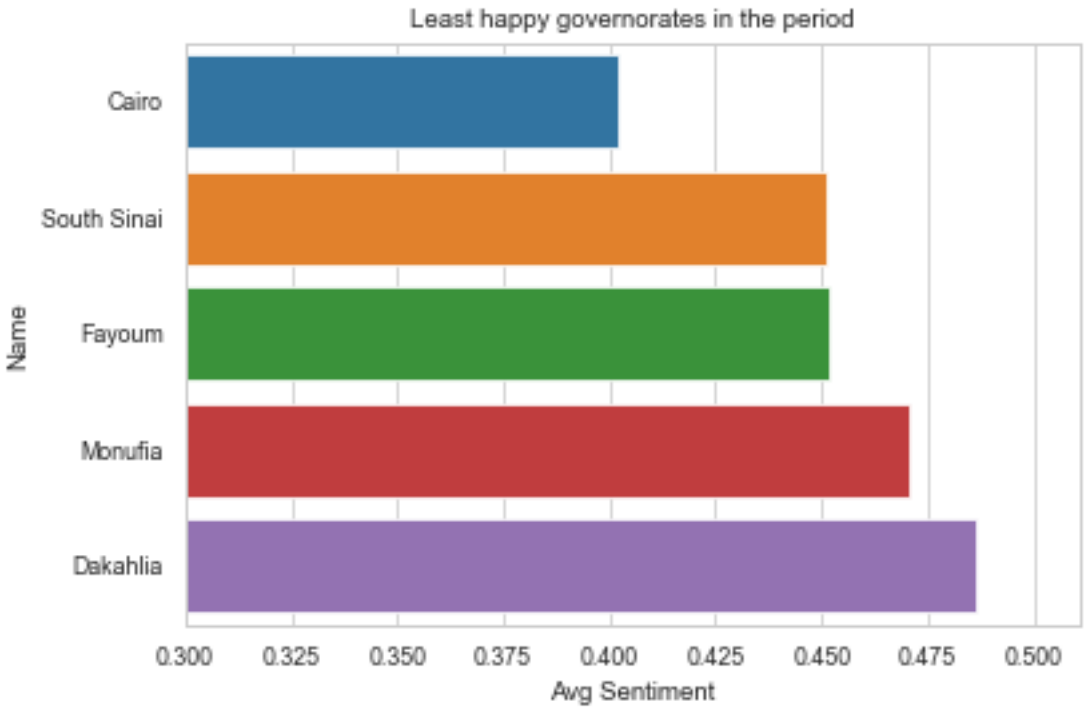

Figure 4.24: Least happy governorates during the specified period

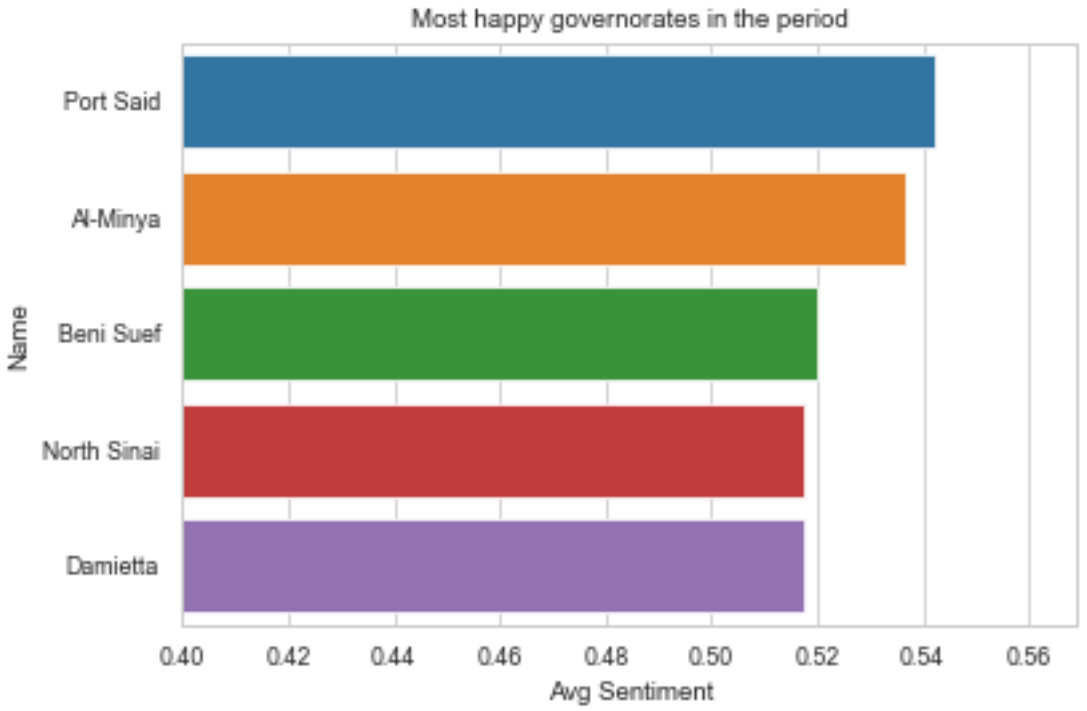

Figure 4.25: Most happy governorates during the specified period

4.4 COVID-19 Case Study in Egypt

We have taken advantage of our COVID-19 lexicon to use in the process of extraction all tweets that are relevant from the corpus regarding the virus. Following the World Health Organization's (WHO) characterization of the COVID-19 virus as pandemic[22], the Egyptian Prime minister has imposed multiple decisions to stay at home in quarantine that had affected the people[1], which encouraged us to investigate the people's social activity in this period.

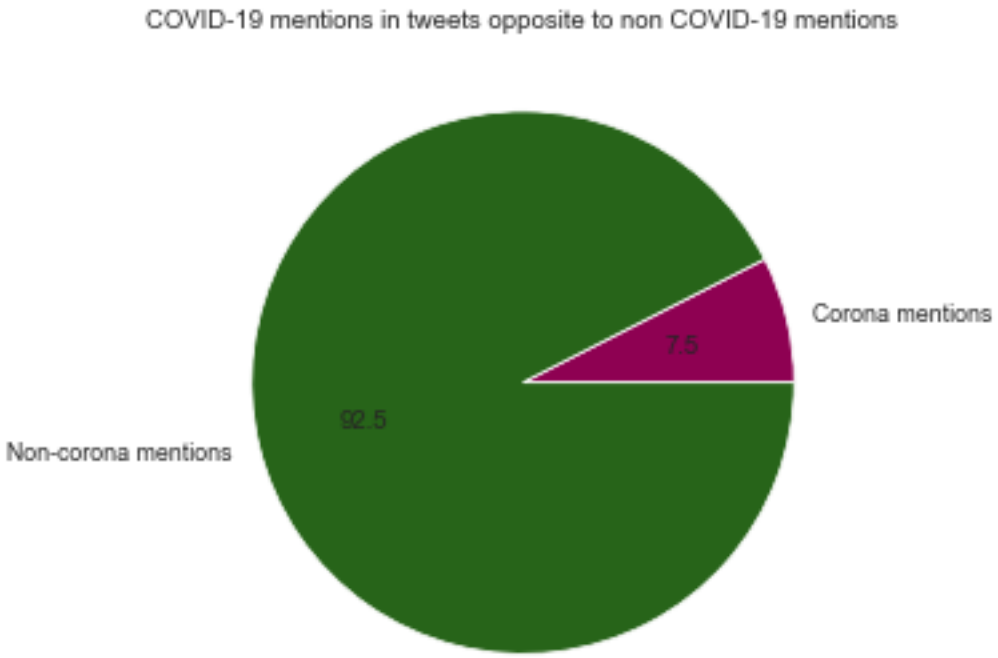

Figure 4.26: Ratio of coronavirus tweets detected to all tweets

Over the 14 day period, we can observe that a significant amount of the tweets from the corpus have mentions regarding the virus making around 65,000 tweets to look closer into. Moreover, we illustrate the most frequent words in the virus' tweets in the following word cloud.

As we are seeing in figure 4.27, there are words associated with the quarantine and the virus, nonetheless there appears to be some country names popping up like Iran, China and Italy, in which this countries were the most cases in a country at the specific period of our tweets.

[1]https://cabinet.gov.eg/Arabic/MediaCenter/CabinetNews/Pages/default.aspx

Figure 4.27: Word cloud of the most frequent words in the coronavirus tweets.

With our governorate annotated tweet corpus, we had the opportunity to analyze the behaviour of the governorates regarding the virus in terms of which governorates is most actively tweeting about the virus and mentioned.

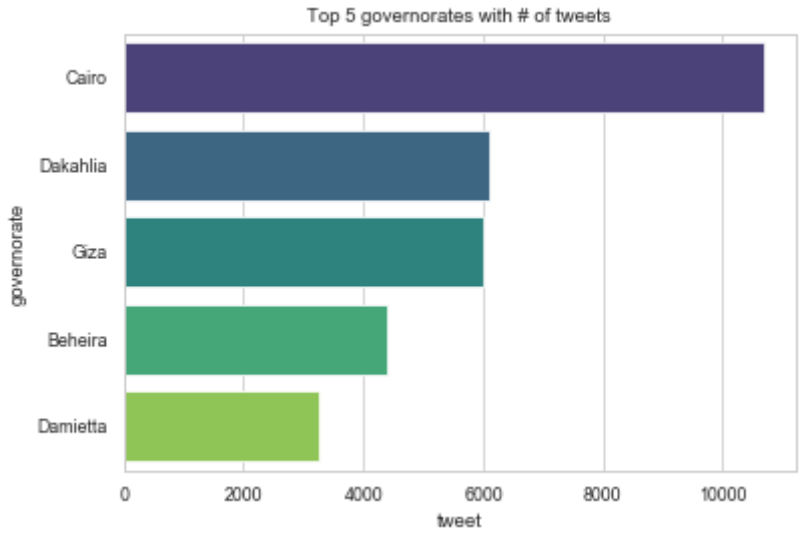

Figure 4.28: Governorates most active in tweeting about coronavirus.

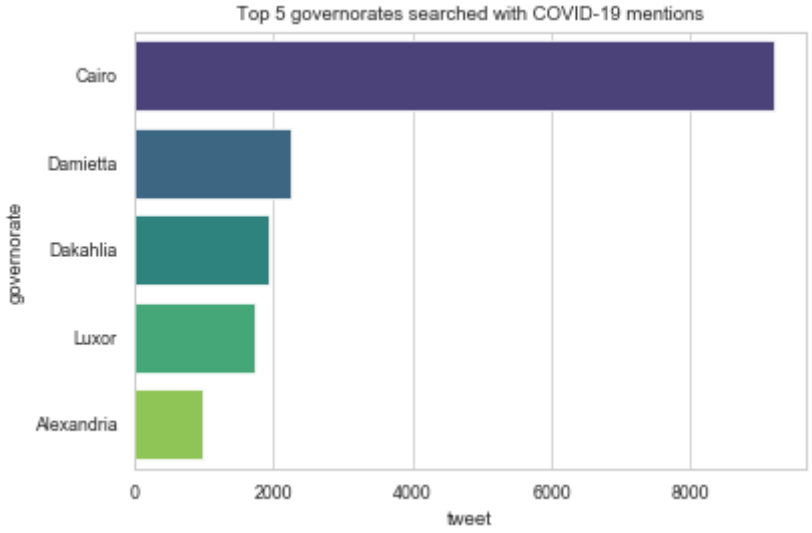

Figure 4.29: Governorates most mentioned in tweets regarding thess coronavirus.

Once again, Cairo is heavily involved in the corona tweets regarding the rest of the governorates. What is interesting is that the circular areas of the geocode actually would

come in handy as it would show the area with most corona tweets regardless of the governorate and would make hotspots over the Egyptian map. Another interesting sight is the high mentions of Luxor in coronavirus tweets, given not having that much activity in the norm, pushed us to investigate the situation where it turned out that Nile boat in Luxor had positive COVID-19 cases on it.

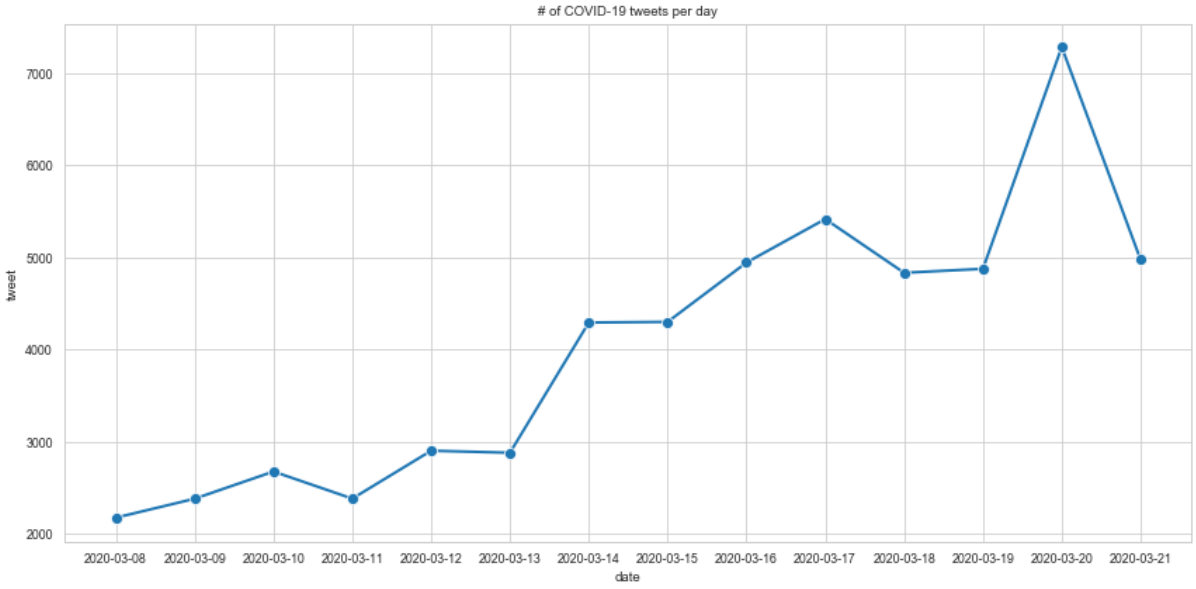

Figure 4.30: Number of coronavirus tweets per day

The number of corona tweets with respect to the graph in figure 4.30 is observed to be increasing by the day as more users and tweets are posted regarding the virus. Also, we can see a significant leap on the 14th of March in the tweets that can be explained by the Prime Minister's decision that day to hold all education systems and impose a stay at home protocols.

During the specified period of the tweets, there are daily reports on the COVID-19 cases and government decisions that all affect the overall mood of the people. We wanted to analyze the sentiment of the people over the days on all tweets regardless whether they are relevant to the virus or not, using our sentimentality annotated corpus.

The findings reveal that there is an overall negative sentiment on most days, with a huge drop from the 11th to the 14th of March. These days where the days COVID-19 was declared pandemic along with once again the decision of the prime minister to quarantine on education schools, further support the effect on the decision with the coronavirus tweet count graph. Furthermore, it appears to be as we approached the weekend of the first week at home, people were starting to adapt to the situation and regain some positivity in their overall mood.

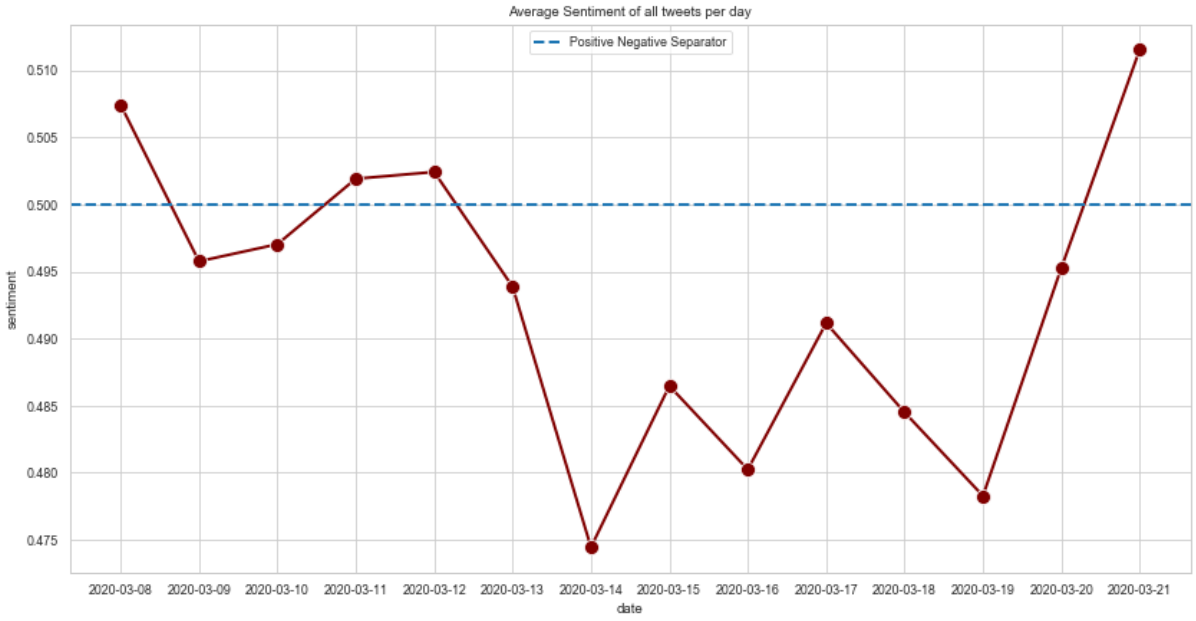

Figure 4.31: The average sentiment per day in all tweets.

Unlike what is expected, there have been significance in the positivity on the tweets regarding the pandemic virus over one third of the tweets, yet the majority is still negative, given the period is at the initial stages of the virus spread.

We also present a timeframe graph for each day holding the average sentiment for each

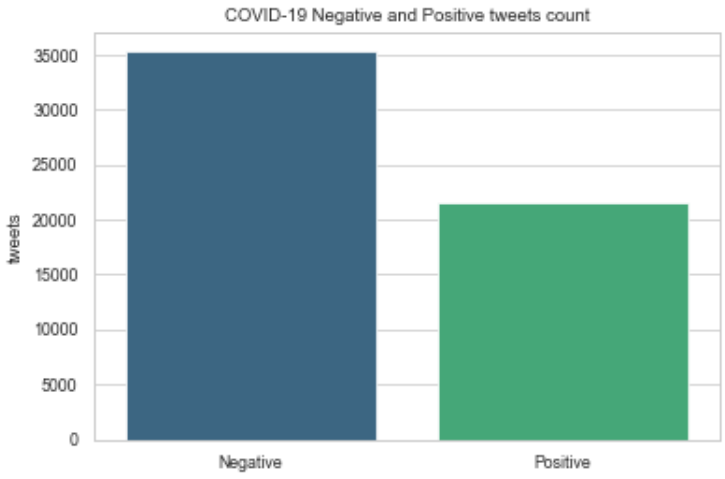

Figure 4.32: Negative and Positive tweets on the coronavirus

day during the period. The graph illustrates how all days are mentioning the coronavirus in a negative sentiment (less than 0.5), however it can be seen as days go by the tweets on the virus are getting more and more positive.

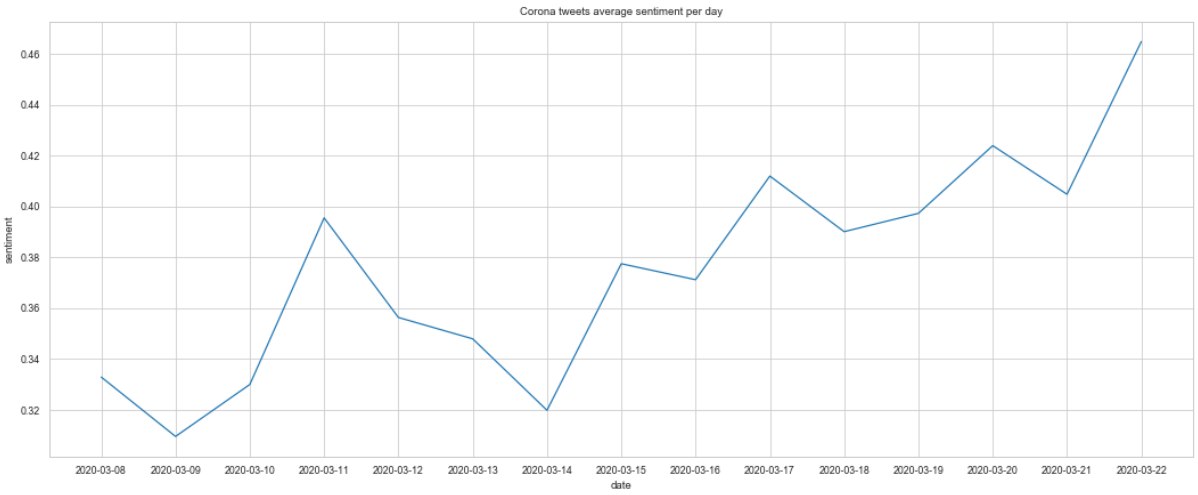

Figure 4.33: The average sentiment per day in coronacirus tweets.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

The field of Sentiment Analysis has developed over the past years, especially with the recent studies on the Arabic language which encourages the opportunity for application on spatio temporal data, given the existence and variance of applications on spatio-temporal data which lacked in the Middle East region. This study investigates various techniques on internet text data mining and preprocessing over microblogging services in order to analyze their performance along with the behaviour of Egyptian governorates given a certain period in terms of both urban analysis and regarding the coronavirus pandemic.

We have collected and combined Egyptian governorates' information into a dataset that we used later for data collection and analysis. In addition, we proposed four approaches to collect spatio-temporal data over Twitter and concatenated them into one corpus labelled to their corresponding governorate and method of collection. Using the ASTD dataset, we trained and tested the model over multiple settings of preprocessing, feature extraction and classifiers in order to use the model in sentimentally annotating the corpus. Finally, we have made a dictionary over the terms regarding the coronavirus and further manually increased for better detection of related coronavirus tweets in the corpus.

Our findings show how much the capital city Cairo has been divergent to the other governorates in terms of the most twitter activity given the 100% urban land use as well as the one with most population exposed to urban areas, nevertheless it had the least sentiment average over the period. Furthermore on the urban analysis, we were able to show a reaction between the percentage of urban land use and the Twitter activity, however a stronger relation appeared when considering the population in urban areas with Twitter activity. On the other hand, we illustrated the rise in the coronavirus mentions over the days in the specified period, noting a significant rise that correlated with the lockdown procedures announcements that also negatively affected the overall mood of the people. Eventually we have shown as time went by during the lockdown, people began regaining their positivity and the coronavirus tweets had more positive content.

Unfortunately, the study falls short in the consistency of its governorates database considering its collection from multiple different resources due to the unavailability of the official databases and dealing with old data. Regarding the introduced data collection methods, we have found problems in the completion of data collection process with the Nearby approach, ending up discarding all of its data from further analysis. A drawback of the Geocode approach's performance was shown in its representation of the geographical areas of the governorates yielding inaccurate results that negatively affected the relations in the urban analysis. Moreover, the Profile Info approach's collection was limited due to the relatively small size of its collected users limited by the resources of used in the study for collection. Finally the sentimentally annotated dataset was small in size and would have had a better accuracy with a larger dataset.

5.2 Future Work

In future work, a collection to an official updated database for the governorates info is an essential fundamental for all spatio-temporal applications in Egypt. A better geographical coordinates' data that encapsulates the exact shape of the governorates and more details about the distribution of the population in the urban lands shall take this study's work into yielding better results.

The Profile Info approach is worth further work for its accuracy of data, with high performance computers and fast access to Twitter servers for much larger collections, given the collected followers were not even 1% of the actual followers count due to its inefficiency. For better sentiment annotation, a large dataset shall be collected within the Egyptian domain and annotated by professional annotators.

An interactive visual representation over a geographical map is needed for better visualization of patterns and the behaviour of the governorates as well as the introduced approaches. Such representation can be integrated in a graphical user interface with time widgets to alter between periods for more insights and serve a crucial analytical tool in the future. Also, integrating topic modelling for the identification of coronavirus tweets might include more tweets in the domain that weren't detected.

Appendix

List of Figures

List of Tables

Bibliography

[1] Nawaf A Abdulla, Nizar A Ahmed, Mohammed A Shehab, and Mahmoud Al-Ayyoub. Arabic sentiment analysis: Lexicon-based and corpus-based. In *2013 IEEE Jordan conference on applied electrical engineering and computing technologies (AEECT)*, pages 1–6. IEEE, 2013.

[2] Hady ElSahar and Samhaa R El-Beltagy. Building large arabic multi-domain resources for sentiment analysis. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 23–34. Springer, 2015.

[3] Aaron J Schwartz, Peter Sheridan Dodds, Jarlath PM O'Neil-Dunne, Christopher M Danforth, and Taylor H Ricketts. Visitors to urban greenspace have higher sentiment and lower negativity on twitter. *People and Nature*, 1(4):476–485, 2019.

[4] Vincent X Gong, Winnie Daamen, Alessandro Bozzon, and Serge P Hoogendoorn. Estimate sentiment of crowds from social media during city events. *Transportation research record*, 2673(11):836–850, 2019.

[5] Amira Shoukry and Ahmed Rafea. Preprocessing egyptian dialect tweets for sentiment mining. In *The Fourth Workshop on Computational Approaches to Arabic Script-based Languages*, page 47, 2012.

[6] Mahmoud Nabil, Mohamed Aly, and Amir Atiya. Astd: Arabic sentiment tweets dataset. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2515–2519, 2015.

[7] Amira Shoukry and Ahmed Rafea. Sentence-level arabic sentiment analysis. In *2012 International Conference on Collaboration Technologies and Systems (CTS)*, pages 546–550. IEEE, 2012.

[8] Lewis Mitchell, Morgan R Frank, Kameron Decker Harris, Peter Sheridan Dodds, and Christopher M Danforth. The geography of happiness: Connecting twitter sentiment and expression, demographics, and objective characteristics of place. *PloS one*, 8(5), 2013.

[9] Mohammed Korayem, David Crandall, and Muhammad Abdul-Mageed. Subjectivity and sentiment analysis of arabic: A survey. In *International conference on advanced machine learning technologies and applications*, pages 128–139. Springer, 2012.

[10] Cornelia Caragea, Anna Cinzia Squicciarini, Sam Stehle, Kishore Neppalli, Andrea H Tapia, et al. Mapping moods: Geo-mapped sentiment analysis during hurricane sandy. In *ISCRAM*, 2014.

[11] Jonathon Read and John Carroll. Weakly supervised techniques for domain-independent sentiment classification. In *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, pages 45–52, 2009.

[12] Peter Sheridan Dodds, Kameron Decker Harris, Isabel M Kloumann, Catherine A Bliss, and Christopher M Danforth. Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *PloS one*, 6(12), 2011.

[13] Morgan R Frank, Lewis Mitchell, Peter Sheridan Dodds, and Christopher M Danforth. Happiness and the patterns of life: A study of geolocated tweets. *Scientific reports*, 3(1):1–9, 2013.

[14] DESA UN. World urbanization prospects: The 2014 revision. *United Nations Department of Economics and Social Affairs, Population Division: New York, NY, USA*, 41, 2015.

[15] Aiman Soliman, Kiumars Soltani, Junjun Yin, Anand Padmanabhan, and Shaowen Wang. Social sensing of urban land use based on analysis of twitter users' mobility patterns. *PloS one*, 12(7):e0181657, 2017.

[16] Nivan Ferreira, Jorge Poco, Huy T Vo, Juliana Freire, and Cláudio T Silva. Visual exploration of big spatio-temporal urban data: A study of new york city taxi trips. *IEEE transactions on visualization and computer graphics*, 19(12):2149–2158, 2013.

[17] Debjyoti Paul, Feifei Li, Murali Krishna Teja, Xin Yu, and Richie Frost. Compass: Spatio temporal sentiment analysis of us election what twitter says! In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1585–1594, 2017.

[18] Bruno Justino Garcia Praciano, João Paulo Carvalho Lustosa da Costa, João Paulo Abreu Maranhão, Fábio Lúcio Lopes de Mendonça, Rafael Timoteo de Sousa Júnior, and Juliano Barbosa Prettz. Spatio-temporal trend analysis of the brazilian elections based on twitter data. In *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 1355–1360. IEEE, 2018.

[19] Vincenza Carchiolo, Alessandro Longheu, and Michele Malgeri. Using twitter data and sentiment analysis to study diseases dynamics. In M. Elena Renda, Miroslav Bursa, Andreas Holzinger, and Sami Khuri, editors, *Information Technology in Bio- and Medical Informatics*, pages 16–24, Cham, 2015. Springer International Publishing.

[20] Janaína Gomide, Adriano Veloso, Wagner Meira Jr, Virgílio Almeida, Fabrício Benevenuto, Fernanda Ferraz, and Mauro Teixeira. Dengue surveillance based on a

computational model of spatio-temporal locality of twitter. In *Proceedings of the 3rd international web science conference*, pages 1–8, 2011.

[21] W John Wilbur and Karl Sirotkin. The automatic identification of stop words. *Journal of information science*, 18(1):45–55, 1992.

[22] World Health Organization et al. Coronavirus disease (covid-19)-events as they happen. 2020, 2020.