# Twitter Reactions on Coronavirus across UK - How British behaved differently to the pandemic across country?

**Data Driven Dissertation Project in Business Analytics (BUSI4374)**
**(SUMMER 19-20)**

Author: Sannia Nasir
Student ID: 20201243
Supervisor: Dr. James Goulding

**Abstract**

COVID-19 has generated a plethora of discussions on social media. Twitter has been extensively used as a platform to discuss facts, share opinions, and express sentiments during the times of crisis. It is imperative to understand that the response to COVID-19 in different areas may be dependent upon local cases, government policies, health infrastructure and multiple socioeconomic factors. As the pandemic has affected everyone's' life on a different scale, it is essential to capture the public's main topics of concern, and how people in different locations have reacted to the situation. In the event of crisis, monitoring people's responses across various events and topics will be extremely necessary for policy makers to be able to devise effective policies. In this dissertation, twitter data is used to perform region-specific exploratory analysis on people's discussion and sentiments about COVID-19 in the UK. Topic modelling and sentiment analysis is used to unravel the intra-region dynamics. It shows that people across different regions discuss the same topics in different contexts. The study also discusses how sentiments of people change over time and are heavily influenced by government policies and information on media. Overall, the proposed framework shows the potential of using twitter analysis to assist the local authorities in designing policies for new interventions.

# Table of Contents

## 1. Introduction

The lives of millions of people globally have been affected due to the coronavirus. In March 2020, the COVID-19 outbreak was declared a pandemic by the World Health Organization (World Health Organization [WHO], 2020). Governments around the world took strict measures such as partial or full lockdown in order to contain the spread of virus. Social distancing measures such as the 2m distance rule and work from home regulations were put in place across the UK as well. Transport systems in some areas such as London were suspended. Various health awareness campaigns were launched to make sure that the general public follows expert health advice as it had a substantial impact on the death rate from the disease. The need to effectively communicate the social distancing policies was felt because ignorance or misinformation could lead to a burden on National Healthcare Services (NHS). Due to varying local challenges such as the disease contraction rate, population density and medical resources, the local government authorities implemented different strategies to curb the spread of disease overtime. The inter-region policies and their timelines varied from region to region which made it crucial to understand the attitudes and emotions of people on the national as well as the local level.

### 1.1 Objectives of Research

This dissertation analyzes the impact of the coronavirus pandemic on the social and behavioral responses of the people in the United Kingdom. It focuses on identifying the themes of discussion related to the coronavirus across different regions in the UK using twitter data. The three main hypothesis tested in this research are:

1) There are differences in the subjects of discussion of people belonging to regions in the UK towards Covid-19.

2) The sentiments of people towards the pandemic change overtime.
3) There are differences in the sentiments of people across different regions in the UK towards Covid-19.

The purpose of conducting regional study is to determine whether the people located in different geographical areas responded differently to Covid-19 or not. The regional differences could be attributed to demographics, number of covid-19 cases and numerous other socio-economic factors which is why this regional study also takes account of the deprivation conditions across areas. It explores the responses of people not only in terms of the discussion topics but also analyzes the sentiments towards covid-19 across Lower Layer Super Output Areas (LSOA) in the UK. The aim of this research is to extract insights which could be useful to develop informational campaigns on a local level and assist the relevant government departments to take intervention decisions objectively.

## 1.2 Research Questions

The focus of this study is to answer three main questions:

1) Are there differences in the opinions and behaviors of people belonging to different regions of the UK towards COVID-19?
2) Are the sentiments of people in the UK changing overtime due to COVID-19?
3) Are there regional differences in the sentiments of the people due to COVID-19? If any, what possibly could be the factors responsible for these differences?

## 2. Literature Review

A wide range of research has been carried out in the area of using social media data to make predictions or judgements regarding individuals' response to pandemics. In a study conducted by Jain and Kumar (2015), an approach based on analyzing tweets via Twitter was proposed to underpin surveillance systems. The research discussed how twitter data could be used to detect early outbreaks of H1NI virus and how these signals could serve as early warning signs. This could then improve the monitoring of disease by the government agencies and can also help in effective resource planning for the medical services. Another research by Kostkova et al. (2014) showed the correlation of self reported cases of swine flu via twitter and the official reported cases. The comparison indicated that twitter data managed to pick up earlier trends of outbreak whereas there was a lag of 1-2 weeks in reporting the confirmed cases officially. This research also shows the dissemination of information via twitter, analyzing which information sources were more widely shared, and whether they were trustable or not. These researches show that twitter data can indeed be an effective source for carrying out social and behavioral studies. Thus, this serves as a foundation for conducting research which aims to use twitter data to interrogate the changing behaviors of different sub groups of population overtime during a pandemic.

Recent advances in statistical analysis have introduced new techniques to aid in analysis of individual and consumer response to health events. A research "*Twitter Catches Flu - Detecting Influenza Epidemics Using Twitter (2011)"* talks about how *natural language processing* (NLP) techniques can now be used to extract only relevant tweets related to a particular topic. In this dissertation, text mining techniques will be used to extract tweets which are relevant to COVID-19 to analyze the change in sentiments of people over time and their social attitudes. This is supported by evidence from studies which have been conducted to identify various topics of discussion during the pandemics using the social media data. A study by Chew et al. (2010) conducted content analysis of tweets during the H1N1 outbreak in 2009. It also collected and filtered out the RSS feeds relating to H1N1 or Swine flu from the top newspaper agencies. Latent Dirichlet Allocation (LDA) was the technique used to divide news headlines to find the most popular news for a particular day, and allowed analysis of the variation in the tweets gathered over an observed time interval. The study also used chi square test to determine if content/links tweeted changed linearly over time, or whether there were any non-linear patterns. Similarly, a paper written by Kim et al. (2015) performs topic based content and sentiment analysis of Ebola virus based on twitter and news data. The research focused on identifying the different topics emerging from each source of information. It was observed that the news focused more on event-related entities such as person, organization and location, whereas Twitter covered more time-oriented entities. It also showed that there was a difference in content coverage and sentiment dynamics of news articles and tweets data. However, this study did not explore the relationship between the news and the responses of people on Twitter. Hence, in this dissertation, the influence of media on people's opinions will also be analyzed via twitter data. In addition, Kim et al. also used topic based sentiment scores to summarize the difference in sentiments of twitter posts and news posts. It also delved into analyzing whether the degree of sentiments change over time. The results showed that there is variation in the emotions of people as time progresses. Hence, in case of covid-19, a similar pattern is likely to be observed.

Recently, due to the spread of coronavirus, a lot of research has poured out, ranging from using sentiment analysis to gauge the emotional responses of people to using topic modelling to understand the emerging discussions around the topic. In a study by Thelwall et al. (2020), word frequency analysis was used to suggest that different genders react differently to the pandemic and that their interests and engagement on topics varied. The results showed that females are more likely to tweet about the virus in the context of family, social distancing and healthcare whereas males are more likely to tweet about sports cancellations, the global spread of the virus and political reactions. It suggested that certain government policies might be beneficial in communicating the seriousness of the issue to one gender but the same policy might have no effect on the other gender. Although this study explored evident gender differences, it did not take the regional aspect into account nor investigated any geographical differences. Another study by Xue et al. (2020) used around 1.9 million tweets to perform topic modelling on COVID-19 communications. 11 distinct topics were identified with underlying themes- such as the outbreaks in China, Korea, New York; some themes also involved the economic impact and preventive measures taken as a result of covid-19. It also showed that during the early stages of the virus, the dominant emotion among all the topics is fear. However, again no breakdown was provided into regional or demographic variations in response.

A more recent study by Yin et al. (2020) used 13 million tweets to analyze the dynamics of people's sentiments towards the disease. The study disclosed that the overall sentiment polarity was positive during the study period, however, the sentiment polarity across topics varied. Although this study effectively pointed out that sentiments across topics can vary, only two weeks of data were analyzed, which did not allow measurement of change in emotions overtime. Different areas of the world have been considered, but generally only in isolation, rather than in comparison. A study by Das et al. (2020) observed the overall sentiment trends during COVID-19 in India. It conducted emotion mining using the "NRC" lexicon of - six basic emotions: joy, sadness, anger, fear, disgust, and surprise. The study used valence shifters to attach sentiment scores to the tweets. The corpus-level analysis showed that positive sentiments are much higher compared to negative sentiments in tweets. It also conducted topic modelling on the negative and positive sentiment tweets' corpuses. Another study by Jang et al. (2020) compared tweets from the US and Canada. LDA was used to identify 5 topics which included social distancing, air travel, hand washing, impact on essential workers and use of preventive measures such as face masks. The study also pointed out that the public health related intervention topics change overtime. It also conducted sentiment analysis using the lexicon based methods, which showed that the ratio between the number of positive and negative occurrences differed from topic to topic. This indicated that if different discussions reflect varied sentiments, the geographic and demographic differences might hold as well. The topics involving Asians, outbreak of disease and misinformation had primarily negative sentiments. However, this study did not show how the sentiments change with respect to the time.

There have also been some first analyses of regional variations, albeit targeted at very specific themes. A study by Feng et al. (2020) analyzed around 650,000 geotagged tweets from 50 states of the USA. The study aimed at measuring the changes in work engagement during different phases across these states. Social sentiment analysis was conducted via facial emojis to measure the general public's emotions on stay-at-home orders, reopening and other local regulations. The study observed that negative moods dominated the public sentiment over key COVID-19 events which showed that although the overall sentiment of people could be positive over time, it might differ if event specific sentiments are analyzed. The pattern was similar across states which implied that there were no geographical differences. The study also conducted topic modelling. LDA model was applied on the TF-IDF corpus to extract 10 latent topics. However, the topic modelling covered the discussions on twitter across all states. The study did not interrogate the difference between topic expressions among states to understand the important factors in each state, and how they varied, which again left a key gap in understanding how discussions across regions vary.

In the past, different studies have analyzed the change in specific actions of people across different geographic regions amidst pandemics. A study by Daughton et al. (2019) analyzed twitter posts regarding the 2015-16 Zika virus outbreak and identified specific behavior changes relevant to the spread of disease—travel cancellation. It showed differences in the demographics of individuals who changed their travel behavior in response to Zika, with significantly more discussion by women than men related to the issue. It also highlighted significant differences between geographic areas in the United States which illustrated that responses between states can differ depending upon the levels of exposure to disease-related information. Another study by Achrekar et al. (2012), showed that twitter data is highly correlated with the illness-like-influenza (ILI) rates across different regions within the USA and can be used to effectively improve the accuracy of disease prediction. It was observed through the comparisons across regions that when the ILI rate peaks later in a particular region than the rest of the country, the tweet reports also peak later and there is a relationship between the decrease in ILI rates and the decrease in tweet reports. This could imply that the number of covid-19 cases and deaths might be an important factor in determining the regional response to the disease. Unlike studies which looked at regional tweets, Nagar et al. (2014) tested Twitter for daily city-level data for New York City in order to predict the outbreak of disease, showing an understanding of the ways in which the public perceives and reacts on risks of emerging contagious diseases. The study indicated that granular Twitter data provides more important spatiotemporal insights. Hence, in case of covid-19, if the regional data is not able to reveal important differences, a granular level of study might uncover variations in behaviors of people towards the disease.

More recently, a few studies have analyzed the geographical differences, albeit targeted at specific behaviors of people in response to covid-19. Kwon et al. (2020) examined peoples' reaction towards social distancing during the COVID-19 pandemic in the US in a spatiotemporal context. The results showed that cities who had a less number of social distancing tweets had an increasing trend with time which corresponded with the rise of COVID-19 cases in the city. It communicated the idea that the overall volume of social distancing tweets can reflect the relative case count in respective locations. Another study by Huang et al. (2020) discussed how people's travel routines were affected by government policies. The country level comparisons showed the discrepancies in responsiveness which was evidenced by the varying mobility patterns in different phases of the pandemic. It also showed that mobility changes correspond with the national announcements of mitigation measures. A key conclusion was that although the influence of the COVID-19 pandemic on mobility was distinct in the US, the impacts varied considerably among states. Despite effectively illustrating how inter-state responses of people on Twitter can vary, this study also limits its focus on a specific activity i.e. mobility.

Based on the existing research, it can be seen that there is significant literature which uses twitter data to predict the outbreak of pandemics and understand its effects on the population. There is considerable research on approaches to extracting the relevant information from tweets in order to analyze the responses of people. However, most of the studies consider only the polarity of sentiments and their change overtime across an overall population. The themes identified as a result of topic modelling are also assumed to be applicable to every region and demographic in that the general population. This is unlikely to be the case, with differing reactions and responses across geographic areas. The studies which did include the regional aspect, focused on very specific and narrow themes related to the reactions of people. This dissertation will hence address the gap related to how regional differences, which include the geographical aspect, as well as the index of multiple deprivation, account for a change in different social behaviors of people over time. It will also explore how topics of discussion vary across people belonging to different subpopulations. The established techniques of topic modelling and content analysis will be used effectively along with geotagging and other demographic features to discern the sociological aspect of Covid-19.

## 3. Methodology

This section explains the framework used to conduct topic modelling and sentiment analysis in order to answer the three research questions mentioned earlier. It is divided into five stages - data collection, explanation of data characteristics, data preprocessing, topic modelling and finally the sentiment analysis.

### 3.1 Data Collection

The data was collected using twitter API from 21st March till 23rd May 2020. A total of 28,841,308 tweets were extracted. The keywords used to identify multiple common ways of referring to the disease were "coronavirus, corona-virus, COVID-19, COVID19, corona". As this dissertation only focused on the UK, the tweets which were specifically related to the UK were selected. When the tweets were filtered, the tweets originating from Isle of Man were also selected. Although Isle of Man is an independent state and is not governed by the UK government, it was included in the analysis because of the status of its citizens who are considered British citizens under the constitution (*Isle of Man Government*, 2020). Another reason for its inclusion was to analyze if geographical location and different government powers could have significant influence on the reactions of people.

For the final data set, the duplicates including multiple retweets as well as identical tweets with different usernames and hashtags were removed because they added little value to the extraction of distinct discussion topics and insights. The process resulted in 548,344 unique tweets from the UK and Isle of Man. Apart from using twitter data, two publicly available data sources were also used. The official website of the government of the UK (www.gov.uk) was used to collect statistics of index of multiple deprivation (IMD) across the Lower Layer Super Output Areas (LSOA). The Office of National Statistics website (www.ons.gov.uk) was used to gather information about the regions of the UK. Both of these supplementary sources were used to classify the tweets according to the geographical regions and the IMD scores.

## 3.2 Data Characteristics

The twitter data collected included information related to the tweet itself i.e. the text, hashtags, retweets, date etc. It also included the profile information such as followers, favorites and location information. Out of the overall UK generated tweets, a relatively small proportion of tweets contained the LSOA information. The overall tweets were used to construct the LDA model and gauge the overall sentiment of the UK's population overtime. On the other hand, the tweets with specific location information were used to conduct the regional study. In order to prepare the data for geographical analysis, the tweets' location was matched against the region that they originated from. The Office of National Statistics' data was used to match the tweet locations against 12 main regions of the UK.

In order to conduct regional analysis based on the socio-economic conditions of the regions, the index of multiple deprivation was used. “The Index of Multiple Deprivation (IMD) is an overall relative measure of deprivation for small areas (Lower-layer Super Output Areas) across England, based on seven domains of deprivation” (*English indices of deprivation*, 2019). These are combined using the following weights to produce the overall Index of Multiple Deprivation:

1. Income Deprivation (22.5%)
2. Employment Deprivation (22.5%)
3. Education, Skills and Training Deprivation (13.5%)
4. Health Deprivation and Disability (13.5%)
5. Crime (9.3%)
6. Barriers to Housing and Services (9.3%)
7. Living Environment Deprivation (9.3%)

For the regional analysis, the variable used to incorporate the element of IMD was deciles. “The deciles are calculated by ranking the 32,844 LSOAs in England from most deprived to least deprived and dividing them into 10 equal groups. LSOAs in decile 1 fall within the most deprived 10% of LSOAs nationally and LSOAs in decile 10 fall within the least deprived 10% of LSOAs nationally” (*English indices of deprivation,* 2019).The other relevant variables used included the text of the tweet itself, its date and the location information. The location variable was important in order to conduct the regional aspect of the study whereas the date variable provided information to analyze the change in sentiments over time.

## 3.3 Data Preprocessing

In order to prepare the data for natural language text processing, the tweets were preprocessed. The punctuations, special characters, extra spaces, single letters and non-alphabets were removed as part of the data cleaning process. The hashtags (#), usernames (@) and website links (http://www) were also eliminated from the text because they did not reflect distinct opinions of people. The words were then converted to lowercase letters in order to avoid unnecessary duplication. The next step involved performing word tokenization to convert every tweet into a sequence of words. In order to ensure good quality of data analysis, the most frequently occurring keywords which added no value to the analysis such as coronavirus, pandemic, covid-19 were removed along with the stop words. After the removal of stop words, POS tagging was implemented to characterize the context in which a particular word was used. This provided the basis for effective lemmatization which converted all the words with inflectional endings to their root word. This served to be the last step of preparing the data for further analysis.

## 3.4 Topic modelling

This dissertation used an unsupervised learning technique i.e. topic modelling to conduct exploratory analysis for large unstructured text data. Topic modelling is commonly used for the extraction of hidden semantic structures. The research question at hand involved exploring and interpreting twitter data to inductively assign meanings to the themes derived. In order to implement this technique, Latent Dirichlet allocation (LDA, Blei et al., 2003) was used which helped to gain a descriptive understanding of different themes. LDA rearranged the distribution of topics within the documents i.e. each tweet. The LDA model was then used to predict the distribution of obtained topics for unseen documents.

A random sample of 5000 tweets from all across the UK was used to construct the LDA model. A study by Kim et al. (2018) suggests that simple random sampling is an efficient technique in obtaining a more representative sample of Twitter data rather than stratified or constructed week sampling. Due to the computational constraints, a random sample of 1% of the tweets data was selected.

The preprocessed data was then used to construct bigrams and trigrams. Genism's Phrases model was used for their implementation. Bigrams were predominantly used in this context because they were able to capture the language structure in a better way than unigrams. For example, bigrams such as contraction rate, dry cough, stay home etc. conveyed the information in a more effective way compared to analyzing words separately in the context of covid-19. The next step was to transform the tweet documents from one vector representation to another. The aim was to make the document representation more compact so that the noise in the data is reduced. In order to achieve this, term frequency- inverse document frequency (TF-IDF) was used. As TF-IDF assigned higher weights to the words which occurred frequently in one particular document but not in other documents, the chances of discovering relationships between the words and interpreting the documents in a more semantic way was increased. The transformation was then applied on the entire corpus to determine the relevance of words in all documents.

Before passing the transformed corpus to the LDA model, it was essential to decide the optimal number of topics for LDA. In order to choose the best topic number, coherence score was used. The LDA model was tested for different values of k topics. Fig 3.4 shows the coherence score across 40 topics. The graph shows that the average score reaches a local maximum as the number of topics are increased to 8 followed by a decreasing trend till topic 20 as the topic number becomes larger. After manual investigation of 8, 10 and 15 topics, the LDA model with 8 topics was chosen as it reflected diverse topics with no redundancy.

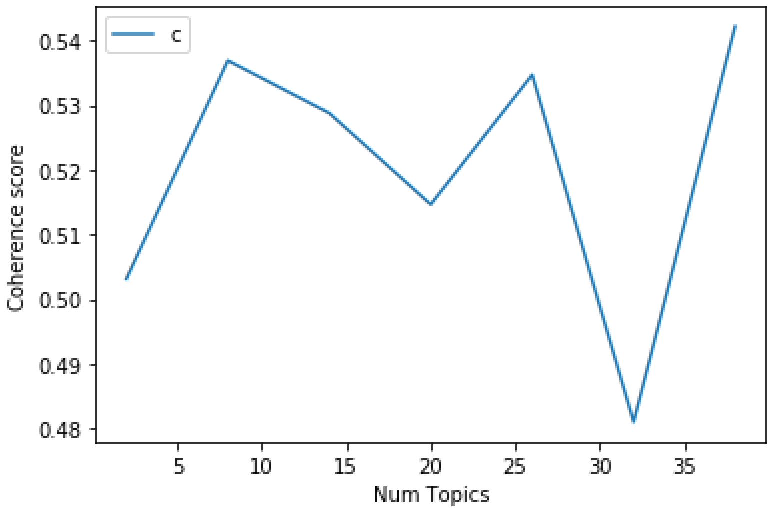


***Fig 3.4***

In order to implement the LDA model, the alpha and beta hyperparameters were optimised. The research suggests that choosing a high alpha-value leads to more topics being included in a single document whereas a low alpha value implies that a document will contain fewer dominant topics. Likewise, "a high beta-value means that each topic is likely to contain a mixture of most of the word and if the value of β is relatively small, it can result in a fine-grained decomposition of the corpus into topics that address specific research areas" (Griffiths, 2004).

The alpha and beta parameters were varied by the values of 0.01, 0.1, 0.5, and 0.8. After multiple iterations and manual inspection of topics, an alpha value of 0.01 was selected. A lower value of alpha facilitated the assignment of tweets to one or two main topics rather than multiple topics. A beta value of 0.1 was chosen. As this parameter controlled the distribution of words per topic, a lower value ensured that there were few dominant words per topic which eased the interpretation of topics. After setting the hyperparameters, the LDA model was applied on the TF-IDF corpus to extract latent topics. The final stage involved inspecting the topics and assigning labels to each topic to simplify the analysis.

The LDA model was then used to predict the distribution of topics across different geographical regions of the UK. A corpus was formed involving the tweets with region specific information and was passed to the LDA model. Each tweet document was assigned topical weights. The topic weights of documents belonging to a particular region were then averaged to obtain the regional distribution of topics. The analysis was further expanded by adding the dimension of deprivation score to the lower layer support areas (LSOA) of the UK. The aim was to investigate how certain topics are more popular or relevant among people belonging to a less deprived or most deprived area. In the final stage, MANOVA was used to test the null hypothesis that there are no differences in the topic distribution across various geographical regions as well as areas classified on the basis of IMD scores.

## 3.5 Sentiment Analysis

In order to answer the second and third research questions, a comprehensive sentiment analysis was conducted. "Sentiment analysis is a computational and natural language processing-based method that analyzes the people's sentiment, emotions and attitudes in given texts" (Beigi, Hu, Maciejewski & Liu, 2016). In this dissertation, a sentiment analysis was performed to gain daily insights on how the reactions of people towards COVID-19 changed overtime across the entire UK and Isle of Man. The second analytical procedure explored the polarity of different geographic regions in terms of their emotions. It was also analyzed how the regional emotions changed over time with respect to different events.

In order to perform emotional mining, VADER was used. “It is a lexicon and rule-based sentiment analysis tool that is specifically designed to analyze the sentiments expressed in social media data” (Hutto, Gilbert, 2014). It can classify the emotions into negative, neutral, and positive categories. It also gives a “compound score as an output which is computed by summing the valence scores of each word in the lexicon and normalized in range (-1,1), where “-1" represents most extreme negative and “1" represents most extreme positive emotion” (Hutto, Gilbert , 2014). In order to implement it, data preprocessing and model training was no required. Raw tweets were used to generate sentiment polarity because VADER performed well with emojis, slangs, and acronyms in sentences.

For this dissertation, the compound score was the main variable of interest to evaluate the changes in sentiments. Instead of classifying the tweets into 3 categories; positive, negative and neutral, and evaluating the emotions on the basis of frequency values, a continuous scale was used. The rationale behind using the continuous scale was used to show the gradual and subtle change of sentiments with respect to time and events. However, in order to facilitate understanding, if the compound score was less than 0, the text was referred to be negative. If the score was closer to 0, the tweet expressed neutral emotions and the farther it moved away from 0 i.e. greater than 0.05, the text polarity was understood as positive. Some examples of tweet sentiment results from VADER are shown in Table 3.5.

| Tweets | Sentiment Score | Sentiment Polarity |
|---|---|---|
| 0.01% is still someone’s kid, stay inside. | {'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound. 0.0} | Neutral |
| "The UK has announced a $252 million package to help @UN @WHO @wfp @IFRC @UNICEF @UNFPA @Refugees many others respond to #COVID19. #UKaid's total contribution now stands at $939 million, making □□ one of the biggest donors to the international response. | {'neg': 0.0, 'neu': 0.94, 'pos': 0.06, 'compound': 0.4019} | Positive |
| 'Interesting point of view from very poor people around the world: they can recover from CoronaVirus, but not from starvation. And alas for some that is the stark choice. #coronavirus' | {'neg': 0.144, 'neu': 0.801, 'pos': 0.055, 'compound': -0.4582} | Negative |

***Table 3.5***

## 4. Overall Results

This section covers 4 main areas of analysis. The first section discusses the most frequently used words in the tweets across the entire UK. The next section talks about the themes discussed in the topics identified as a result of LDA. The third section examines the geographic as well as IMD based topic results. The final section then discusses the results of overall as well as regional sentiment analysis and how the sentiments change over the period of 3 months.

### 4.1 Word Frequency Analysis

Fig 4.1.1 is the visual representation of the most frequently occurring words in the tweets. It reflects that people are concerned about the spread of disease and show acknowledgment of the ongoing crisis. It also suggests that many tweets might discuss the lockdown and government policies including the need to stay and work from home.

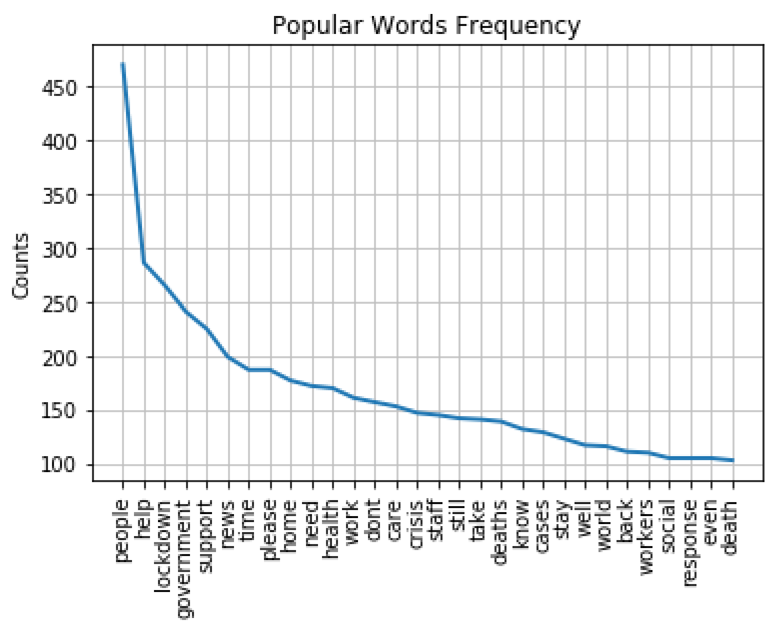


***Fig 4.1.1***

Fig 4.1.2 also shows the representative words which were widely included in the majority of the tweets. The size of the word signifies its frequency and importance. It can be observed that words like “people, help, government, news and lockdown” are very prominent. This implies that most of the people potentially talked about the impact of the disease, and the support needed to tackle the crisis situation.

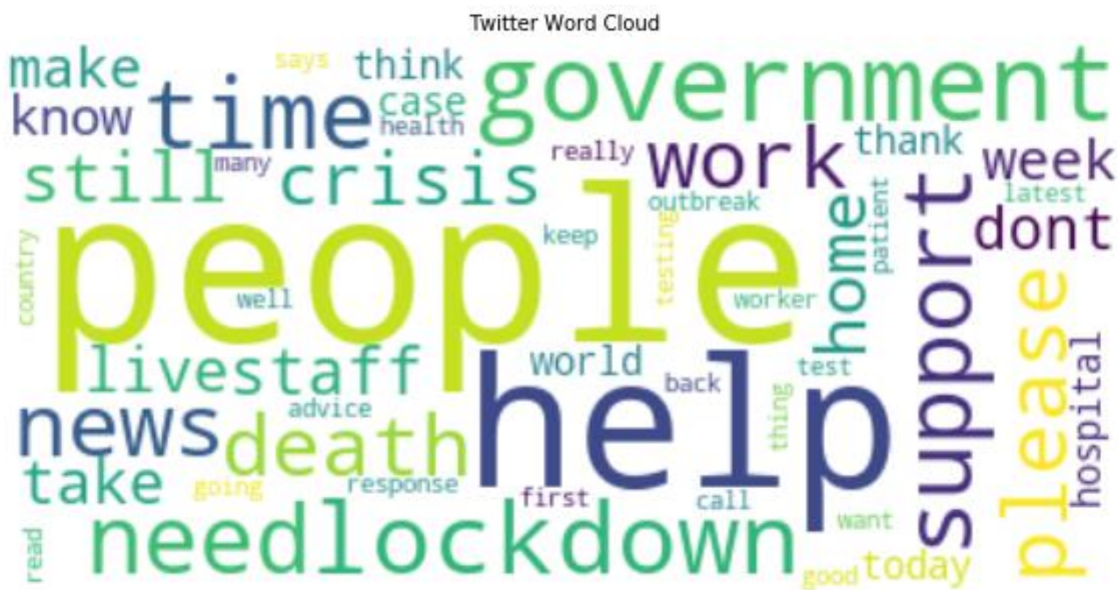


***Fig 4.1.2***

## 4.2 Topic Analysis

The implementation of LDA led to the extraction of 8 latent topics. The discovered topics reflect on facts and information about the disease, concern of people over increasing number of cases, adjustments to the life in lockdown and opinions of people on government policies. Fig 4.2.1

shows the most representative words belonging to each topic. Table 4.2.2 on the other hand, includes the theme of each topic along with its top 15 relevant words.


***Fig 4.2.1***

| **Topics** | **Keywords** |
|---|---|
| Topic 1<br>**Lifestyle Changes** | next, join, change, write, currently, disease, create, episode, rest, madness, negative, fine, transport, cancel, quarantine |
| Topic 2<br>**Facts and Figures** | death, report, case, total, fund, least, staff, rate, test, nurse, possible, survey, statistic, apply, age |
| Topic 3<br>**Situational Outcomes** | begin, fuck, critical, sale, broadband, cope, lad, enjoy, webinar, animal, confirmed_case, class, fine, date, count |

| | |
|---|---|
| Topic 4<br>**Media and Information** | thank, evening, city, scheme, step, murder, late, shut, eye, release, message, turn, charity, extend, pile, |
| Topic 5<br>**Family** | family, child  help, protect, people, love, contract, reach, experience, nation, act, together, go, crisis, carer, |
| Topic 6<br>**Government Policies & Impact** | policy, approve, wrong, people, death, work, government, die, country, good, live, political, strategy, frontline, true |
| Topic 7<br>**Emotions & Reactions** | kill, hero, wish, happy, order, urge, cost, hate, proud, maybe, moment, walk, pension, release, recovery |
| Topic 8<br>**Initiatives for Recovery** | support, school, sign, interview, campaign, kid, distance, save, walk,  useful, stay, info, show, control, worth |

***Table 4.2.2***

Topic 1 talks about the disease itself and how it has changed the way in which people are interacting with each other. It also sheds some light on how people are adjusting to change. The keywords such as “rest, quarantine, cancel” might indicate the different activities undertaken by the people during the lockdown period. The example tweets in Table 4.2.3 show how businesses, universities and other institutions have changed their way of operations following the outbreak.

| **Topic** | **Relevant Tweets** |
|---|---|
| Topic 1<br><br>Lifestyle Changes | '@McrBikeHire @Chris_Boardman □□ thank you Pavol and team at @McrBikeHire for providing free bikes; repairs to our #NHSheroes in #GreaterManchester. This means they can travel to work without fear of catching or spreading #COVID19. And get a daily dose of exercise at the same time. #ThatCounts @GmMoving □ https://t.co/CEnIvaH753'<br><br>'A big thanks @BoothsCountry team for brilliant Click & Collect service - Social Distancing; no contact handled perfectly - brings shopping out to the car, pops it in the open boot...brilliant! □□ #SocialDistancing #COVID19'<br><br>"An exciting opportunity from @ManchesterUniv ! They are presenting a number of Lockdown Lectures! Please see this website for the full schedule! https://t.co/OQDIRQjbwF We can't wait! @StWilfridsRCcol" |

***Table 4.2.3***

Topic 2 focuses on objective tweets which include information about the spread of covid-19. The keywords such as "cases, deaths, and test" might indicate that people are quoting facts and figures from different news sources or websites. The examples tweets are shown in Table 4.2.4.

| Topic | Relevant Tweets |
|---|---|
| Topic 2<br>Facts and Figures | 'SIX Premier League players and staff at three clubs test positive for coronavirus Watford Football Club confirms that three people have tested positive for the Covid virus following testing at the training ground over the past hours'<br><br>'Belarus 10 day forecast of 240 dead. This wave may last ~163 days and see 647,207 cases and 9,061 deaths. Cases double every 5.1 days □: https://t.co/EqHo3zn00r\n\n□ 17/04 #CovidBelarus<br><br>'THIS! In a country where 26,000 +/- die each year (of all ages) of flu of all variants including the coronas, , knowing how this strain is impacting the figures is vital and probably why we are not being told! #coronavirus' |

***Table 4.2.4***

Topic 3 is related to the outcomes of the pandemic. It discusses the impact of covid-19 on individuals, businesses and on a societal level. The tweets in Table 4.2.5 show the discussions of people on how the situation has impacted their lives and how they are reacting to it.

| Topic | Relevant Tweets |
|---|---|
| Topic 3<br>**Situational Outcomes** | 'How should corporate boards be responding to #Covid19? https://t.co/QozWbKXUOj'<br><br>'Coronavirus pandemic could delay graduation for thousands of nursing students |

| | |
|---|---|
| | as America faces shortage'<br><br>'Ireland 'has undergone a Deep social conditioning programme,could you imagine this thirty yrs ago? Or the pubs closed St Patrick's day? Virus or not,all hell would've broke loose,but today.........? https://t.co/J2uPoethPL' |

***Table 4.2.5***

Table 4.2.6 shows that topic 4 involves media and information release. The keywords "city, message, release" might suggest that the tweets belonging to this topic usually quote media resources and the news updates related to coronavirus. The daily press releases by government officials and the discussions surrounding it are included in this topic.

| **Topic** | **Relevant Tweets** |
|---|---|
| Topic 4<br>**Media and Information** | "□ LIVE: Environment Secretary George Eustice updates the nation on the latest #Covid19 developments.Watch it on The Telegraph's YouTube channel.<br><br>'Men much more likely to die from coronavirus - but why? |

***Table 4.2.6***

Topic 5 highlights family and community related issues. The keywords "family, child, help, protect" might suggest that the people are calling for unity in following the orders to stay at home and protect the loved ones from the disease. The manual investigation of tweets **(**Table

4.2.7**)** also shows that this topic also includes discussions relating to raising funds for the community and the people in need.

| **Topic** | **Relevant Tweets** |
|---|---|
| Topic 5<br>**Facts and Figures** | "Today we're launching @BhamCityCouncil Community Grants for community, voluntary faith groups that support Children & Families through #Covid19.<br><br>'What part of "do not leave your home for any reason" do you not understand. It's clear to the whole country that he broke the rules that he helped devise. Children died alone and afraid because families were forced to follow the rules. Think on that when you try and sleep tonight'<br><br>Coronavirus has laid bare the fact that our health isn't a lottery, and that deep inequalities within the UK are being brought to the forefront during a time of self-isolation. |

***Table 4.2.7***

Topic 6 relates to the government policies and measures taken to contain the virus. It also talks about how people have responded to government rules. The related words include "political, strategy, people, country" which may indicate the steps taken by the government to combat the disease as well as limit its negative consequences for the economy and people of the country. The representative tweets are shown in the table below.

| Topic | Relevant Tweets |
|---|---|
| Topic 6<br>**Government Policies & Impact** | 'From flip flopping to poorly managing TFL and failing to protect our key workers, 'How Khan has been playing politics during the coronavirus pandemic'. @CamillaTominey #Covid19UK #SadiqKhan #London https://t.co/mnCRi4A4fC'<br><br>'We are acutely aware of the huge impact the #coronavirus pandemic is having on people and businesses throughout the country, and have already provided £2.3 billion in support to businesses and livelihoods through this difficult time.<br><br>'UK facing deepest recession on record, says Bank of England' |

***Table 4.2.8***

Topic 7 primarily discusses the emotions of people. The analysis of tweets relating to this topic show that the reactions are not limited to the disease itself but also regarding the statements of government officials and the new rules in place. The top words include "hate, proud, happy, wish". The representative tweets (Table 4.2.9) in this topic reflect a wide range of emotions from hope and mistrust to anger and frustration.

| Topic | Relevant Tweets |
|---|---|

| | |
|---|---|
| Topic 7<br>**Emotions & Reactions** | ‘When This Is Over I wanna see everyone at Strongroom and get silly drunk and give everyone a massive hug’<br><br>'It’s ok if you’re feeling anxious, sad or stressed right now. Getting outdoors and staying active (while keeping a physical distance from others) can help. #COVID19 outbreak’<br><br>'And he has lied again! This is a one-trick pony government. It still wants to do the only trick it can do, but can’t twig it’s not the trick we now want.' |

***Table 4.2.9***

The last topic i.e. topic 8 discusses the government's roadmap to economic recovery which leads to people voicing concerns over the government actions. The representative tweets also show different initiatives taken by individuals and organizations to support the community, businesses and those in need.

| **Topic** | **Relevant Tweets** |
|---|---|
| Topic 8<br>**Initiatives for Recovery** | 'Businesses in #WinchesterUK please visit the @bizsupportgovuk website for details of the business support schemes available from the Government to help with the impact of #Coronavirus on your business<br><br>"Getting back to normality quicker than the competition is not the challenge you face. Normal got us into this mess, and normal won't get us out of it. To stop the slide into irrelevance we need to stop asking, “How can we rebuild what Corona |

| | |
|---|---|
| | took away?"<br><br>"At @scottishmusic we're building a new resource for musicians, freelancers & music businesses- links to sources of practical; financial help, work from home, health/wellbeing tips & ways to help support our music sector. |

***Table 4.2.10***

## 4.3 Regional Analysis Results

This section is divided into two parts. The first section talks about the geographical differences in the discussions of topics, and the second section talks about the differences in regions based on index of multiple deprivation.

### 4.3.1 Geographical Results

When the distribution of topics across all regions was analyzed, it was observed that topic 5, 6 and 8 were the most popular areas of discussion among all the regions. This implied that the discussions across the country mainly revolved around people's concern for their family and the impact of government actions on their lives. 55% of the discussions in all regions consisted of these 3 main topics.

Another main observation (Fig 4.3.1) showed that the East of England and Isle of Man discussed family and community more than the other regions.

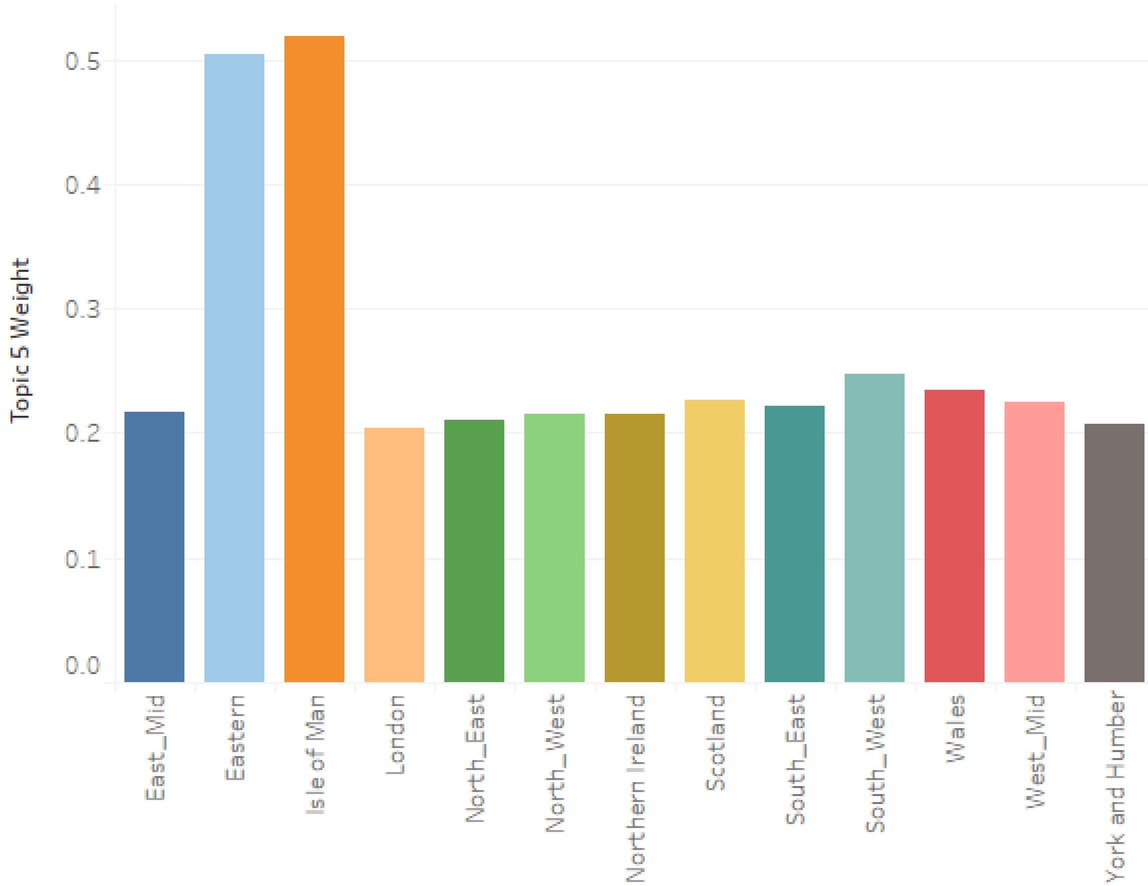


***Fig 4.3.1***

The results also showed that these two regions not only had the highest proportion of topic 5 discussion as compared to other regions, but this topic also had the highest weight compared to other topics as shown in Fig 4.3.2.

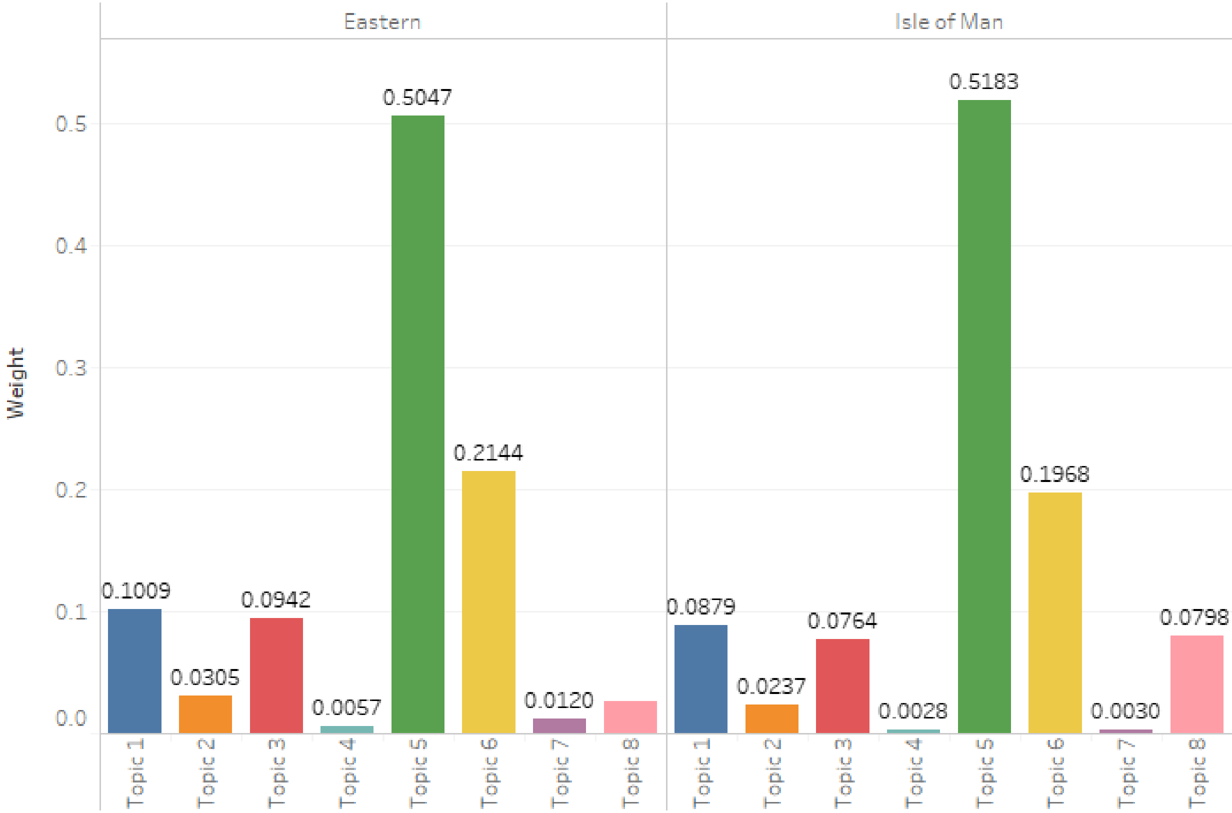


***Fig 4.3.2***

On the other hand, the proportion of each topic discussed across the remaining 11 regions of the UK was almost the same. Table 4.3.3 highlights that the inter-region variation relative to the

| | Topic 1 | Topic 2 | Topic 3 | Topic 4 | Topic 5 | Topic 6 | Topic 7 | Topic 8 |
|---|---|---|---|---|---|---|---|---|
| **count** | 13.000000 | 13.000000 | 13.000000 | 13.000000 | 13.000000 | 13.000000 | 13.000000 | 13.000000 |
| **mean** | 0.096659 | 0.095477 | 0.066292 | 0.060128 | 0.265292 | 0.196749 | 0.075344 | 0.133557 |
| **std** | 0.009423 | 0.030976 | 0.012042 | 0.026968 | 0.109901 | 0.017078 | 0.031659 | 0.037976 |
| **min** | 0.078355 | 0.023679 | 0.048352 | 0.002809 | 0.204061 | 0.169186 | 0.002957 | 0.026618 |
| **25%** | 0.091501 | 0.100130 | 0.059937 | 0.055651 | 0.215149 | 0.186394 | 0.074846 | 0.135572 |
| **50%** | 0.097528 | 0.108418 | 0.062198 | 0.066985 | 0.222262 | 0.196828 | 0.086434 | 0.149495 |
| **75%** | 0.104793 | 0.110270 | 0.074469 | 0.073425 | 0.235168 | 0.202902 | 0.091503 | 0.153226 |
| **max** | 0.110376 | 0.121968 | 0.094197 | 0.092833 | 0.518324 | 0.235271 | 0.105288 | 0.157142 |

***Fig 4.3.3***

topics is not significant. The highest variation is noted in Topic 5 whereas Topic 1 had the lowest difference in weights across regions.

In order to confirm whether there are differences in the weights of topics across regions, MANOVA was implemented. The null hypothesis was that there are no differences in the distribution of topics across regions. The test failed to reject the null hypothesis at the alpha value of 0.05 as well as 0.01. This implies that there are no regional differences in the topic distribution.

```
                  Multivariate linear model
==============================================================

--------------------------------------------------------------
       Intercept         Value   Num DF Den DF F Value  Pr > F
--------------------------------------------------------------
          Wilks' lambda   0.0012 7.0000 5.0000 587.3156 0.0000
         Pillai's trace   0.9988 7.0000 5.0000 587.3156 0.0000
 Hotelling-Lawley trace 822.2419 7.0000 5.0000 587.3156 0.0000
    Roy's greatest root 822.2419 7.0000 5.0000 587.3156 0.0000
--------------------------------------------------------------

--------------------------------------------------------------
           Regions         Value  Num DF Den DF F Value Pr > F
--------------------------------------------------------------
             Wilks' lambda 0.4237 7.0000 5.0000  0.9714 0.5321
            Pillai's trace 0.5763 7.0000 5.0000  0.9714 0.5321
    Hotelling-Lawley trace 1.3599 7.0000 5.0000  0.9714 0.5321
       Roy's greatest root 1.3599 7.0000 5.0000  0.9714 0.5321
==============================================================
```

## 4.3.2 IMD Results

In order to inspect the differences across regions based on the index of multiple deprivation, the topic weights across eight deciles were analyzed. The areas in decile 2 were classified as the most deprived regions whereas the areas part of decile 9 comprised the most well off locations.

The results showed that there is no variation between areas belonging to different deciles. The MANOVA test also confirmed that there is no difference in the distribution of topics across deciles. An interesting observation however was that topic 7 - emotions, was relatively discussed more in deprived areas. 9% of the discussion of decile 2 was focused on this topic as compared to 7% of decile 9. Although the difference does not seem stark, it might have significant implications if the tweets are analyzed in greater depth.

The analysis also showed that topic 6 which talked about government policies and its impacts was equally discussed between the most and least deprived areas. Fig 4.3.21 shows that it formed around 22% of the discussions of areas belonging to decile 2 and 4. On the other hand, decile 9 also had a significant discussion of around 21% around this topic. Similarly, topic 8 which was related to government plans on economic recovery, and initiatives by different organizations to support people was also discussed relatively more in deprived areas i.e. decile 2 and 3 as compared to decile 8 and 9 (Fig 4.3. 22).

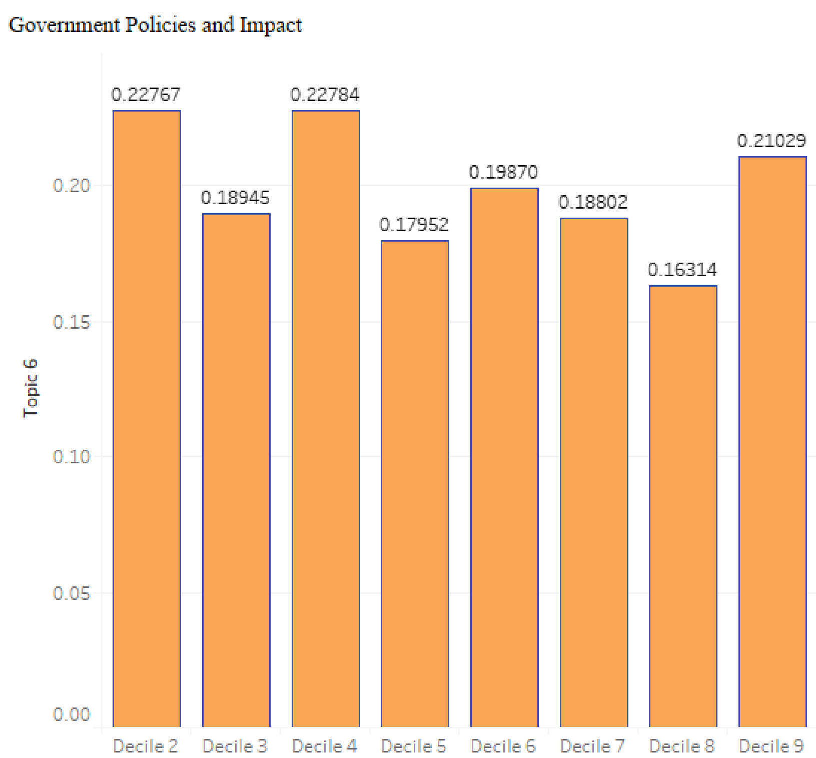


*Fig 4.3.21*

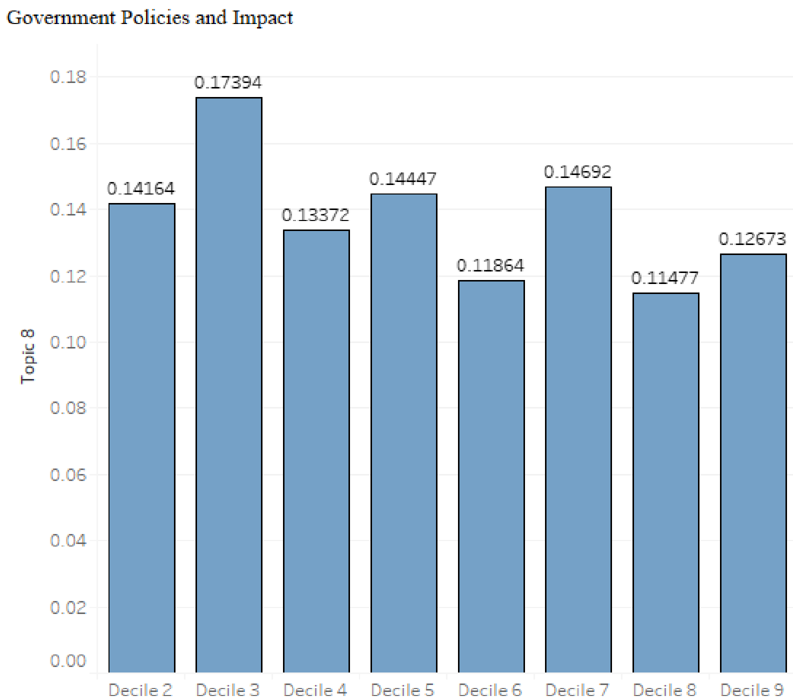


***Fig 4.3.22***

## 4.4 Sentiment Analysis Results

The sentiment analysis showed that there is variation in the emotional reactions of people over time. (Fig 4.4.1) shows that the compound score was the highest in the month of March, and it fell gradually in the month of April. It was the lowest in the month of May. However, the average sentiment still remained relatively positive. This indicated that the sentiments may change over time.

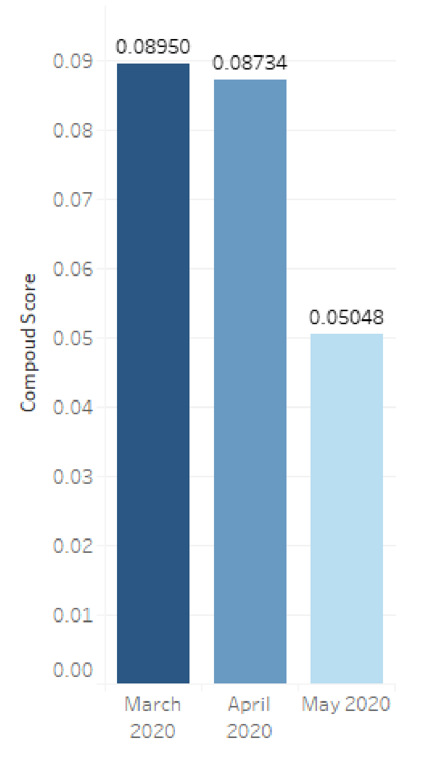


***Fig 4.4.1***

Fig 4.4.2 further illustrates the weekly average sentiment of tweets over three months. It shows that on average, the daily sentiments of people are positive. However, it does show a few dips in the last week of March and early May. The daily sentiment analysis suggested that the average sentiment score became negative in the last week of March specifically 28th March, 2020. It can also be observed that the second wave of negative sentiments was around mid-May. The highest positive sentiment observed was on 16th April.

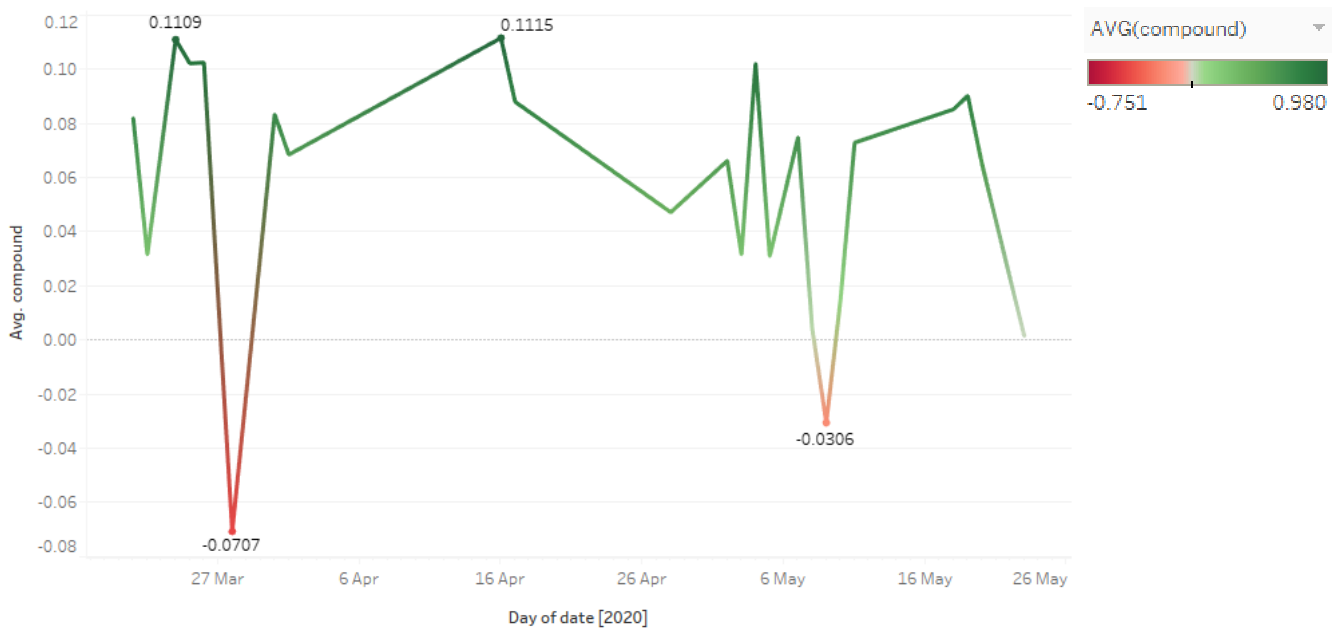


***Fig 4.4.2***

## 4.4.1 Geographic Sentiment Analysis Results

This section focuses on answering the last research question which tends to analyze if there are any regional differences in the emotions of people. In the first part of the study, the results revealed that there are no significant differences in the discussion of topics across regions. To further investigate any differences, the sentiments of the geographic regions were considered. It was observed that over the course of 3 months, the intensity of emotions of people varied across regions. Fig 4.4.3 shows that the East of England on average was dominated by negative sentiments. On the other hand, Isle of Man predominantly had the highest positive emotions on the scale. This is an interesting observation as Isle of Man is not governed by the UK government but is in close proximity to the UK. The figure also highlights that in the UK, Northern Ireland ranks the highest in terms of positive sentiments.


***Fig 4.4.3***

### 4.4.2 Monthly Geographic Sentiment Analysis

The monthly analysis (Fig 4.4.4) shows that the East of England experienced negative emotions in the months of March and May, however, the emotions were relatively positive in the month of May. On the other hand, the remaining regions in the UK, gradually converged towards neutral compound score over the course of months. This could be because as the number of cases increased overtime and the social and economic impacts were felt overtime which led to the change in emotions. In case of Isle of Man, it was clearly observed that the sentiments of people improved overtime.


***Fig 4.4.4***

Fig 4.4.5 shows the change in daily sentiments across regions. An interesting observation was that the daily sentiments across regions were influenced by specific events. All the regions in the UK, except West Midlands faced negative emotions on 28th March when the UK death toll rose to 1000. Similarly, the majority of regions exhibited negative emotions on 9th May as well. This again reinforced the idea that the sentiments stemmed from major events. All the regions followed the same distribution of emotions apart from East of England which predominantly displayed negative emotions.

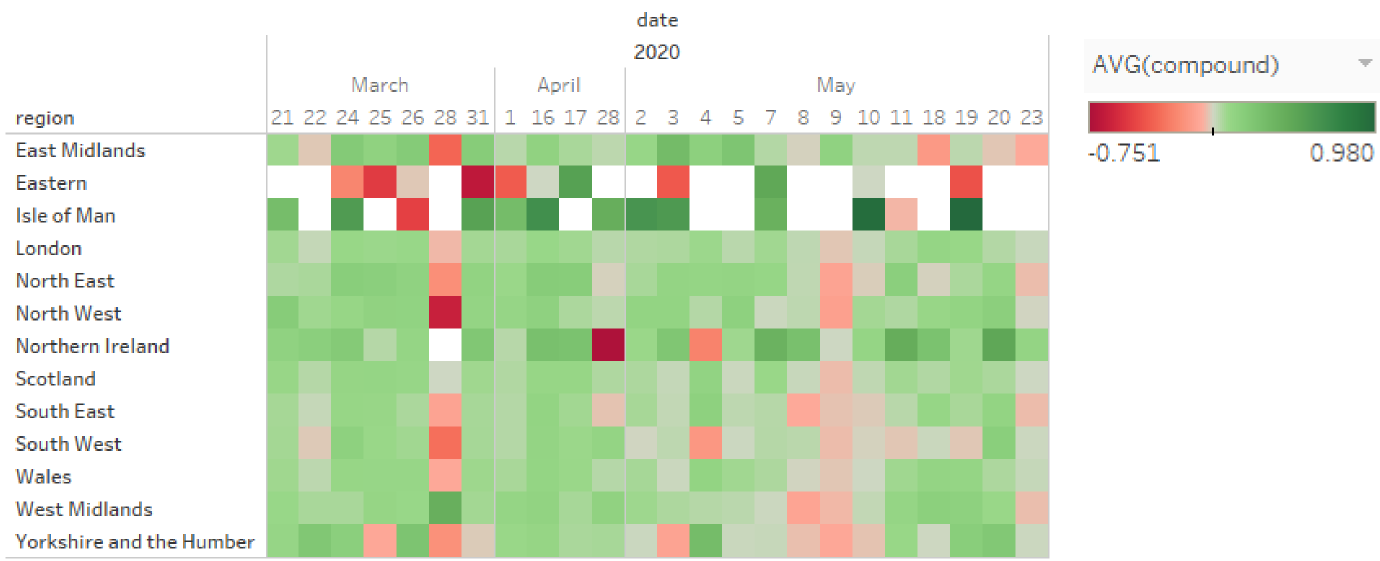


***Fig 4.4.5***

## 5. Discussion

The findings in this dissertation facilitated an understanding of public thoughts and reactions to the outbreak of COVID-19. It shed light on several interesting regional similarities in terms of themes discussed, combined with the differences in sentiment polarity, which contributed to the understanding why certain areas might react differently on same topics.

The results indicated that all the regions in the UK as well as Isle of Man concentrated more than half of their discussion around 3 main topics which were related to the government policies and the impacts of their decisions on personal and social lives of people. The reason for the prevalence of these topics could be explained by the timeframe of the study. During the months of March, April and May, the seriousness of the disease was realized by the people of the UK, and the government announcements on lockdown and containment of disease were widely discussed.

According to the results, East of England's major chunk of discussion was related to the impact on personal life as a result of government policies. This was the highest among all the other regions. The Opinions and Lifestyle Survey by the Office of National Statistics suggested that during the month of April, the people in East England struggled the most in terms of acquiring groceries, medication and essentials (*Coronavirus and the social impacts,* 2020). Apart from this, only one in three workers in the East of England worked from home because of the COVID-19 pandemic which was the lowest among all the regions other than East Midlands (*Coronavirus and the social impacts*, 2020). This could explain the reason of concern for most of the people belonging to this region as most of them had to travel to work which could have increased their chances of contraction. It could also explain why this region specifically had the highest negative sentiment across the UK.

The results revealed that the distribution of topics across regions was the same which rejected the initial hypothesis that there are differences in the responses of people towards covid-19. Although there were some minor differences which were spotted, they were not significant. In order to further confirm this analysis, the index of multiple deprivation was used as a means to classify the regions. It revealed that despite the similar distribution of topics across deciles, there were a few topics which indicated some differences between the less deprived and more deprived regions. For example, it was observed that people belonging to more deprived areas talked more about emotions relative to the well-off areas. The reason for the differences could be that people residing in poor areas might be hit harder in terms of financial distress, medical facilities and living standard in general as compared to well off areas. These people might have communicated the emotional distress through twitter which is reflected in the analysis. Similarly, the deprived regions are more likely to suffer from economic strain which could reflect their interest in support schemes. This might explain the rationale behind a greater proportion of discussions related to topic 8 i.e. initiatives for recovery.

It is also important to note that a few topics such as topic 6 i.e. government policies and its implications were equally discussed across deciles, however, the context of discussions seemed to be extremely different. Table 5.1 illustrates some example tweets from decile 2 and decile 9 which clearly show that the discussions related to the government were carried out in relatively different contexts. In decile 2, the people primarily talked about the government policies in context of how they were affecting the deprived people, and what could be done to minimize the negative consequences for those who were already economically challenged. On the other hand, the example tweets from decile 9 show that people used government policies to invoke discussions on national level issues i.e. economy as a whole. This could be because their lives might not have been as significantly impacted as those who were already struggling even before the pandemic.

| **Deciles** | **Relevant Tweets** |
| --- | --- |
| Decile 2 | 'Some research indicates that poorer students will be further disadvantaged by online learning'<br>'The coronavirus rescue packages should be used to rescue the needy from the coronavirus Nothing more nothing less' |
| Decile 9 | 'Govt announces tariff on new imported trucks from Jan More crippling cost for hauliers moving essential goods will hit recovery warned'<br><br>'IMPORTANT if your business is waiting to hear from Cornwall Council about the k Small Business Grant Fund or the Retail Hospitality and Leisure |

| | |
|---|---|
| | Grant up to k then please see the important thread below Pls RT'<br><br>'V bad situation facing universities gtgt Education hit hardest as coronavirus batters UK economy' |

***Table 5.1***

The sentiment analysis of COVID-19 further contributed to the understanding of the dynamics of online users' concerns and emotions during the pandemic, which helped answer the second research question. The results suggested that the sentiments change overtime, and usually on a daily basis as well, depending on several factors. An interesting observation showed that the overall sentiments were negative on 28th March. This was potentially because on 28th March, the number of corona cases in the UK reached 1000 (*Number of UK coronavirus deaths*, 2020). This implied that during this period, the negative emotions such as concern, fear and anxiety prevailed which was reflected in the results. Similarly, the negative emotion prevailed in mid-May. This was the time when the government started to consider easing the lockdown. The government changed the slogan "Stay Home" to "Stay Alert" (*8 tweets to explain #StayAlert, The Independent, 2020*). The analysis showed that this was not well received by the twitter users. Table 5.2 exhibits example tweets from the date which clearly shows the emotional state of people.

| **Date** | **Relevant Tweets** |
| --- | --- |
| 9th May | '100s if not 1000s of people, many of whom survived WWII have died from #COVID19 because of @BorisJohnson & his useless Govt. meanwhile this lot appear hell bent on partying whilst grieving families can't even go to their loved one's funerals□ #VEDay75 |
| 12th May | "A further 627 deaths. That's worrying. Best to lockdown again. Boris needs to be careful with his decisions. The numbers are up again! Excess deaths will be huge too. We're now over 32,000 deaths. It will be wise to #StayHome advice rather than #StayAlert seriously! #coronavirus"<br><br>"All those screaming for Britain to ease its lockdown for economic reasons need to take a long hard look at our real #coronavirus 'excess death' toll of over 60,000 people. This is an unfurling catastrophe & we're no nearer beating the virus now than we were two months ago." |

***Table 5.2***

The observation of negative sentiments on particular dates suggested that the emotions of many people could be driven by the government announcements and news to a certain extent. As the daily lives of people are heavily influenced by government decisions during the pandemic, it could be inferred that the information via news and media could be a very important factor in controlling the emotions of people. This is also supported by a study conducted by Zou et al

(2020) which showed correlations between the public responses towards the COVID-19 on Twitter and the government policies related to the COVID-19 in the United Kingdom and the United States.

Another instance was 16th April when the highest positive sentiments were recorded. Further investigation showed that this could potentially be driven by two main events. Foreign Secretary Dominic Raab, and Chancellor Rishi Sunak joined the nation in paying tribute to all the doctors, nurses and all staff  on the frontline of the fight against coronavirus" (*UK thanks key coronavirus workers*, 2020). Apart from this, Tom Moore hit his 12 million pounds mark for gathering donations for the NHS which was applauded by the general public. Table 5.3 shows the sample tweets from the data which contributed to the positive sentiment score. This again reinforces the idea that the sentiments of people during the pandemic are quite dynamic to changes in events, news and government policies.

| **Date** | **Relevant Tweets** |
|---|---|
| 16th April | "£12m raised for NHS to fight #coronavirus - Captain Tom Moore completes final lap for NHS as donations soar past £12m! Big respect! □□<br><br>‘Tonight @RishiSunak and I joined the nation in paying tribute to all the heroes on the frontline of the fight against #coronavirus – especially our brave carers and NHS workers. #ClapForOurCarers #ThankYouNHS’ |

***Table 5.3***

The results also showed that two regions can have the same distribution of topics along with similar topic weights however, their sentiments might be completely different. In order to answer the first research question, investigation of the regional differences from the lens of sentiment changes was carried out. The inspection revealed some interesting information. It was analyzed that East of England and Isle of Man have similar topic weights and follow similar distribution, however, the sentiments over time are completely opposite. While East of England primarily displayed negative sentiments, Isle of Man had the highest average positive sentiment. It must be noted that the Isle of Man is not governed by the UK government but is geographically closely located to the UK. There could be several reasons for the difference in sentiments. As previously discussed, east of England faced several challenges due to the pandemic which could have led to the generation of predominantly negative sentiments. On the other hand, the sentiments of the Isle of Man improved overtime. This could be attributed to the fact that the number of cases in the region were 336 with only 24 deaths in the given months (*Isle of Man Coronavirus,* 2020). The sentiments improved in May because the active cases reduced in the month of May (Fig 5.4).

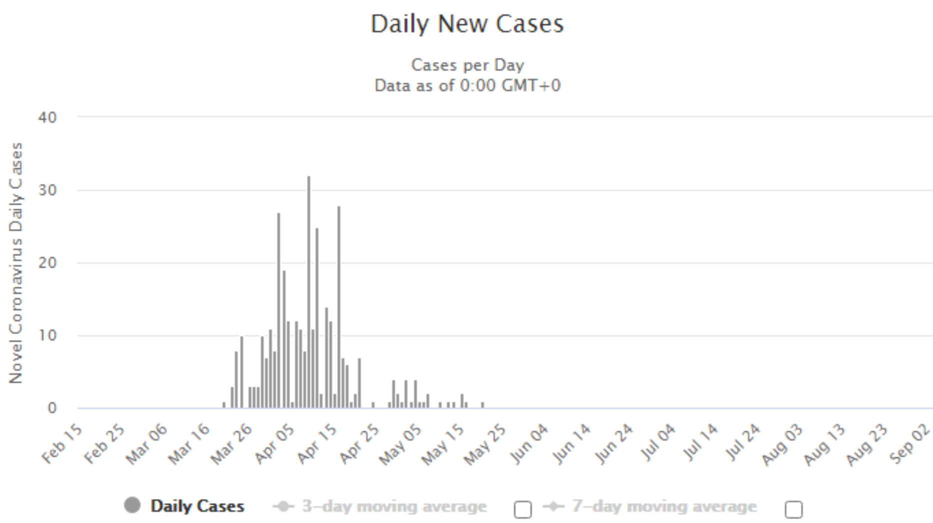


***Fig 5.4 - Worldometer (Isle of Man - Coronavirus Cases 2020)***

Furthermore, as the government policies and their implications are different, this could be one of the biggest reasons for the difference in sentiments. This leads to the identification of a very important link between the three research questions and the results. The results reflect that the most dominant topics of discussions revolve around the government strategies and related news which in turn also influence the sentiments of the people. The regional topic differences might seem insignificant i.e. majority of the areas discuss the same topics however, in many cases their experience of the crisis is very different. The difference in the experiences leads to variation in sentiments overtime. Same policies and information can be interpreted differently across different groups. This also explains the theory of selective attention in which people tune out unimportant details and focus on what matters to them (Dayan et al., 2000). For example, deprived areas will focus more on policies related to support schemes than well off areas which might pay more attention to economic issues.

## 6. Conclusion

In this dissertation, the exploratory results of topic modeling and sentiment analysis on COVID-19 across different regions of the UK were presented. The importance of different topics across different areas was compared. It was revealed that despite similar discussions across the entire UK, the context and premise used to discuss a theme varied among areas. The results showed that the majority of the discussions stem from government plans and intervention to combat the disease. This also influenced the sentiments of people over time as they were directly or indirectly impacted by the policies.

The real time dynamic change in responses of people following major news and policy changes has several important implications. In future, the government could use twitter user's reactions about a certain topic before implementing the policies to make sure that the policy is generally well received by the masses specifically those who are directly impacted by it.

Despite the similar distribution of topics across regions, this study pinpointed a few differences among sentiments of people across regions. It was important to understand that regional patterns can be hard to obtain because even neighboring local authorities in the same region can have entirely different experiences to the pandemic. This study provided an understanding that the tweets could also be used to inform the local health authorities about the real and perceived concerns of the people located in a particular region. This could not only improve the dissemination of relevant information on LSOA level but also help the local institutes develop their strategies accordingly and provide better guidance to the public. It might also prevent the spread of negative emotions and provide an opportunity to deal with the local crisis such as access to necessities, lost jobs etc. in an effective and timely manner.

The limitation of this work is that there were a limited number of tweets with location information. This did not allow the use of the whole dataset while conducting regional analysis. Apart from this, the majority of the tweets did not have precise location information i.e. information regarding the Lower Layer Super Output Areas. This prevented the study of regions on a granular level. Due to lack of information regarding the LSOA, the tweets belonging to multiple LSOA with different IMD scores were grouped together and the average IMD score was used. This would have potentially led to loss of some insights. This is because different areas in the same region can have completely different IMD scores. For example, one area in London can belong to decile 9 (least deprived area) and the other can belong to decile 2 (most deprived area).

When the LSOA information was missing, the tweet was assigned an average score of the region it belonged to. In order to make effective use of IMD scores, it is important that geo-tagged tweets have precise location information while conducting future studies in this domain.

Apart from this, due to time and computational constraints, only a small sample of tweets were used to construct the LDA model. A larger number of observations could be used to train the models in future studies to improve the model performance. Furthermore, in case of sentiment analysis, a lexicon based approach was used which is a rather simplistic technique for emotion is mining. However, during a crisis such as covid-19 pandemic, it might be difficult to gauge the complex emotions of people. In order to classify the sentiments, a machine learning approach could be used in future.

Another limitation was the missing data in the month of April. The data was only available for 4 days of April. It involved tweets from a couple of days in the first week of April and a couple of days from mid-April. This could mean that tweets about several events and their relevant impact might have been missed. In future study, this could be avoided by ensuring consistency of tweets' volume over the period of study.

It must also be remembered that twitter users are not representative of the whole population. The tweets are only indicative of the views of a proportion of online users. Other popular mediums of social media such as Facebook, Instagram etc. can also be used in future studies to have a holistic picture of how people have reacted to COVID-19. However, twitter does remain a valuable source of gauging the opinions of people related to the disease and it has been used in the past for conducting several studies related to the pandemics.

This dissertation sets the baseline for future studies to identify regional differences in the UK after conducting granular studies of the LSOA. The twitter data can be used in conjunction with government statistics related to local cases and death statistics to unpack local contexts. Apart from this, tweets identifiable with LSOA can be accurately matched against their IMD scores to generate a better picture of the social responses of people towards covid-19.  Future studies could also use event specific local news and government plans along with the tweets to gauge the change in sentiments of people over time.

# 7. References


1. Achrekar, H., Gandhe, A., Lazarus, R., Yu, S.-H. and Liu, B. 2012. Twitter Improves Seasonal Influenza Prediction. *In*: *Healthinf.*, pp.61–70.
2. Aramaki, E., Maskawa, S. and Morita, M. n.d. Twitter catches the flu: detecting influenza epidemics using Twitter. , p.9.
3. Beigi, G., Hu, X., Maciejewski, R., & Liu, H. (2016). An overview of sentiment analysis in social media and its applications in disaster relief. In Sentiment analysis and ontology engineering (pp. 313-340). Springer, Cham.
4. Blei, D.M. and Ng, A.Y. n.d. Latent Dirichlet Allocation David M. Blei, Andrew Y. Ng and Michael I. Jordan Variational Extensions to EM and Multinomial PCA Wray Buntine Summary by Jaakko Peltonen, Feb 17, 2004.
5. Chew, C. and Eysenbach, G. 2010. Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1 Outbreak. *PLOS ONE*. **5**(11), p.e14118.
6. Coronavirus and the social impacts on the countries and regions of Britain - Office for National Statistics. [Accessed 10 September 2020]. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/healthandwellbeing/bulletins/coronavirusandthesocialimpactsonthecountriesandregionsofbritain/april2020.
7. Das, S. and Dutta, A. 2020. Characterizing public emotions and sentiments in COVID-19 environment: A case study of India. *Journal of Human Behavior in the Social Environment*. **0**(0), pp.1–14.

8. Daughton, A.R. and Paul, M.J. 2019. Identifying Protective Health Behaviors on Twitter: Observational Study of Travel Advisories and Zika Virus. *Journal of Medical Internet Research*. **21**(5), p.e13090.
9. Dayan, P., Kakade, S. and Montague, P.R. 2000. Learning and selective attention. *Nature neuroscience*. **3**(11), pp.1218–1223.
10. English indices of deprivation 2019. *GOV.UK*. [Online]. [Accessed 7 September 2020]. Available from: https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019.
11. Feng, Y. and Zhou, W. 2020. Is Working From Home The New Norm? An Observational Study Based on a Large Geo-tagged COVID-19 Twitter Dataset. *arXiv:2006.08581 [cs]*.
12. Gilbert, C.H.E. and Hutto, E. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text *In*: *Eighth International Conference on Weblogs and Social Media (ICWSM-14). Available at (20/04/16) http://comp. social. gatech. edu/papers/icwsm14. vader. hutto. pdf*., p.82.
13. Griffiths, T.L. and Steyvers, M. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*. **101**(suppl 1), pp.5228–5235.
14. Huang, X., Li, Z., Jiang, Y., Li, X. and Porter, D. 2020. Twitter, human mobility, and COVID-19. *arXiv:2007.01100 [cs, stat]*.
15. 'If you need 8 tweets to explain #StayAlert then it probably isn't clear enough': No 10 mocked over 'meaningless' Twitter thread | The Independent | The Independent. [Accessed 10 September 2020]. Available from: https://www.independent.co.uk/news/uk/home-news/coronavirus-number-10-boris-johnson-twitter-thread-stay-alert-response-a9508561.html.

16. Isle of Man Government - Constitution. [Accessed 7 September 2020]. Available from: https://www.gov.im/about-the-government/departments/cabinet-office/external-relations/constitution/.

17. Isle of Man Coronavirus: 337 Cases and 24 Deaths - Worldometer. [Accessed 10 September 2020]. Available from: https://www.worldometers.info/coronavirus/country/isle-of-man/.

18. Jain, V.K. and Kumar, S. 2015. An Effective Approach to Track Levels of Influenza-A (H1N1) Pandemic in India Using Twitter. *Procedia Computer Science*. **70**, pp.801–807.

19. Jang, H., Rempel, E., Carenini, G. and Janjua, N. 2020. Exploratory Analysis of COVID-19 Related Tweets in North America to Inform Public Health Institutes. *arXiv:2007.02452 [cs]*.

20. Kim, E.H.-J., Jeong, Y.K., Kim, Y., Kang, K.Y. and Song, M. 2015. Topic-based content and sentiment analysis of Ebola virus on Twitter and in the news: *Journal of Information Science*.

21. Kim, H., Jang, S.M., Kim, S.-H. and Wan, A. 2018. Evaluating Sampling Methods for Content Analysis of Twitter Data. *Social Media + Society*. **4**(2), p.2056305118772836.

22. Kostkova, P., Szomszor, M. and St. Louis, C. 2014. #swineflu: The Use of Twitter as an Early Warning and Risk Communication Tool in the 2009 Swine Flu Pandemic. *ACM Transactions on Management Information Systems*. **5**(2), pp.1–25.

23. Kwon, J., Grady, C., Feliciano, J.T. and Fodeh, S.J. 2020. Defining Facets of Social Distancing during the COVID-19 Pandemic: Twitter Analysis. *medRxiv*., 2020.04.26.20080937.

24. Nagar, R., Yuan, Q., Freifeld, C.C., Santillana, M., Nojima, A., Chunara, R. and Brownstein, J.S. 2014. A Case Study of the New York City 2012-2013 Influenza Season With Daily Geocoded Twitter Data From Temporal and Spatiotemporal Perspectives. *Journal of Medical Internet Research*. **16**(10), p.e236.

25. Number of UK coronavirus deaths rises above 1,000. *BBC News*. [Online]. [Accessed 10 September 2020]. Available from: https://www.bbc.com/news/uk-52077997.

26. Thelwall, M. and Thelwall, S. 2020. Covid-19 Tweeting in English: Gender Differences. *arXiv:2003.11090 [cs]*.

27. UK thanks key coronavirus workers in nationwide Clap For Our Carers. *ITV News*. [Online]. [Accessed 10 September 2020]. Available from: https://www.itv.com/news/2020-04-16/uk-clap-for-our-carers-nhs-coronavirus-lockdown.

28. WHO | Novel Coronavirus – China. WHO. [Online]. [Accessed 9 September 2020]. Available from: http://www.who.int/csr/don/12-january-2020-novel-coronavirus-china/en/.

29. Xue, J., Chen, J., Chen, C., Zheng, C. and Zhu, T. 2020. Machine learning on Big Data from Twitter to understand public reactions to COVID-19. *arXiv:2005.08817 [cs]*.

30. Yin, H., Yang, S. and Li, J. 2020. Detecting Topic and Sentiment Dynamics Due to COVID-19 Pandemic Using Social Media. *arXiv:2007.02304 [cs]*.

31. Zou, C., Wang, X., Xie, Z. and Li, D. 2020. Public Reactions towards the COVID-19 Pandemic on Twitter in the United Kingdom and the United States. *medRxiv*., 2020.07.25.20162024.