An Exploratory Study of COVID-19 Information on Twitter in the Greater Region

Ninghan Chen\textsuperscript{a}, Zhiqiang Zhong\textsuperscript{a}, Jun Pang\textsuperscript{a,b,*}

\textsuperscript{a}Faculty of Sciences, Technology and Medicine, University of Luxembourg, L-4364 Esch-sur-Alzette, Luxembourg
\textsuperscript{b}Interdisciplinary Centre for Security, Reliability and Trust, University of Luxembourg, L-4364 Esch-sur-Alzette, Luxembourg

Abstract

The outbreak of the Coronavirus disease (COVID-19) leads to an outbreak of pandemic information in major online social networks (OSNs). In the constantly changing situation, OSNs are becoming a critical conduit for people in expressing opinions and seek up-to-the-minute information. Thus, social behaviour on OSNs may become a predictor or reflection of reality. This paper aims to study the social behaviour of the public in the Greater Region (GR) and related countries based on Twitter information with machine learning and representation learning methods. We find that tweets volume only can be a predictor of outbreaks in a particular period of the pandemic. Moreover, we map out the evolution of public behaviour in each country from 2020/01/22 to 2020/06/05, figuring out the main differences in public behaviour between GR and related countries. Finally, we conclude that tweets volume of anti-contiguous measures may affect the effeteness of the government policy.

Keywords: COVID-19, online social media, spatio-temporal analysis, topic modelling, pandemic information, human behaviour

\textsuperscript{*}Corresponding author

Email addresses: ninghan.chen@uni.lu (Ninghan Chen), zhiqang.zhong@uni.lu (Zhiqiang Zhong), jun.pang@uni.lu (Jun Pang)
1. Introduction

On January 20th 2020, the World Health Organisation (WHO) declared a global health emergency over the COVID-19 outbreak. Later, on March 12th 2020, WHO announced the COVID-19 outbreak as a pandemic. The outbreak of the COVID-19 Coronavirus leads to an outbreak of pandemic information in major online social networks (OSNs), including Twitter, Facebook, Instagram, and YouTube. In the middle of a massive COVID-19 outbreak and constantly changing situation, OSNs are becoming a critical conduit for people to seek up-to-the-minute and local information. Moreover, due to physical isolation and social distancing, people spend much more time on OSNs — engaging in expressing opinions, encouraging others, openly lambasting mismanagement, and voicing vitriol, etc. On the one hand, social behaviour on OSNs may become a predictor or reflection of reality. On the other hand, the related information diffusion over OSNs can strongly influence peoples behaviour, and thus have an impact on the effectiveness of control and protective measures deployed by the governments.

There is a growing body of research that links OSNs activities to COVID-19. Some existing results have already shown that OSNs conversations can be a leading indicator of COVID-19 cases, discussions on OSNs can be categorised into multiple specific topics and OSNs may help to design more efficient pandemic models for social behaviour and to implement more responsive government communication strategies. However, there are three main problems with the existing research. First of all, researches with geographic data are conducted through rough processing of the location information. Second, the current topic modelling study is mostly focused on a relatively long period (weeks or months), which does not provide a precise representation of how topics change from day to day and the existing studies mainly focus on general characteristics of user behaviour. Third, shared information on OSNs at a global or country level are coarser in terms of geographic dividing. When analysing the COVID-19 information on Twitter by geographic locations, it cannot be ignored that COVID-19 performs in a way which is inescapably spatial. Existing studies have shown that whether it is Spanish flu, Ebola or COVID-19, the geographic changing of contagion reveals that economic, logistical, and flow-oriented relationality is intrinsically linked to the transmission pattern

1https://bit.ly/39EMOtY
of infectious diseases [15]. Hence, research results focusing only on political sovereign states would be biased.

To fill this gap, we innovatively introduce relational urbanisation, the efflorescent idea from human geographic, that city orientation is influenced by the network of materials, capital, information and culture, and reinforced by financialised capitalism [16] to determine the study area, and the Greater Region (GR), a typical relational urbanisation product with Luxembourg at its centre and including adjacent regions of Belgium, Germany and France (i.e., Wallonia, Saarland, Lorraine, Rhineland-Palatinate and the German-speaking Community of Belgium) is chosen as the representative in our case study and we define the neighbouring countries mentioned above as the related countries of GR.

GR has the highest number of cross-border commuters in Europe, with approximately 250,000 commuters per day[2]. This makes GR a particular and classic example: virus spreads with high mobility, while the whole business model in GR requires a large number of cross-border workers to sustain. With the implementation of a set of policies including border closures and the progression of the pandemic, the people living in the GR area are affected in economy, daily life, travel, and other aspects. This study is divided into two parts (Sections 4 and 5) to address following three main questions:

RQ1 Can the Twitter posts volume be a predictor of COVID-19 daily cases during a long-term period in GR and its related countries?

RQ2 Whether there are certain patterns of social behaviour on OSNs at different periods of the pandemic and how GR, as a relational urbanisation product, differs from other countries in terms of social behaviour on OSNs during the pandemic?

RQ3 Whether more attention to the pandemic and prevention measures in the early stages of the pandemic will impact the effectiveness of control and protective measures deployed by the governments?

To answer these questions, we collected 51,966,639 tweets from Twitter, which are posted by 15,551,266 Twitter users from 2020-01-22 to 2020-06-05 globally. Among them are 1,643,308 posts posted by 41,690 users in GR and its related countries. To investigate RQ1, basic reproductive rate $R_0$ and effective reproductive rate $R(t)$ in epidemiology [17] are introduced to
slice the pandemic periods, and correlations between tweets volume and daily cases in each period are calculated by Pearson Correlations (PC). A novel topic modelling method combing Bidirectional Encoder Representations from Transformers (BERT) [18] and the Latent Dirichlet Allocation (LDA) topic modelling method [19] is utilised, and a supervised Support Vector Machine (SVM) [20] for classifying topics into given categories is trained to study RQ2 and RQ3.

The main contributions in this paper are fourfold.

(I) We generate a novel Twitter dataset from 2020-01-02 to 2020-06-05 which contains users with locations labelled in GR, and related countries including Luxembourg, France, Germany and Belgium, and the COVID-19 related tweets and conversations posted by the users. This dataset will be shared with the public to advance related research.

(II) A Spatio-temporal analysis is carried out to figure out how the COVID-19 cases are correlated with the Twitter posts during a long-term period. We find that tweets volume only can be a predictor of outbreaks during the early period of the pandemic. Before \( R(t) \) value peaks, there is a spike of public concern about the pandemic, which is the best time to conduct the pandemic precaution advocacy [21].

(III) We find that for countries with a long interval (average 27.3 days) between the date of the first case and the date of an outright outbreak, the appearance of the first case did not attract enough attention of the public to anti-contagion and treatment measures. Furthermore, GR and Luxembourg showed a greater concern for anti-contagion and treatment measures before COVID-19 reaches its peak, and exhibited a higher level of interest in policy and daily life before \( R(t) < 1 \) than that of Germany, France and Belgium.

(IV) Discussions about anti-contagion and treatment measures on OSNs before the COVID-19 reaching its peak may have an influence on government protective policies and shorten the period for the policies to take effect.

This study sheds light on how the public reacts differently over time in GR and related countries through an interdisciplinary approach. It may, therefore, be useful to understand changes in social behaviour on OSNs during the pandemic, and in particular, the distinction of social behaviour between relational urbanisation region with high mobility and political sovereign states. Identifying when messages work best on public not only helps to generate
policy support but also to ensure individuals actions needed to combat the spread of the virus.

2. Related Work

Some existing results have already shown that social media conversations can be a leading predictor of a new pandemic cases [2, 22, 3], and in many countries tweets increase in volume before the number of confirmed cases increases. Studies have shown that anti-contagion policies can significantly and substantially reduce the spread of COVID-19 [23, 24, 25], and the effect of policies on the mitigation of spread varies, influenced by factors including culture, demographic information, socio-economic status and national health systems, where changes in public knowledge may affect the impact of the policies. If the public adjusts their behaviour in response to other new information not related to the policies, such as from online sources, this may change the spread of COVID-19 [23].

Researches of public behaviour patterns of the pandemic have been conducted based on data from smart devices [8], search index [26, 27], and COVID-19 related conversations on Twitter. Bento et al. [9] mention that, there is a spike in searches for basic information about Covid-19 when the first case was announced in each state in the United States, but the first case report does not trigger discussions about policy and daily life. Topic modelling, an unsupervised approach that detects latent semantic structure [4] is widely used. Cinelli et al. [1] extract topics with word embedding on a global scale, making the conclusion that social media may help to design more efficient epidemic models for social behaviour and to implement more efficient communication strategies. The LDA model is used by Medford et al. [5] and Ordun et al. [6] to analyse the topics in early period of the pandemic. Sharma, et al. [7] use character embedding [28] and Term Frequency Inverse Document Frequency (TF-IDF) word distribution with manual inspection for topic modelling. However, LDA, a bag-of-words approach, which is widely used to identify latent subject information in a large-scale document collection or corpus, has some drawbacks: it needs large corpus to train, ignores contextual information and performs mediocely in handling short texts [29]. As a result, these studies extract the topic over certain time periods, and the time granules are too coarse to accurately reflect the trend of the topics.
3. Data Description

In this section, we briefly describe how to collect COVID-19 tweets and COVID-19 daily cases information for GR and the related countries.

3.1. Twitter data collection

Twitter, one of the most prominent online social media platform, has been used extensively during the pandemic. In this study, 51,966,639 tweets posted by more than 15 million Twitter users are retrieved. The data collection consists of the following steps. First of all, we collect posts with COVID-19 related keywords from 2020/01/22 to 2020/06/05 and the user-id and location of users who posted them based on Chen et al.’s work [30] with the Twitter Streaming API. Secondly, as the user location information we collected so far is user-defined, it is not necessarily a true location, nor machine-parseable, so the fuzzy location context is processed into real location information by leveraging geocoding APIs, Geopy[3] and ArcGis Geocoding[4]. After getting the valid geographic location of each user, we select users which are located in the GR, Luxembourg, France, Germany, and Belgium to form the dataset. Table 1 gives an example of the final dataset, and Table 2 shows the summary of the collected tweet data of GR, Luxembourg, France, Germany, Belgium and the global.

| Attribute         | Description               | Example                                                                 |
|-------------------|---------------------------|-------------------------------------------------------------------------|
| Tweet_id          | A unique identifier for a Tweet | 12319668395******                                                      |
| Full_text         | Text of a tweet           | RT @*****: The Diamond princess is a UK ship managed by the US. UK should Be Responsible. #DiamondPrincess #coronavirus |
| User_id           | Unique identifier for this user | u9181074902******                                                      |
| User_geo_orginal  | User-defined location information | Moselle                                                                   |
| User_geo          | Geocoded user location     | Moselle, Lorraine, France                                               |

Table 1: A sample of our COVID-19 Twitter dataset

[3]https://bit.ly/3gfW2PP
[4]https://bit.ly/3f90UDA
| Region/Country | Tweets volume | User volume |
|----------------|---------------|-------------|
| Global         | 51,966,639    | 15,551,266  |
| GR             | 35,329        | 7,894       |
| Luxembourg     | 7,512         | 1,545       |
| Belgium        | 119,467       | 31,446      |
| France         | 1,050,312     | 288,009     |
| Germany        | 430,688       | 87,796      |

Table 2: Summary of our COVID-19 Twitter dataset

3.2. COVID-19 data collection

For the COVID-19 cases data, the dataset published by European Center for Disease Prevention and Control allows us to obtain COVID-19 data including daily cases, deaths and locations for the country we selected. As there is no official COVID-19 data published for GR, which is composed of Luxembourg, Wallonia in Belgium, Saarland and Rhineland-Palatinate in Germany and Lorraine in France, when counting daily cases and deaths in the GR, we add up all the data for the cities and regions mentioned above from the datasets published by corresponding countries as the final GR region data. It should be noted that in France, the number of daily new cases is not available at the region level, and deaths, hospitalisations, departures data have been published only since March 18, 2020. So the data for Lorraine is counted as zero until March 18, 2020, and the sum of hospitalisations, hospital departures and deaths is considered as the total number of cases on that particular day.

4. Correlation between COVID-19 Cases and Tweets Volume

To investigate whether tweets volume can be a predictor of daily cases during the pandemic (RQ1), we introduce basic reproductive rate $R_0$ and effective reproductive rate $R(t)$ in epidemiology to slice the period as the research covers a long duration, and a Spatio-temporal analysis of correlations between tweets volume and daily cases in each period is conducted by Pearson Correlations (PC).

5. https://bit.ly/3jYhefx
6. https://bit.ly/2ErDii7,https://bit.ly/3gaGGMm,https://bit.ly/33c8CM8
On 3rd June, France published a revision of data that lead to a negative number of new cases, see [https://bit.ly/33c8CM8](https://bit.ly/33c8CM8) for the original news.
4.1. $R(t)$-based time division

$R_0$ is the expected number of cases arising directly from a single case in a population where all individuals are susceptible to infection [17] and $R(t)$ represents the average number of new infections caused by an infected person at time $t$. If $R(t) > 1$, the number of cases will increase, e.g. at the beginning of an epidemic. When $R(t) = 1$, the disease is endemic, and when $R(t) < 1$, the number of cases will decrease. For the calculation of real-time $R(t)$, we use a Bayesian approach [31] with Gaussian noise to calculate the time-varying $R(t)$ based on daily new cases, which is also the official method for calculating $R(t)$ in Luxembourg[7] In this cases, While the study of $R_0$ of COVID-19 is still ongoing, in this research we use the $R_0$ estimated by WHO[8] which ranges between 1.4 and 2.5.

The results of time-varying $R(t)$ for GR, Luxembourg, Belgium, France,
and Germany are shown in Figure 2. Here, we divide the pandemic into four periods, which are: Pre-peak period (30 days before $R(t) < 2.5$); Free-contagious period ($1.4 < R(t) < 2.5$); Measures period ($1 < R(t) < 1.4$); Decay period ($R(t) < 1$). It is necessary to note that during the Free-contagious period, $R(t)$ is based on $R_0$, taking into account the results of the prevention measures [32], so during this period the measures taken to prevent the spread of the pandemic have not shown results or have failed to stop the spread of the pandemic, and COVID-19 still spreads during this period with an $R_0$ reproductive rate. The precise time duration of these pandemic periods for each country and region are summarised in Table 3.

|        | Pre-peak   | Free-contagious | Measures period | Decay period |
|--------|------------|-----------------|----------------|-------------|
| GR     | 2/14 - 3/15/2020 | 3/15 - 3/21/2020 | 3/21 - 4/17/2020 | 4/17 - 6/05/2020 |
| Luxembourg | 2/19 - 3/20/2020 | 3/20 - 3/24/2020 | 3/24 - 4/01/2020 | 4/01 - 6/05/2020 |
| Belgium | 2/04 - 3/05/2020 | 3/05 - 3/25/2020 | 3/25 - 4/18/2020 | 4/18 - 6/05/2020 |
| France  | 2/05 - 3/06/2020 | 3/06 - 3/30/2020 | 3/30 - 4/23/2020 | 4/23 - 6/05/2020 |
| Germany | 1/29 - 2/28/2020 | 2/28 - 3/24/2020 | 3/24 - 4/02/2020 | 4/02 - 6/05/2020 |

Table 3: Time duration of the four pandemic periods for GR, Luxembourg, Belgium, France and Germany

The exact numbers of days of each pandemic period are shown in Figure 3 for the countries and GR. The Free-contagious period in Luxembourg and GR is particularly short (4& 6 days) compared to the other countries (24-20 days). Existing research indicates that the differences in periods in each region and country are influenced by a variety of factors including government
policy [33, 34], population density [35], mobility [36, 37] and so on. In this paper, we will analyse it from the impact of the social behaviour on OSNs in Section 5 (RQ3).

4.2. Research question RQ1

To answer RQ1, we hypothesise that daily tweets volume can be predictive of daily cases and we calculated the relationship between them by PC, where a PC with a large absolute value means greater relation strength. The results are shown in Table 4. A lag refers to the tweets occurring after the cases; a Lag = -5 days means that we match the daily cases with the tweets volume from five days earlier, in other words, a 5-days lead.

**Pre-peak period.** As shown from Table 4, there is a clear trend of strong correlation (PC > 0.8, p < 0.05) with lags during the Pre-peak period, reaching its maximum at -5 or -6 days, indicating that tweets volume may be a predictor of outbreaks at an early period of the pandemic. Although the current research on the incubation period of COVID-19 is inconclusive, several studies have suggested that the incubation period of COVID-19 is on average 5-6 days [38, 39, 40]. In this regard, we propose that the 5-6 day lead may be related to the lag between infection and the onset of symptoms to be detectable and confirmed.

**Free-contagious period.** There is no clear trend of correlation with lags except the value of Luxembourg, indicating that tweets volume cannot be an indicator to predict the daily cases in the Free-contagious period. The period only lasted for 4 days in Luxembourg, which is too small to make PC a reflection of the correlation. However, the PC values show a highly negative correlation between Tweets volume and daily cases, which reflects there is a short period of a downward trend in the discussion of the pandemic after it reached its peak, even though the number of cases continued to rise rapidly. This result validates the conclusion of Smith et al. [41] from our dataset, they noted that public awareness of disease declines sharply after the peak, even though the infection rates remain high. In other words, the public’s interest in the pandemic declines after a period of heightened attention during the Pre-peak period.

**Measures period.** There is a clear trend of correlation with lags, tweets volume begins to level off, with a 0 or 1-day-lag moderate correlation (0.8 > PC > 0.3, p < 0.05) to the daily cases. Tweets volume cannot act as a predictor for daily cases here, it fluctuates with the number of cases on the current
Table 4: $PC$ (Pearson’s correlation) between Tweets Volume and COVID-19 daily cases with different lags

or previous day. It is worth noting that Pearson’s coefficient is sensitive to outliers and does not have robustness. With too few dates included, a single outlier can change the direction of the coefficients. This period existed for only 8 days in Luxembourg, resulting in an anomaly value ($PC = -0.903$). Tweets during this time may be related or relevant to the topics of the day and the day before.

12
Decay period. The correlations between tweets volume and daily cases occur in two ways here — one is weakly correlated and the other is that although there is a correlation, the trend of correlation with lags is not significant. Both demonstrate that it is not possible to predict daily cases by tweets volume during this period.

In summary, with the Spatio-temporal analysis of the correlation between tweets volume and COVID-19 new cases during the four period of the pandemic, we find that tweets volume only can be a predictor of outbreaks during the Pre-peak period of the pandemic. Regardless of the time at which $R(t)$peaks, there is a 5-6 day lead between tweets volume and COVID-19 daily cases, which may be related to the lag between infection and the onset of symptoms. What’s more, Before a pandemic strikes, there is a high level of public concern about the pandemic, and the Pre-peak period is the perfect time to conduct the pandemic precaution advocacy [21] as public concern about the pandemic will decrease after enter the Free contagious period. On the particularity of GR, we find that the Free-contagious period in GR and Luxembourg are exceedingly short, while Luxembourg is similarly short in the Measure period. The reasons for this will be explored further in Section 5 from the perspective of social behaviour on OSNs (RQ2).

5. Topic Modelling and Classification of Tweets

In order to have an actual understanding of the situation, and gain further insights into the behaviour of the public in GR and related countries on social media, the tweets posted by users in GR and related countries are analysed with BERT [18] and the LDA [19]. We extract the main daily topics on tweets, and categorise the generated topics, figuring out whether there are certain patterns of social behaviour on OSNs at the periods of the pandemic (RQ2). More importantly, we investigate how GR, as a relational urbanisation product, differs from other countries in terms of of public concern on OSNs during the pandemic (RQ2) and whether more attention to the pandemic and prevention measures in the early stages of the pandemic will shorten the Free-contagious period duration (RQ3). The overall workflow of our topic modelling and classification tasks is shown in Figure 4.

5.1. Text prepossessing and topic modelling

Text prepossessing. Prior to topical modelling, the tweets data needs to be preprocessed. Particularly, missing delimiters are detected according to the
uppercase letter, all text are lower-cased, while URLs, mentioned usernames and ‘RT’ are removed as well. Besides, punctuation and numbers are filtered out, typos are corrected by Symspell\textsuperscript{9} and stop words are removed. Then since the tweets are based on the keyword search, we removed keywords such as ’coronavirus’, ’koronavirus’, ’corona’, ’covid-19’, ’covid’ from the text to avoid bias. At last, Natural Language Toolkit (NLTK) is used for tagging the part-of-speech, stemming and tokenization. Nouns, verbs, adjectives and adverbs are selected.

\textit{Topic modelling.} Aiming to identify the latent topics of the tweets posted by the public in GR and related countries, we adopt the general structure of contextual topic embedding method (CTE)\textsuperscript{10} in this paper, to extract daily topic data and get a more accurate picture of topic trends. CTE mainly consists of two components, LDA and BERT, to extract different information from sentences to embedding. LDA, a bag-of-words approach which is widely used to identify latent subject information in a large-scale document collection or corpus has some drawbacks: it needs large corpus to train,

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{workflow.png}
\caption{Workflow of topic modelling and classification}
\end{figure}

\textsuperscript{9}https://github.com/wolfgangarbe/symspell
\textsuperscript{10}https://bit.ly/3hUQjzf
ignores contextual information and performs mediocly in handling short texts [29]. BERT utilises bidirectional transformers for pre-training on a large unlabelled text corpus, taking both left and right context into account simultaneously, which compensates for the shortcoming of LDA. And BERT is a method available for sentence embedding, thus we concatenate the generated tokens of each tweet as input sentences for BERT to obtain tweet representation. CTE combines the sentence embedding vector generated by BERT with the probabilistic topic assignment vector generated by LDA with a hyper-parameter $\gamma$.

Besides, as our data are multilingual, some words appear less frequently than in English which is predominantly spoken and are easily overlooked in the topic modelling, hence we adopt the TF-IDF model to determine word relevance in the documents [42]. And further feed the generated corpus by TF-IDF to LDA, instead of sample bag-of-words corpus. After obtaining the concatenated vector in high-dimensional space, CTE uses an autoencoder to learn a low-dimensional latent space representation of the concatenated vector with more condensed information. Then $k$-means [43] is implemented for clustering, the number of clusters $k$, that is, the number of topics, reserved as a hyper-parameter. We extract the word frequency in each cluster, sort and then take the top ten as the representation topic of that cluster. In terms of visualisation, Uniform Manifold Approximation and Projection (UMAP) [44] is used for low-dimensional latent space degradation, which is the state-of-the-art visualisation and dimension reduction algorithm.

| Country   | Coherence score | Silhouette score |
|-----------|-----------------|------------------|
| GR        | 0.432           | 0.893            |
| Luxembourg| 0.474           | 0.894            |
| France    | 0.351           | 0.590            |
| Belgium   | 0.377           | 0.864            |
| Germany   | 0.336           | 0.655            |

Table 5: Average coherence score and average silhouette score of CTE

Average coherence score and average silhouette score are utilised as the metrics of CTE. We calculated an average coherence score by calculating the topic coherence for each topic individually and averaging them. And the average silhouette score is the mean of the silhouette score for each day. Topic modelling is conducted on daily tweets of GR and the related countries, We
finetune the topic models and arrive at the optimal $n = 7$ and $\gamma = 0.5$ with highest average coherence score [45, 46] and average silhouette score [47]. The results are shown in Table 5 and a sample of clustering result from UMAP is shown in Figure 5. It can be observed from Table 5 and Figure 5 that the results generated by CTE are coherent and can be observed as well-separated clusters.

5.2. Topic classification

After getting 4,763 topics from topic modelling, we then randomly selected 2,435 topics and classified manually into the following 7 categories:

1. ‘Wuhan & China’: Topics about Wuhan and China-related issues.
2. ‘Measures’: Topics about basic information including symptom, anti-contagion and treatment measures of COVID-19.
3. ‘Local news’: Topics about local COVID-19 news and daily new cases, deaths, etc.
4. ‘International news’: Topics about international COVID-19 news
5. ‘Policy and daily life’: Topics about COVID-19 related policies encompass lockdown, Closure of borders, limits on public gatherings and the impact of the policies on daily life.
6. ‘Racism’: Topics about racism.
7. ‘Other’: Other topics.

These manually classified topics are used to train a Support Vector Machine (SVM) \cite{20} for supervised classification. In particular, words of each topic are converted to word frequency vectors with TfidfVectorizer\cite{11} and country are encoded with Label Encoder\cite{12}. The feature vector is made up with these two elements. Since our manually labelled dataset is class-imbalance, Synthetic Minority Oversampling Technique \cite{48}, is utilised for oversampling imbalanced the dataset and mitigate imbalances. The dataset is split 80\% as the training dataset and 20\% as the test dataset. Grid search with 10-fold cross-validation is deployed on training dataset to find the optimal hyperparameter, and the final SVM model is obtained with the entire training set. Table \ref{tab:6} shows the precision, recall, F1 score, support and Macro-average F-Score of the trained classifier for each topic category. Then, the obtained SVM model is utilised to classify the rest of the topics. Table \ref{tab:7} shows the number of topics of each category for each country. The categories with the high percentages are identified as Wuhan & China and policy and daily life. In general, the number of topics about policy and daily life is much higher in Luxembourg (56.6\%) than in other countries (ave = 33.0\%). France, on the other hand, shows a high level of interest in local news (30.2\%), relative to other countries (9.4\%). In terms of the overall data of GR, however, it does not show particular differences from other countries. Note that as there may be cases where the cluster for a topic contains no more than two tweets, we treat such topics as the invalid topic and remove them. This leads to a different total number of topics in each country. Next, we introduce dates to plot the changes in categories over time.

Figure \ref{fig:6} shows the tweets volume contained in each category as a percentage of the total tweets volume on that day (CR), with the darker red representing the higher CR. The interval shaded in white represents the period from 22 January to Pre-peak period, the shaded regions in different colours indicate, in order, Pre-peak period, Free-contagious period, Measures

\begin{table}[h]
\centering
\caption{Precision, recall, F1 score, support and Macro-average F-Score of the trained classifier for each topic category.}
\begin{tabular}{|c|c|c|c|c|}
\hline
\textbf{Category} & \textbf{Precision} & \textbf{Recall} & \textbf{F1 score} & \textbf{Support} & \textbf{Macro-average F-Score} \\
\hline
Policy and daily life & 0.80 & 0.75 & 0.78 & 100 & 0.78 \\
Racism & 0.70 & 0.65 & 0.68 & 50 & 0.68 \\
Other & 0.50 & 0.45 & 0.48 & 25 & 0.48 \\
\hline
\end{tabular}
\end{table}

\begin{table}[h]
\centering
\caption{Number of topics of each category for each country.}
\begin{tabular}{|c|c|c|c|c|}
\hline
\textbf{Country} & \textbf{Policy and daily life} & \textbf{Racism} & \textbf{Other} & \textbf{Wuhan & China} \\
\hline
Luxembourg & 56.6\% & 33.0\% & 9.4\% & 56.6\% \\
France & 30.2\% & 9.4\% & 9.4\% & 33.0\% \\
\hline
\end{tabular}
\end{table}

11 https://bit.ly/30bA8Ye
12 https://bit.ly/39E05kK
| Category | Precision | Recall | F1-score | Support |
|----------|-----------|--------|----------|---------|
| 1        | 0.89      | 0.77   | 0.82     | 163     |
| 2        | 0.92      | 0.93   | 0.93     | 166     |
| 3        | 0.80      | 0.79   | 0.80     | 155     |
| 4        | 0.74      | 0.86   | 0.80     | 155     |
| 5        | 0.73      | 0.68   | 0.71     | 149     |
| 6        | 0.99      | 1.00   | 0.99     | 157     |
| 7        | 0.97      | 1.00   | 0.98     | 142     |
| Macro avg| 0.86      | 0.86   | 0.86     | 1,087   |

Table 6: Metrics of the classification results

| Category | GR  | Luxembourg | Belgium | France | Germany | Total |
|----------|-----|------------|---------|--------|---------|-------|
| 1        | 245 | 168        | 287     | 202    | 315     | 1,217 |
| 2        | 64  | 34         | 48      | 65     | 41      | 252   |
| 3        | 99  | 44         | 109     | 285    | 110     | 647   |
| 4        | 134 | 77         | 114     | 52     | 167     | 544   |
| 5        | 353 | 525        | 370     | 250    | 295     | 1,793 |
| 6        | 23  | 7          | 23      | 31     | 15      | 99    |
| 7        | 41  | 72         | 15      | 60     | 23      | 211   |
| Total    | 959 | 927        | 966     | 945    | 966     | 4,763 |

Table 7: Topic volume for each category/country (region)

period, and Decay period. The black dotted line illustrates the date on which the first case appeared. With the exception of the GR, there is an interval of time between the date of the first case and the date of consecutive cases every day. The solid black line indicates the date that new cases appear every day since that date. For ease of discussion, we have named the day as ‘outbreak day’ (OD).

5.3. Research question RQ2

In this section, we aim to answer the RQ2: whether there are certain patterns of social behaviour on OSNs at different periods of the pandemic; and how GR, as a relational urbanisation product, differs from other countries in terms of social behaviour on OSNs during the pandemic.

This behavioural pattern obtained from Figure 6 is slightly different in our cases compared with the conclusion of Bento et al. [9]. In France, Germany
and Belgium, the appearance of the first case triggered only a small amount of discussion about the protective measures, and discussion about them does not start to increase until OD. In other words, the public did not really heed the pandemic until OD, when the virus was already spreading. This may explain by the existence of a large (average = 27.3 days) interval between the date of the first case and OD in France, Germany, and Belgium. During this interval, sporadic cases may not attract enough public attention, and the public’s attention was still focused on China-related news. What’s more, the report of first case does not stimulate discussions about policies and daily

Figure 6: Topic categories in GR and related countries
life as well, and discussion about it did not emerge frequently until OD.

The early picture in Luxembourg and GR is different. Figure 6 shows that the public in Luxembourg and GR started to have discussions about measures 1-2 days before the first case appeared. This may be explained by the late occurrence of the first case in Luxembourg and GR, where the other three countries have already passed OD, the outbreak in other countries may have attracted public attention in GR and Luxembourg. Interestingly, in Luxembourg, the discussion about policies and daily life persisted before the first case was announced and increased immediately after then. A word cloud of the topics from 22 January to 1 March (date of the first case) of Luxembourg is depicted in Figure 7, it shows that the topics are mainly travel-related. This may be explained by the fact that the proportion of foreign residents in the Luxembourg region is 47.4%13 and residents are more concerned about travel-related policies in Luxembourg and other countries.

As a relational urbanisation product, however, GR exhibits a high level of interest in policy and daily life with tweets volume for 47.1% of total tweets volume during the Free-contagious and the Measures period, while Luxembourg, the central region of GR, this percentage is 66.1%. Figure 8(a) shows boxplots of the distribution of the CR on policy and daily life during the Free-contagious and the Measures period. This shows that the public is more responsive to policies as a region that relies on foreign labour and has high mobility than Belgium, France and Germany. Figure 8(b) shows that during the Free-contagious and the Measures period, public interest in local news was similar in all regions except for France. Moreover, during the Decay period, while there is a downward trend ($p < 0.05$) in the total daily tweets volume, there is a upward trend ($p < 0.05$) in the CR of policy and daily life, except in Luxembourg, where the rate is consistently high.

5.4. Research question RQ3

As can be seen from Figure 6 and Figure 3, the discussion volume about measures during the Pre-peak period may affect the length of the Free-contagious period ($p < 0.05$). Table 8 lists the anti-contagion policies in different countries, the countries have implemented highly similar policies including lockdown, gatherings limitation and school closures. During the Pre-peak period, the CR of measures was much higher in GR (3.41%) and

13https://bit.ly/3fdhgVj
Luxembourg (7.62%) than in France (1.90%), Belgium (1.84%) and Germany (0.0%). It should be noted that the discussion of measures is not actually non-existent in Germany, but the tweets volume may be too small to be recognised as separate topics during the topic modelling process. From the results of topical classification, it is evident that there may be an underestimation of the severity of the pandemic by the public in Germany, France and Belgium during the Pre-peak period. The underestimation can be interpreted as optimism bias, which make people believe their exposure to disease is low [49]. During a pandemic, people often exhibit an optimism bias, a cognitive bias that causes someone to believe that they will be less likely to get involved in negative events [50]. In our results, it shows that even though the public in Belgium, France and Germany have shown sustained and long-
term concern about COVID-19 occurring in China on OSNs, optimism bias emerged when COVID-19 appeared, causing the public to ignore the emergence of the cases and to show low support for anti-contagious measures and government policies \[51\]. Thus, tweets about anti-contagion or treatment measures may serve as an indicator of whether the public are experiencing an optimism bias, and to some extent, shorten the *Free-contagious* period and public knowledge may affect the impact of the policies.

| Country  | Lockdown   | Maximum gatherings | School closures |
|----------|------------|--------------------|-----------------|
| Luxembourg | 3/15/2020  | No gatherings      | 3/13/2020       |
| Belgium  | 3/18/2020  | No gatherings      | 3/13/2020       |
| France   | 3/17/2020  | No gatherings      | 3/13/2020       |
| Germany  | 3/22/2020  | 2 individuals      | 3/13/2020       |

Table 8: Anti-contagion policies in different countries

6. Conclusion and Discussion

In this paper, we have studied the COVID-19 related content on Twitter, and innovatively introduced the idea of relational urbanisation and chosen the Greater Region and its related country as the research area. Our analyses focused on three research questions: can the Twitter posts volume be a predictor of COVID-19 daily cases during a long-term period? (RQ1), whether there are certain patterns of social behaviour on OSNs at different periods of the pandemic and how GR differs from other countries? (RQ2), and whether more attention to the pandemic and anti-contagious measures in the early stages of the pandemic will impact the effectiveness of control and protective measures deployed by the governments? (RQ3).

With the spatio-temporal analysis of the correlation between tweets volume and COVID-19 new cases during the four periods of the pandemic, our general answer to RQ1 is that tweets volume only can be a predictor of outbreaks during the Pre-peak period of the pandemic and the lead between tweets volume and COVID-19 daily cases may be related to the lag between infection and the onset of symptoms. For RQ2, we have shown a certain pattern in the main categories of public discussion on social media during the pandemic. In the early stage, the public focused on news related to China. If the first case occurs in a country and there is no rapid outbreak, the public
concern will not be diverted to anti-contagious and treatment measures, policies and local news until a complete outbreak, that cases begin to occur daily in that region. GR, as a region with a large number of cross-border workers, has shown high interest in policies since the first case, even if no lockdown policy has been implemented at that time. At the same time, Luxembourg, which has a foreign resident population of 47.4%, has shown a great concern for policies including travel from the beginning of the pandemic, which is not found in other regions. For RQ3, our answer is that the discussion volume about measures during the Pre-peak period may affect and shorten the length of the Free-contagious period.

Our results in this paper can be used to understand social behaviour on OSNs during the pandemic, and the differences of the social behaviour in relational urbanisation regions when facing the pandemic. We identify when is the perfect time to conduct the pandemic precaution advocacy which help to generate policy support.

There are still some limitations of our study. First, in our dataset, we did not take into account bots that post misleading information, which can lead to a possible bias in topical modelling and classification. For our initial exploration of topic categories, we chose SVM to build a baseline method for topic classification. We will utilise other state-of-the-art text classification methods to refine the classification in further study. Second, our cases study has some statistical limitations, and data from more countries will be included in future studies to ensure the statistical significance of the conclusions. Third, more research can be performed based on our dataset. For example, in future, we will conduct sentiment analysis on the tweets of different categories at each pandemic period to find out if there is a certain pattern in the public’s sentiment about the pandemic and how it differs from GR to other countries. And for RQ3, multi-class sentiment analysis with BERT will be conducted to figure out whether and to what extent people are optimistic or pessimistic about being affected by a pandemic during the Pre-peak period. Finally, during the writing of this article, the second wave of COVID-19 began to appear in Luxembourg and some countries. In a future study, we will conduct a comparative study focusing on the regions where the second wave occurred. Sentiment analysis and text classification with the state-of-the-art method will be deployed to investigate whether OSNs information that may reflect public attitude and behaviour. We will attempt to identify social behaviour that may lead to the second wave, such as laxity or resistance to policies and anti-infection measures. Such timely indicators
are potentially useful for appropriate policy changes to avoid a new pandemic outbreak.

Acknowledgements. This work was partially supported by Luxembourgs Fonds National de la Recherche, via grant COVID-19/2020-1/14700602 (PandemicGR).

References

[1] M. Cinelli, W. Quattrociocchi, A. Galeazzi, C. M. Valensise, E. Brugnoli, A. L. Schmidt, P. Zola, F. Zollo, A. Scala, The covid-19 social media infodemic, arXiv preprint arXiv:2003.05004 (2020).

[2] L. Singh, S. Bansal, L. Bode, C. Budak, G. Chi, K. Kawintiranon, C. Padden, R. Vanarsdall, E. Vraga, Y. Wang, A first look at COVID-19 information and misinformation sharing on Twitter, arXiv preprint arXiv:2003.13907 (2020).

[3] K. Jahanbin, V. Rahmanian, Using Twitter and web news mining to predict COVID-19 outbreak, Asian Pacific Journal of Tropical Medicine 13 (2020) 26–28.

[4] C. Wang, B. David M., Collaborative topic modelling for recommending scientific articles, in: Proceedings of the 2011 International Conference on Knowledge Discovery and Data Mining (SIGKDD), ACM, 2011, pp. 448–456.

[5] C. Ordun, S. Purushotham, E. Raff, Exploratory analysis of covid-19 tweets using topic modelling, UMAP, and digraphs, arXiv preprint arXiv:2005.03082 (2020).

[6] R. J. Medford, S. N. Saleh, A. Sumarsono, T. M. Perl, C. U. Lehmann, An ”Infodemic”: Leveraging high-volume Twitter data to understand public sentiment for the COVID-19 outbreak, medRxiv (2020).

[7] K. Sharma, S. Seo, C. Meng, S. Rambhatla, Y. Liu, COVID-19 on Social Media: Analyzing Misinformation in Twitter Conversations, arXiv preprint arXiv:2003.12309 (2020).

[8] S. Gupta, T. D. Nguyen, F. L. Rojas, S. Raman, B. Lee, A. Bento, K. I. Simon, C. Wing, Tracking public and private response to the COVID-19
epidemic: Evidence from state and local government actions, Tech. rep., National Bureau of Economic Research (2020).

[9] A. I. Bento, T. Nguyen, C. Wing, F. Lozano-Rojas, Y. Y. Ahn, K. Simon, Evidence from Internet search data shows information-seeking responses to news of local COVID-19 cases, Proceedings of the National Academy of Sciences of the United States of America (PNAS) 117 (21) (2020).

[10] C. E. Lopez, M. Vasu, C. Gallemore, Understanding the perception of COVID-19 policies by mining a multilanguage Twitter dataset, arXiv preprint arXiv:2003.10359 (2020).

[11] M. Thelwall, S. Thelwall, Retweeting for COVID-19: Consensus building, information sharing, dissent, and lockdown life, arXiv preprint arXiv:2004.02793 (2020).

[12] P. Adler, R. Florida, M. Hartt, Mega regions and pandemics, Tijdschrift voor Economische en Sociale Geografie 111 (3) (2020) 465–481.

[13] S. H. Ali, R. Keil, Networked disease: Emerging infections in the global city, Wiley-Blackwell, 2009.

[14] C. Connolly, R. Keil, S. H. Ali, Extended urbanisation and the spatialities of infectious disease: Demographic change, infrastructure and governance, Urban Studies (2020).

[15] M. Hesse, M. Rafferty, Relational cities disrupted: reflections on the particular geographies of COVID-19 For small but global urbanisation in Dublin, Ireland, and Luxembourg City, Luxembourg, Tijdschrift voor economische en sociale geografie 111 (3) (2020) 451–464.

[16] T. Sigler, K. Martinus, I. Iacopini, B. Derudder, The role of tax havens and offshore financial centres in shaping corporate geographies: An industry sector perspective, Regional Studies (2019).

[17] J. A. P. Heesterbeek, K. Dietz, The concept of Ro in epidemic theory, Statistica Neerlandica 50 (1) (1996) 89–110.
[18] J. Devlin, M. W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (HLT), Vol. 1, ACL, 2019, pp. 4171–4186.

[19] D. M. Blei, A. Y. Ng, M. T. I. Jordan, Latent dirichlet allocation, Journal of Machine Learning Research 3 (2003) 993–1022.

[20] C.-C. C. Chang, C.-J. J. Lin, LIBSVM: A library for support vector machines, Transactions on Intelligent Systems and Technology 2 (3) (2011) 1–27.

[21] P. M. Sandman, Crisis communication best practices: Some quibbles and additions, Journal of Applied Communication Research 34 (3) (2006) 257–262.

[22] C. St Louis, G. Zorlu, Can Twitter predict disease outbreaks?, British Medical Journal 344 (7861) (2012).

[23] S. Hsiang, D. Allen, S. Annan-Phan, K. Bell, I. Bolliger, T. Chong, H. Druckenmiller, L. Y. Huang, A. Hultgren, E. Krasovich, Others, The effect of large-scale anti-contagion policies on the COVID-19 pandemic, Nature (2020) 1–9.

[24] C. Courtemanche, J. Garuccio, A. Le, J. Pinkston, A. Yelowitz, Strong social distancing measures in the United States reduced The COVID-19 Growth Rate, Health Affairs (2020) 10–1377.

[25] T. Dergiades, C. Milas, T. Panagiotidis, Effectiveness of government policies in response to the COVID-19 outbreak, SSRN (2020).

[26] D. Hu, X. Lou, Z. Xu, N. Meng, Q. Xie, M. Zhang, Y. Zou, J. Liu, G. P. Sun, F. Wang, More effective strategies are required to strengthen public awareness of COVID-19: Evidence from Google trends, Journal of Global Health 10 (1) (2020).

[27] M. Effenberger, A. Kronbichler, J. I. Shin, G. Mayer, H. Tilg, P. Perco, Association of the COVID-19 pandemic with Internet search volumes: A Google trendsTM Analysis (2020).
[28] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, T. Mikolov, Fasttext. zip: Compressing text classification models, arXiv preprint arXiv:1612.03651 (2016).

[29] X. Yan, J. Guo, Y. Lan, X. Cheng, A biterm topic model for short texts, in: Proceedings of the 22nd International Conference on World Wide Web, 2013, pp. 1445–1456.

[30] E. Chen, K. Lerman, E. Ferrara, COVID-19: The First Public Coronavirus Twitter Dataset, arXiv preprint arXiv:2003.07372 (2020).

[31] L. M. L. M. A. Bettencourt, R. M. Ribeiro, Real time Bayesian estimation of the epidemic potential of emerging infectious diseases, PLoS One 3 (5) (2008) e2185.

[32] J. Wallinga, P. Teunis, Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures, American Journal of Epidemiology 160 (6) (2004) 509–516.

[33] T. Hale, A. Petherick, T. Phillips, S. Webster, Variation in government responses to COVID-19, Blavatnik School of Government Working Paper 31 (2020).

[34] Y. Fang, Y. Nie, M. Penny, Transmission dynamics of the COVID-19 outbreak and effectiveness of government interventions: A data-driven analysis, Journal of Medical Virology 92 (6) (2020) 645–659.

[35] J. Rocklöv, H. Sjödin, High population densities catalyse the spread of COVID-19, Journal of travel medicine 27 (3) (2020) 038.

[36] S. Engle, J. Stromme, A. Zhou, Staying at home: mobility effects of covid-19, Available at SSRN (2020).

[37] M. U. G. Kraemer, C.-H. Yang, B. Gutierrez, C.-H. Wu, B. Klein, D. M. Pigott, L. Du Plessis, N. R. Faria, R. Li, W. P. Hanage, Others, The effect of human mobility and control measures on the COVID-19 epidemic in China, Science 368 (6490) (2020) 493–497.

[38] World Health Organization, Coronavirus disease 2019 (COVID-19) Situation Report 73, Tech. rep., World Health Organization (2020).
[39] S. A. Lauer, K. H. Grantz, Q. Bi, F. K. Jones, Q. Zheng, H. R. Meredith, A. S. Azman, N. G. Reich, J. Lessler, The incubation period of Coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: Estimation and application, Annals of Internal Medicine 172 (9) (2020) 577–582.

[40] S. Lei, F. Jiang, W. Su, C. Chen, J. Chen, W. Mei, L.-Y. Zhan, Y. Jia, L. Zhang, D. Liu, Others, Clinical characteristics and outcomes of patients undergoing surgeries during the incubation period of COVID-19 infection, EClinicalMedicine (2020) 100331.

[41] M. C. Smith, D. A. Broniatowski, M. J. Paul, M. Dredze, Towards real-time measurement of public epidemic awareness: Monitoring influenza awareness through twitter, in: Spring Symposium on Observational Studies Through Social Media and Other Human-generated Content, Vol. 20052, 2016, p. e198.

[42] J. Ramos, Others, Using TF-IDF to determine word relevance in document queries, in: Proceedings of the 2003 instructional conference on machine learning (ICML), Vol. 242, 2003, pp. 133–142.

[43] K. Wagstaff, C. Cardie, S. Rogers, S. Schrödl, Others, Constrained k-means clustering with background knowledge, in: Proceedings of the 2001 International Conference on Machine Learning (ICML), Vol. 1, Citeseer, 2001, pp. 577–584.

[44] L. McInnes, J. Healy, J. Melville, Umap: Uniform manifold approximation and projection for dimension reduction, arXiv preprint arXiv:1802.03426 (2018).

[45] D. O’callaghan, D. Greene, J. Carthy, P. Cunningham, D. O’Callaghan, D. Greene, J. Carthy, P. Cunningham, An analysis of the coherence of descriptors in topic modelling, Expert Systems with Applications 42 (13) (2015) 5645–5657.

[46] D. Newman, J. H. Lau, K. Grieser, T. Baldwin, Automatic evaluation of topic coherence, in: Proceedings of the 2010 Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, ACL, 2010, pp. 100–108.
[47] S. Aranganayagi, K. Thangavel, Clustering categorical data using silhouette coefficient as a relocating measure, in: Proceedings of the 2007 International Conference on Computational Intelligence and Multimedia Applications, Vol. 2, 2008, pp. 13–17.

[48] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, Journal of Artificial Intelligence Research 16 (2002) 321–357.

[49] M. C. Smith, D. A. Broniatowski, M. J. Paul, M. Dredze, Using social and behavioural science to support COVID-19 pandemic response, Nature Human Behaviour 4 (5) (2020) 1–12.

[50] T. Sharot, The optimism bias, Current Biology 21 (23) (2011) 941–945.

[51] H.-J. J. Paek, K. Hilyard, V. S. Freimuth, J. K. Barge, M. Mindlin, Public support for government actions during a flu pandemic: lessons learned from a statewide survey, Health Promotion Practice 9 (4) (2008) 60–72.