engineering proceedings

MDPI

Proceeding Paper

Forecasting the Spread of the COVID-19 Pandemic Based on the Communication of Coronavirus Sceptics †

Melinda Magyar *, László Kovács and Dávid Burka

Department of Computer Science, Corvinus University of Budapest, Fővám tér 13-15, 1093 Budapest, Hungary; laszlo.kovacs2@uni-corvinus.hu (L.K.); david.burka@uni-corvinus.hu (D.B.)

* Correspondence: melinda.magyar@uni-corvinus.hu

† Presented at the 7th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 19–21 July 2021.

Abstract: The COVID-19 pandemic has left a mark on nearly all events since the start of the year 2020. There are many studies that examine the medical, economic, and social effects of the pandemic; however, only a few are concerned with how the reactions of society affect the spread of the virus. The goal of our study is to explore and analyze the connection between the communication of pandemic sceptics and the spread of the COVID-19 pandemic and its caused damages. We aim to investigate the causal relationship between communication about COVID-19 on social media, anti-mask events, and epidemiological indicators in three countries: the USA, Spain, and Hungary.

Keywords: COVID-19; sceptics; social media; Twitter; sentiment; VAR; Granger causality; government stringency

Citation: Magyar, M.; Kovács, L.; Burka, D. Forecasting the Spread of the COVID-19 Pandemic Based on the Communication of Coronavirus Sceptics . *Eng. Proc.* **2021**, *5*, 35. https://doi.org/10.3390/engproc2021005035

Academic Editors: Ignacio Rojas, Fernando Rojas, Luis Javier Herrera and Hector Pomare

Published: 1 July 2021

1. Introduction

Coronavirus is the latest of many infectious diseases affecting humanity throughout history that have reached the state of a pandemic. Pandemics, by definition, affect large regions across continents or even the whole world; thus, even in case of a low mortality rate, the number of casualties can reach millions in a relatively short time period. The COVID-19 outbreak is among the deadliest pandemics of the last hundred years, only outdone by HIV/AIDS (human immunodeficiency virus infection and acquired immune deficiency syndrome) [1].

However, the COVID-19 pandemic is the first to occur since social media became widespread. The swine flu (H1N1) outbreak, being the most recent one, happened between 2009 and 2010 [2], but at that point, Facebook had just started its rise in popularity, and other platforms that are well known today (i.e., Twitter, Reddit, Instagram) had barely started to gain popularity [3]. HIV is an exception, as it still costs around 800 thousand lives per year because of its high mortality rate, but it has infected far fewer people than the other mentioned pandemics [4]. Additionally, HIV was the focus of attention in the 1980s and 1990s, but it has not been covered in the media too often in recent years.

This means that COVID-19 is the first pandemic about which an immense volume of online written communication exists, which can be analyzed with the help of different text mining solutions. Never before has the opportunity been presented to examine the opinion of the masses regarding such events; thus, this is a completely new field of research, and in this relatively short time period, there have not been many investigations exploiting its potential. There are many studies about social communication during the pandemic, including false news and its impact on the pandemic and vice versa [5,6]; however, these usually focus on a single conspiracy theory, a set of news, or a small group of events instead of long-running time series.

Our research aims to examine the connection between social responses and pandemic-related events in the USA, Spain, and Hungary. We examined the most prevalent social

platforms of each country and collected a large volume of COVID-19-related comments and their timestamps. Sentiment analysis was used to process this text-based data source; thus, it was possible to create a sentiment time series for each language group.

Reliable corona-related pandemic data are available on the *Our World in Data* (OWID) site in a research-friendly form [7]. Regarding the activity of deniers, we manually collected a list of significant demonstrations and assemblies from different news sources. We only considered "offline" events as these are the ones that could have directly influenced the number of infections.

We compared the sentiment time series with the events and corona-related time series by applying an augmented vector autoregression (VAR) model according to the Toda–Yamamoto procedure [8] on the examined time series in each country separately. Granger causality models have been successfully applied in order to assess the economic and financial effects of the COVID-19 pandemic, for example, by [9] and [10]. We show that the volume of the online comments and the sentiment index had a significant mutual relationship with the official epidemiological indicators. The characteristics of these relationships differed along countries and waves of the pandemic. In Spain, the antimask events had a significant effect on the volume of comments during the first wave and on sentiment in the second wave.

2. Data Sources

For constructing sentiment time series, we need textual data obtained from representative sources. Every target country has some preferred social media sites, such as forums, microblogging sites, or even comment sections of their leading news sites. The most important social media site is Facebook, and Twitter is also in the top 20 in every country except in Hungary, according to Similarweb [11]. The contents of these platforms could be a good starting point to examine social reactions about pandemic events and vice versa. As the most widely used search service in the world, Google cannot be ignored either: not only do the topics searched show an increased interest in the COVID-19 pandemic, but they can give us an idea of the focal points of interest. These platforms together are appropriate sources for text mining research studies, which can transform human sentiments into data, map the topics, and find the most influential ones.

In the examined countries, for data source, the common ground could have been Facebook [11]. However, Facebook is not an easy option for text mining research studies since the Cambridge Analytica scandal [12], so Twitter was chosen as a source for mining sentiments for the English and Spanish languages. Because Twitter is not so popular in Hungary, gyakorikerdesek.hu (hereinafter referred to as FAQ) was used for this country as a text mining source. This is a Q&A-type website, which is the 31st most visited site in Hungary.

Twitter provides an API for researchers under friendly conditions, and there is a project named Twitter Stream Grab by Archive Team that allowed us to download all tweets for the examined period [13]. FAQ does not provide API for grabbing data, so we developed an application for scraping purposes. During scraping, the software collects questions and answers from two relevant categories: health and politics [14].

A series of corona-sceptic events were collected manually based on the collections of national Wikipedia pages related to coronavirus and on the Google Labs search terms related to coronavirus [15].

From the times series published on the website OurWorldInData.org, three are used to describe the pandemic situation. The first time series is the rate of positive coronavirus tests. It is used to describe the spread of the virus. This is in line with WHO recommendations [16]. The severity of the pandemic is described by the daily number of deaths per million people. The daily values of the government stringency index are also considered to examine whether the sentiment of the online public is reacting to government measures or vice versa. The index is calculated by the Oxford Coronavirus Government Response Tracker (OxCGRT) project. This is a composite measure based on nine response indicators,

including school closures, workplace closures, and travel bans, rescaled to a value between 0 and 100 (100 = strictest) [17].

The time periods examined were different for each country to ensure an adequate level of variance in each time series as the start of the pandemic differed for each examined country. For example, in the US, the number of deaths was 0 on most days until 13 March 2020, and testing data were only available since 7 March 2020, so 13 March 2020 was used as a starting point. The number of daily deaths per million people was quite scarce for Spain. There were two negative values on 25 May and 12 August that were imputed as 0. There was a weekly seasonality for 0 entries. Therefore, we took a 7-day moving average of daily new deaths per million people for Spain. The end point for all these time series was 31 December 2020, as the focus of our investigation was the past year.

Descriptive statistics for each examined time series are available in Table 1. To check for outlier effect, a mean trimmed off the bottom and upper 10% was used. Outliers had no great effect on the examined time series.

Sentiment in US tweets was the most negative on an average day with low standard deviation, while the mean sentiment in Spain seemed to be the highest, though still a negative value. Hungary had the greatest standard deviation in its sentiment index.

Table 1. Descriptive statistics for all of our examined time series.

Variables	No of Obs.	Mean	St. Dev.	Tr. Mean
Positive rate USA	282	0.08	0.04	0.07
Deaths per million USA	282	3.61	2.30	3.33
Stringency USA	282	68.63	5.41	69.15
Entry count USA	282	3580.59	2166.54	3199.61
Sentiment USA	282	−0.46	0.12	−0.46
Positive rate ESP	282	0.06	0.04	0.06
Deaths per million ESP	246	2.59	2.41	2.31
Stringency ESP	246	66.37	9.48	66.39
Entry count ESP	246	644.04	269.15	627.54
Sentiment ESP	246	−0.09	0.11	−0.09
Positive rate HUN	246	0.08	0.09	0.06
Deaths per million HUN	284	3.47	5.51	2.26
Stringency HUN	284	59.45	12.58	59.62
Entry count HUN	284	88.25	60.45	81.93
Sentiment HUN	284	−0.13	0.19	−0.14

3. Methods

The data on Twitter Stream Grab are available on a monthly basis, and there is one compressed JSON file for every minute, so to examine a whole year, more than half a million files must be processed. A time frame between 01/03/2020 and 31/12/2020 was chosen according to the availability of pandemic data from OWID. Datasets contained time data, text, detailed user data, and language index. There were two important limitations: we did not have data about the specific followers for a given user, and there was no precise location data; we could only rely on user-supplied information. In order to reduce the data size, we filtered out relevant tweets based on a few selected keywords, which were grabbed from Google Labs Corona search terms [15]. English-language tweets were narrowed down to the United States based on user-defined location, and a 10% random sample was taken for Spanish-language tweets. The extracted data were transformed into comma-separated files, which can be easily imported into other systems. The texts scraped from FAQ for Hungarian-language analysis did not needed further preprocessing, as the scraper software was designed specifically for this research and had taken the necessary steps.

After extracting tweets and comments, the texts were cleaned and prepared for sentiment analysis. For stemming and lemmatization, the *hunspell* package was utilized, which is a spell checker and morphological analyzer originally designed for the Hungarian lan-

guage, but it performs well in English and Spanish also [18]. For examining sentiments regarding the pandemic, collected text entries should be labelled with polarity: negative or positive. To do this, a dictionary-based sentiment analysis was applied.

There could be some structural breaks in each time series due to the different characteristics of the first and second waves of the pandemic. Therefore, we should identify possible structural breaks in each examined time series that best separated the first and second waves of the pandemic.

When investigating Granger causality, it is advisable to fit a model separately on sections defined by structural breaks to ensure stability [8]. To identify structural breaks, the breakpoint function from the *strucchange* R package was utilized [19]. If we assume that the number of breakpoints in a linear trend for a time series is b, then the breakpoint function estimates the location of b breakpoints by minimizing the residual sum of squares (RSS) of a linear model where the slope of the trend can change b times. The optimal b is chosen by the Bayes–Schwarz information criterion (BIC) as this IC prefers the sparsest models. This was preferable for us as we had a relatively small number of observations for each country already, so we should avoid overparameterization.

After determining the breakpoints, we fit VAR models for each country and each wave separately to discover the Granger causality between the time series in both waves of the pandemic. A vector autoregression (VAR) process with k endogenous and m exogenous variables can be considered a system of equation with k equations. Model parameters are estimated by OLS. See [20] for details. Maximum lag of the endogenous variables is denoted by p.

However, the typical Granger causality test based on the classical VAR model cannot be relied on when one or both time series are nonstationary, which could lead to spurious causality [21]. Thus, an augmented Dickey–Fuller (ADF) test was employed. Besides, a Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test, in which the null hypothesis is stationarity, was also conducted as a cross-check. To handle the possible integration in our time series, the VAR models were set up according to the Toda–Yamamoto (TY) procedure [8] using the levels of the data without differencing and adding q extra lags if the maximum order of integration was q. The advantage of the TY procedure is it saves the cointegration test and prevents pretest bias. However, there was a need to ensure that the VAR models of each country were specified in a way that there was no serial correlation in the residual values. This was tested by the portmanteau test.

In the optimal VAR models, Wald tests of Granger causality were applied. The null hypothesis is that the coefficients of the first p lagged values of endogenous variables in each equation are 0 after being tested. The reason for including the coefficient of the lags from $p + 1$ to q is that the additional lagged values are to fix the asymptotic so that the Wald test statistics under the null hypothesis follow asymptotical chi-square distribution. Rejection of the null hypothesis of the Wald test implies a Granger causality.

4. Models

For our investigations, three countries were considered. The English-language tweets were narrowed down to tweets originating from the USA, so epidemiological and government stringency indicators of the US were considered here. For the Spanish-language tweets, the indicators of Spain were considered as during the first wave of the pandemic, Spain was the hardest-hit Spanish-speaking country. By 30 June 2020, the cumulative number of deaths per million was 606 in Spain and 297 and 215 in Chile and Mexico, respectively. During the second wave, the pandemic situation in Latin America became more serious, so the effects of COVID-related tweets from other Spanish-speaking countries could act as confounders. Managing these issues is part of our further research. The indicators of Hungary were considered for the Hungarian language.

Sentiment dictionaries were gathered from different sources: Bing for English, TASS for Spanish, and PrecoSenti for Hungarian [22–24]. Further processing was performed with R using the *tidytext* and *dplyr* packages.

Figure 1 shows that the basic sentiment in Spanish-language tweets was more positive than in English-language ones. The daily count of tweets followed the usual trend of a scandal: at the beginning of the pandemic, we could experience a large volume of comments about corona-related topics, and the numbers started to fall during the year even at the time of the second wave.

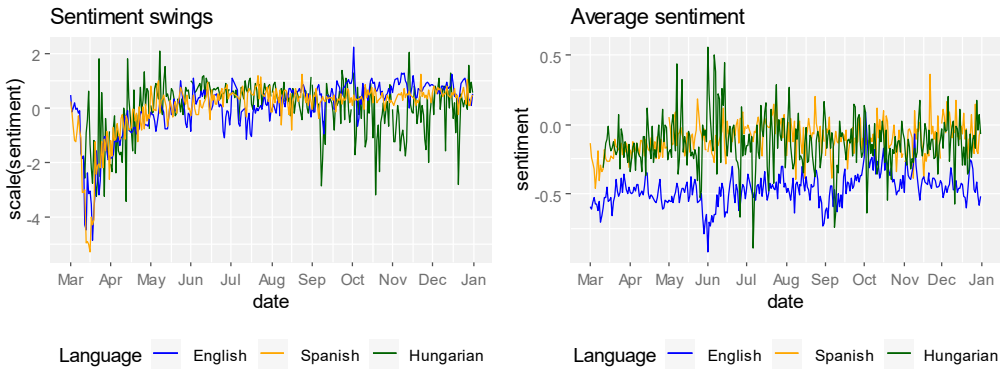

Figure 1. The tendencies in sentiment time series are relatively similar in the three examined datasets; however, the average sentiment is higher in Spanish- and Hungarian-language tweets than English-language contents.

The breakpoint function from the *strucchange* package identified two to four breaks in the time series based on the *BIC*. These breakpoints needed to be narrowed down, as four breakpoints would partition our sample into parts with very small sizes. To select the breakpoints that best separated the two waves, the breakpoints of the positive rate in each county were examined in more detail as this was the measure describing the spread of the pandemic in line with WHO recommendations [16].

The breakpoints of the positive rate in each county are examined in more detail in Figure 2 to define sections on which the Granger causality between the time series is examined by fitting VAR models.

We can see that in Spain and Hungary, we could easily select the structural breakpoint that best separated the start of the second wave of the pandemic. It is also noticeable that Hungary had quite a long period in the summer where the positive rate stagnated on a lower level before the second wave started in September. However, we did not wish to separate this period from the first wave as three breakpoints would result in small subsamples. That is why we also ignored the break that marked the peaking of the second wave. In Spain, the second wave started around the middle of summer, much earlier than in Hungary. We disregarded the other breakpoints marking different periods in the first and second waves as splitting along these would result in small subsamples just like in the case of Hungary.

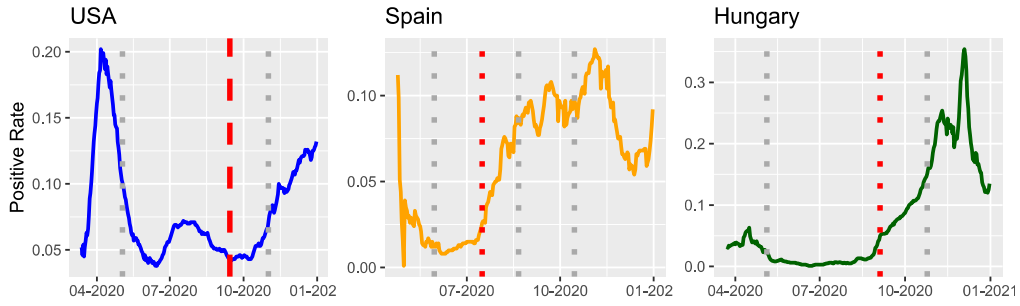

Figure 2. The positive rate for the three examined countries. Structural breakpoints are marked with dotted lines. The breaks marked with red are the ones that best separate the first and second waves of the pandemic. In the US, a custom breakpoint is added to separate the two waves marked with a dashed red line.

The case of the US was more complicated as it had a short flare of the pandemic in the middle of summer and the second wave started in late October. To preserve the sample size, we considered the short flare in positive rate in the summer as an aftershock of the first wave and defined a custom breakpoint on 20/09/2020, marked by the dashed red line in Figure 2. We separated every examined time series into two parts, representing the first and second waves of the pandemic according to the country-specific breakpoints selected as shown in Figure 2.

As we had five time series for each country, we had $k = 5$ endogenous variables. Dummy variables were used as exogenous variables to account for day-of-the-week effect. One more dummy exogenous variable represented whether there was an antimask event with at least 100 participants at time t for each country, making $m = 6 + 1 = 7$. The number of p lags will be chosen later.

Based on the results of the ADF and KPSS tests, taking the first difference of each time series mostly eliminated the unit root. The only exceptions were the stringency time series in the US and the positive rate for Spain and Hungary during the first wave, according to the KPSS test, but only on $\alpha = 10\%$, not on $\alpha = 5\%$. The ADF test rejected the H_0 of the unit root on all common significance levels in these cases. Thus, the maximum order of integration was set to 1.

The VAR models were set up according to the TY procedure to account for the first-order integration. First, we determined the appropriate lag length for the endogenous variables. Based on the Akaike information criterion, Hannan–Quinn information criterion, Bayes–Schwarz criterion, and final prediction error, lags $p = 1$ and $p = 2$ were recommended.

From the results of a portmanteau test controlling for dynamic stability, it was observed that lag 2 removed residual serial autocorrelation at 1% for all VAR models except for Hungary during the first wave. As accepting the H_0 of no serial correlation in the residuals was not convincing on all common significance levels, adding more lags could be considered, but we already had a larger parameter–sample size ratio with the dummies and the two lags for each variable (17 + 1 parameters for each equation, which is slightly less than fifth of the number of observations (circa 160 and 120 for each wave) in all three countries). The VAR models could be considered stable, again except for Hungary during the first wave, as all roots of the characteristic polynomials were inside the unit circle. Detailed diagnostic results for each VAR model are shown in Table 2.

Table 2. Model diagnostic results for the examined VAR(1) and VAR(2) models.

Setup	Lag = 1		Lag = 2	
	Portmanteau Test *p*-Value	**Range of Roots of Characteristic Polynomials**	**Portmanteau Test *p*-Value**	**Range of Roots of Characteristic Polynomials**
USA-1st wave	0.0213	0.508–0.940	0.0596	0.196–0.948
USA-2nd wave	0.8364	0.565–0.902	0.9043	0.038–0.901
Spain-1st wave	0.8667	0.154–0.971	0.9108	0.129–0.962
Spain-2nd wave	0.0369	0.095–0.945	0.0849	0.189–0.936
Hungary-1st wave	0.0005	0.093–1.014	0.0001	0.070–0.992
Hungary-2nd wave	0.1067	0.053–0.980	0.2559	0.094–0.959

Lag $p = 2$ was chosen for the VAR models, and one more lag into each variable was added to every equation, given that the maximum order of integration was 1. Therefore, the augmented VAR models proposed by the TY procedure were constructed, and the Granger causality tests were executed.

5. Results

Results of the Granger causality tests are shown in Table 3. Granger causality in Hungary during the first wave was not investigated as the underlying VAR model was not stable, and it had significant residual serial autocorrelation.

Table 3. Significant Granger causalities found in each examined VAR(2). For each causal relationship, the most significant lag in the appropriate VAR equation and the sign of this lag's coefficient are given in brackets.

Setup	Significant Granger Causalities
USA-1st wave	Stringency -> entry count * (lag = 1; sgn = +) Sentiment -> entry count * (lag = 2; sgn = +) Positive rate -> deaths per million ** (lag = 2; sgn = +) Entry count -> deaths per million ** (lag = 1; sgn = +)
USA-2nd wave	Positive rate -> stringency ** (lag = 2; sgn = +)
Spain-1st wave	Deaths per million -> sentiment * (lag = 1; sgn = −) Deaths per million -> entry Count * (lag = 1; sgn = +) Entry count -> deaths per million ** (lag = 1; sgn = +) Deaths per million -> stringency *** (lag = 2; sgn = +)
Spain-2nd wave	Entry count -> stringency * (lag = 2; sgn = +)
	Entry count -> deaths per million ** (lag = 1; sgn = −)
Hungary-1st wave	-
Hungary-2nd wave	Entry count -> stringency ** (lag = 1; sgn = +)
	Deaths per million -> entry count ** (lag = 1; sgn = −)

* Significant at 10%, ** significant at 5%, *** significant at 1%.

Table 3 shows that more significant Granger causal relationships could be found during the first wave of the pandemic than during the second. This is not surprising as the novelty of the virus posed more challenge during the first wave as decision makers and health professionals had to operate under limited information. Therefore, it is logical that we can find a higher number of relationships between our examined time series during the first wave. Unfortunately, owing to lack of a well-specified model for Hungary, this conclusion can only be made for Spain and the US.

In the US, the two most significant relationships were those between Twitter entry or post count and deaths per million and between positive rate and deaths. It seems that if the test positive rate increased, mortality usually followed 2 days later. This relationship was not significant at any of the common significance levels during the second wave, which suggests that the situation had improved by that time. During the second wave, we could also find that the increase in the rate of positive tests caused a stricter government response. This suggests that by the second wave, the US government started to react faster to changes in the pandemic situation. In the first wave, an increase in government stringency caused the count of Twitter entries to rise a day later. This can confirm that the US population was quite concerned with government response, so the measures were debated on Twitter. This finding is further supported by the fact that 2020 was election year in the US, so it is natural that government actions were under more scrutiny. These debates happened during the hardest days of the pandemic in the US, which is reflected in the significant Granger causality of Twitter entry count on mortality. Lastly, we observed that an increase in Twitter sentiment caused an increase in the number of posts 2 days later. It can be theorized that some positive messages about the pandemic could spread fast in the US, where the population grew frustrated with the lockdowns [25]. The antimask event exogenous variable had no significant effect on any of the endogenous time series in the US.

In Spain, during the first wave, the most significant Granger causality was the one that showed government stringency increasing 2 days after the deaths per million people increased. Therefore, the Spanish government reacted based on mortality, not on the rate of positive tests as the US government did. The less significant relationships showed that the number of tweets increased, and Twitter sentiment declined 1 day after an increase in mortality. Therefore, the increase in government stringency can be also considered an indirect reaction to public sentiment. This seems to suggest that in Spain, the public had some effect on stringency measures, namely, triggering a stricter response. The significant Granger causality of Twitter entry count on mortality suggests that the increased Twitter traffic happened during the hardest days of the pandemic in Spain, similar to the US. These findings seem to confirm the findings of [26,27], who suggest that public opinion had a part in reintroducing strict government measures during the summer of 2020. During the second wave, the effect of Twitter entry count on government stringency remained with a lag of 2 days, although the rest of the Granger causalities in the first wave had become insignificant except for the relationship of Twitter entry count and deaths per million. However, the directions of this relationship changed. It now shows the decrease of deaths per million a day after the number of tweets increases. This might be because Twitter activity concentrated on the peak of the second wave, after which mortality decreased somewhat. In Spain, antimask events had an echo on Twitter, as their exogenous variable had a significant positive effect on Twitter entry count in the first wave and a significant negative effect on Twitter sentiment in the second wave—however, in both cases only at 10%. Therefore, it can be theorized that during the first wave, the increased Twitter entry count that had a significant effect on mortality was partly due to these antimask events.

We only had a stable and well-specified VAR model for Hungary during the second wave, so only the results of this model are discussed. We had two significant Granger causalities—both effects significant at 5%, but not at 1%. The number of posts on Hungary's FAQ page seemed to be followed by an increase in government stringency a day later. This effect is something similar experienced in Spain, as public opinion was critical of the late government response during the second wave in Hungary [28]. We also found that there was a decrease in the number of FAQ posts a day after deaths per million increased. This is something similar to Spain's second wave: posting activity was concentrated on the peak of the second wave where mortality was highest, after which posting activity somewhat decreased. The antimask event exogenous variable had no significant effect on any of the endogenous variables in Hungary.

These VAR models can also be used to make short-period forecasts for any of the endogenous time series based on the other variables in the model. Therefore, for example, government stringency and mortality in Spain can be estimated based of Twitter entry counts of the previous day. However, this direction was not investigated further due to page limits.

6. Summary

Based on our results, the relationships between social media communication and epidemiological indicators were stronger during the first waves of the pandemic than during the later ones.

The US results were heavily influenced by the presidential election throughout the whole year, as the volume of Twitter comments reacted to government stringency in the first wave, but the sentiment did not seem to be affected. By the second wave, government stringency started to react to changes in the positive rate.

During the first wave of Spain, government stringency along with Twitter volume and sentiment all reacted to changes in the mortality rate. Government stringency lagged 2 days behind the changes, while the Twitter events followed only 1 day late. During the second wave, this relationship was reduced to government stringency reacting to Twitter traffic with a delay of 2 days. It is important to note though that around the second wave of Spain, the first wave of Mexico started as well; thus, Spanish Twitter comments might

reflect this. Antimask events also had some influence on Twitter traffic, but mainly around the first wave.

In Hungary, our model was not stable for the first wave. However, the discovered relationships were very similar to what we experienced in Spain, as government stringency reacted to the volume of comments on the Hungarian FAQ. The reason for the lack of stable results in the first wave was probably the fact that even though there was a huge media hype in the spring of 2020, the number of confirmed cases was considerably lower than in the other waves.

A number of opportunities for further development have been identified. We would like to achieve greater heterogeneity across source platforms in order to reduce the effects of Twitter's typical "telegram" style. As an effect of abbreviated and compressed tweet texts, inaccuracies resulting from dictionary- and word-based text mining methods are presumably present. Another problem with Twitter is the unbalanced age distribution: only 10% of Twitter users are above 50 years [29]. It follows from all of this that it would be advisable to conduct the research based on the content of the much more widely used Facebook platform, or if it is not possible, then additional country-specific sources need to be utilized.

To identify corona topics and conspiracy theories, the utilized tool should be topic modelling; then social network analysis (SNA) can be performed along with topic modelling results. With SNA, we will examine how these topics spread. Finally, it will be possible to compare the results with the official WHO data collected during the pandemic; thus, we can analyze the impact of society on the pandemic and the impact of the pandemic on society.

Author Contributions: Conceptualization, M.M.; L.K. and D.B.; methodology, M.M.; L.K. and D.B.; software, M.M.; L.K. and D.B.; validation, M.M. and L.K.; formal analysis, M.M. and L.K.; investigation, M.M. and D.B.; resources, M.M.; L.K. and D.B.; data curation, M.M.; writing—original draft preparation, M.M. and L.K.; writing—review and editing, M.M.; L.K. and D.B.; visualization, M.M. and L.K.; supervision, D.B.; project administration, D.B.; funding acquisition, D.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Corvinus University of Budapest, grant number EFOP-3.6.3.-VEKOP-16-2017-00007.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Corona-related pandemic data are collected from the site Our World in Data (OWID). https://ourworldindata.org/coronavirus (accessed on 2 February 2021). Tweets had been collected from the Twitter Stream Grab project by Archive Team. The project provides twitter data in compressed downloadable format. https://archive.org/details/twitterstream (accessed on 31 May 2021). Hungarian-language analyzed texts had been collected from the Hungarian FAQ page. This is a Q&A-type website, which is the 31st most visited site in Hungary. Politics and Health categories had been scraped. https://www.gyakorikerdesek.hu/ (accessed on 31 May 2021). Corona-related search terms had been collected from Google Trends Datastore. http://googletrends.github.io/data/ (accessed on 21 May 2021). Corona-sceptic events had been collected from especially national Wikipedia collections. https://en.wikipedia.org/wiki/Protests_over_responses_to_the_COVID-19_pandemic (accessed on 31 May 2021); https://es.wikipedia.org/wiki/Categor%C3%ADa:Manifestaciones_y_protestas_en_2020 (accessed on 31 May 2021).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Taskinsoy, J. The Great Pandemic of the 21st Century: The Stolen Lives. 2020. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3689993 (accessed on 30 June 2021).
2. Centers for Disease Control and Prevention: Estimated Global Mortality Associated with the First 12 Months of 2009 Pandemic Influenza A H1N1 Virus Circulation: A Modelling Study. 2012. Available online: https://www.cdc.gov/flu/spotlights/pandemic-global-estimates.htm (accessed on 31 May 2021).

3. Ortiz-Ospina, E. The Rise of Social Media. Our World in Data. 2019. Available online: https://ourworldindata.org/rise-of-social-media (accessed on 30 June 2021).
4. Centers for Disease Control and Prevention: Statistics Overview. 2020. Available online: https://www.cdc.gov/hiv/statistics/overview/index.html (accessed on 31 May 2021).
5. Douglas, K.M. COVID-19 conspiracy theories. *Group Process. Intergroup Relat.* **2021**, *24*, 270–275. [CrossRef]
6. Vaezi, A.; Javanmard, H.J. Infodemic and Risk Communication in the Era of CoV-19. *Adv. Biomed. Res.* **2020**, *9*, 10. [CrossRef] [PubMed]
7. Roser, M.; Ritchie, H.; Ortiz-Ospina, E.; Hasell, J. Coronavirus Pandemic (COVID-19). Our World in Data. 2020. Available online: https://ourworldindata.org/coronavirus (accessed on 2 February 2021).
8. Toda, H.Y.; Yamamoto, T. Statistical inference in vector autoregressions with possibly integrated processes. *J. Econom.* **1995**, *66*, 225–250. [CrossRef]
9. Ding, D.; Guan, C.; Chan, C.M.; Liu, W. Building stock market resilience through digital transformation: Using Google trends to analyze the impact of COVID-19 pandemic. *Front. Bus. Res. China* **2020**, *14*, 1–21. [CrossRef]
10. Gherghina, Ș.C.; Armeanu, D.Ș.; Joldeș, C.C. Stock Market Reactions to COVID-19 Pandemic Outbreak: Quantitative Evidence from ARDL Bounds Tests and Granger Causality Analysis. *Int. J. Environ. Res. Public Health* **2020**, *17*, 6729. [CrossRef] [PubMed]
11. Similarweb: Top Websites Ranking. 2021. Available online: https://www.similarweb.com/top-websites/ (accessed on 21 May 2021).
12. Bruns, A. After the 'APIcalypse': Social media platforms and their fight against critical scholarly research. *Inf. Commun. Soc.* **2019**, *22*, 1544–1566. [CrossRef]
13. Archive Team: The Twitter Stream Grab. Available online: https://archive.org/details/twitterstream (accessed on 31 May 2021).
14. GyakoriKerdesek Homepage. Available online: https://www.gyakorikerdesek.hu (accessed on 31 May 2021).
15. Google Trends Datastore. Available online: http://googletrends.github.io/data/ (accessed on 21 May 2021).
16. World Health Organization. Overview of Public Health and Social Measures in the Context of COVID-19: Interim Guidance, 18 May 2020. (No. WHO/2019-nCoV/PHSM_Overview/2020.1). World Health Organization. Available online: https://apps.who.int/iris/handle/10665/332115 (accessed on 8 April 2021).
17. Hale, T.; Petherick, A.; Phillips, T.; Webster, S. Variation in Government Responses to COVID-19. Blavatnik School of Government Working Paper 31. 2020. Available online: https://www.bsg.ox.ac.uk/research/publications/variation-government-responses-covid-19 (accessed on 8 April 2021).
18. Hunspell Homepage. Available online: https://hunspell.github.io/ (accessed on 31 May 2021).
19. Zeileis, A.; Kleiber, C.; Kraemer, W.; Hornik, K. Testing and Dating of Structural Changes in Practice. *Comput. Stat. Data Anal.* **2003**, *44*, 109–123. [CrossRef]
20. Stock, J.H.; Watson, M.W. *Introduction to Econometrics, Third Update, Global Edition*; Pearson Education Limited: London, UK, 2015.
21. He, Z.; Maekawa, K. On spurious Granger causality. *Econ. Lett.* **2001**, *73*, 307–313. [CrossRef]
22. Hu, M.; Liu, B. Mining and summarizing customer reviews. In Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, WA, USA, 22–25 August 2004.
23. TASS: Workshop on Semantic Analysis at SEPLN. Available online: http://tass.sepln.org/ (accessed on 25 May 2021).
24. Szabó, M. Experiences of Creation of a Hungarian Sentiment Lexicon. Conference "Nyelv, kultúra, társadalom". Precognox, Budapest. 2014. Available online: http://publicatio.bibl.u-szeged.hu/8791/12/cikk_mszny_2015.pdf (accessed on 8 April 2021).
25. Deane, C.; Parker, K.; Gramlich, J. A Year of U.S. Public Opinion on the Coronavirus Pandemic. 2021. Available online: https://www.pewresearch.org/2021/03/05/a-year-of-u-s-public-opinion-on-the-coronavirus-pandemic/ (accessed on 8 April 2021).
26. Royo, S. Responding to COVID-19: The Case of Spain. *Eur. Policy Anal.* **2020**, *6*, 180–190. [CrossRef]
27. Oliver, N.; Barber, J.X.; Roomp, K.; Roomp, K. Assessing the Impact of the COVID-19 Pandemic in Spain: Large-Scale, Online, Self-Reported Population Survey. *J. Med. Internet Res.* **2020**, *22*, e21319. [CrossRef] [PubMed]
28. Szakacs, G.; Dunai, M. Orban Given Special Powers as Hungary Locks Down against COVID Surge. 2020. Available online: https://www.reuters.com/article/uk-health-coronavirus-hungary-casualties-idUKKBN27Q2MZ\T1\textquoteright (accessed on 8 April 2021).
29. Tankovska, H. Statista-Distribution of Twitter Users Worldwide as of January 2021, by Age Group. Available online: https://www.statista.com/statistics/283119/age-distribution-of-global-twitter-users/ (accessed on 31 March 2021).