Monitoring COVID-19 through text mining and visualization

M.Besher Massri
Jožef Stefan Institute, Slovenia
besher.massri@ijs.si

Joao Pita Costa
Quintelligence, Slovenia
joao.pitacosta@quintelligence.com

Andrej Bauer
University of Ljubljana, Slovenia
andrej.bauer@andrej.com

Marko Grobelnik
Jožef Stefan Institute, Slovenia
marko.grobelnik@ijs.si

Janez Brank
Jožef Stefan Institute, Slovenia
janez.brank@ijs.si

Luka Stopar
Jožef Stefan Institute, Slovenia
luka.stopar@ijs.si

ABSTRACT

The global health situation due to the SARS-COV-2 pandemic motivated an unprecedented contribution of science and technology from companies and communities all over the world to fight COVID-19. In this paper, we present the impactful role of text mining and data analytics, exposed publicly through IRCAI's Coronavirus Watch portal. We will discuss the available technology and methodology, as well as the ongoing research based on the collected data.

KEYWORDS

Text mining, Data analytics, Data visualisation, Public health, Coronavirus, COVID-19, Epidemic intelligence

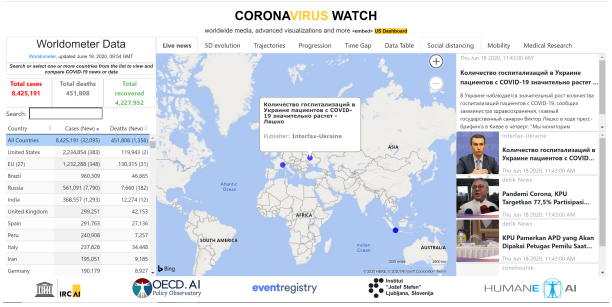

Figure 1: Coronavirus Watch portal

1 INTRODUCTION

When the World Health Organization (WHO) announced the global COVID-19 pandemic on March 11th 2020 [25], following the rising incidence of the SARS-COV-2 in Europe, the world started reading and talking about the new Coronavirus. The arrival of the epidemic to Europe scaled out the news published about the topic, while public health institutions and governmental agencies had to look for existing reliable solutions that could help them plan their actions and the consequences of these.

Technological companies and scientific communities invested efforts in making available tools (e.g. the GIS [1] later adopted by the World Health Organisation (WHO)), challenges (e.g. the Kaggle COVID-19 competition [13]), and scientific reports and data (e.g. the repositories medRxiv [15] and Zenodo [27]).

In this paper we discuss the Coronavirus Watch portal [12], made available by the UNESCO AI Research Institute (IRCAI), comprehending several data exploration dashboards related to the SARS-COV-2 worldwide pandemic (see the main portal in Figure 1). This platform aims to expose the different perspectives on the data generated and trigger actions that can contribute to a better understanding of the behavior of the disease.

2 RELATED WORK

The many platforms that have been made publicly available over the internet to monitor aspects of the COVID-19 pandemics are mostly focusing on data visualization based on the incidence of the disease and the death rate worldwide (e.g., the CoronaTracker [3]). The limitations of the available tools are potentially due to the lack of resolution of the data in aspects like the geographic location of reported cases, the commodities (i.e., other diseases that also influence the death of the patient), the frequency of the data, etc. On the other hand, it was not common to monitor the epidemic through the worldwide news (with some exceptions as the Ravenpack Coronavirus News Monitor [21]).

The Coronavirus Watch portal suggests the association of reported incidence with worldwide published news per country, which allows for real-time analysis of the epidemic situation and its impact on public health (in which specific topics like mental health and diabetes are important related matters) but also in other domains (such as economy, social inequalities, etc.). This news monitoring is based on state-of-the-art text mining technology aligned with the validation of domain experts that ensures the relevance of the customized stream of collected news.

Moreover, the Coronavirus Watch portal offers the user other perspectives of the epidemic monitoring, such as the insights from the published biomedical research that will help the user to better understand the disease and its impact on other health conditions. While related work was promoted in [13] in relation with the COVID-19, and is offered in general by MEDLINE mining tools (e.g., MeSH Now [16]), there seems to be no dedicated tool to the monitoring and mining of COVID-19 - related research as that presented here.

3 DESCRIPTION OF DATA

3.1 Historical COVID-19 Data

To perform an analysis of the growth of the coronavirus, we need to use the historical data of cases and deaths. This data is retrieved from a GitHub repository by John Hopkins University[4]. The data source is based mainly on the official data from the World Health Organization (WHO)[24] along with some other sources, like the Center for Disease and Control[2], and Worldometer[26], among others. This data provides the basis for all functionality that depended on the statistical information about COVID-19 numbers.

Information society '20, October 5–9, 2020, Ljubljana, Slovenia

3.2 Live Data from Worldometer

Apart from historical data, live data about the COVID-19 number of cases, deaths, recovered, and tests are retrieved from the worldometer website. Although the cases might not be as official as the one provided by John Hopkins University (which is based on WHO data), this source is updated many times per day providing the latest up-to-date data about COVID-19 statistics at all times.

3.3 Live News about Coronavirus

The live news is retrieved from Event Registry [10], which is a media-intelligence platform that collects news media from around the world in many languages. The service analyzes news from more than 30,000 news, blogs, and PR sources in 35 languages.

3.4 Google COVID-19 Community Mobility Data

Google's Community Mobility [11] data compares mobility patterns from before the COVID-19 crisis and the situation on a weekly basis. Mobility patterns are measured as changes in the frequency of visits to six location types: Retail and recreation, Grocery and pharmacy, Parks, Transit stations, Workplaces, and Residential. The data is provided on a country level as well as on a province level.

3.5 MEDLINE: Medical Research Open Dataset

The MEDLINE dataset [14] contains more than 30 million citations and abstracts of the biomedical literature, hand-annotated by health experts using 16 major categories and a maximum of 13 levels of deepness. The labeled articles are hand-annotated by humans based on their main and complementary topics, and on the chemical substances that they relate to. It is widely used by the biomedical research community through the well-accepted search engine PubMed [19].

4 CORONAVIRUS WATCH DASHBOARD

The main layout of the dashboard displayed in figure 1 consists of two sides. It is split into the left table of countries, where a simple table of statistics is provided about countries along with the total numbers of cases, deaths, and recovered. On the right side, there is a navigation panel with tabs, each representing a functionality. Each functionality answers some questions and provides insights about a certain type of data.

4.1 Coronavirus Data Table

The data table functionality is a simple table that shows the basic statistics about the new coronavirus. It's taken from Worldometer as it's the most frequently updated source for coronavirus. The data table comes in two forms, one that is a simplified version which is the table on the left, and one contains the full information in a separate tab.

4.2 Coronavirus Live News

The second functionality is a live news feed about coronavirus from around the world. The feed comes from Event Registry, which is generated by querying for articles that are annotated with concepts and keywords related to coronavirus. The user can check for a country's specific news (news source in that country) by clicking on the country name on the left table. As seen in figure 1.

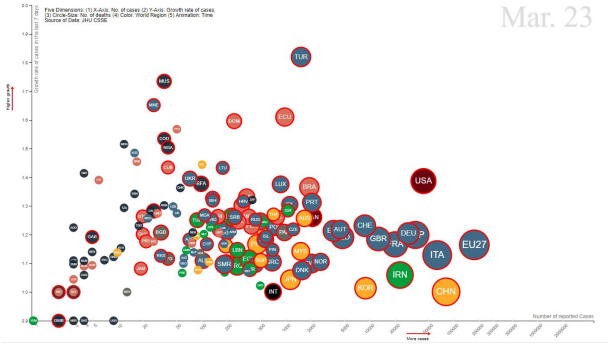

Figure 2: A snapshot of the 5D Visualization on March 23rd. Countries that were at the peak in terms of growth are shown high up like Turkey. Whereas countries that mostly contained the virus are shown down like China.

4.3 Statistical Visualizations

The following set of visualization all aims at displaying the statistics about COVID-19 cases and deaths in a visual format. While they all provide countries comparison, each one focus on different perspective; Some are more complex and focus on the big picture (5D evolution), and some are simple and focus on one aspect (Progression and Trajectory). Besides, all of them have configuration options to tweak the visualization, like the ability to change the scale of the axes to focus on the top countries or the long tale. Or a slider to manually move through the days for further inspection. Furthermore, the default view compares all the countries or the top N countries, depending on the visualization. However, it's possible to track a single country or a set of countries and compare them together for a more focused view. This is done by selecting the main country by clicking on it on the left table and proceeding to select more countries by pressing the ctrl key while clicking on the country.

4.3.1 5D Evolution. 5D Evolution is a visualization that displays the evolution of the virus situation through time. It is called like that since it encompasses five dimensions: x-axis, y-axis, bubble size, bubble color, and time, as seen in figure 2. By default, it illustrates the evolution of the virus in countries based on N. cases (x-axis), The growth factor of N. Cases (y-axis), N. Deaths (bubble size), and country region (bubble color) through time. In addition, a red ring around the country bubble is drawn whenever the first death appears. The growth rate represents how likely that the numbers are increasing with respect to the day before. A growth rate of 2 means that the numbers are likely to double in the next day. The growth rate is calculated using the exponential regression model. At each day the growth rate is based on the N. cases from the previous seven days. The goal of this visualization to show how countries relate to each other and which are exploding in numbers and which ones managed to "flatten the curve", since flattening the curve means less growth rate. It's intended to be one visualization that gives the user a big picture of the situation.

4.3.2 Progression. The progression visualization displays the simple Date vs N. cases/deaths line graph. It helps to provide a simplistic view of the situation and compare countries based on the raw numbers only. The user can display the cumulative

numbers where each day represents the numbers up to now, or daily where at each date the numbers represent the cases/deaths on that day only.

4.3.3 Trajectory. While the progress visualization displays the normal date vs N. cases/deaths, this visualization seeks to compare how the trajectory of the countries differ starting from the point where they detect cases. This visualization helps to compare countries' situations if they all start having cases on the same date. The starting point has been set to the day the country reaches 100 cases, so we would compare countries when they started gaining momentum.

4.4 Time Gap

The time gap functionality tries to estimate how the countries are aligned and how many days each country is behind the other, whether that is in the number of cases or deaths. This assumes that the trajectory of the country will continue as it with taking much more strict/loose measurements, which is a rough assumption. It helps to estimate how bad or good the situation in terms of the number of days. To see the comparison, a country has to be selected from the table on the left. However, not all countries are comparable as they have very different trajectories or growth rates.

The growth of each country is represented as an exponential function, the base is calculated using linear regression on the log of the historical values (that is, exponential regression). Based on that, the duplication N. days, or the N. days the number of cases/deaths will double is determined. two countries are comparable if they have a reasonable difference is the base or doubling factor. If they are comparable, we see where the country with the smaller value fits in the historical values of the country with the larger numbers, with linear interpolation if the number is not exact, hence the decimal values.

4.5 Mobility

The mobility visualization is based on google community mobility data that describe how communities in each country are moving based on 6 parameters: Retail and recreation, Grocery and pharmacy, Parks, Transit stations, Workplaces, and Residential. The data is then reduced to 2-dimensional data while keeping the Euclidean proximity nearly the same. The visualization can indicate that the closer the countries are on the visualization, the similar the mobility patterns they have. The visualization uses the T-SNE algorithm for dimensionality reduction [23], which reduces high dimensional data to low dimensional one while keeping the distance proximity between them proportionally the same as possible. The algorithm works in the form of iterations, at each iteration, the bubbles representing the country are drawn. We used those iterations to provide animation to the visualization.

4.6 Social Distancing Simulator

The Social Distancing simulator is displayed in figure 3. Each circle represents a person who can be either healthy (white), immune (yellow), infected (red), or deceased (gray). A healthy person is infected when they collide with an infected person. After a period of infection, a person either dies or becomes permanently immune. Thus the simulation follows the Susceptible-Infectious-Recovered-Deceased (SIRD) compartmental epidemiological model.

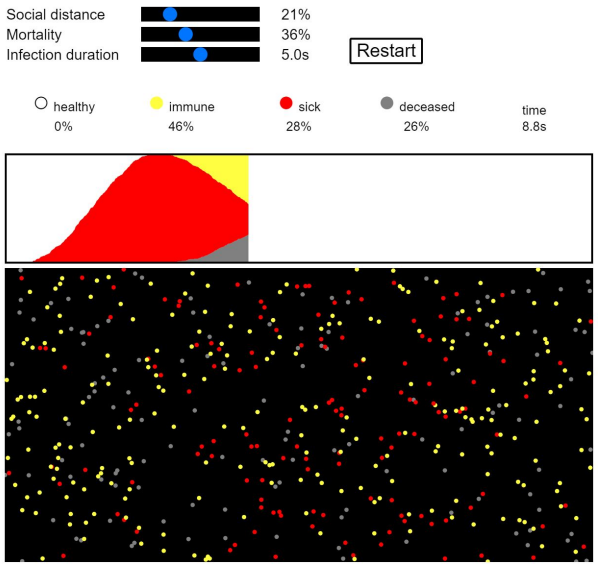

Figure 3: A snapshot of the Social Distancing Simulator. The canvas show a representation of the population. with red dots representing sick people, yellow dots representing immunized people, and grey dots represent deceased people.

The simulator is controlled by three parameters. First, Social distancing that controls to what extent the population enforces social distancing. At 0% there is no social distancing and persons move with maximum speed so that there is a great deal of contact between them. At 100% everyone remains still and there is no contact at all. Second, mortality is the probability that a sick person dies. If you set mortality to 0% nobody dies, while the mortality of 100% means that anybody who catches the infection will die. Finally, infection duration determines how long a person is infected. A longer time gives an infected person more opportunities to spread the infection. Since the simulation runs at high speed, time is measured in seconds.

4.7 Biomedical Research Explorer

To better understand the disease, the published biomedical science is the source that provides accurate and validated information. Taking into consideration a large amount of published science and the obstacles to access scientific information, we made available a MEDLINE explorer where the user can query the system and interact with a pointer to specify the search results (e.g., obtaining results on biomarkers when searching for articles hand-annotated with the MeSH class "Coronavirus").

To allow for the exploration of any health-related texts (such as scientific reports or news) we developed an automated classifier [5] that assigns to the input text the MeSH classes it relates to. The annotated text is then stored in Elasticsearch [18], from where it can be accessed through Lucene language queries, visualized over easy-to-build dashboards, and connected through an API to the earlier described explorer (see [8], [20] and [17] for more detail).

The integration of the MeSH classifier with the worldwide news explorer Event Registry allows us to use MeSH classes in the queries over worldwide news promoting an integrated health news monitoring [9] and trying to avoid bias in this context [7]. An obvious limitation is a fact that the annotation is only

available for news written in the English language, being the unique language in MEDLINE.

5 CONCLUSION AND FUTURE WORK

In this paper, we presented the coronavirus watch dashboard as a use-case of observing pandemic. However, this methodology can be applied to other kinds of diseases given the availability of similar data. For further development, we plan to implement a local dashboard for other countries as well which would provide local data in the local language. In addition, given the existence of more than seven months of historical data, we would like to build some predictive models to predict the number of cases/deaths in the next few days.

Moreover, we are using the StreamStory technology [22] in order to: (i) compare the evolution of the disease between countries by comparing their time-series of incidence; (ii) investigate the correlation between the incidence of the disease with weather conditions and other impact factors; and (iii) analyze the dynamics of the evolution of the disease based on incidence, morbidity, and recovery. This technology allows for the analysis of dynamical Markov processes, analyzing simultaneous time-series through transitions between states, offering several customization options and data visualization modules.

Furthermore, following the work done in the context of the Influenza epidemic in [6], we are using Topological Data Analysis methods to understand the behavior of COVID-19 throughout Europe. In it, we examine the structure of data through its topological structure, which allows for comparison of the evolution of the epidemics within countries through the encoded topology of their incidence time series.

ACKNOWLEDGMENTS

The first author has been supported by the Knowledge 4 All foundation and the H2020 Humane AI project under the European research and innovation programme under GA No. 761758), while the second author was funded by the European Union research fund 'Big Data Supporting Public Health Policies', under GA No. 727721. The third author acknowledges that this material is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-17-1-0326.

REFERENCES

[1] ArcGIS. 2020. ArcGIS who covid-19 dashboard. https://covid19.who.int/. (2020).
[2] CDC. 2020. Center for Disease Control and Prevention. https://www.cdc.gov/coronavirus/2019-ncov/index.html. (2020).
[3] CoronaTracker. 2020. CoronaTracker. https://www.coronatracker.com/analytics/. (2020).
[4] CSSE. 2020. Covid-19 data repository by the center for systems science and engineering (csse) at johns hopkins university. https://github.com/CSSEGISandData/COVID-19. (2020).
[5] J. Pita Costa et al. 2020. A new classifier designed to annotate health-related news with mesh headings. *Artificial Intelligence in Medicine*.
[6] J. Pita Costa et al. 2019. A topological data analysis approach to the epidemiology of influenza. In *Proceedings of the Slovenian KDD conference*.
[7] J. Pita Costa et al. 2019. Health news bias and its impact in public health. In *Proceedings of the Slovenian KDD conference*.
[8] J. Pita Costa et al. 2020. Meaningful big data integration for a global covid-19 strategy. *Computer Intelligence Magazine*.
[9] J. Pita Costa et al. 2017. Text mining open datasets to support public health. In *WITS 2017 Conference Proceedings*.
[10] EventRegistry. 2020. Event Registry. https://eventregistry.org. (2020).
[11] Google. 2020. Google COVID-19 Community Mobility Report. https://www.google.com/covid19/mobility/. (2020).
[12] IRCAI. 2020. IRCAI coronavirus watch portal. http://coronaviruswatch.ircai.org/. (2020).
[13] Kaggle. 2020. Kaggle covid-19 open research dataset challenge. https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge. (2020).
[14] MEDLINE. 2020. MEDLINE description of the database. https://www.nlm.nih.gov/bsd/medline.html. (2020).
[15] medRxiv. 2020. medRxiv covid-19 sars-cov-2 preprints from medrxiv and biorxiv. https://connect.medrxiv.org/relate/content/181. (2020).
[16] MeSHNow. 2020. MeSHNow. https://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/MeSHNow/. (2020).
[17] MIDAS. 2020. MIDAS COVID-19 portal. http://www.midasproject.eu/covid-19/. (2020).
[18] Elastic NV. 2020. Elasticsearch portal. https://www.elastic.co/. (2020).
[19] PubMed. 2020. PubMed biomedical search engine. https://pubmed.ncbi.nlm.nih.gov/. (2020).
[20] Quintelligence. 2020. Quintelligence COVID-19 portal. http://midas.quintelligence.com/. (2020).
[21] Ravenpack. 2020. Ravenpack coronavirus news monitor. https://coronavirus.ravenpack.com/. (2020).
[22] Luka Stopar. 2020. StreamStory. http://streamstory.ijs.si/. (2020).
[23] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9, (November 2008), 2579–2605.
[24] WHO. 2020. WHO Coronavirus portal. https://www.who.int/emergencies/diseases/novel-coronavirus-2019. (2020).
[25] WHO. 2020. World Health Organization who director-general's opening remarks at the media briefing on covid-19 - 11 march 2020. https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020. (2020).
[26] WorldoMeters. 2020. WorldoMeters. https://www.worldometers.info/coronavirus/. (2020).
[27] Zenodo. 2020. Zenodo coronavirus disease research community. https://zenodo.org/communities/covid-19/. (2020).