**DATA624 - DATATHON #5**

**GROUP 2 – MINKI LEE, RYAN LITTLEWOOD, ADRIAN MITCHELL**

**MARCH 10, 2021**

**STUDIES OF DIGITAL SURVEILLANCE OF PUBLIC HEALTH**

**Introduction**

Digital platforms include search engines and social media. More than half of the world are using online platforms, and every day, the number of online platform users increases (Pérez-Escoda et al., 2020; Abad et al., 2021). People share health information on online platforms, which became the first source of information for many people. Therefore, digital data have been a popular subject in public health surveillance studies. Our main interest is to observe different platforms that researchers used for data collection and investigate the issues examined globally. Research on these issues may reflect the differences in the burden of health issues around the world.

The guiding questions that we wanted to explore, and their rationale, were as follows:

1. *Which online platform has been used the most by researchers around the world to examine health issues?*

There are many different types of social media platforms people use and they differ by country. Some countries don’t have access to certain platforms. Our team was interested to find out different platforms used for social media analysis for public health around the world.

1. *Looking at the top 3 platforms used by researchers, what were the top issues examined globally?*

By limiting the number of platforms, it will help us to properly determine what is being investigated on the platforms. Our team will be aiming to determine the main topics being researched on these platforms, which will help us to provide a better analysis.

1. *Looking at the counts of research studies on the top Health Event Under Surveillance, can we see which countries have the most research, and how have the research counts varied over time?*

We were interested in examining how researchers globally are studying online surveillance of the top health events. Not only would this highlight the varying levels of support for research into these issues by country, but also different countries’ interest in the health events themselves – possibly giving insight into the distribution of the severity of the event under surveillance.

Additionally, looking at how the research counts have varied over time, by country, could highlight how the event (or interest in the event) has travelled the globe temporally. This examination should also serve to highlight how online surveillance, or the research in online surveillance, has spread globally in recent years.

**Datasets**

“Communicable\_Diseases\_Surveillance.csv”

This dataset contains 22 columns of data which span 756 rows. The columns within the dataset included fields we explored including: # of Authors, Author #1 Country, Platforms, Date of Publication, Health Event under Surveillance, and Substance Drug/Disease/Name. Since this dataset was so large in nature, it was very important to generate specific research questions that would limit the amount of data used in analysis, allowing for much more specific results. Different sets of variables were used for parts of the analysis, because they gave us the ability to create a story with our data using multiple combinations of variables.

“CDC USA Influenza.csv”

This dataset contains two columns, one for the year that flu season ended (2010 for example was the end of the 2009-2010 flu season), and one for the estimated number of cases of symptomatic influenza in the USA. This data was found on the Centers for Disease Control and Prevention (CDC) webpage (U.S. Department of Health & Human Services, 2020).

**Analysis**

We investigated digital public health surveillance data. Data cleaning and wrangling were performed using python. Calculations and visualisations were performed using Tableau.

**Global Map with Pie Charts**

The public health surveillance dataset was filtered using Python to investigate the top 5 main platforms that the researchers used for data collection. There were missing values and incorrect inputs in the number of platforms column and inconsistent data input in the co-author's country columns in the original dataset. Therefore, we focused on the lead author’s country and the main platforms. Unnecessary columns were dropped, and only the columns of interest, such as “Author’s country #1” and “Platform #1”, were extracted. Each row was assigned a new value of one in the number of platforms column. The top 5 platforms remained unchanged regardless of adding the number of additional platforms that researchers used. The cleaned dataset was then imported into Tableau to display a global map with pie charts, a line chart, and a bar chart. Pie charts illustrate the numerical proportion of the main platforms, a line chart shows the number of publications per year, and a bar chart indicates the proportion of ‘yes’ or ‘no’ responses regarding the usefulness of platforms.

**Sunburst**

The dataset was read into Python, which allowed us to filter the data and prepare it for analysis. There were many unnecessary columns in the dataset which were dropped, as they offered no value for the analysis our group was performing. The columns that were left were only the columns of interest: “Platform #1”, “Health Event Under Surveillance”, “Sub Surveillance”, “Substance drug disease/name”, and “Author Country”. Once this was done, a column was added to the end of the remaining data frame labelled “Total” which was created to help perform the visual analysis on the remaining data. The last step before visualizing the data was to filter the data frame so it only represented the top 3 platforms found in the analysis from question 1: “Twitter”, “Facebook” and “Google Trends”.

A path was created in Python using the data, and a sunburst plot was created. This plot was used to better help analyse the proportions within the data. From this we found that the top platform was Twitter, and the most referenced event was Communicable Disease.

**Choropleth**

Counts for research into online surveillance on the top Health Event Under Surveillance (Communicable Diseases, see Findings) were filtered out of the main dataset using python. Entries for all platforms and Author’s Country were merged with a country code table for use with the Plotly package for python. A global choropleth for research count by country was generated and published to the internet for visualisation in a Tableau dashboard.

**Stacked Bar Chart**

Including “Year of Publication”, “Platform”, and “Substance Drug/Disease/name” in the same source table as used for the Choropleth allowed for visualisation of the data in a stacked bar chart in Tableau. The data was ordered by year, from 2006 to 2019, with the counts of research for each country stacked in ascending order according to total research count by authors from each country.

“Platform” and “Substance Drug/Disease/name” were enabled as filters, so that specific research on those data fields could be visualised over the time period. The stacked bar chart was interactive, so that countries could be highlighted individually, and so that values for each section, or block, of a stacked column could be visualised by hovering the mouse over the block.

**Bar and Line Plot**

This visualisation included bars for the number of studies of online surveillance of influenza by authors from the US (top health event studied and top country, see Findings). Overlayed on top of the bars were estimated numbers of cases of influenza in the USA from 2010-2019. This plot serves to demonstrate the domestic relevance of the research being undertaken by showing how the research count reflects (or does not reflect) the domestic prevalence of the disease.

**Findings**

**Top Platforms Researched**

The top 5 main platforms that researchers used for data collection were: 1. Twitter, 2. Google Trends, 3. Facebook, 4. Specific Websites, 5. Weblogs. The US published the greatest number of research papers, followed by the UK, Australia, Italy, and Canada, and Twitter was the main data source. A line chart shows the total number of publications per year, and we can see that there is an increasing trend from 2005 to 2019. In 2019, the greatest number of papers were published. More than 94 % of researchers found the platforms useful and among those who did not find the platforms useful for the objectives of the study, interestingly, Twitter was the least useful platform.

**Online Surveillance Research on Twitter, Google Trends and Facebook**

Of the top 5 platforms the main points of interest became 1. Twitter, 2. Google Trends, and 3. Facebook. Twitter was the most used platform of those top 3. By looking at Twitter as the only platform we discovered that the most referenced health event was Communicable Diseases, with Behavioural Risk factors being a close second. By limiting the search to only Communicable Diseases, we then determined that the most referenced Substance/Drug/Disease name is ILI/Influenza. None of these findings are surprising when we put into perspective the findings discussed from the Top Platforms Research.

**Research on Online Surveillance of Communicable Diseases**

Globally, research into online surveillance of communicable diseases has been undertaken by researchers in the USA far more than any other country. From 2006 to 2019 there were 97 studies on online surveillance of communicable diseases undertaken by researchers from the USA. Of those, 27 were about ILI/Influenza. In the last seven years captured by this data, four or more studies on digital surveillance of ILI/Influenza were undertaken by US researchers, with the peak year being 2015, when there were six studies on the subject. The first seven years covered by this data only included a total of six US studies on the subject. This may reflect the trend of increased adoption of online platforms for mass use by the public. All three of the top platforms used for research on online surveillance were only launched for public use in 2006.

To better understand the domestic relevance of the research being undertaken, numbers for US symptomatic cases of influenza were displayed against counts of research. The two data show a similar trend of increasing with time. Considering that influenza has continued to increase from 2010 to 2019, the research undertaken likely has greater relevance to US public health.

**Limitations**

To truly understand how public health events have spread across the globe, numbers of symptomatic cases for each country, measured with similar reporting requirements, should be examined. Research counts may in-fact be a better reflection of the availability of research funding and resources around the world.

**Conclusion**

We explored the data collected from research papers that have been published from 2005 to 2020 on social media analysis for public health. The accessibility of social platforms differs among countries, which led to differences in the main platforms used for research. Moreover, according to the National Science Board (2019), “specialization in scientific fields differs among countries, with the United States, the EU, and Japan more specialized in health sciences and China and India more specialized in engineering”. The number of scientific publications on public health increases every year, as well as social platform usage. Since online users share various information related to health on digital platforms, health care professionals can use the information to engage with the public to promote and educate health behaviours and create innovative health care policies and programs.

**References**

Abad, Z. S. H., Kline, A., Sultana, M., Noaeen, M., Nurmambetova, E., Lucini, F., ... & Lee, J. (2021). Digital public health surveillance: a systematic scoping review. *npj Digital Medicine*, *4*(1), 1-13.

National Science Board, National Science Foundation. 2019. Publication Output: U.S. Trends and International Comparisons. *Science and Engineering Indicators 2020*. NSB-2020-6. Alexandria, VA. Available at <https://ncses.nsf.gov/pubs/nsb20206/>.

Pérez-Escoda, A., Jiménez-Narros, C., Perlado-Lamo-de-Espinosa, M., & Pedrero-Esteban, L. M. (2020). Social Networks’ Engagement During the COVID-19 Pandemic in Spain: Health Media vs. Healthcare Professionals. *International journal of environmental research and public health*, *17*(14), 5261.

U.S. Department of Health & Human Services, 2020. *Centers for Disease Control and Prevention.* [Online]   
Available at: https://www.cdc.gov/flu/about/burden/past-seasons.html  
[Accessed 9 March 2021].