CURBING THE SPREAD

AN EXPLORATORY STUDY OF THE COVID-19 INFODEMIC IN THE UNITED STATES

MASTER THESIS | INFORMATION STUDIES

DATE	7th OCTOBER 2020
AUTHOR	ANDREAS WINDFELD
SUPERVISOR	FLORIAN MAXIMILIAN MEIER

CHARACTERS	191.787
PAGES	79.9

AALBORG UNIVERSITY
DENMARK

ABSTRACT

This thesis explores the COVID-19 infodemic in the United States. The Coronavirus has severely impacted the US, and misinformation is a potential contributing factor making it an exciting area of research. The study is executed by analysing search query logs from the Microsoft Bing search dataset for Coronavirus intent (2020). The main problem statement concerns misinformation search behaviour from January to August 2020. The study explores overall distributions of misinformation, including state distributions, popular misinformation types, and potential impacting factors. Various methods are used throughout the study, including manual content-coding, supervised machine learning for automatic text classification, and exploratory analysis. Findings are validated by using data from Google Trends. The top five misinformation searches were "*Qanon*", "*herd immunity*", "*hydroxychloroquine coronavirus*", "*Bill Gates coronavirus*", and "*malaria drugs for coronavirus*". Initially, people were mostly searching for misinformation related to the origin of the virus. Later this changed to include miracle-cures, alternative treatments, and conspiracy theories. Wyoming was observed as having a significantly higher misinformation level both relative to total queries from the state and population size. A trend was observed of misinformation moving from the states with the largest US cities towards rural states in recent months. No association was found between state-level implemented COVID-19 policies, political orientation and misinformation levels. This study has provided insight into the changes in misinformation during COVID-19. It is suggested that further research continues the work by researching causal connections, misinformation types through topic modelling and unsupervised learning, and utilizing a similar approach to investigate the infodemic in other countries.

Keywords: Infodemic, misinformation, automatic text classification, search query logs

TABLE OF CONTENTS

1 INTRODUCTION 3

1.1 PROBLEM STATEMENT 4
1.1.1 Research Questions & Flow 4
1.2 CONTRIBUTION 4

2 LITERATURE REVIEW 5

2.1 SEARCH PROCESS 5
2.2 BACKGROUND - INTRODUCING INFODEMICS 7
2.3 COVID-19 INFODEMICS & SOCIAL MEDIA 8
2.4 INFODEMIOLOGY – IMPROVING PUBLIC HEALTH WITH DATA 12
2.5 RESEARCHING EPIDEMIC OUTBREAKS WITH SEARCH QUERY LOGS & GOOGLE TRENDS 13
2.6 SEARCH QUERY LOGS 18
2.7 FINDINGS 19

3 METHODOLOGY 20

3.1 RESEARCH DESIGN 20
3.2 DATA 21
3.2.1 Bing Search Dataset for Coronavirus Intent 21
3.2.2 CoronaNet Research Project Data 24
3.2.3 Google Trends 24
3.3 SOFTWARE 25
3.4 PROCEDURE 26
3.4.1 Data Cleaning & Preparation 26
3.4.2 Keyword Index 27
3.4.3 Sampling 30
3.4.4 Manual Coding 31
3.4.5 Building a Classification Model 32
3.4.6 Exploratory Analysis 37
3.5 RELIABILITY & VALIDITY 37
3.6 ETHICAL CONSIDERATIONS 39

4 ANALYSIS & RESULTS 41

4.1 RESULTS OF MANUAL CODING 41
4.2 RESULTS OF AUTOMATIC TEXT CLASSIFICATION 43
4.2.1 Evaluating Model Performance 43
4.2.2 Selecting the Final Model 47
4.3 EXPLORATORY ANALYSIS 50
4.3.1 Overall Distributions 50
4.3.2 Search Queries by State 54
4.3.3 Top Search Queries 62
4.3.4 Exploring Other Data Sources (CoronaNet and Election) 70
4.4 IS THE DATA REPRESENTATIVE? 75

5 DISCUSSION & CONCLUSION 79

6 REFERENCES 83

7 APPENDICES 95

1 INTRODUCTION

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), which causes the coronavirus disease COVID-19, was first observed in China in December 2019. It was a previously unknown strain of the respiratory disease, and it has been spreading rapidly throughout the world in 2020 (Danish Health Authority, 2020). The World Health Organization declared COVID-19 a global pandemic on March 11, 2020. Globally, more than 32.7 million COVID-19 cases and almost one million deaths have been reported (WHO, 2020). The Americas, including both south- and north America, are severely impacted and accounts for more 50% of all new cases and 55% of all recorded deaths. The United States has the highest number of recorded deaths at 207.072 and is also among the countries with the highest number of deaths per capita (Statista, 2020). Compared to many other countries, the United States has continually failed to implement measures to limit the spread of the coronavirus. Under President Trump, the government has downscaled measures to counter potential epidemics, including the Centers for Disease Control and Prevention and defunding of epidemiological researchers in China. The United States has had an empty seat at the WHO executive board for two years, just recently filling the position in May - months into the pandemic (Yong, 2020). The virus has not been taken seriously by the leadership of the United States, which have resulted in high numbers of infected and diseased.

While the above factors have certainly played a part in the rapid spread, a recent study from Cornell University found that President Trump is "*likely the largest driver of the COVID-19 misinformation infodemic*" (Evanega et al., 2020). *Infodemic* is a concept introduced by the World Health Organization, concerned with misinformation in the context of an epidemic. Donald Trump has spread crucial misinformation throughout the outbreak, especially by downplaying the severity of the virus and outright presenting misinformation about miracle cures as scientific facts. Misinformation has potentially played an essential part in the United States' inability to contain the virus, as it impedes the distribution of helpful evidence-based information. In April 2020, the World Health Organization called for action in response to the COVID-19 infodemic, as misinformation was considered a severe threat to the containment of the outbreak (Tangcharoensathien, 2020). One of the suggested actions was to explore data to measure the impact and trends of the infodemic, which has been further backed by related research (Hua & Shaw., 2020, Leitner et al., 2020) and the Pan American Health Organization (paho.org, 2020).

This thesis will explore the COVID-19 infodemic of the United States by analysing search query logs from the Microsoft Bing search dataset for Coronavirus Intent (2020). Search queries can be considered a proxy for the public interest, but they do not necessarily represent the opinions of the users (Bento et al., 2020). This thesis will define the overall extent of misinformation in the United States, explore differences between states, and investigate the types of misinformation they represent. It lays the groundwork for an assessment of why the United States has been impacted so severely by COVID-19, and findings can be useful in future infodemiological studies. At the time of writing, few studies are concerned with COVID-19 misinformation, as represented in search query logs,

making it very relevant to the current state of the United States. This thesis responds to the call for action by the World Health Organization by performing exploratory analysis based on the following problem statement.

1.1 Problem Statement

How has COVID-19 prompted misinformation seeking behaviour changed during the recent pandemic in the United States?

The problem statement is concerned with identifying misinformation in search query logs and exploring how this has changed over time. This will be done through an exploratory analysis based on the research questions below. The United States was selected due to the severity of the COVID-19 outbreak there and a large amount of available data from the country.

1.1.1 Research Questions & Flow

- RQ1. How can search queries related to COVID-19 misinformation be identified?
- RQ2. What is the overall extent of expressed interest in COVID-19 misinformation in the United States?
  - How does this differ across states?
- RQ3. What are the top search queries related to misinformation, and how has this changed over time?
- RQ4. What are some possible explanations for variations between states?

The first research question (RQ1) is vital to the subsequent findings of the study. This will be answered by creating a keyword index of misinformation, extracting queries for manual coding, and using results for supervised machine learning. When the entire dataset is classified, the remaining research questions will be answered in the exploratory analysis. RQ4 will introduce some perspectives on how findings can be used, but further research is necessary if the objective is to establish causality between variables.

The research scope is limited to misinformation from the United States on the Microsoft Bing platform, but selected findings are validated by exploring related data from Google Trends. Several steps of the study, including keyword index, classification, and perspectives, is limited to work with data from the United States. Each country has different languages and types of misinformation, so the process would have to be altered for similar research within other countries.

1.2 Contribution

This study has the unique opportunity of researching misinformation while the coronavirus is still ongoing. At the time of writing, few studies are released about this topic. Similar research uses Google Trends, while this study provides insight from a different platform. Findings can potentially inform future actions aimed at containing the epidemic in the United States. Finally, this study provides a foundation for further research investigating causality.

2 LITERATURE REVIEW

An essential element of any research is that it needs to be informed by existing knowledge in a subject area (Rowley & Slack, 2004). Given the focus on the new COVID-19 virus in this thesis, it was essential to stay up to date on the most recent research, introduced daily during the process. To structure the search process and the following review; information search was divided into thematic categories or concepts which covered the main themes of the research questions. The main concepts covered in the study are:

- *Infodemics*. This concept was the primary motivation for the research in this thesis and was covered in the initial searches.
- *Infodemiology*. Given the novel nature of infodemics, past research about information in the context of a pandemic was investigated.
- *Search query logs analysis*. Including literature that analysed search query logs was an overarching focus in the entire review. A brief section was dedicated to exploring the history, methodological foundation, benefits and limitations of working with data in this format.

2.1 SEARCH PROCESS

A primary objective during the search process was to locate research that would allow me to figure out where this thesis fits into the current body of research. This was done by considering the contribution, impact, logic, and thoroughness of the literature as described by Webster & Watson (2002). While not addressing these categories systematically, they were a deciding factor in the selection of the final literature. The search process was iterative and took place during the entire writing process – the primary objective being to find literature that would allow me to answer the research questions. It was decided not to do an initial systematic review, as new research was continually added, which would be easier identified using other methods (Hammersley, 2019).

The primary literature on the methodological foundation was selected based on previous knowledge obtained during my studies as well as literature obtained through personal interests. These also include literature providing the technical foundation for the analysis for examples *Social Research Methods* (Bryman, 2012), *Handbook of Research on Web Log Analysis* (Jansen et al., 2009), *Text Mining with R* (Silge & Robinson, 2020), and *R for Data Science* (Wickham & Grolemund, 2017). The remaining literature, both including books and scientific articles, were located through literature search and inspiration from the supervisor. Given the context of this thesis, investigating a pandemic that is still ongoing, a lot of scientific articles discovered in the literature search are still in pre-print or accepted but not yet officially published in a scientific journal. The primary platforms used in the search process were Google Scholar[1] and the online platform of the Aalborg University Library[2] (AUB). Both platforms search

[1] https://scholar.google.com/
[2] https://www.en.aub.aau.dk/

across many different databases, making it possible to get a broad view of the current research without being restricted to a specific research discipline. The comprehensive search across multiple databases was important as several of the most important related work came from very different types of scientific journals which could potentially be missed if the inquiry was too restricted. Some notable journals that were responsible for revealing several of the articles used in this paper were *ACM Digital Library* (Downey, 2008)*, JMIR Publications (Journal of Medical Internet Research)* (Rovetta & Bhagavathula., *2020b:* Tangcharoensathien et al., 2020)*, Cornell University* (Suh et al., 2020)*,* or even *The American Journal of Tropical Medicine and Hygiene* (Islam et al., 2020).

The search strategy consisted of several different steps, including brief searches, building blocks and citation pearl growing (Rowley & Slack, 2004. Cronin & Ryan, 2008). Initially, brief searches were used to get a broad understanding of the disciplines in play as well as identifying some of the most cited sources related to the topic. Examples of the initial searches included the search terms "*infodemics*", "*information AND epidemic*", "*infodemiology*", *"infodemiology and misinformation*" and *"infodemiology AND SARS/Ebola/Zika/H1N1/COVID-19*". The initial search on infodemics revealed that most articles were related to COVID-19 and misinformation. The term was defined by the World Health Organization who just recently introduced the concept concerning COVID-19. Searches on *information AND epidemic* revealed that infodemics are closely related to the term infodemiology, which has been the dominant method for years in research of information and disease outbreaks. The further search focused explicitly on studies of misinformation as well as infodemiological studies covering previous global pandemics. This process can be viewed as a building blocks approach, as each search builds on top of the last. The search is modified by the knowledge gained in the previous search results.

Outside of the brief search and building blocks methods, the primary method of the literature search was *citation pearl growing* or *snowballing* (Rowley & Slack, 2004). This method starts with a few selected papers and builds on those by investigating both sources in the articles as well as other documents citing them. A benefit of the method is that it is simple to use, while also being able to lead the user to sources that, for one reason or another, would be challenging to find in the scientific databases. Further supported by Greenhalgh and Peacock (2005), who found pearl growing to be very time efficient as well as being able to find sources traditional systematic methods were not able to. The technique can essentially go on forever, or at least until every single relevant paper is identified. For this thesis, the process was repeated throughout the entire process of the project, as new knowledge and exciting insight frequently emerged. The review is structured according to the main themes as described above. Another popular approach is to review literature in *chronological order*, which can be useful to determine how research has changed and revealing new findings (Randolph, 2009). While the main idea of this review was to introduce the main concepts, it is structured in *reverse chronological order*. Starting from the new term infodemics and working backwards to understand how the term came to be and why it is crucial now.

2.2 Background - Introducing Infodemics

The COVID-19 pandemic introduced a phenomenon described by the World Health Organisation as an " "Infodemic" - "*an overabundance of information – some accurate and some not – occurring during an epidemic. It makes it hard for people to find trustworthy sources and reliable guidance when they need it. Even when people have access to high-quality information, there are still barriers they must overcome to take the recommended action. Like pathogens in epidemics, misinformation spreads further and faster and adds complexity to health emergency response"* (WHO, 2020). While misinformation is undoubtedly not a new concept in the age of the World Wide Web, an infodemic relates to the specific problems caused by the spread of misinformation and fake news in the context of a pandemic. According to the WHO director-general Tedros Adhanom Ghebreyesus the infodemic is spreading rapidly alongside the COVID-19 epidemic, and both are important to contain the spread of the virus (Zarocostas, 2020). A surge of new information in the wake of an epidemic is a known phenomenon that can be traced to the Middle Ages. However, in the age of the World Wide Web, and especially social media, the impact of misinformation is much more severe as it can spread quickly on a global scale (Bode & Vraga., 2018)., Shahi et al., 2020). The spread is more challenging to contain in the current age given the wide range of different media for news consumption, which was previously mostly restricted to selected media like TV, radio or newspapers. Social media has proven to be a severe source of misinformation and is one of the primary targets in the battle against global infodemics (Cinelli et al., 2020). To contain the virus and limit its spread, people must have the right information on how to act in their daily search activities. WHO has launched the new platform *WHO Information Network for Epidemics* (EPI-WIN) who are in contact with several of the larger social media platforms such as Twitter, Facebook, Tencent, and TikTok. When a surge of misinformation is identified, the topic is sent to WHO's technical risk communications team and, when possible, they provide evidence-based answers (Zaracostas, 2020). Another initiative by the WHO is the introduction of information boxes containing advice from reliable sources (e.g. WHO, Ministry of Health or Centre for Disease Control). On Google, these were implemented to make sure that the first information the user meets is from trusted sources and covers frequently requested info on COVID-19 related themes. On April 7 and 8, 2020, the WHO Information Network for Epidemics arranged a global virtual conference addressing the COVID-19 infodemic[3], with the primary objective of crowdsourcing ideas to establish an infodemic response framework. Invitations were sent to key partners from multiple different professional disciplines including risk communication, health information systems, research and science, policy analysis, evidence synthesis, digital health, community response, and humanitarian response (Tangcharoensathien et al., 2020). More than 1400 individuals signed up for the conference, representing 111 countries and multiple professional sectors. To define a framework that could sufficiently cover the interdisciplinary nature of the infodemic; different thematic categories were established. The output of the virtual conference was further examined by Tangcharoensathien et al. (2020), who used narrative analysis to gather the 594 collected suggestions into five thematic categories. Most of the recommendations were concerned with the amplification and reach of evidence-based and credible information (44%). The distribution of the remaining

[3] https://www.who.int/teams/risk-communication/infodemic-management/1st-who-infodemiology-conference

four thematic categories was: scanning and verifying evidence (18%), explaining the science (20%), measuring the infodemic and assessing trends and impacts (12%), and coordination of governance (6%).

2.3 COVID-19 Infodemics & Social Media

Infodemics is a new term coined by the WHO in response to the global surge of misinformation following the COVID-19 pandemic. However, the concept of information negatively impacting the spread and consequences of pandemics is no new concept. In the 2019 Ebola outbreak in the Democratic Republic of Congo, misinformation resulted in increased violence, mistrust, social disturbance and attacks on healthcare workers. The SARS outbreak in China, caused fear and anxiety across the population, which resulted in certain demographic groups becoming stigmatized (Person et al., 2004). This impacted the pandemic negatively as the stigmatized social groups potentially wouldn't seek medical assistance if needed, which in turn could spread the disease further than if medical attention was sought immediately (Islam et al., 2020). These are examples of how misinformation can negatively impact the spread of diseases. Furthermore, it introduces several different categories that each play a part in the context of a global pandemic, including rumours, stigma and conspiracy theories. Each of which has occurred during the COVID-19 outbreak (Islam et al., 2020). Social media such as Facebook, Twitter and the many online newspapers are all excellent sources for monitoring current trends within the area of infodemics (Kouzy et al., 2020). Reviewing social media data plays an important part in the understanding of global reactions to the COVID-19 outbreak and the potential impact on public health (Allington et al., 2020). Islam et al. carried out a study of data from several online media platforms, with the primary objective of identifying the impact these platforms have had on public health during COVID-19 (2020). They gathered an interdisciplinary team of social scientists, medical doctors, and epidemiologists to collect the data and review the content. Their data was collected between December 31, 2019, and April 5, 2020, on a global scale, meaning subscribing to several international online news media, fact-checking sources, and using data from multiple social media platforms. The data was split into the three categories: rumour, defined as information the is not yet verified and can be found either true, false or fabricated; stigma, defined as a socially constructed phenomenon that assigns certain ideas and actions to a certain social group making them devalued in society, and finally conspiracy theories which are defined as certain individuals or groups working in secret towards reaching a negative outcome (Islam et al., 2020). Based on themes defined by WHO, four additional subcategories were added, including the cause of disease, illness, treatment, interventions, and violence. All data were coded according to the three top-level categories and corresponding theme. 2.311 different reports of misinformation were gathered from 87 countries and 25 different languages. The reports were split into the three different main categories, 89% belonging to rumours, 7.8% to conspiracy theories and 3.5% to the stigma category. In terms of the thematic subcategories 24% were related to illness, 21% to policy interventions implemented in the relevant country, 19% to treatment and cure, 15% related to the cause of the disease, 1% to the violence category and 20% were labelled as miscellaneous. 82% of the reviewed reports were found to be false, and 9% were correct, 8% were misleading, and 1% were not verified. Most of the misinformation came from The United States, India, China, Spain, Indonesia, and Brazil (Islam et al., 2020). As mentioned, the

rumours category had the most categorized reports, with 89%. The most dominant subcategory was illness and mortality, and examples of rumours were eating garlic, keeping the throat moist, avoiding spicy food, taking vitamin C and D, and spraying chlorine. Further, more extreme examples were also collected, including mixing sodium chloride solution and citric acid, directly consuming bleach and alcohol and drinking various forms of animal urine. Examples from the rumours category also included self-diagnosing theories like holding one's breath for 10 seconds to identify infection. Most of the stigma classified reports related to the COVID-19 origin in China. The virus has several times been referred to as "China virus" or "Wuhan virus", especially in The United States (Vazguez, 2020). This has led to several incidents of stigma towards individuals of Asian heritage or people who have visited Asia in the recent past. During COVID-19, there have been several reported examples of violence toward stigmatized groups. One example from Ukraine was a bus with individuals evacuated from Wuhan, being held up and attacked by locals throwing stones (Islam et al., 2020). Finally, there have been examples of self-stigmatization, essentially meaning the guilt that infecting someone else can carry. In India, this led to a man killing himself, as he was worried that he had infected family and friends and feared how his surroundings would perceive this. The final primary category includes reports related to conspiracy theories. Several of these have spread globally, but predominantly find their origin in the United States, the United Kingdom, Russia, China, and Iran (Islam et al., 2020. Ahmed et al., 2020). These theories include speculations of COVID-19 being a bioweapon manufactured in an international collaboration to impact China and their economic growth and global impact. The opposite has also been suggested that the virus was manufactured in China as part of their bioweapon program. Other theories suggest that the virus already has a cure but was allowed to spread further to increase vaccine sales, or the pandemic being a scheme created to control the population. The rumoured lockdown resulted in people going out and panic-buying masks, food, hand sanitizer, and toilet paper. This resulted in the prices going up heavily as well as negatively impacting the spread of the virus, as several people were not able to buy masks and hand sanitizers which could potentially have meant that infected individuals have been walking around spreading the disease further (Islam et al., 2020). These are only a few examples of the severe impact misinformation can have in times of a pandemic (Tasnim et al., 2020).

Their study relates to the category of quantifying impact suggested as a primary point of interest at the WHO virtual conference. This thesis has a similar goal, making the study by Islam et al. highly relevant (2020). They also work with text data and uses it to identify types of misinformation relevant during the COVID-19 pandemic. Especially relevant to this thesis is the three top-level categories rumours, stigma, and conspiracy theories, as well as the corresponding thematic subcategories. These can be used to identify key terms, which can be used in the filtering and pre-processing of search query logs. The study provides an excellent starting point for anyone working with misinformation on the internet during COVID-19, especially given their multidisciplinary research team that made it possible to manually classify information that might be beyond the knowledge of social scientists alone.

A study by Cinelli et al. (2020), researched the diffusion of information during the COVID-19 outbreak. Several social media platforms were explored to investigate user interest and engagement in information related to the

novel coronavirus. The spreading pattern was compared to those of existing epidemic models that measure disease reproduction numbers. Furthermore, they investigated the spread of misinformation compared to correct information on social media platforms. Given the current technological paradigm that makes most information readily available on social media platforms which rely heavily on user preferences and personalization algorithms determined by user actions, it can prove problematic in times of a pandemic. An example of this could be multiple interactions with specific users on Facebook, resulting in these users frequently being featured in the user's feed. If these users are perceived as credible by the user, information is more likely to be shared on the user's personal Facebook wall, despite it being potential misinformation (Chen & Sin, 2014). As most social media platforms work differently, Cinelli et al. (2020) investigated the five different platforms Twitter, Instagram, YouTube, Reddit, and Gab, to determine differences in information diffusion. Epidemic models were used to measure the way information spreads on each platform. Essentially this means counting the average number of secondary cases an individual that starts posting about COVID-19 will generate. In the context of a pandemic, the same measure is used to determine how many individuals an individual infected with a virus will reach and potentially infect. The mathematical metric is called *R-naught* (R0), and to simplify it if R0 < 1 the disease spread is shrinking, R0 = 1 indicates that things are stable, that the disease is spreading, but not at an alarming rate and finally R0 > 1 meaning the disease is dangerous and spreading at a rate corresponding to the R0 value. If the R0 value equals 3, each infected individual is expected to infect three other people (Fisher, NY Times, 2020). The study analysed more than 8 million social media posts collected over 45 days during the COVID-19 outbreak (January 1st – February 14th). 1.35 million of these were original postings whereas the rest were comments, coming from approximately 3.7 million individual users. All the different social media platforms were found to have an R0 value above 1, meaning that each individual is likely to "infect" other individuals with their post. To determine the number of posts from unreliable sources and their growth compared to posts from reliable sources; links were checked and tagged according to data from the fact-checking organization Media Bias/Fact Check[4]. Most of the social media platforms were found to follow a similar growth pattern among questionable posts as the pattern observed among reliable posts. Most of the social media had a small percentage containing unreliable sources, including Reddit (5%), YouTube (7%), and Twitter (11%). However, a significant difference was observed on the Gab platform, where 70% of the volume of reliable posts were found to contain information from questionable sources. This number is caused by a difference in approach to misinformation by each platform, meaning some platforms reduce the impact of misinformation by removing content (YouTube) and some media, like Gab, amplifies them (Cinelli et al., 2020). The study also finds significant increases in post behaviour and interaction on specific dates corresponding to critical dates related to COVID-19. One example is a spike in activity on the 20th of January 2020, the day WHO issued their first public situation report on COVID-19. This study provides valuable insight into the way information spreads on the internet, and especially how platforms manage misinformation. While this thesis is not directly working with social media data, making it more challenging to measure the diffusion of

[4] https://mediabiasfactcheck.com/

information the same way, similar approaches can be taken here by investigating the popularity of specific search queries, and how they spread across states or countries.

Another study was carried out by Li, Bailey, Huynh, and Chan (2020) in their research of popular YouTube videos related to the coronavirus. Rather than focusing on a large selection of social media and news platforms, the top 150 COVID-19 related videos from March 21 2020, were analysed. It is known from the previous H1N1, Ebola and Zika outbreaks that YouTube can be a significant source of misinformation (23%-26% in previous research. The goal of their study was to identify the current state of COVID-19 information on YouTube as compared to previous pandemics. The top 75 videos from two different searches on the terms "*coronavirus*" and "*COVID-19*" were selected and manually coded according to source, content and characteristics. The study is relevant according to the goal set by WHO of amplifying correct information, as YouTube videos from credible and reputable sources were found to be under-represented, both in this and previous studies (Li et al., 2020). Of the original 150 search results, 81 were dropped due to not being in English, duplicates, exceeding 1-hour playtime, live-streams, or without any audio. Sixty-nine videos remained with a total amount of views of more than 257 million and represented various types of news platforms including entertainments news, network news or internet news platforms. Of the 69 videos, 27.5% included non-factual information and had approximately 62 million YouTube views (Li et al., 2020). Between videos containing misinformation and factual videos, no significant difference was found when compared by the number of views, likes, dislikes, or duration. The misinformative videos were mainly found to come from entertainment or internet news, where all the videos from government channels or other professional videos were found to be factual. Government and professional videos were by far the least represented in the selected videos, which could be a potential area of interest if the main goal is to amplify correct information. Like Islam et al. (2020), the study divides the various videos into thematic categories, including rumours, stigma, and conspiracy theories. Rather than labelling the first category, rumours Li et al. (2020) created their scale based on previous work within public health emergencies. This scale was referred to as CSS (COVID-19 Specific Score), a 5-point scale used to assess the level of evidence-based information in the video. The scale cover areas such as transmission, symptoms, prevention strategies, treatment and epidemiology, which are all closely related to the themes covered by the rumours category by Islam et al. (2020). According to the study, YouTube is an untapped platform by most health professionals and government organisations, and this should be corrected in order to decrease the negative impact of misinformation in times of a pandemic. This is further supported in a study by Basch et al. (2020), researching the existence of recommended preventive behaviours, as defined by Center for Disease Control and Prevention (CDC), in YouTube videos. The study finds that only a third of the top 100 COVID-19 related videos (January 2020) included information on preventive measures. While not directly related to misinformation, the study suggests that professional health information is not communicated well on YouTube, and this is something that could be improved in future pandemics (Basch et al., 2020).

2.4 Infodemiology – Improving Public Health with Data

The related work up until this point has dealt with information and misinformation in the context of social media interactions during the COVID-19 outbreak. Infodemics, as previously mentioned, is a term coined by the World Health Organization in the context of the coronavirus. While the term covers several aspects of online behaviour during a pandemic, it is mostly concerned with the negative impact of misinformation and how it can affect human knowledge and behaviour. Infodemics is a combination of the words information and pandemic and relates to a particular context which has been relevant during the current COVID-19 outbreak. However, a similar term introduced much earlier in 2002 by the German researcher Gunther Eysenbach, is *Infodemiology; The Epidemiology of (Mis)information* (Eysenbach, 2002). Infodemiology, blends the two words information and epidemiology, and is essentially a method that uses the available user-generated internet content related to health, and attempts to use this data to improve the general public health. Epidemiology defined as …" *The study of the distribution and determinants of health-related states or events in specified populations, and the application of this study to the control of health problems*" (CDC, 2020)[5]. Research within epidemiology plays an important part when the population is attacked by a pandemic, especially when the disease has previously been unknown. Epidemiologists research and share their data with the government, which in turn creates new policies and guides the population. This has proven especially true during the COVID-19 outbreak that has seen an unprecedented degree of policies implemented both on a national and global scale (Statens Serum Institut, 2020)[6]. Combining current epidemiological methods with other disciplines such as information studies provides new insight into the monitoring and guidance of public health. It can assist policymakers in their initiatives. The early work by Eysenbach (2002), described the different approaches to infodemiology in the then-current research. Several studies were found to be mostly descriptive, for instance, reporting on the percentages of selected websites that were found to be reliable, or whether certain health issues were better covered than others on the internet. Rather than merely using the method for descriptive studies, Eysenbach argues for an analytical approach that uses technical (or formal) markers to predict accurate content. In the context of general web credibility, this could be the presence of sources, suggesting that the website is credible (IFLA, 2020). Historically infodemiology has been defined in two different ways. Initially, the discipline focused on current health information available online and assessing the quality of this information, now often referred to as *supply-based infodemiology*. The concern was that a lot of information on the internet was of a low quality which could potentially impact public health. Later the term expanded to include analysis of human needs and behaviour as expressed online, as well as monitoring of health information-seeking behaviour, also known as *demand-based infodemiology* (Eysenbach, 2009). Both concepts share a similar approach when working with the data and have been defined in various ways during their lifetime. In current infodemiological research, the *supply* side of the concept often refers to data sources from web 2.0 services such as social media, online discussion fora and similar. In contrast, *demand*-based research primarily uses web 1.0 services such as search engines or *Google Trends*

[5] https://www.cdc.gov/csels/dsepd/ss1978/lesson1/section1.html
[6] https://www.ssi.dk/aktuelt/nyheder/2020/nyt-samarbejde-mellem-statens-serum-institut-og-forskerservice-om-covid-19-data-til-forskning

(Zeraatkar & Ahmadi, 2018). When infodemiology was still in its' early stages during the late 1990s and the early 2000s, the method was primarily used to investigate offline or past content but has since expanded to include real-time monitoring of online health-related information. Real-time monitoring or surveillance of current online trends is also referred to as *infoveillance* (Eysenbach, 2009 & 2011). This is useful to identify sudden spikes in certain information or misinformation, making it possible for health professionals to respond to this. For example, during the COVID-19 outbreak, rumours have circulated that it is dangerous to wear masks as they can deprive the body of oxygen, cause carbon dioxide poisoning, and harm the immune system (BBC, 2020)[7]. These claims are all false, and by identifying this as currently trending misinformation, health professionals can respond to these, for example, through popular news media. In the earlier work by Eysenbach (2002, 2006) this concept was not yet referred to as infoveillance, but rather syndromic surveillance. This entails health-related data that precede a diagnosis which could signal a potential upcoming influenza epidemic (Eysenbach, 2006). The whole point is to monitor online trends to catch diseases before they become a pandemic. In 2006, Eysenbach expanded on his previous work on the potential and research methods of infodemiology, by investigating the correlation between data from the Canadian flu season 2004/2005 and internet data from the same time period. A strong correlation between the two was found, and Eysenbach concludes the study as successful while underlining the potential of using search query data to identify early signs of disease outbreaks. He continued his work by expanding to monitor and research social media data, and development of a system that could automatically find, collect and analyse data from these (Eysenbach, 2011). Another impactful study in moving towards a data-driven approach to epidemiology was carried out by Ginsberg et al., in their research of detection of epidemics using search query logs (2009). By researching Google search query logs from 2007-2008 and comparing these to results of studies by the CDC (Centres for Disease Control), they were able to show the impact of using online data rather than traditional methods. Consistently, they were able to draw conclusions on epidemic patterns 1-2 weeks ahead of the CDC, and their studies are speculated to have played an important part in Google opening their Google Trends platform (Jun, Yoo & Choi, 2017).

2.5 Researching Epidemic Outbreaks with Search Query Logs & Google Trends

In early infodemiology, one of the primary concerns was how to collect large-scale datasets covering large parts of the world. This was all made much more comfortable in 2006 when Google introduced the *Google Trends*[8] platform (Jun et al., 2017). Google trends make it possible to look up individual queries and see their popularity over a given time period and how it compares to other popular terms. Search queries are collected in any location where the Google search engine is used, which makes it a valuable tool for the analysis of trending searches on a global scale. Furthermore, it is possible to view the most trending terms of the day, as well as trending terms in real-time, i.e. which terms are the most important of all the recent search queries (Google, 2020)[9]. During COVID-19, a specific

[7] https://www.bbc.com/news/53108405
[8] https://trends.google.com/trends/?geo=US
[9] https://support.google.com/trends/answer/6248105?hl=da&ref_topic=6248052

site has been established dedicated to bringing up-to-date information on how people search for information related to the virus (Google, *Coronavirus Search Trends*, 22.08.2020). Google trends have proven to be an essential platform for research since its' implementation and have assisted in moving infodemiological research beyond surveillance and monitoring towards forecasting (Jun et al., 2017). Multiple studies have studied previous epidemics and pandemics based on data from Google trends, including Cook et al. (2011) and their studies of Google trends performance during the Influenza A (H1N1) virus in 2009. They compared data from Google with data from the *U.S outpatient influenza-like illness surveillance network* (ILINet), which tracks the number of patients that visited doctors or hospitals, but were not admitted for further treatment (CDC, 2020)[10]. The study found a strong correlation between the two data sources divided into four different timeframes; pre H1N1, Summer H1N1, Winter H1N1, and H1N1 total (Cook et al., 2011). A similar study (Bragazzi et al., 2017) investigated the public online reaction to the 2015 outbreak of the Zika virus, transmitted from infected *Aedes* mosquitos (*Zika Virus*, WHO, 24.08.2020). The study analysed almost 4 million tweets, 300.000 Wikipedia visits, YouTube content, Google News, Google Trends, and epidemiological data to determine the global interest and reaction to the Zika virus outbreak. The study covered data from January 2004 to October 2016 and was extracted based on keywords such as *Zika, ZIKV*, or *Zika virus*. The most massive spike in online activity, with the selected terms, was found in late 2015, corresponding to the largest recorded outbreak of the virus (Bragazzi et al., 2017). Until late 2015 the normalized value of interactions for each of the social media platforms remained <10, then the online interest started growing before peaking in February 2016 at >90. Strong correlations were found between all social media except between YouTube and Twitter/Wikipedia. Furthermore, the study found that the largest spikes in search activity occurred right after important events such as WHO acknowledging the problem by establishing an emergency committee, the declaration of Zika virus as a public health emergency, and the first cases of the disease in The United States (Bragazzi et al., 2017). The study provides insight into the analysis of epidemic-related online information by using multiple data sources and comparing these with certain vital events related to the epidemic.

Google Trends has remained a stable of infodemiological research since its' introduction in 2006. Although the concept is initially intended for tracking information as it relates to epidemics, it has also been used in other areas of the health industry. These include tracking seasonal online interest in obesity (Basteris, Mansourvar & Will., 2020) or the detection of seasonal patterns of internet searches on mental health (Soreni et al., 2019). The recent outbreak of the coronavirus has resulted in a considerable amount of new research being released, a lot of this relying on data from Google Trends. One of these studies by Terefe, Rovetta, Rajan, and Awoke investigated search behaviour in Ethiopia during the early stages of COVID-19 (2020). The study was mainly exploratory, and by using Google Trends, they were able to identify the most critical search queries in Ethiopia. Rather than looking at the overall trending words on a national level, keywords were extracted from popular Ethiopian news sources. These were translated and investigated further in Google Trends, before finally being classified into various categories related to the theme of the search query, for example, symptoms, mortality, or world news. The study

[10] https://www.cdc.gov/coronavirus/2019-ncov/covid-data/covidview/08282020/percent-ili-visits.html

introduces an interesting use of Google Trends by working from keywords found in national news media and measuring their use in the Google search engine. Although not the focus of the study which was mostly exploratory and focused on search behaviour surrounding COVID-19, an approach like this could provide insight to national news channels when evaluating their performance and relevancy to the public. The is especially relevant if the goal is to amplify correct information.

Italy was one of the countries that were initially hit very hard by the coronavirus and have consequently been the target of a lot of research both within the field of infodemiology/infodemics and several other research disciplines. One of these studies was released by Rovetta and Bhagavathula and presented their research on COVID-19 related search behaviour and infodemics in Italy (2020b). Like the previously mentioned studies, they researched data from Google Trends ranging from January 21, 2020, to March 24, 2020, as well as mining article titles from the most famous Italian newspapers. The main objective of their exploratory analysis was to identify so-called "infodemic monikers" meaning information that was critically wrong and caused harm for example in the form of spread of fake news and misinterpretations or increased racism (Rovetta & Bhagavathula, 2020b). The infodemic monikers were coded into categories corresponding to the type of attitude they conveyed to the receiver; superficial (unclear communication about COVID-19), misinformative (included words that could lead to false information), racist (stigmatizing a specific group of people), or definitive (clear and correct communication). Google trends allow two different search methods when investigating keywords, as it can either show search results as terms or topics. Searching terms is language-specific and will deliver results that include the search word selected and different combinations using that exact word. Searching for issues related to the search term will output concepts related to the search term in any language, meaning the word is not necessarily explicitly mentioned, but a part of the concepts returned in the results (Google, *Compare Trends Search Terms - Trends Help*, 27.08.2020)[11]. The relevant search terms were investigated further in Google Trends, both on a national scale as well as individually for selected regions of Italy. The top terms in Italy were found to be "novel coronavirus", "China coronavirus, "COVID-19", "2019-nCOW", and "SARS-COV-2". It should be noted that Google Trends uses normalized data to display current trends. This is done by dividing each data entry with the total searches of the relevant location and time period. If this were not done, it would be challenging to compare search terms, as the ones from places with the most Google users would always score higher in popularity. Furthermore, this makes it possible to see how terms compare to the total amount of searches, rather than viewing a count with limited context[12]. Comparing the terms relative to each other "coronavirus" reached the highest volume of searches with a value of 59. This term was dropped from further investigation, making the search query "China coronavirus" the most frequently used at a relative volume of 38. The study also investigated queries related to health and, in the early stages of COVID-19, found significant information spikes about symptoms, followed by information on face masks and disinfectants. Generally, large increases in searches were found at significant points in time in Italy, such as the initial massive

[11] https://support.google.com/trends/answer/4359550?hl=en
[12] https://support.google.com/trends/answer/4365533?hl=en

breakout and when WHO declared COVID-19 a global pandemic. Another purpose of the study was to identify which type of infodemic monikers, characterized various regions. The search queries popular in each area was assigned to one of the previously mentioned categories. It was found that some regions showed very superficial attitudes towards the coronavirus (Basilicata, Umbria, and Emilia Romagna), some areas had a lot of misinformation (Basilicata and Umbria), and certain regions conveyed racist and stigmatizing attitudes through their searches (Campania and Friuli Venezia Giulia) (Rovetta & Bhagavathula, 2020b). An interesting, and perhaps expected consequence of each other, is the relationship between a superficial attitude and the presence of misinformation in search behaviour. If an individual has a very shallow attitude towards a particular topic, it can be speculated that less critical thinking will be applied to their online search behaviour, which in turn leads to searches on potentially misinformative topics. The top trending words related to racism all stemmed from that fact that the virus had initially started in China, and included variations of China, Chinese or Wuhan. While certain regions were shown to contain more racist or stigmatizing attitudes, Rovetta & Bhagavathula believes this is a general problem in all of Italy, due to the national rate of information related to these concepts (2020b). The study provided valuable insight into the exploratory analysis of search query logs and was useful in informing the research design of this paper.

While the studies mentioned in the above section, primarily uses Google Trends to explore search behaviour, an obvious concern is that Google is not the only search platform out there. While Google is certainly the most dominant search platform with 91,5% of the market share worldwide (*Search Engine Market Share Worldwide*, 30.08.2020), other search platforms do exist including Microsoft's Bing platform and Yahoo. In order to fully understand online search behaviour, it is important to cover other platforms as it might provide insight that Google does not. This was done in a recent study by a Microsoft Research team[13], investigating human needs, as expressed through search query logs, during the COVID-19 pandemic (Suh, White, Horvitz, and Althoff., 2020). Based on Maslow's hierarchy of needs (1943), they attempt to define a computational framework to identify population-wide changes in needs during the pandemic. This was done by researching data from Microsoft's Bing search engine consisting of more than 35 billion search interactions across 36.000 zip codes, all from The United States. The search queries are classified into five basic human needs classes known from the work of Maslow, which are self-actualization, cognitive needs, love and belonging, the need for safety, and physiological needs - an additional 79 subcategories were also defined (Suh et al., 2020). The data was collected across 14 months, which made it possible to determine the expressed needs of the query logs before and after the outbreak of the coronavirus. Early in the pandemic, physiological needs were the primary category expressed in the search queries, with a specific focus on health condition questions, toilet paper purchases, and various health measurement equipment. After the US declared a national emergency on March 13 and subsequently implemented mandatory lockdown on March 21, a large increase was seen in queries related to the cognitive needs queries. These included searches on, and visits to, online educational websites. The categories *self-actualization* and *love and belonging* have the highest number

[13] https://www.microsoft.com/en-us/research/

of interactions around April 11-13. Specific subcategories also saw significant decreases in search interactions, namely searches related to various "normal" life activities such as purchases (wedding, apparel, rental), job searches, housing questions, or outdoor questions (Suh et al., 2020). These are only a few examples of the many changes seen after the outbreak of the pandemic, as covering all of them would be beyond the scope of this thesis. Using Maslow's hierarchy of needs introduces an interesting and alternative way to analyse search query logs, especially relevant if the outcome is tracking changes in human needs (Cerbara et al., 2020). The study finds that people generally search for information to cover their basic needs, consistent with the main premise of Maslow's pyramid – basic needs must be covered before moving further up towards self-actualization (1943). Overall, the largest increases were seen in subcategories related to physiological and safety needs, while searches related to growth, positive outlook, and opportunities have decreased significantly. The combination of a large decrease in job-related queries and an increase of 30 times the normal in unemployment as well as very little interest expressed in education or general life goals is an example of the severe impact the pandemic has had on the United States (Suh et al., 2020). Furthermore, behavioural changes were also investigated on a state level while the previous examples were done nationally in the U.S. A shelter-in-place policy was introduced at different times across the states and essentially means that people are ordered to stay in the building they are currently occupying. Given the states' individual choice of when to enforce the shelter-in-place policy, the number of time people was restricted differed immensely between states. An example of possible implications of prolonged enforced shelter is provided in the study in a subcategory related to mental health status. Inhabitants of Mississippi expressed 33.2% less negative mental health concerns in their search queries, while Oregon saw an increase of 27.2% in negative mental health expressions. Comparatively, Mississippi was under the shelter-in-place policy for 24 days, whereas it remained enforced for 88 days in Oregon. While tendencies like these can have several different causes, a possible explanation could be that the shelter-in-place policies negatively impact the mental health of people. This, however, does not explain the decrease in mental health concerns, meaning further research is necessary to understand these changes fully. Another interesting point made in the study concerns the increase in cases of domestic violence. Results showed that search queries expressing needs concerning domestic violence had dropped by 36.7% compared to before the pandemic. It is known that domestic violence increases during times of crisis (UN Nations, 31.08.2020)[14], and in the case of COVID-19 lockdown, measures form a paradoxical situation where the victim is essentially trapped at home with their abuser (Bradbury-Jones & Isham, 2020). This has led to a significant increase in reports of domestic abuse across the world, including France which reported a 32-36% increase, The United States ranging from 21-35% increases, and the UK with a 25% increase (Usher et al., 2020). This is an example of the importance of understanding the data we are researching. Simply concluding that domestic abuse is decreasing based on fewer related search queries would be false, as the victims are potentially trapped at home under lockdown

[14] https://www.un.org/en/coronavirus/un-supporting-%E2%80%98trapped%E2%80%99-domestic-violence-victims-during-covid-19-pandemic

and have less access to computers than they had before the pandemic. Sufficiently answering these connections would require extensive domain-specific knowledge.

2.6 Search Query Logs

While several of the previously reviewed literature uses data collected from online search engines, this section will be dedicated to briefly introduce core concepts of log data, as well as its benefits and limitations. The analysis of search engine queries falls into the category of transaction log analysis. It can be defined as "*an electronic record of interactions that have occurred between a system and users of that system*" (Jansen, Taksa & Spink, 2009. In Jansen, 2009). Transaction log analysis covers a vast area of related categories including the study of weblogs, blog or social media analysis, or as is the case of this project, search engine queries or logs. Collecting log data for analysis, is unobtrusive, meaning it does not require the user to participate in the delivery of data actively. Transaction log analysis is situated within the paradigm of behaviourism, as it attempts to explain behaviour as expressed through the various types of transaction logs. It should be noted that traditional behaviourism is only concerned with the outward actions of thought while not considering the cognitive workings motivating the behaviour. This understanding is slightly expanded in the context of log analysis, also to include consideration of the internal aspects that motivate behaviour. Research within in behavioural science, which search log analysis can be considered a part of, considers the following elements in research questions including the terms; *who (actor), what (behaviours), when (temporal), where (contexts), and why (cognitive)* ((Jansen et al., 2009). In terms of search query logs, we can have various degrees of identifying information, but often the *who* is named purely by an ID or an IP-address. In this context, the *what*, or the behaviour is whatever is expressed in the search query. While this is not a physical action that can be observed, it is still considered behaviour in this context (Jansen et al., 2009). Query logs include temporal data such as data and timestamps addressing any *when* questions. The *where* category provides information on the context of the relevant user, which can be supplied through geographical details on different levels (country – state – city). Finally, the *why*, which can mean several different things, depending on the context of the query. As previously mentioned, in work by Suh et al. (2020), a drop in expressed interest in domestic violence information does not necessarily mean that people are no longer interested in it, but rather that other elements are affecting the behaviour of people that would usually search for this information. These causal connections can sometimes be explained by analysing the data at hand, but it might be necessary to include elements from other data sets, methods, or even research disciplines. Transaction log data is often referred to as *trace data*, which is usually split into two different types: erosion (wearing away) and accretion (building up). Both processes leave some kind of trace behind automatically, which is one of the benefits of collecting this kind of data – it does not impact the behaviour of the subjects, as the data is simply being traced without them noticing it (Jansen et al., 2009). This is one of the significant benefits of search query data, especially compared to other qualitative methods such as surveys or focus groups. There is no interviewer bias or social dynamics that may prevent honest answers, to account for (Scharkow & Vogelgesang, 2011). This does, however, also reflect one of the weaknesses of trace data, as it is impossible to follow up with the subject and learn more about their situation and motivations.

2.7 FINDINGS

The literature review has described relevant related literature. It responds to the suggested actions proposed by WHO and is concerned with measuring the infodemic and assessing trends and impacts. Using mixed methods research the current acceptance by the public should be monitored, for example, through sociobehavioural research and analysis of digital information from online communities (Tangcharoensathien et al., 2020). Several mentioned studies were instrumental in the research approach. Islam et al. and Li et al. (2020) introduced popular misinformation types during COVID-19 as represented on YouTube and a wide range of online media. Cinelli et al. described how misinformation is handled differently by social media platforms (2020), which motivated the later external validity check using Google Trends data. Implementing Google Trends was further supported by the wide range of studies that used it in their research (Jun et al., 2017. Cook et al., 2011. Bragazzi et al., 2017). Rovetta and Bhagavathula used a very similar research approach as this thesis but gathered their keywords from popular Italian newspapers rather than using internet articles (2020b). One study used data from Microsoft Bing (Suh et al., 2020), with the objective of identifying behaviour changes during COVID-19. Their study was especially important to the interpretation of findings when working with search query logs. Finally, the benefits and limitations of log data were briefly described (Jansen, 2009).

3 METHODOLOGY

3.1 RESEARCH DESIGN

A basic workflow when working with either supply-, or demand infodemiology is the selection and filtering of areas of interest that exists in a sizeable textual dataset, researching the semantic qualities of the data using natural language processing methods, investigating geographic differences, and using descriptive and statistical methods to further understand the data for instance by identifying clusters and trends (Eysenbach, 2009).

This thesis uses modified elements of a framework suggested by Nelson (2020) called computational grounded theory, combined with traditional grounded theory methods (Strauss & Corbin, 1990). The approach is used for qualitative text data and was defined in response to frequent discussion among social scientists about how to approach data of this type. Traditionally, grounded theory has been used in the analysis of search query logs, which includes discovering theories and models from working with the data, both of which are grounded in observations of the world (Jansen, 2009). The concept involves inductive reasoning as observations from the qualitative data is used to form theories and rules that can be generalized (Bryman, 2012). One significant difference of computational grounded theory is implied in its modified name – the use of computers to assist regular grounded theory methods such as manual content coding. Nelson suggests replacing the manual coding with topic modelling or unsupervised machine learning, as a means of eliminating the problem of personal bias in the coding process. The output of the topic modelling is then evaluated and labelled by human researchers. This was initially considered but modified slightly as the main objective was to identify the extent of misinformation and less so the types of misinformation. Instead, computer-assisted pattern recognition was used in the search query analysis. In the second step of grounded theory, the outcome of the initial topics (whether manually coded or computer-generated) is evaluated by researchers. This should be considered part of an iterative process of identifying patterns, evaluating them and, if necessary, going back and modifying the initial patterns (Nelson, 2020). This thesis used a primary keyword index of regular expressions to identify misinformation search queries, supervised machine learning to manually label the entire dataset, reflecting on the output, and going back to modify the machine learning models in order to achieve a more accurate outcome. The third step of the computational grounded theory approach is called pattern confirmation and uses supervised machine learning or natural language processing methods to confirm the findings of previous efforts. Much of the analysis of this paper resides within this final step, combined with the human interpretation and evaluation from the second step.

Overall, the research design, based on grounded theory, can be considered as exploratory research as the primary purpose of the thesis is to identify misinformation and explore how it is represented in search queries. The analysis consists of mixed methods using both elements from qualitative and quantitative research disciplines (Bryman, 2012). The initial qualitative data and qualitative content analysis are subjected to quantitative measures such as descriptive statistics which can support the generalizability of the findings (Creswell & Clark, 2006).

While this research does touch on associations between variables, it does not attempt to define causality or perform hypothesis- and significance testing. It is an exploratory thesis that can be used as a foundation for further quantitative research. This could include the examination of causality and variable correlations through causal modelling (Gencoglu & Gruber, 2020), as well as modifying the first step by using topic modelling or unsupervised machine learning to investigate the different types of misinformation.

3.2 DATA

The following section will introduce the different data sets and sources used during this master's thesis project. The data was used in various degrees, and for different purposes. The main objective in the data collection was to gather enough data to be able to make informed predictions and analyses as well as investigating various data sources to ensure validity.

3.2.1 Bing Search Dataset for Coronavirus Intent

The Bing search dataset for Coronavirus Intent[15] is a large dataset containing search query logs from Microsoft's Bing search engine[16]. The data collection by Microsoft started in January 2020, and new data were added monthly. All the collected search queries contain expressed interest in topics related to the Coronavirus. This interest or intent can either be explicit or implicit, which will be defined in the following section. The dataset only includes queries that were performed many times by several users, but the exact criteria for selection are not described further by Microsoft. The dataset used in this thesis contains observations from January 1st, 2020, to August 31st, 2020. The entire dataset, including every country, has 3.83 million observations, and the overall search query frequency can be seen below in Figure 3.1. This contains data from every country in the dataset, and provides a broad view of global interest, as expressed in search queries, in COVID-19. It should be noted that this might vary for individual countries, which will be addressed later in the analysis. The graph shows a massive spike in interest around March and April, corresponding to the time WHO declared COVID-19 as a global pandemic, steady levels in May and June, going up again in July, before finally decreasing in August. See Appendix A 1.2 for a table showing specific numbers.

[15] https://github.com/microsoft/BingCoronavirusQuerySet
[16] https://www.bing.com/

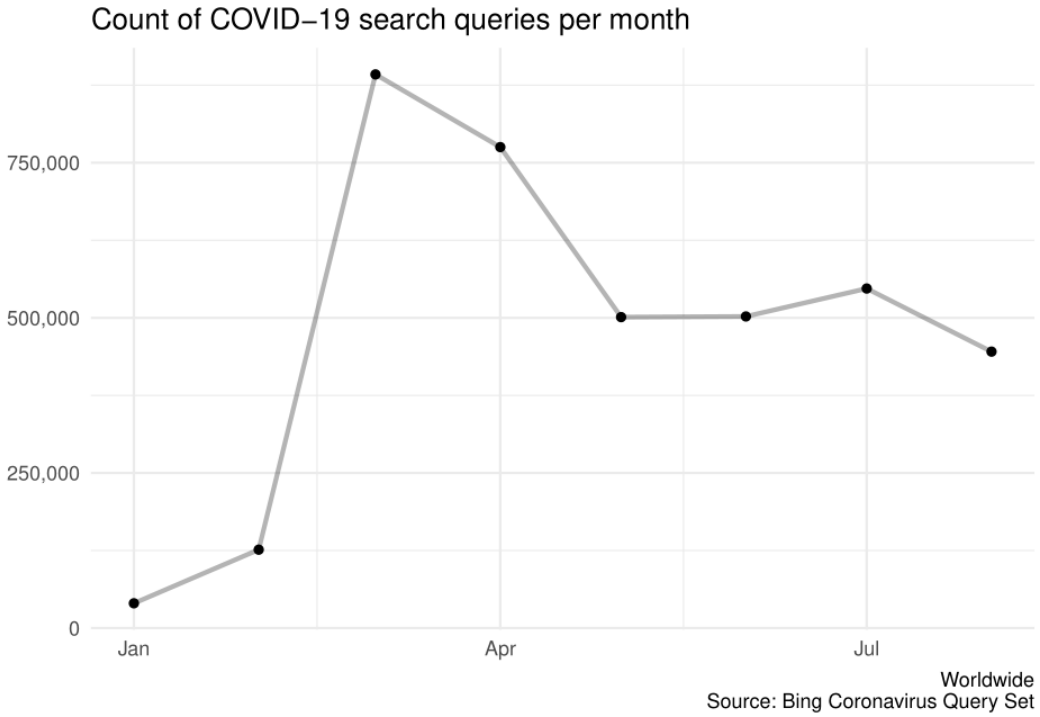


Figure 3.1 - Unique search queries per month (Global)

The United States makes up nearly half the dataset with 1.75 million observations, the United Kingdom has a little more than half of that at around 800.000, France is represented by almost 200.000 of the queries, and that number rapidly decreases going further down the list. Table 1 shows the top 10 countries ordered by the number of search queries.

Country	# Queries
United States	1751769
United Kingdom	805201
France	197275
Italy	172992
Germany	166362
Canada	144528
Japan	78967
Spain	73105
Australia	71794
India	49231

Table 3.1 - Country distribution in the Bing dataset

The dataset contains six different variables, which are briefly described in the following. The first variable *Date* is a string that contains temporal information about the day, month, and year. The second variable *Query* contains text strings of the individual search queries. This is the variable that will be used for the later text classification of search queries related to misinformation. Query length varies from single words of down to 2 characters, and some contain multiple sentences of up to 537 characters[17]. The string variables Country and State has information on country and region. *PopularityScore* is an integer variable including details on the popularity of the search term in a specific country or state on a given day. One would indicate the least used query for the day in the relevant region, and 100 is assigned to the most popular search query of the day and country/state. Finally, *IsImplicitIntent,* which relates to the way the data set was created. The boolean variable returns FALSE if the query specifically mentions the terms *COVID, coronavirus*, or *sarsncov2*. If none of those words occurs, the query is considered implicitly related to COVID-19 and returns TRUE. An example of implicit intent could be search queries pertaining to *toilet paper*, which under normal circumstances would not be related to any kind of virus, but in the context of the early shopping sprees at the beginning of the outbreak could be considered related to the coronavirus. The implicit intent is decided by a method known as *random walks on the click graph* (Craswell & Summer, 2007). Essentially the model uses click logs to determine which terms and topics usually are searched and clicked following each other or in the same session. If a user searches for *coronavirus, toilet paper,* and *COVID-19* in the same session, the method will calculate the probability of the terms being connected. This is similar to the way search engines suggest related queries to the one entered by the user (Hiemstra et al., 2020).

All the variables in the Bing dataset have been presented in the above section. For further clarification, a random sample of 10 queries was extracted from the dataset, including information on whether the query carries explicit or explicit intent regarding COVID-19, and the popularity score. The sample can be seen in table 3.2 and were all extracted from the United States. The queries directly related to the coronavirus are labelled false in the boolean IsImplicitIntent, and examples of these are "*coronavirus USA*", "*New Mexico coronavirus*", and "*symptoms of covid-19*". Examples of queries labelled as implicit are "*Jefferson Parish*", which is a state in Louisiana, USA. This was most likely determined to be related as other users have searched for coronavirus in the same area, although that information is not supplied. "*Irs stimulus check*" and "stimulus checks" relates to the financial support provided by the US government as a result of COVID-19, and especially the lockdowns which impacted the job market in the US heavily (*Effects of the Coronavirus COVID-19 Pandemic (CPS)*, 26.08.2020). Related to the above financial implications of COVID-19, the query "*Pfizer stock*" was also labelled as being implicitly related, potentially due to interest in the effects of COVID-19 in Big Pharma companies and their work towards providing a vaccine (Rao, Reuters, 2020). The popularity score in the sample provides information on how widespread the specific query was for the given day and state. While most of these are very low ranging from 1-8, one query stands out at 23

[17] Three queries contained 791, 999, and 14787737 characters, but they either did not have meaning or were the same thing copy-pasted multiple times.

("*coronavirus in Nebraska*"), meaning that specific query was quite popular in Nebraska on that day, although not being near the most popular query which would have a value of 100.

Query	IsImplicitIntent	PopularityScore
coronavirus usa	FALSE	1
john hopkins coronavirus	FALSE	2
jefferson parish	TRUE	5
new mexico corona virus	FALSE	1
coronavirus in nebraska	FALSE	23
florida coronavirus update	FALSE	5
irs stimulus check	TRUE	2
pfizer stock	TRUE	1
symptoms of covid 19	FALSE	2
stimulus checks	TRUE	8

Table 3.2: Example of search queries, intent and popularity score

3.2.2 CoronaNet Research Project Data

The *COVID-19 Government Response Event Dataset* (Cheng et al., 2020), is maintained by *the CoronaNet Research Project* lead by researchers from *New York University*, *Yale*, and *Hochschule für Politik* in Munich. It is an open science initiative who wants to make the data available to anyone who might have an interest. Outside of the primary research group, researchers within social, political, public health, and medical science from all over the world contributes to the dataset. In total, more than 500 researchers have contributed to the project so far, and it provides extensive information on political developments during COVID-19. This includes information on the political interventions introduced, which level of government that is implementing the policy, specific areas affected, who and what the policy addresses and whether the system requires mandatory or voluntary action (Cheng et al., 2020). The dataset comes in two different versions; a primary core version which includes the above-mentioned information and an extended version that contains data from other sources providing information on tests carried out cases and deaths, and country information such as GDP, democracy scores, and more. Only selected variables will be used in this thesis, to provide further insight into the potential causes and implications of misinformation search activity. These include the overall policy type variable and the compliance variable.

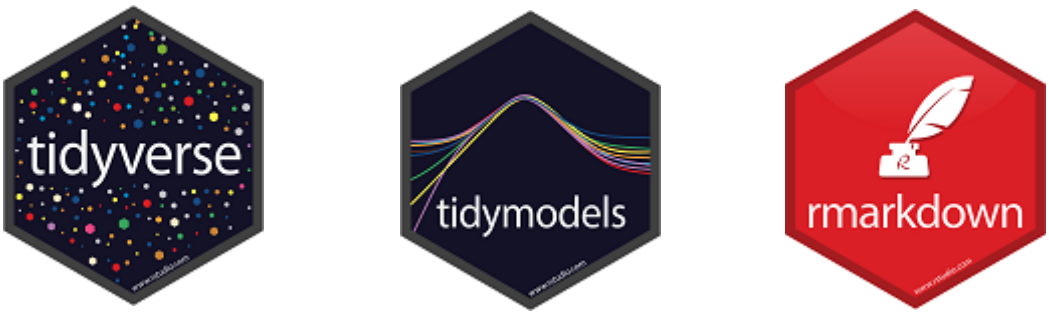
3.2.3 Google Trends

As previously mentioned in section 2.5, Google Trends is a valuable resource when researching search query logs. It is by far the most popular search engine, and the Google Trends platform provides access to samples of search queries that make it possible to investigate trending search terms and topics across the world. In this thesis, the platform is used to compare results from the Microsoft Bing and Google engines. Specifically, selected plots from the exploratory analysis will be recreated using data from Google Trends.

3.3 Software

This section will briefly describe the software used for the analysis in this thesis, as well as essential packages and extensions. The selected software should cover a wide range of methods, including data manipulation and wrangling, text classification and machine learning, descriptive statistics, and data visualization. The programming languages Python, SQL, and R are among the most popular in modern data science, and generally, people agree that they all work well, each with their advantages (Mitchell, 2019)[18]. Due to knowledge gained from previous courses in my studies, personal interests and experience from work, *R* was selected as the primary language. *R* is a language, and programming environment created for statistical computing and graphics (*R: What Is R?* 09.08.2020)[19] and consists of practices known from both functional- and object-oriented programming. R is especially useful for people who do not come from a computer science background, but rather an experience in analytics and modelling (Silge & Kuhn, 2020). All code used in the data pre-processing, wrangling, machine learning, and visualizations was written in the *RStudio* IDE (integrated development environment)[20], and can be found attached in Appendix A. RStudio is an open-source product that is primarily targeted towards *data science, scientific research, and technical communication* (*About RStudio*, 06.08.2020).

While the R language comes with a lot of base syntax and functionalities, working in R studio is a modular process generally including download of multiple different packages. Each package is targeted at a specific use-case and comes with a library of functions and options. The libraries work as building blocks each assisting the overall process in the different steps. A project like this, which includes several different methods of analysis, will consist of many different building blocks. Working with data usually involves spending a large amount of time on data processing, wrangling and preparation tasks. The tidy data structure attempts to simplify this process by keeping data in a format that works across multiple different libraries (Wickham, 2014). This makes the workflow much more streamlined and keeps the data wrangling tasks to a minimum.



[18] https://towardsdatascience.com/programming-languages-for-data-scientists-afde2eaf5cc5
[19] https://www.r-project.org/about.html
[20] https://rstudio.com/

The two primary packages used in this project were *the tidyverse*[21] and *tidymodels*[22]. Both are umbrella-packages or collections of a wide range of connected libraries. They all follow the tidy data structure, which makes them work well together. Notably, the tidyverse includes the *Dplyr*[23] package for data wrangling, *Stringr*[24] for working with strings, and *Ggplot2*[25] for data visualization. *Tidymodels* include the packages *Recipes*[26] for preprocessing matrices used in machine learning models, *Rsample*[27] to set up and evaluate resamples such as cross-validation, *Parsnip*[28] to use a wide range of machine learning algorithms while maintaining the same syntax, and Yardstick[29] to estimate model performance. They each include several packages that were not mentioned here, but the ones mentioned were found to be especially useful during the work with the data. Finally, *Rmarkdown*[30] was used throughout the entire coding process. Markdown makes it possible to easily share both code, output, errors, messages, and visualizations by merely compiling the script to HTML or pdf. Appendix A, which contains the code and output, was created using *Rmarkdown*.

3.4 Procedure

The following section introduces the various methods used in the project. It explains the overall methodological flow of the thesis and provides step-by-step information on each step used in the analysis. The initial phase of the procedure was data collection, which has been previously covered in section 3.1. The following five areas defined the project and were each important to reach credible and reliable conclusions. While an effort was made to keep personal bias at a minimum, some areas are near impossible to cover without some bias. Potential personal preference will be described in each section if relevant.

3.4.1 Data Cleaning & Preparation

After the initial data collection, the data was loaded into RStudio. As the dataset includes countries from all over the world, and my primary focus was on The United States, all other countries were filtered out. As the data already adhered to the three tidy data principles, it was not necessary to perform any data wrangling at this step. The three tidy data principles are, as described by Hadley Wickham (2014, 2017):

- *Each variable must have its own column.*
- *Each observation must have its own row.*

[21] https://www.tidyverse.org/
[22] https://www.tidymodels.org/
[23] https://dplyr.tidyverse.org/
[24] https://stringr.tidyverse.org/
[25] https://ggplot2.tidyverse.org/
[26] https://recipes.tidymodels.org/
[27] https://rsample.tidymodels.org/
[28] https://parsnip.tidymodels.org/
[29] https://yardstick.tidymodels.org/
[30] https://rmarkdown.rstudio.com/

- *Each value must have its own cell* (Grolemund & Wickham, 2017).

One of the significant benefits of working with tidy data is that the workflow throughout the entire process gets more comfortable, as it is not necessary to go back and forth and change the data structure continually. This also ensures that the data structure works for the relevant libraries used (section 3.3), as they all require a tidy data structure. The *Date* variable contains information on day, month, and year but, in order to simplify some later exploratory analysis, a dedicated *Month* variable was created using data from the *Date* variable. As selected data had to be used for manual coding, and the data did not contain a dedicated ID variable, one was created using row numbers. This was vital as it would allow joining the coded data back to the original dataset without any identification problems. The Microsoft Bing data set was already in a tidy data structure, so no additional data cleaning and preparation was necessary at this step.

3.4.2 Keyword Index

In order to determine the extent of search queries expressing interest in topics related to COVID-19 misinformation, a preliminary word search was performed on the dataset. This initial search used words determined by personal knowledge about certain myths, misconceptions, and conspiracy theories. Examples of this search included terms such as *"qanon"," hydroxychloroquine", "herd immunity", "malaria drug",* or *"bill gates".* The main objective of this process was to get a sense of the overall representation of misinformation. For instance, a search on "*qanon*", revealed that 4685 queries included that word (table 3.4, Appendix A, 2), which is a small fraction of the total 1.75 million search queries from the United States. The process was repeated for different related words, all of which were only present in a small section of the overall observations. This meant that selecting a random sample of the data for manual coding would most likely results in very few queries related to misinformation. This could potentially be problematic for supervised text classification purposes, as the sample would be too unbalanced. The entire premise of supervised machine learning is that a ground truth or preexisting knowledge about the data already exists (Aggarwal, 2018). Hence, it was essential to extract data for manual coding, which would make it possible for a classification algorithm to distinguish between the different categories. Essentially a personal ground truth was established by creating an index of regular expressions related to COVID-19 misinformation. To accomplish this; misinformation in the context of COVID-19 was investigated on the internet. As covered in the literature review, misinformation has been a paramount concern during the recent outbreak (i.e. Islam et al., 2020. Cinelli et al. 2020). The literature review was used combined with several other COVID-19 misinformation reviews found on the internet. The World Health Organisation (WHO) has created the *Mythbusters* resource[31] as a part of their *Advice for the public* section on the WHO website. This section contains information on some of the most prevalent rumours or misconceptions related to the incubation time, infection sources, sickness process, and recovery from the coronavirus. It does however not include misinformation pertaining to conspiracy theories such as Bill Gates being blamed for spreading COVID-19 in order to be able to vaccinate everyone, and in the process

[31] https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public/myth-busters

install microchips meant to track and control individuals worldwide[32]. In order to cover these, additional COVID-19 misinformation resources were located, including covid19misinfo.org[33], NewsGuard.com[34], and even Wikipedia[35]. During the outbreak, a tendency has been identified of blaming people of a particular ethnicity (Islam et al., 2020. Vazquez, 2020. Rovetta & Bhagavathula, 2020a), namely people of Chinese or Asian origin as well as people returning home from China. On top of the two primary misinformation categories concerned with rumours and conspiracy theories, claims related to racial stigma and racism were also considered in the keyword selection.

The different misinformation claims were combined in an excel sheet, including a description and source. Each of the claims was assigned keywords that reflected the content. A few examples of these can be seen in Table 3.3, and the full spreadsheet is attached in Appendix B. It should be noted that keyword selection can be quite tricky in this context, as people might have used different tenses of the word, different spellings, or even spelled the words completely wrong. However, in most cases, the spelling is correct, so most queries should appear when using the

Example of Claims, Myths and Keywords (Full list in Appendix B)		
Claim / Myth	Source	Keyword(s)
Studies show hydroxychloroquine does not have clinical benefits in treating COVID-19	https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public/myth-busters	hydroxychloroquine, hydroxy, chloroq
Drinking alcohol does not protect you against COVID-19 and can be dangerous	https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public/myth-busters	alcohol
5G cell phone technology is linked to the coronavirus outbreak.	https://www.newsguardtech.com/covid-19-myths	5g, cell phone, radio waves
Bill Gates plans to use COVID-19 to implement a mandatory vaccine program with microchips to surveil people.	https://www.newsguardtech.com/covid-19-myths/	bill gates, microchip, chip
Gargling vinegar and rose water or vinegar and salt may kill the virus in throat	Islam et al., 2020. Table 2	vinegar, rose water
Eating bat soup is the source of the (COVID-19) outbreak	Islam et al., 2020. Table 2	bat soup
Wearing a mask dangerously raises carbon dioxyde levels and causes hypoxia.	https://www.poynter.org/coronavirusfactsalliance/	carbon dioxide, hypoxia
QAnon about the coronavirus. There are several different conspiracy theories introduced by qanon in relation to the coronavirus. One being that COVID-19 is a chinese bioweapon created in collaboration between China and the Democrats to stop Donald Trump	https://theconversation.com/qanon-conspiracy-theories-about-the-coronavirus-pandemic-are-a-public-health-threat-135515	qanon, hoax, lie, deep state, elite, bioweapon

Table 3.3: Example of misinformation claims (Appendix B)

[32] https://allianceforscience.cornell.edu/blog/2020/04/covid-top-10-current-conspiracy-theories/
[33] https://covid19misinfo.org/
[34] https://www.newsguardtech.com/covid-19-myths/
[35] https://en.wikipedia.org/wiki/Misinformation_related_to_the_COVID-19_pandemic

correct spelling of a word (Mavragani & Ochoa, 2019). As all the words were specified in English, and the search queries were from the United States, it was expected that the vast majority of search queries would be in English and thus would not need a translation. Another critical reflection in the keyword selection process was the placement of white spaces and related words. Examples of this include *5g* which would return nothing if it was written as *5 g* or *bat soup*, which had to be searched together as a combination of words as the word *soup* alone would not necessarily be pointing towards misinformation. It was an iterative process with a lot of trial and error before finding keywords that made provided accurate results. Each word was individually searched, and the responses were read through in order to determine whether the results did express interest in areas of misinformation. Some words provided new results if they were split into smaller pieces, for example, *hydroxychloroquine* which returned 2076 queries, while merely using *chloroq* (chloroquine) yielded 4117 observations. In this case, the full word was used, and two additional keywords were created: *hydroxy* and *chloroq*. It was especially important to consider the implications of word order when including terms related to stigma and racism. Simple searching for the word *China* would return 11337 observations, but they would not be directly associated with misinformation. However, searching for *China virus*, which could be argued to express racist or at least xenophobic views (Vazquez, 2020), would return 902 results. The same principle was relevant for filtering on *Wuhan* or *Wuhan virus*. All keyword filtering was set to ignore word cases to ensure that keywords were not missed because of capitalization. The search and evaluation process was repeated for every single claim and corresponding keywords, and the words used in the final misinformation keyword index are featured in the keyword column of Appendix B.

While the process of keyword selection is fundamental, especially for the validity of the results (Mavragani & Ochoa, 2019), covering every single claim of misinformation would be beyond the scope of this paper. The covid19misinfo.org initiative has manually reviewed more than 4000 misinformation claims related to COVID-19, and they are continuously expanding their work (*Misinformation Watch* -, 15.08.2020). This shows the extent of attempting to completely cover all the many aspects of COVID-19 infodemics, as new claims are continually surfacing. It is believed that the majority of the top trending misinformation claims were included in my work, but there might be examples of missing allegations, or even new claims that became popular after the keyword index was finished. A final consideration was to select keywords that were relevant to the information in the United States. The primary objective of the keyword index was to sample data that could be used for manual coding and later for training machine learning models. The best machine learning model would be used to classify the entire dataset, so it was important that the information was relevant to the US. The selected keywords were all found to be applicable to the United States, but should the approach be repeated for other countries, additional country-specific keywords would have to be collected. Table 3.4 shows the most frequently occurring words after filtering by the keyword index.

Query	n
qanon	4685
herd immunity	1641
hydroxychloroquine coronavirus	1402
face masks made in usa	1400
bill gates coronavirus	991
masks made in usa	685
malaria drug for coronavirus	679
wuhan virus	662
diy hand sanitizer	505
bat soup	402

Table 3.4 - Top misinformation queries after keyword filtering

3.4.3 Sampling

While research within infodemiology and search queries, especially within the Google Trends environment, has become increasingly popular in the last decade, it is still a relatively new area of research, which tend to lack a unified way of gathering insight and reporting results. One suggested approach is to use manual coding of sample data in order to be able to generalize on more massive datasets (Mavragani & Ochoa, 2019). Not only is manual coding suggested in related research, but it is also a necessary step when working with supervised learning and unlabeled data. Before the manual coding process can begin a sample of the data has to be extracted.

A common concept when working with sample selection within qualitative research is to use *purposive sampling* (Bryman, 2012). As the name suggests, purposive sampling is the process of extracting samples with a purpose - in this case; the objective is to answer the research questions. Purposive samples can have different levels of sampling, depending on the research questions on hand. One example of a popular approach in qualitative research is to first sample on geographical location and then later selecting participants within that area. Furthermore, it should be noted that a research project does not need to strictly adhere to one specific sampling method, as the different levels of sampling might require or apply different approaches (Bryman, 2012). The first step of sampling in this thesis was to select the search queries from the appropriate geographical location. As the primary problem statement, and all research questions, relates to the United States, the data was sampled to only include queries from the US population. This introduces a limitation to the level of generalization, as the sample restricts findings to be relevant only within the selected area With the initial country-based sampling done, the second level of sampling was to extract data for manual coding, which would later be used for classification model training. However, this proved challenging as the initial searches by keywords revealed that most of the data were regular information, so simply selecting a random sample from the entire dataset would not ensure that any queries containing misinformation would be included. To work around this limitation; a keyword index of regular expressions (section 3.4.2) was created. The index was used to label all the observations and made it possible to extract samples that would include misinformation queries. The main goal of this sample collection was to end up

with a balanced dataset which could be trained in a classifier without having issues of one factor level being overrepresented. In that sense, this step of the sampling could be referred to as purposive sampling. However, the actual process of selecting observations was *simple random sampling*, as known from quantitative research (Bryman, 2012). A random sample of 1000 observations was extracted from both the misinformation and regular information category. The two were merged into one long dataset with a total of 2000 observations, and the rows were scrambled in order to remove bias in the coding process. To further remove any information that could affect the coding and introduce bias, the misinformation column created by the keyword index was removed. The code for the process can be found in Appendix A 2.1.2.

3.4.4 Manual Coding

A popular approach when working with unstructured text data, like for example, search query logs, is content analysis, otherwise referred to as manual coding. It is a method used in several different research disciplines and is especially useful in identifying clusters of information in otherwise unstructured data (Lazar, Feng & Hochheiser, 2017). Typically, this is an elaborate process of getting to know the data extensively and identifying themes, characteristics, and other relevant variables in the data (Bryman, 2012). However, in this thesis, the primary objective of the coding process is to identify cases of misinformation. In a sense, the coding process functions as an evaluation of the manual keyword index, with the primary objective of identifying observations that were correctly classified, as well as identifying cases of misinformation that were classified as search queries related to regular or credible information. Traditional content analysis utilizes either *emergent coding* or *a priori coding*. Emergent coding depends on concepts that emerge in the coding process. As the coding process proceeds, more and more coding concepts appear until a final model includes all the different aspects of the relevant data. This approach is especially suitable when no pre-existing theory or hypothesis about the data exists (Lazar et al., 2017). A priori coding operates based on pre-existing ideas or premise, and the content is coded according to these. In this case, the main problem is to identify which observations relate to misinformation and which does not, the hypothesis being that the data can be divided into one of the two categories. This choice is heavily based on the underlying research questions, as they determine the coding method (Bryman, 2012). This project attempts to define the *extent* of expressed interest in COVID-19 misinformation related topics but had the primary goal been to identify the *types* of misinformation appeared, emergent coding could potentially have been a better choice. However, the process could still have used a priori coding, for example, based on misinformation categories defined by related research such as Islam et al. (2020) or covid19misinfo.org. Furthermore, the objective of the coding process was to be able to train a classification algorithm to distinguish between the two categories and then use the outcome of the classifier for further exploratory analyses. Based on previous experience (Windfeld, 2019, 9th-semester paper), it can prove challenging to use emergent coding if the goal is to use automatic text classification and machine learning to classify an extensive dataset. This due to the massive imbalances that can occur between classes and the difficulties of using significantly underrepresented classes for model training.

The sample data were coded according to the various misinformation claims specified in Appendix B, as well as pre-existing knowledge of the area. The observations were coded in a binary variable and assigned a value of 1 if related to misinformation. Observations that were not associated with misinformation was either coded as 0 or kept blank. A secondary coder was recruited and instructed in the problem area and coding process. As one of the primary objectives of the manual coding step was to evaluate own keyword index and knowledge, it was important to recruit an individual with a certain level of expertise on misinformation in the context of COVID-19. The secondary coder is a student of political science at the University of Copenhagen. Outside of personal interests, he has worked with misinformation projects at the DIPLOFACE[36] and Digital Disinformation research groups at the University of Copenhagen, as well as working with me at the COVID-19 Snapshot Monitoring Denmark project (COSMO)[37].

3.4.5 Building a Classification Model

Working with text data from the internet; there are generally two different approaches to machine learning: *supervised* or *unsupervised*. Unsupervised learning is concerned with the discovery and clustering of data without labels, usually through processes such as clustering by k-means or topic modelling using algorithms such as Latent Dirichlet Allocation (LDA) (Aggarwal, 2020). This was the approach taken by Suh et al. in their study of behavioural changes in the US during the coronavirus outbreak (2020), which also included labelling the clusters at a later stage. However, this thesis utilizes supervised learning based on the labels created through manual coding, meaning the labels already exists when the machine learning algorithms are tested. The motivation behind this choice is to be able to distinguish between queries related to regular information and misinformation accurately. Supervised learning involves training a smaller sample of the data, testing the understanding on a test set, before finally applying the optimal model to generalize on a more extensive collection of unlabeled data (Kuhn, 2008). The initial training data is crucial to the overall success of the trained model, as all subsequent results and labels will be determined by the success of the training process (Aggarwal, 2018).

The data was split into a training and test set at an 80/20 ratio. 1601 observations were used in the training set, and 399 observations were saved as a test set for the final model. The function `set.seed` was used in order to ensure reproducibility (Appendix A, 3.2). This was done in all subsequent code that included elements of randomization. If a more extensive training set is available, another split can be done in order to create a validation set, which can be used for model optimization before fitting it to the final test set. In case the data sample is limited in size, this process can be substituted by using cross-validation. Cross-validation splits the training set up into a set amount of equally sized segments, one of these is considered a virtual testing set and the remaining components are used for training. This process can be continued a defined number of times, each time a new segment is left out for testing (Aggarwal, 2018). All machine learning testing was performed, using 10-fold cross-validation and

[36] https://politicalscience.ku.dk/research/projects/diploface/
[37] http://copsy.dk/cosmo/

the splits were set to be stratified according to the original partition, meaning the ratio of misinformation to regular information was kept the same across all cross-validation splits. Initially, when trying out different models, the 10-fold cross-validation was set to repeat one time. At later stages, repeats were increased to determine their impact on model performance.

3.4.5.1 Pre-processing and feature extraction

Before starting the machine learning experiments, a few steps had to be taken to prepare the data. The *Date* column was changed to a character format after importing the CSV file and was changed back to the proper date format. As mentioned, the response variable was decided to be the results of coder 2, and this was renamed to *misinfo*. Furthermore, it was transformed into a factor variable containing the levels *yes* and *no*. As the tidymodels library always predicts the first level of a factor as positive, the levels were checked. These were in the wrong order, which would make the classification algorithm predict regular information as yes, which was not the intended outcome. The levels were reordered, so misinformation would be the level to predict as positive (yes) (Appendix A, 3.2).

Response variable and all predictors are defined in the initial stage of a tidymodels workflow, otherwise referred to as the *recipe* handled by the recipes package in the tidymodels environment. Initially, it must be decided which variables should be used as predictors, and which should be left out. In the early experiments *Date* and *State* was used as predictors alongside the text from the Query variable. However, these were left out for later training, as it was decided that they did not make much sense for the relevant classification problem at hand. Investigating whether a particular date contained more misinformation was an interesting theory, but it proved remarkably unreliable in early testing. Given the small size of the split used for classification (2000) when compared to the overall size of the data set (1.75 million), meant that the classifier would identify patterns that were only present in the training set, but indicated that several queries were misclassified when applied to new data. The same was evident when using *State* as a predictor. Rather than treating these variables as predictors in the classification process, they were used for later exploratory analysis.

It was decided only to use the *Query* variable containing the text of the search queries as the predictor, which was stored as a character variable within *R*. An initial step in natural language feature generation is to convert all text content into a numeric representation. This makes it possible for machine learning algorithms to understand and perform training on them. This process is referred to as tokenization (Hvitfeldt & Silge, 2020). The final output and method are often known as a *document-term matrix* or *bag-of-words* model. Each word is converted to a vector represented in the columns, each row is a document (in this case the individual queries), and each value is determined by the number of times the word occurred in the text corpus (Silge & Robinson, 2017). While this process would usually involve several steps and the use of other libraries (tidytext, tm), the tidymodels framework does all of it in the step_tokenize function. An example of the text unnesting process and final vector representation can be seen in Figure 3.2. In this process, all text material was also converted to lowercase. In some cases, it can be necessary to find and remove punctuation or strange characters, but none of these was found when exploring the data, so this was left out of the pre-processing. Search queries are generally shorter than other text

material from the internet such as forum- or social media posts which would likely require more cleaning. The next step in the feature generation was the removal of common words that generally do not carry any semantic meaning, and in turn, should not be considered for classification training. Outside of the primary purpose of removing insignificant words, it also reduces the sparsity of the document-term matrix, which can become extensive when working with text. This was done using the step_stopwords function which uses the stopwords() package. The *Snowball* stopword index, created initially by Porter (2001), was used. It is quite a bit smaller than some of the other lists such as SMART or ISO, but it was decided that it would be better to allow a word with no semantic value than accidentally removing essential words. Choosing the list was important, as several stopwords lists have proven to be of a low quality which would end up eliminating words that could be significant to the document (Hvitfeldt & Silge, 2020). The list was reviewed to make sure no essential words were removed in the process, and an additional custom stopwords list was created. Creating the custom stopwords list was an iterative process of training the various models and reviewing the variable importance metric and adding words to the list if they were not relevant to predicting misinformation. Examples of words manually added to stopwords are "*d*", "*can*", "*dr*", and "*George Floyd*" (Appendix A, 3.2). The next step in the recipe step_ngram defined the number of words that could be considered as a unit for classification purposes. This allows the use of both unigrams and

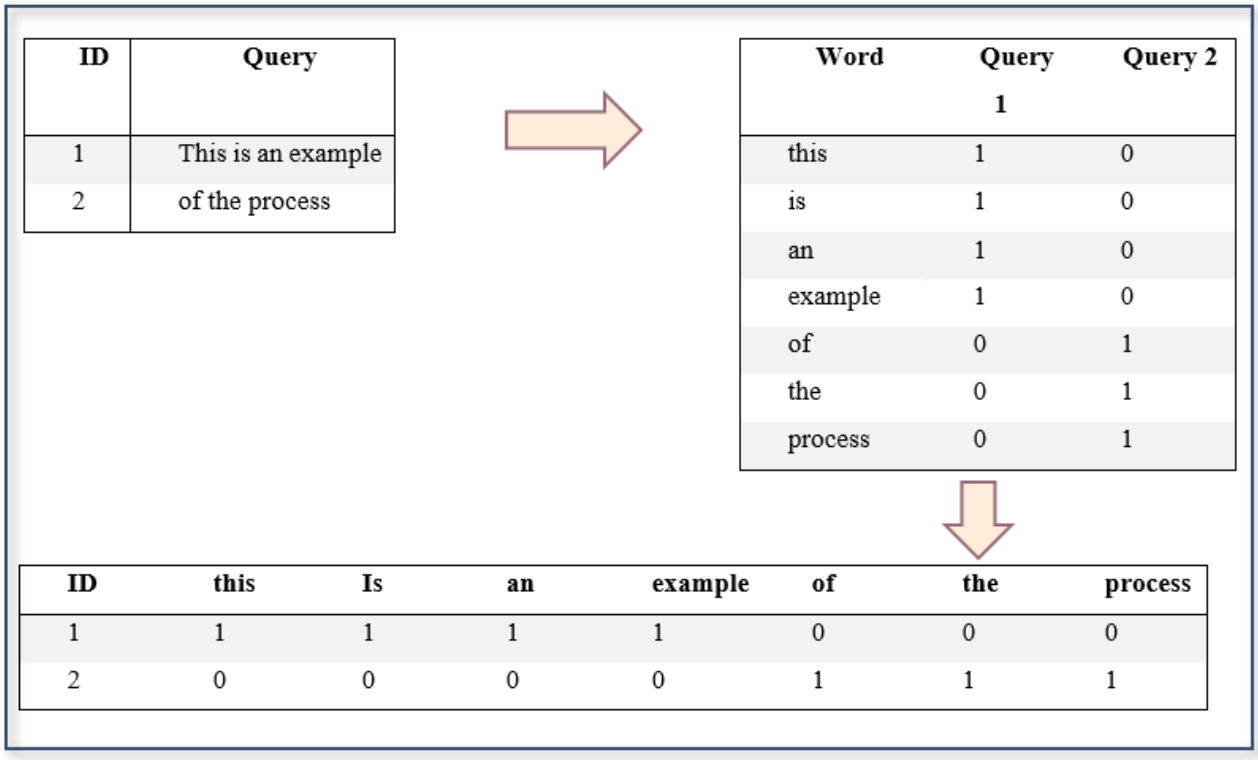


Figure 3.2 - Example of text preprocessing

bigrams in the same classification training process, which is convenient when compared to other workflows where models would be trained on document-term matrices of the two variations separately. The recipe was assigned to consider a maximum of two words together (bigrams), and a minimum of 1 word (unigram). The next step includes various filtering options such as the number of times tokens must appear before being removed from further

predictions, as well as the number of total tokens to consider for classification. The `max_tokens` function was used to set an integer range between 20 and 210, which allowed the model to find the optimal number of tokens to consider without being restricted. The max number of tokens was discovered in the model training as several cross-validation folds returned errors when attempting to select more than 200 tokens/features for prediction. A common question when working with text data is how to determine what the document is about. This can either be done by looking at the overall *term frequency* (tf), the frequency of occurrence of each word. However, sometimes the frequency might not be the best approach, as insignificant words could appear more than essential words. Instead, each word can be assigned a value by *inverse document frequency* (idf). The two measures can be combined in the tf-idf score, which measures the importance of each word in their document (query) in a larger corpus (the collection of queries) (Silge & Robinson, 2017). Essentially the *tf-idf* scores account for word frequency by decreasing the value of frequently occurring words and increasing the value of words that occur less in the full dataset. It was already known, from the preliminary word searches, that some words were much more frequent, and for this reason, *tf-idf* scores were used for model training (Serrano et al., 2020). This was the final step of the recipe and was executed with the step_tfidf function from the *textrecipes*[38] package.

3.4.5.2 Model Selection

After the recipe was defined, the next step was to decide which algorithms to test and defining the (hyper)parameters for each of them. While fully understanding the concepts behind each of the algorithms was both beyond both the scope of this thesis and personal capability, it was still necessary to select models that have proven to work well for text data. The first model chosen was a logistic regression model set to classification mode. The model was used from the *glmnet* package and included *LASSO regularization,* which uses both variables and regularization to find the best model and decrease the chance of overfitting (Qin et al., 2020). A benefit of the *glmnet* model and all other selected models is the possibility of extracting variable importance, i.e. the most important words used by the algorithms in their classification choices (Silge, 2018). Furthermore, it is a model that works well with sparse data, as is the case with text (Boehmke & Greenwell, 2019). The model was set to tune the *penalty* parameter between the values -4, and 0. The next model was a random forest algorithm from the *ranger* package. *Ranger* is an exceptionally fast random forest type (Wright & Ziegler, 2017), which was beneficial as the algorithm had to be run on a personal desktop computer with limited resources compared to the powerful machine learning servers that are often used for these cases. Due to the high-dimensional nature of text data, random forest algorithms using default parameters are usually not that well-suited for text classification (Aggarwal, 2018). It is, however, possible to get good results of the hyperparameters are tuned well. Various tests with different hyperparameter settings, but the final values used were 500 trees, a range of 10-20 randomly sampled features considered in each decision (mtry) and the number of trees set to the range 2-8. The next model was a boosted tree algorithm from the *XGBoost* package, using default parameters. Finally, a support vector machine (*SVM*) model was trained with a polynomial kern and default parameters. SVM models have proven very efficient when

[38] https://www.rdocumentation.org/packages/textrecipes/versions/0.3.0

applied to text data (Nguyen et al., 2016), and it was a natural choice to include an SVM variation in the model comparison.

As the primary goal of the initial model training and comparison was to be able to determine which model most accurately distinguished between regular- and misinformation words, the models were set to collect the same metrics which could be used for comparison. The extracted metrics were *sensitivity*, *specificity* and *ROC-AUC* (Receiver Operating Characteristic – Area Under the ROC curve). Both sensitivity and specificity are metrics used to evaluate the performance of binary classification tasks and are used frequently in the field of diagnostic medicine (Florkowski, 2008). The sensitivity value is a measure of how many true positives that were correctly predicted, whereas the specificity value measures how many of the true negatives were correctly predicted. There are different approaches to defining these metrics within different research disciplines. In the context of information retrieval, other metrics such as precision and recall would likely be preferred. Precision, otherwise known as *positive predictive value,* is the number of true positives within all positive predictions. Recall measures how many of the total positives were predicted as being positive. Sensitivity and recall are essentially the same measures with different names depending on the scientific context (Ting, 2010). Deciding which metric to optimize is highly dependent on the situation, as the consequences of false positives/false negatives vary (Koehrsen, 2018). The main goal of the classification process was to correctly identify all queries related to misinformation, making the most essential optimization metric sensitivity/recall. The final models were compared by the ROC-AUC score, which measures the probability of the model's ability to distinguish between the classes. The score is calculated between 0 and 1, the higher the value, the better the model is. A value of 0 means that the model is essentially predicting the opposite as true, meaning it is wrong a 100% of the time. This was a problem in the initial training as the order of the response variable was wrong but was quickly corrected by reordering the factor levels. A value of 0.5, often marked by a diagonal line on the plot, means that the model is correct 50% of the time and is not able to separate the two outputs at all. In this case, each prediction is essentially a guess, and the differences between the classes were not learned in the training and testing process.

Finally, a few notes on model complexity in the context of text classification; While more complex algorithms may perform better in many cases, they can become problematic when working with a simple binary classification problem based on text. Due to the sparsity of text, individual words might not mean a lot to the overall prediction as it is only meaningful in the context of all the features. Some classifiers, such as random forest, use sequential decisions which can often lead to overfitting and results that can't be trusted. For a binary classification task, linear classification models that can use all features at the same time (*glmnet*) often provide better results than more complex algorithms (Aggarwal, 2018). Additionally, the presence of a word in the model is often more important than the absence of the same word, meaning the word should occur before something is classified as misinformation – something should not be classified as misinformation simply because of some words not appearing. This is especially relevant when applying the findings from a small training set to a large new dataset, as this could potentially lead to many wrongly classified observations and overfitting. Finally, if models have similar

performance, it can be beneficial to use a simpler model, especially if they are trained and applied in local desktop environments with limited resources. Complex models require extensive computing power, which was a concern in the context of this thesis.

3.4.6 Exploratory Analysis

The exploratory analysis was divided into different sections corresponding to the research questions. The first research question, about identifying misinformation queries, was covered by the previous sections on manual coding and automatic text classification. The second section of the analysis explores overall distributions of the classified dataset in order to determine the overall extent of misinformation across the United States. This section uses descriptive statistics to describe and visualize the data and describes distributions across the entire time period as well as monthly distributions. The second part of the exploratory analysis describes differences between the states and visualizes these in scatterplots and map visualizations of the United States. This provides an overall picture of misinformation across the US and makes it possible to identify states that are significant outliers. This is also done both for the entire time period and per month. The following section explores the most popular search queries of the time period, both when measured by term frequencies and weighted log odds. In order to evaluate the initial results of the supervised machine learning experiments, and identify interesting patterns within the data, a bigram network was created. This made it possible to identify related words and clusters and identify potential outliers or misclassified queries. Finally, other data sources were explored, namely, CoronaNet and Google Trends. CoronaNet was used to investigate possible explanations for the findings, and Google Trends was used to establish external validity, which will be covered in the next section. Three different potential reasons for differences between states were explored: The overall *number* of policies implemented in each state, the *type* of policies implemented, and finally, the *political orientation* of the states.

3.5 Reliability & Validity

Several measures were taken in order to increase the reliability and validity of the work. Generally, social research is concerned with the three concepts reliability, replication, and validity (Bryman, 2012).

Reliability is concerned with the overall consistency or quality of the research, and whether the results can be repeated. In this case, the reliability was mostly a concern when defining what makes a search query related to misinformation. Replicability is closely associated with reliability, and it is valued highly especially by social researchers working with quantitative methods. The ability to reproduce the work is sometimes referred to as *external reliability* (Bryman, 2012). This was achieved by creating a sheet of misinformation queries and using the content of the sheet as the basis of the manual coding process. If the same sheet was used by other researchers, results should be reproducible. There is, however, a problem in terms of reproducibility and the creation of a keyword index. If researchers were to read the same claims and define their own keywords, other words might very well come up, which would change the results of the latter classification, analysis, and results. This relates to

the other half of the concept – *internal reliability*. In order to ensure internal reliability, a secondary researcher was involved in the coding process, and the intercoder reliability was found to be very high. However, additional coders could have been recruited to ensure representation by different backgrounds, capabilities or research disciplines. The two coders in this thesis were very similar in academic interests, knowledge of the area, and experience, so recruiting beyond that, and still achieving intercoder agreement, could have improved the overall internal reliability (Lazar et al., 2017). In terms of the data pre-processing, wrangling, text classification, and visualization, everything was made fully reproducibly. All the code is shared in Appendix A, and the full R data file, containing the individual objects, is also attached. Furthermore, the `set.seed` function was used in any chunk of code that involves randomization. This means that running the *R* code in the reader's own environment will produce the same results presented in the thesis.

One of the essential concepts in all research is *validity* which is concerned with the conclusions generated by the study (Bryman, 2012). Like reliability, the idea of validity also involves several different aspects, including *external-* and *internal validity*. Internal validity is concerned with any conclusions suggesting causal relationships between variables. These variables are frequently defined as the *independent variable* (the factor the causes something to happen) and the *dependent variable* (the effect of the previous factor) (Bryman, 2012). Care was taken to separate the concept of causality from this observational research, which is more concerned with finding associations between variables, rather than concluding that one caused the other (Zweig & DeVoto, 2015). However, one kind of causality existed in the research design, which impacted the internal validity. The defined keyword index was used for manual coding (factor) and later caused the corresponding labelled output of the machine learning process (effect). The primary way this impacted validity was through the quality of the selected keyword index. While measures were taken to handle this with care, the overall process is prone to be personally biased, which might have impacted the internal validity negatively. This process could have been strengthened by having the secondary coder assign keywords as well, or alternatively use keywords from ongoing COVID-19 misinformation studies such as covid19misinfo.org. *External validity* concerns how much the relevant findings can be generalized to other areas outside of the specific context the research was performed within (Bryman, 2012). In this case, Microsoft Bing search queries from the United States. In order to improve the validity of the study, selected results were compared to similar queries from the Google search engine using Google Trends. As Google holds a considerable part of the market share, it was relevant to review their results and compare them to my own findings. This improves external validity in terms of similarity across different search engines. However, an additional layer of validity existed in terms of language and population. As the keyword index was created in English, and several of the pre-processing steps used in classification (i.e. stopwords) are optimized to work well in English, the process would have to be adopted to other countries. The algorithms were trained and tested on other English-speaking countries, but the number of observations was much smaller in these, and it was not possible to thoroughly review external validity in this context. For this to be accurately tested, additional observations would have to be collected.

3.6 Ethical Considerations

Automatic collection of individual trace data or transaction logs has sparked several ethical debates through the history of the method. Examples of this are issues of privacy and anonymity, data ownership, and whether consent should be sought if the data is to be used both in business and research (Penniman, 2009. In Jansen, 2009). It should be noted that demographic data such as age and gender are usually unknown in the collection of search query logs, which does help the process of ensuring user anonymity (Hawkey, 2009. In Jansen, 2009). As everything becomes more connected between social media platforms and Google representing such a significant share of the market, larger companies likely already have this information. If data is released for research, it should be carefully considered how much demographic information should be presented to the public, as this makes it easier to re-identify individuals even though the data might seem anonymized (Rocher, Hendrickx, & Montjoye, 2019).

The data used in this thesis contains no demographic information, such as age, gender, or ethnicity. Neither does it have any identifying names or pseudonyms which could be used to identify the sender. Furthermore, all the search queries were only included if they were performed by many different users, and as such, did not contain an individual ID. For these reasons, no further actions were taken to anonymize the data.

There was an additional layer of ethics to consider in terms of using the data for machine learning purposes. When classifying concepts, the consequences of the outcome should be considered. There have been several examples of machine learning leading to problematic outcomes. A study from the University of Washington found that Google's AI, created to recognize hate speech in online environments, classified content as hate speech if it was written in African American English. This meant that the algorithm was essentially racially biased, and twice as likely to label content of this type as toxic or offensive (Lu, 2019). Another problem was found in an AI trained by Amazon.com to match potential job candidates with businesses with the tech-domain. The machine learning process was trained on resumes over a range of 10 years. The problem was that, due to the tech industry being male-dominated, it was taught to identify and prioritize wordings used by men, while penalizing words more commonly used by women (Dastin, 2018). This led to an algorithm that predominantly featured men as most qualified for jobs. When I initially trained the machine learning algorithms, the variable state was also included. The primary motivation for comprising states was the hypothesis that some states would be more prone to search for misinformation. While this hypothesis could be true or false, it was problematic to introduce this kind of bias into the training process, especially given the small sample size used for training. If the model found individual states to be especially interested in misinformation, it could have a massive impact when labelling the entire 1.75 million queries. Furthermore, as previously mentioned, it just didn't make sense for the classification of search queries to also use states, as the primary goal is to research the extend of misinformation in the search queries. The states were instead used in the exploratory analysis and were purely observational rather than driven by personal bias.

Finally, as the data used for classification were manually coded, bias could have impacted the process. A great effort was made to only include misinformation claims that had been proven to be myths or conspiracy theories

by credible sources, and all claims were critically reflected on in the coding process, as well as being coded by a secondary coder. Generally, an effort was made to reduce any kind of personal bias and remain objective.

4 Analysis & Results

4.1 Results of Manual Coding

In order to determine a ground truth on misinformation status, the results of the coding process had to be evaluated. The results were collected and combined into one dataset consisting of the original data variables and two new variables corresponding to coding results. As the datasets were mostly coded with a 1 for misinformation, and regular information kept blank, all missing values were replaced with 0. The secondary coder was labelled *Coder_1*, and my own coding results are found in the *Coder_2* variable (Appendix C). Comparing the individual count of the two coders, Coder 1 labelled 829 observations as related to misinformation and 1171 as regular information. Coder 2 marked 831 as misinformation and 1169 as standard information (Appendix A, 2.1.3). Interrater reliability was calculated using Cohen's Kappa with a value of 0.946 (Appendix A, 2.1.4). While the agreement between the two coders is indeed very high, almost near identical, additional information is required to evaluate the coding results. Some queries might be coded differently, and this should be investigated in order to determine which variable to use in further classification model training. A total of 52 observations were found to be coded differently across 35 unique search queries. The differences are visualized in Figure 4.1 (Appendix, 2.1.5), with coder 1 (secondary) in orange and coder 2 (own) in purple.

One of the most significant differences is seen in queries related to Dr Fauci, which was coded as relating to misinformation by coder 1. Dr Fauci is the head of the National Institute of Allergy and Infectious Diseases (NIAID), and while his name is frequently mentioned when people are discussing myths about mask-wearing, Dr Fauci himself is not a source of misinformation. On the other hand, he has been contributing to debunking claims about masks being dangerous to wear[39]. Secondly, coder 1 labelled several DIY mask related queries as misinformation. However, even DIY masks can have an impact and reduce the spread of COVID-19, and the Centers for Disease Control and Prevention (CDC) have their own guide to making homemade masks[40]. Three different claims related to pet animals were coded as misinformation by coder 1. Some claims have attributed the

[39] https://www.msn.com/en-us/health/medical/dr-fauci-says-there-is-no-truth-at-all-to-this-common-mask-myth/ar-BB16SbRz
[40] https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/how-to-make-cloth-face-covering.html

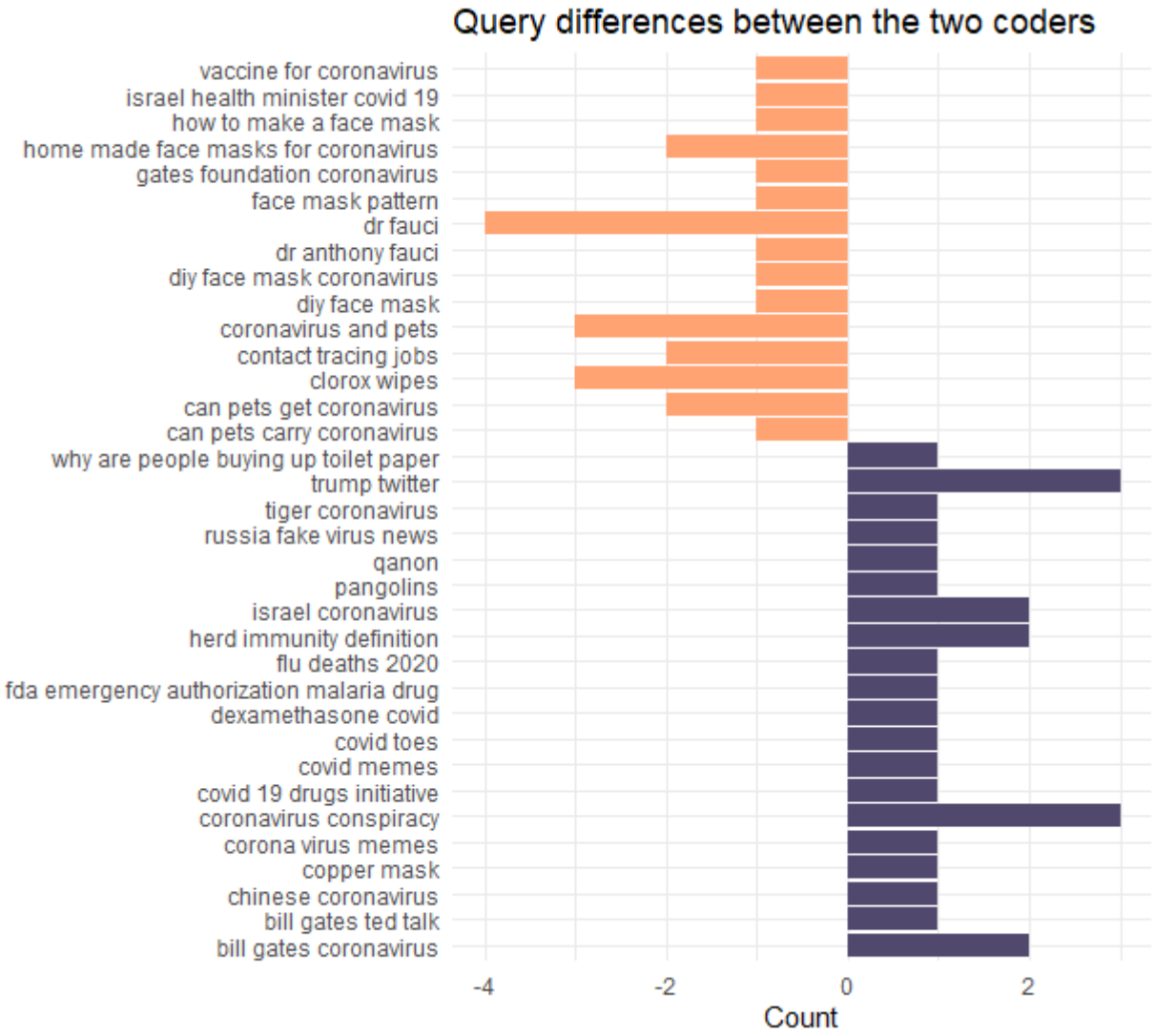


Figure 4.1 - Queries coded differently by the two coders

spread of COVID-19 to animal pets, but there is no evidence that they do so in a significant way when compared to humans. There have been rare examples of animals being infected, and queries related to this topic should not be considering misinformation[41]. Coder 2 caught several prominent misinformation queries that were missed by coder 1, for example relating to QAnon, Bill Gates, herd immunity, and malaria drugs as treatment for COVID-19. Trump twitter was coded as misinformation, and while this may seem like a stretch, there are several examples of direct misinformation coming from Donald Trump (Evanega et al., 2020). He has been responsible for contributing to several false claims such as immunity among children, injecting disinfects or bleach[42], and using UV light inside the body to kill the virus[43]. Several of Donald Trump's social media posts have been labelled false or even directly removed on both Facebook and Twitter[44]. Some queries related to Bill Gates were labelled as

[41] https://www.cdc.gov/coronavirus/2019-ncov/community/veterinarians.html

[42] https://www.theguardian.com/us-news/2020/apr/24/trump-disinfectant-bleach-coronavirus-claims-reaction
[43] https://www.theguardian.com/us-news/2020/apr/24/trump-disinfectant-bleach-coronavirus-claims-reaction
[44] https://www.bbc.com/news/election-us-2020-53673797

misinformation by coder 2, and this is considered correct due to his presence in several popular COVID-19 misinformation claims. Additional false allegations related to the origination and spread of the virus were correctly labelled by coder 2. These include falsely proposing *pangolins* as the source of the virus (Frutos et al., 2020) and *Chinese coronavirus* displaying stigma or xenophobia (Islam et al., 2020. Vazquez, 2020). Some queries were mislabeled like for instance *dexamethasone* which has indeed shown preliminary success in the treatment of critical patients[45], and *tiger coronavirus* which most likely is a result of a small COVID-19 outbreak among tigers at the Bronx Zoo in New York[46]. Generally, the observations coded by me (Coder 2) was found to be the most accurate, especially considering the correct labelling of major misinformation topics such as QAnon, Bill Gates, herd immunity, and malaria drug treatment. Herd immunity was considered misinformation due to the various conspiracy theories about vaccinations which have gained popularity in the last years (Jolley & Douglas., 2017). All these frequently appeared in the preliminary word index search (section 3.4.2), and were considered necessary, especially when training a model to identify these as related to misinformation accurately. For this reason, my own coding results were used as the response variable in the following classification process.

4.2 Results of Automatic Text Classification

This section presents the analysis and results of the different classification models. Each of them is compared on overall performance; a final model is selected and used to label the entire dataset. The models are compared on the metrics sensitivity, specificity and area under the curve (roc-auc). The area under the curve is especially relevant as it describes the overall model performance and ability to separate the levels of the prediction class (yes/no).

4.2.1 Evaluating Model Performance

The four models were initially compared by ROC-AUC score and were visualized in Figure 4.2. Each line represents a different fold of the cross-validation process. The jagged representation of the lines is most likely caused by the relatively small sample-size in conjunction with the binary classification problem at hand. Furthermore, the predictions are discrete or categorical rather than continuous, which can also explain the jagged lines[47]. The performance is generally very high, especially using logistic regression classification, random forest and support vector machine (SVM) models. A significant outlier is the boosted tree model (XGB) at the bottom left. The XGB model was not able to achieve the level of performance of the other three, which were all able to achieve sensitivity and precision values above 90%. Compared to the others, XGBoost is a more complex model and has likely been learning patterns the others were not learning. One of these patterns could be the previously mentioned absence of specific words, which in this case could have hurt overall performance (Aggarwal, 2018).

[45] https://www.who.int/news-room/detail/16-06-2020-who-welcomes-preliminary-results-about-dexamethasone-use-in-treating-critically-ill-covid-19-patients
[46] https://www.nationalgeographic.com/animals/2020/04/tiger-coronavirus-covid19-positive-test-bronx-zoo/
[47] https://www.quora.com/What-does-it-mean-when-an-ROC-curve-is-not-smooth

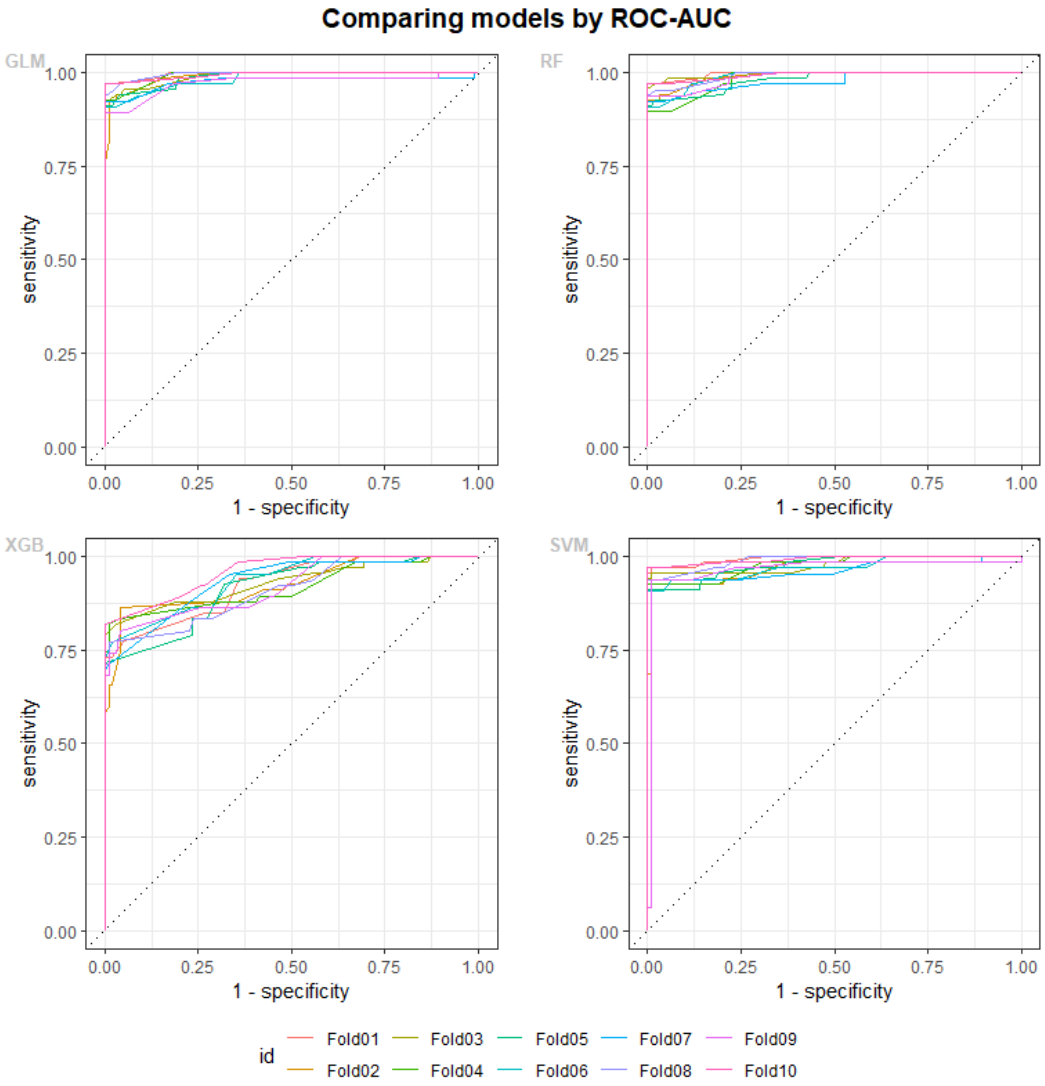


Figure 4.2 - Comparison of machine learning algorithms (AUC-ROC)

Overall, logistic regression and random forest models perform the best, and their results can be further examined using the `collect_metrics` function from the tune package in tidymodels. One of the essential tuning parameters was the max number of tokens considered in each prediction, set to test values between 20 and 210. As mentioned in section 3.4.5.1, the max value was determined after several training experiments which returned notifications of ~200 being the maximum number of tokens available. Even if this number was increased, no difference was observed, as no more than around 200 tokens/features were available for selection. The impact of the max number of tokens used in each prediction can be seen visualized in Figure 4.3 (glm) and Figure 4.4 (random forest). In the glm visualization, the colour represents the number of retained tokens (tokens used in predictions), and in the random forest visualization the lines are coloured by minimal node size and retained tokens are shown in facets. The x-axis in the glm figure shows the impact of the amount of regularization, determined by the penalty parameter specified earlier (-4 to 0). Regularization is used to penalize complexity, which can help prevent overfitting. Essentially it smoothes out and simplifies the model, and eventually, the model will be too simple to

make accurate predictions. This is evident in the figure that shows relatively stable curves until the regularization is too high, and the model can no longer accurately distinguish between the response classes.

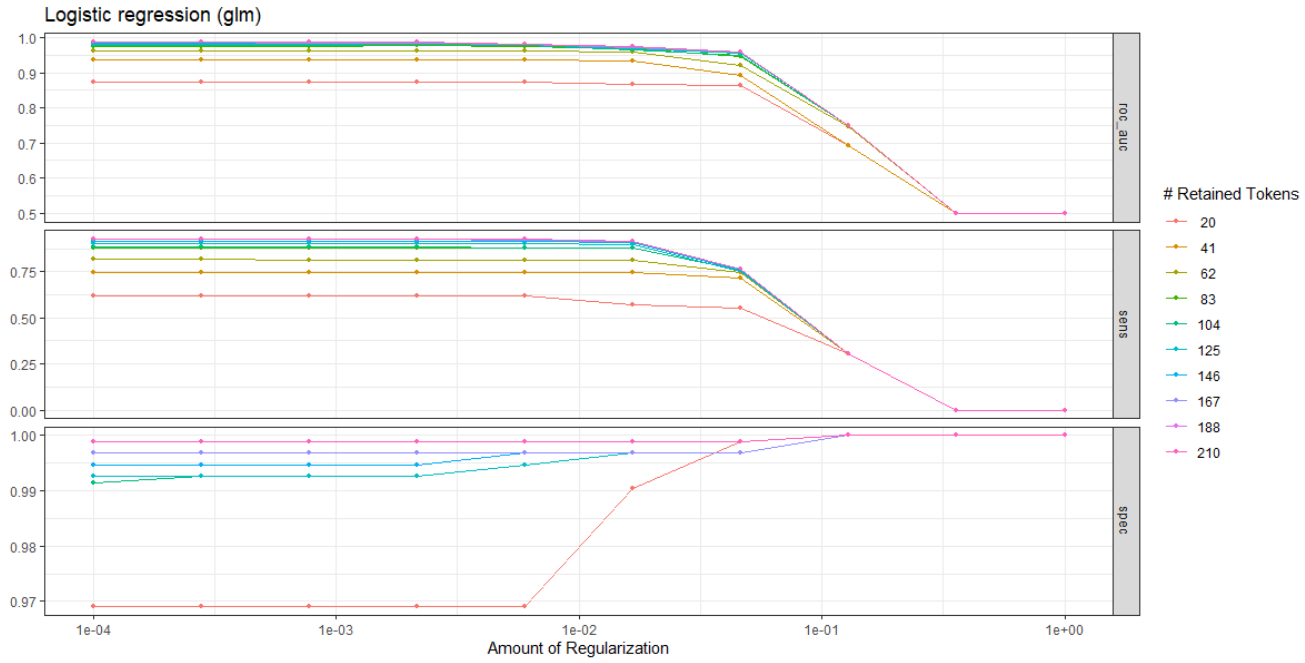


Figure 4.3 - GLM performance with regularization

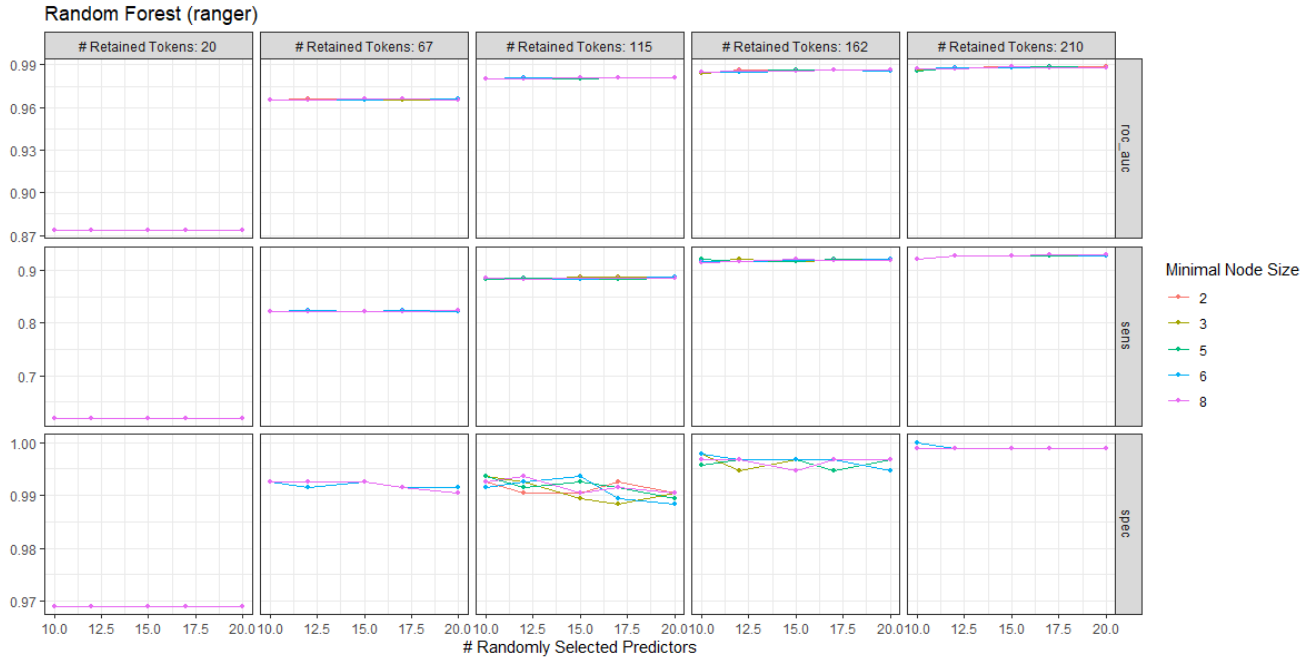


Figure 4.4 - Random Forest performance (ranger)

Generally, both models have better performance using a larger number of retained tokens in their predictions. Both models perform best with all the available tokens selected and perform the worst with the smallest number

of tokens (20). This is to be expected, as the models learn more about the classification problem when more tokens are introduced. The random forest model shows similar performance across the different minimal node sizes, with specificity showing the most considerable differences. It should be noted that the precision scale ranges from .97 to 1, so it is fluctuating within a tiny margin, and is generally very high. The number of randomly selected predictors has a low impact, although the higher value (20) shows a minor performance increase at the higher number of retained tokens. Similar visualizations and tables for XGBoost and SVM can be found in Appendix A, 3.8. Additionally, the top-performing models can be extracted using the *show_best* function and assigning which metric to measure by, in this case, roc-auc. These are shown in Table 3.1 and confirms the findings of the previously mentioned roc-auc curves (Figure 4.2). Both glm and random forest perform best with the max-tokens set high (188 and 210), which uses all the available tokens for prediction. Glm performs best with a low penalty (0.0001), giving a top roc-auc score of 0.987. Random forest performs best with a higher number of randomly selected predictors (17) with a roc-auc score of 0.988.

Top 5 glm by ROC AUC

penalty	max_tokens	.metric	.estimator	mean	n	std_err	.config
0.0001000	188	roc_auc	binary	0.9877355	10	0.0025368	Recipe09_Model01
0.0001000	210	roc_auc	binary	0.9877355	10	0.0025368	Recipe10_Model01
0.0002783	188	roc_auc	binary	0.9876552	10	0.0025237	Recipe09_Model02
0.0002783	210	roc_auc	binary	0.9876552	10	0.0025237	Recipe10_Model02
0.0007743	188	roc_auc	binary	0.9873142	10	0.0024595	Recipe09_Model03

Top 5 random forest by ROC AUC

mtry	min_n	max_tokens	.metric	.estimator	mean	n	std_err	.config
17	5	210	roc_auc	binary	0.9886889	10	0.0020064	Recipe5_Model14
15	8	210	roc_auc	binary	0.9886742	10	0.0020917	Recipe5_Model23
17	3	210	roc_auc	binary	0.9886360	10	0.0019968	Recipe5_Model09
15	2	210	roc_auc	binary	0.9884967	10	0.0020519	Recipe5_Model03
20	2	210	roc_auc	binary	0.9884592	10	0.0021131	Recipe5_Model05

Table 4.1 - GLM and Random Forest top 5 by ROC-AUC

4.2.2 Selecting the Final Model

When the optimal model for each algorithm was defined, it was implemented in the final workflow in tidymodels and applied to the testing set. The ROC-AUC curves of the last models are shown in Figure 4.5 and the

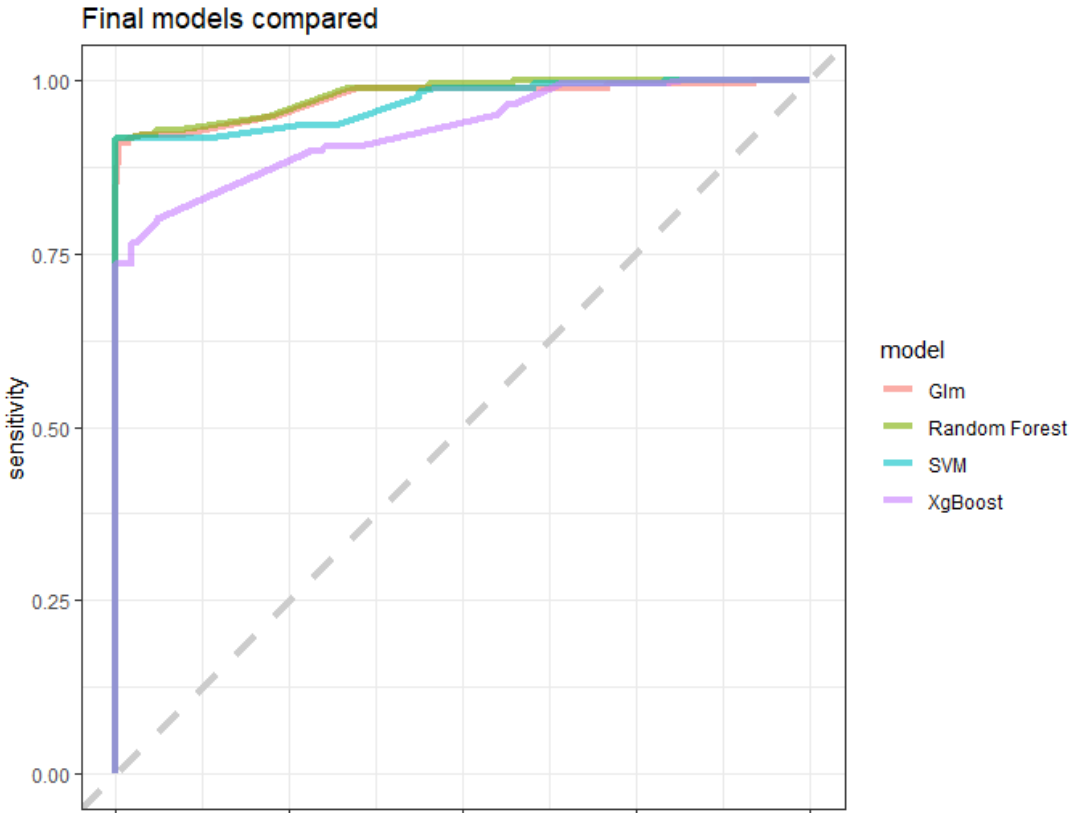


Figure 4.5 - Comparison of final models by ROC-AUC

corresponding metrics in Table 4.2. As we are especially interested in a high recall, the sensitivity score is essential, as well as the overall roc-auc score. The performance was excellent across all models except XGBoost, which had a very low sensitivity score of 45%. Glm, random forest, and SVM all had sensitivity scores above 90%, and roc-auc scores above 96% (Appendix A, 3.9.2). Specificity, or the number of true negatives located by the models, was very high for all the models. Glm had the lowest specificity, but even that was at 99%, and the remaining models had a specificity of 100%.

Model	Sensitivity	Specificity	ROC-AUC
Glm	0.910	0.991	0.974
Random Forest	0.916	1	0.979
XGBoost	0.452	1	0.928
SVM	0.916	1	0.967

Table 4.2 - Metrics of final models

An interesting observation found when looking at the immediate jump in sensitivity on the roc-auc curves (Figure 4.2), suggesting some variables might be critical to the overall sensitivity. There are several ways to investigate this, one being gain curves, which is often used in business analytics, and especially within marketing (Jurczyk, 2020). In this case, it can be used to understand the overall importance of using all the queries related to misinformation rather than smaller samples (Figure 4.6). The graph shows that using only 37-38% of the dataset will find ~90% of the true positives, which suggests that some features are significant to the classification problem at hand. The remaining gain curves can be found in Appendix A, 3.9.4.

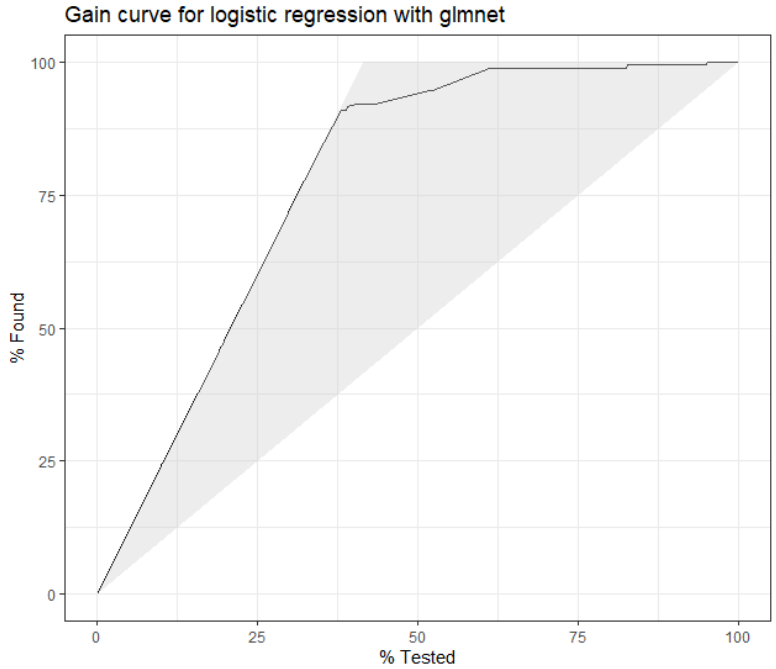


Figure 4.6 - Gain curve for GLM algorithm

This can be further investigated by looking into the importance of the individual features (queries), otherwise known as variable importance (VI). Figure 4.7 shows the most important features used for classification by the glm model. The tokens (both unigrams and bigrams) important to positive predictions (misinformation) are all relevant when compared to the results of the coding process and keyword index. *Bill gates* is the most important variable with twice the importance of the second token *herd* (related to herd immunity). One token having such a significantly higher value could explain the behaviour observed in the gain curve (Figure 4.6). Especially considering that several of the tokens with higher importance values were also among the most frequently occurring words in the keyword filtering process (Section 3.4.2, Table 3.4). All the tokens in the positive column are related to general misinformation topics, suggesting that the model can accurately identify misinformation queries. Tokens important to negative predictions are all generic, and none of them relates directly to misinformation.

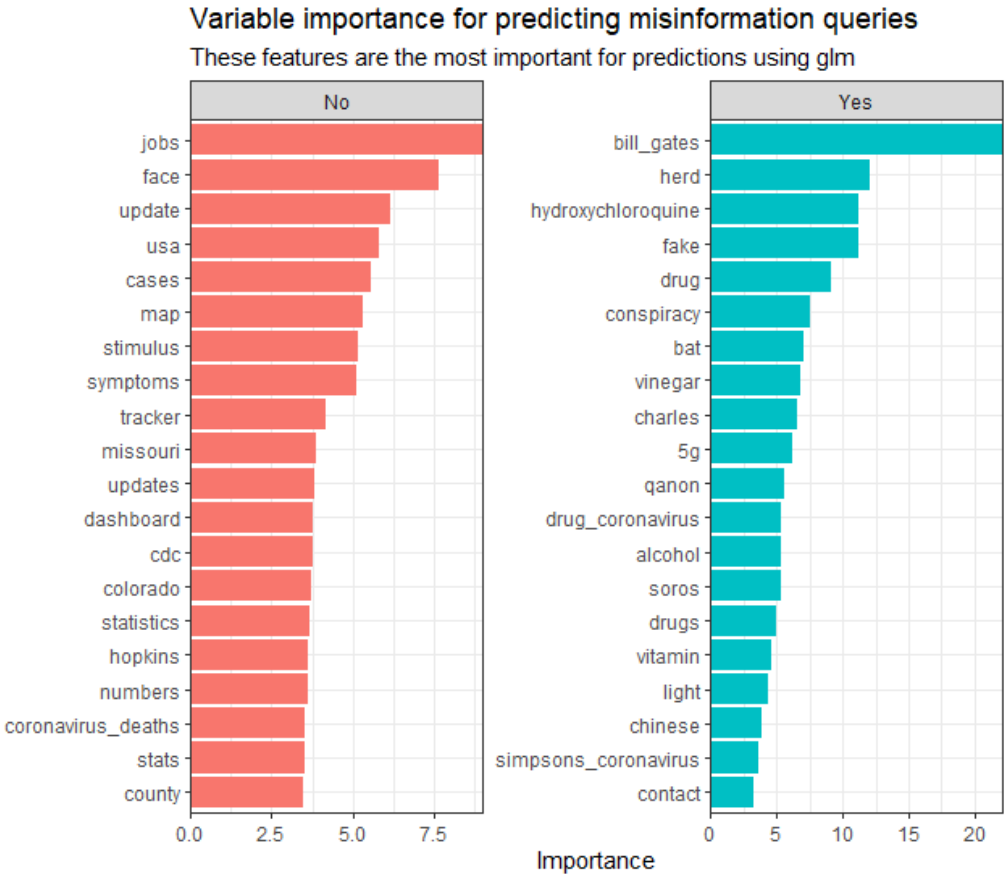


Figure 4.7 - Variable importance using GLM

The next step was to select a model to use for labelling the remaining and data. The two best models, according to the roc-auc curves (Figure 4.2), were glm and random forest. While SVM had similar performance when looking at sensitivity and specificity, it had a lower roc-auc score and was not used for the final classification. XGBoost was also dropped due to very low sensitivity scores. The choice was between glm and random forest. Given the very similar performance between the two, glm was selected as it is the simpler model. If a simple model can have equal performance to a more complex model, it can be the better choice, especially given the lower computational requirements (Aggarwal, 2018). The data was also classified using the random forest model, mostly for comparison purposes. This will be further reflected on in the discussion. The data used in the exploratory analysis were all labelled with the logistic regression model (classification mode) from the *glmnet* package. The glm model returns a labelled dataset as well as two columns with the corresponding positive and negative predictive probability values. The describes the strength of the prediction, i.e. how certain the model is of the predicted class. The probability is assigned a value between 0 and 1 split between the two outcomes. Converted to a percentage, the glm model had 99,5% probability of correctly classifying misinformation and 94,5% probability of correctly classifying regular information.

Response variable	Prob. Yes (mean)	Prob. No (mean)
Misinformation (.pred_Yes)	99,5%	0,5%
Regular information (.pred_No)	5,5%	94,5%

Tabel 4.3 - Average prediction probabilities for GLM

4.3 Exploratory Analysis

4.3.1 Overall Distributions

This section explores the extent of misinformation in search queries, both on a national scale in the US, as well as on a state level. The initial step of the exploratory analysis is to explore the outcome of the labels created by the glm classifier. A total of 17.288 queries were classified as misinformation, and the remaining 1.73 million observations were classified as regular information. Translated to percentages; ~1% of the search queries were related to misinformation claims (Table 4.4). This can be further investigated by looking into the distributions by month (Figure 4.8 and Figure 4.9). Relatively few queries were related to COVID-19 early on in January (17.338) and February (51.080), then a significant increase in overall queries during March (436.027) and April (384.403), a drop to almost half in May (215.382) and June (213512), a slight spike in July (251.636), and finally a decline in August (182.391) (Appendix A, 4.2).

.pred_class	n	percentage
Yes	17288	0.0098689
No	1734481	0.9901311

Table 4.4 - Distribution after classification

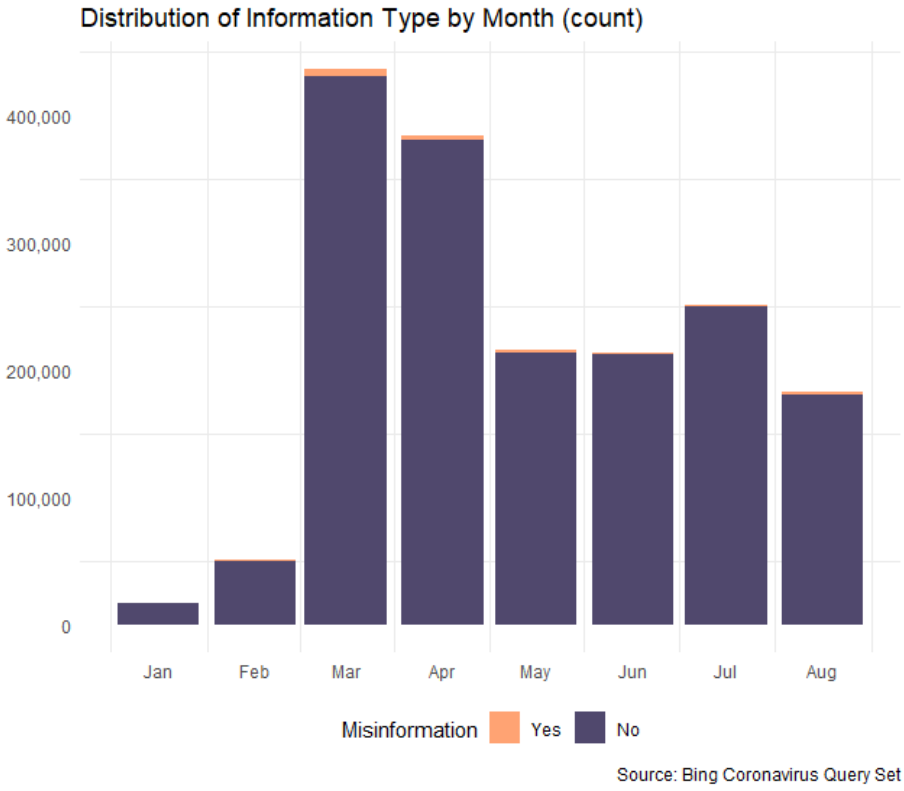


Figure 4.8 – Query frequency by month

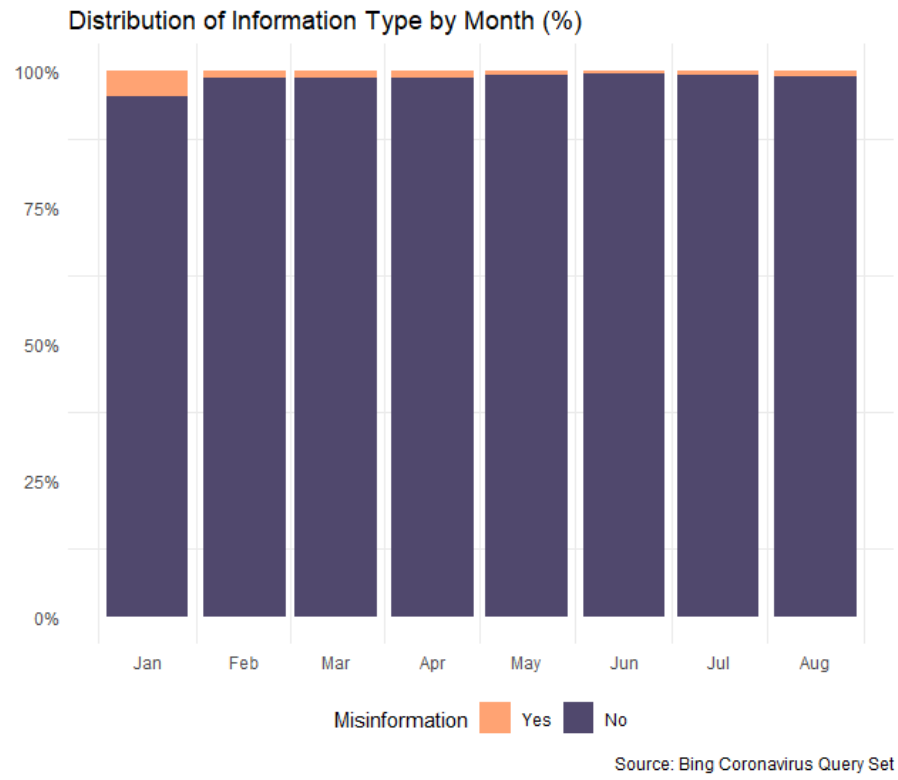


Figure 4.9 - Percent distribution by month

Looking at misinformation, it is undoubtedly higher in terms of overall frequency in the months of March and April (Figure 4.8). However, that is to be expected as the overall count of those months is much higher than the mean frequency (218.000). Instead, misinformation can be visualized as a percentage relative to the total queries of the month (Figure 4.9). In January, 4.5% of all queries were related to misinformation. This drops to 1.3 – 1.1 % of all queries from February to April, gradually dropping further to 0.6% in June, before increasing slightly towards ~1% through July and August (Table 4.5 & Appendix A, 4.2). While the above graphs give a particular indication of the overall distribution, it does so based on the overall frequency/percentage of queries. Alternatively, the number of *distinct* queries can be examined, here visualized in a line graph (Figure 4.10). The lines are coloured by misinformation status. Queries related to regular information follows the same trajectory, suggesting the two are correlated, i.e. if the total number of queries increases, the number of distinct queries increases as well and vice versa. The number of unique search queries related to misinformation is low across all the months, with minor increases in March and April. A second line chart was Made to investigate the unique misinformation queries further.

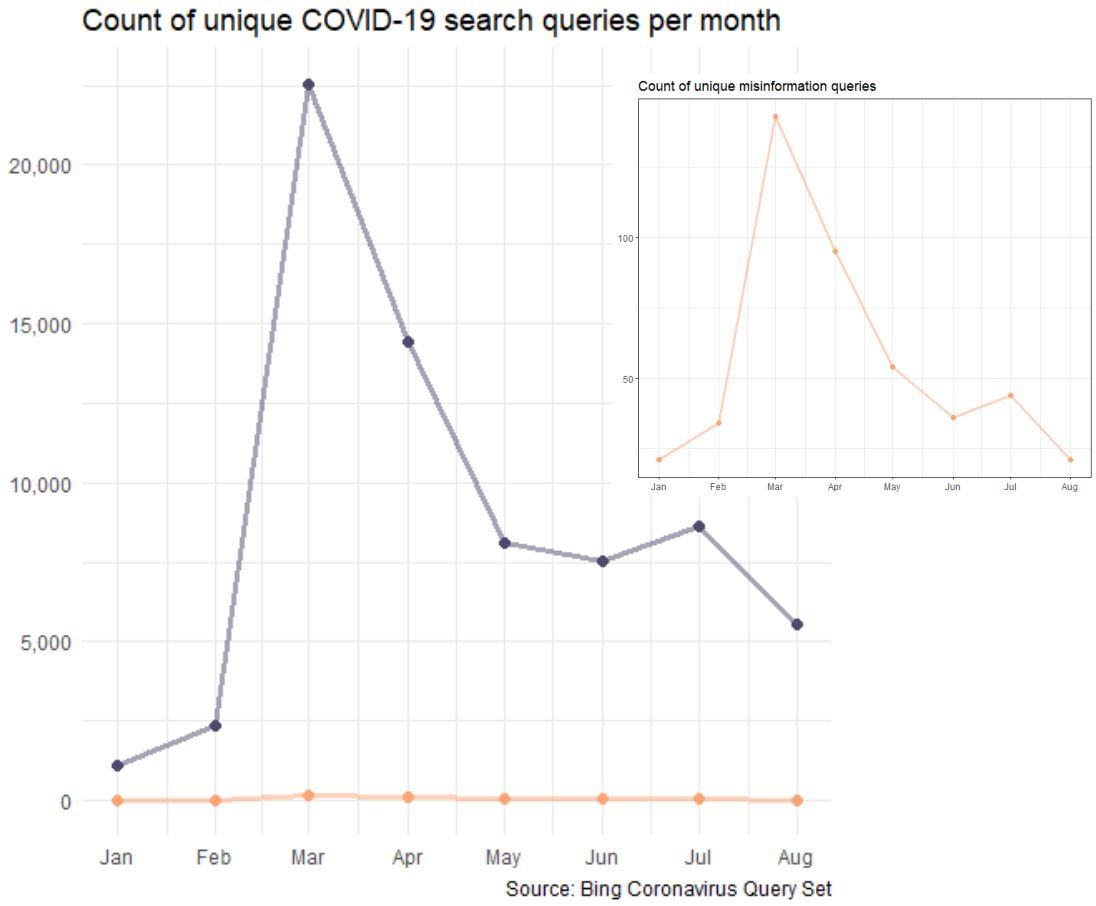


Figure 4.10 - Unique queries by month

This shows that the number of unique queries ranges from around 10 (January and August) to ~150 at the peak in March. Comparing the number of unique regular queries to the overall frequency of regular queries in March, ~22.500 is unique out of the total 430.810 (~5%). Doing the same for misinformation, 150 are unique out of the total of 5217 (~3%) (Appendix A, 4.2). This trend is relevant through all the months; the relative percentage of unique queries is slightly higher for regular queries than misinformation (Table 4.5). One exception is seen in February where misinformation had slightly more unique queries relative to the total (5.5%). This could be expected, as a higher number of overall queries has the potential for more variation, i.e. more unique queries. The individual queries will be explored further in a later section. Finally, the relationship between the two information types can be visualized using a logarithmic scale, making it possible to visually compare their overall development over time (Figure 4.11). This graph was created using the total counts of the two information types and shows them following the same path across the months. One major difference can be spotted in February, where the

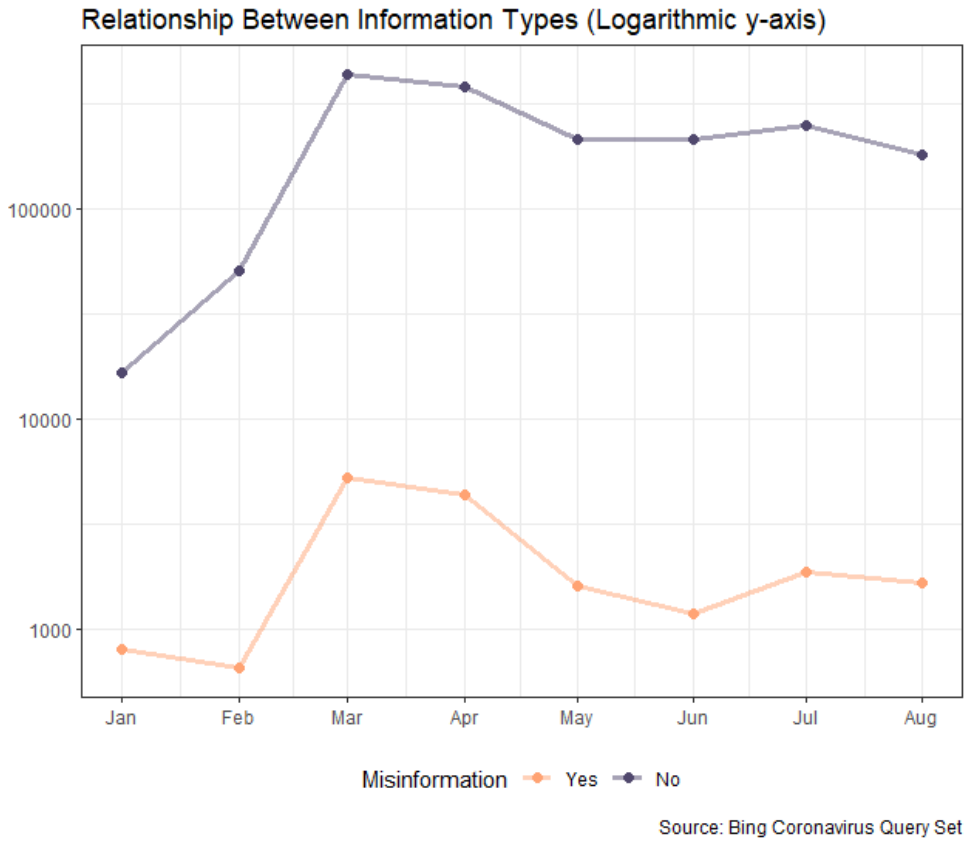


Figure 4.11 - Comparing information types on a logarithmic y-axis

total amount of misinformation dropped slightly, and regular information increased. While the relative differences fluctuate a bit more for misinformation, the overall direction of the line is the same as standard information. The correlation between the two information levels, using Spearman's correlation coefficient, is very strong (.99). Table 4.5 includes the total frequencies, percentage of misinformation as well as numbers for unique queries (total and relative % of all queries).

	Reg. Total	Misinfo. Total	% Misinfo	Reg. Unique	Misinfo. Unique	% Reg. Unique	%. Misinfo Unique
Jan	16.542	796	4.8%	1.090	21	6.6 %	2.6%
Feb	50.426	654	1.3%	2.346	34	4.7%	5.5%
Mar	430.810	5.217	1.2%	22.544	143	5.2%	2.7%
Apr	380.091	4.312	1.1%	14.446	95	3.8%	2.2%
May	213.785	1.597	0.7%	8.129	54	3.8%	3.4%
Jun	212.325	1.187	0.6%	7.522	36	3.5%	3%
Jul	249.758	1.878	0.8%	8.648	44	3.5%	2.4%
Aug	180.744	1.647	0.9%	5.570	21	3.1%	1.3%

Table 4.5 - Numbers by month, including the unique relative percentage.

4.3.2 Search Queries by State

The initial step in exploring the overall distribution of search queries by state was to create a bar chart (Figure 4.12). California is by far the state with most queries with 140.414 total queries, followed by Texas (105.702) and New York (100.018). The states with the least search queries were Wyoming (5.054) and Vermont (4.866). The full distributions can be found in Appendix A, 4.6). The average total number of queries is 34.803 across all states, 34.459 for regular queries, and 344 in misinformation. California has 1.522 queries related to misinformation, Texas has 1.218, and New York has 1.122. At the opposite end of the scale, Vermont has 17 misinformation queries, and Delaware has 28. The total number of queries by state seems correlated to the overall population of the states. This was further

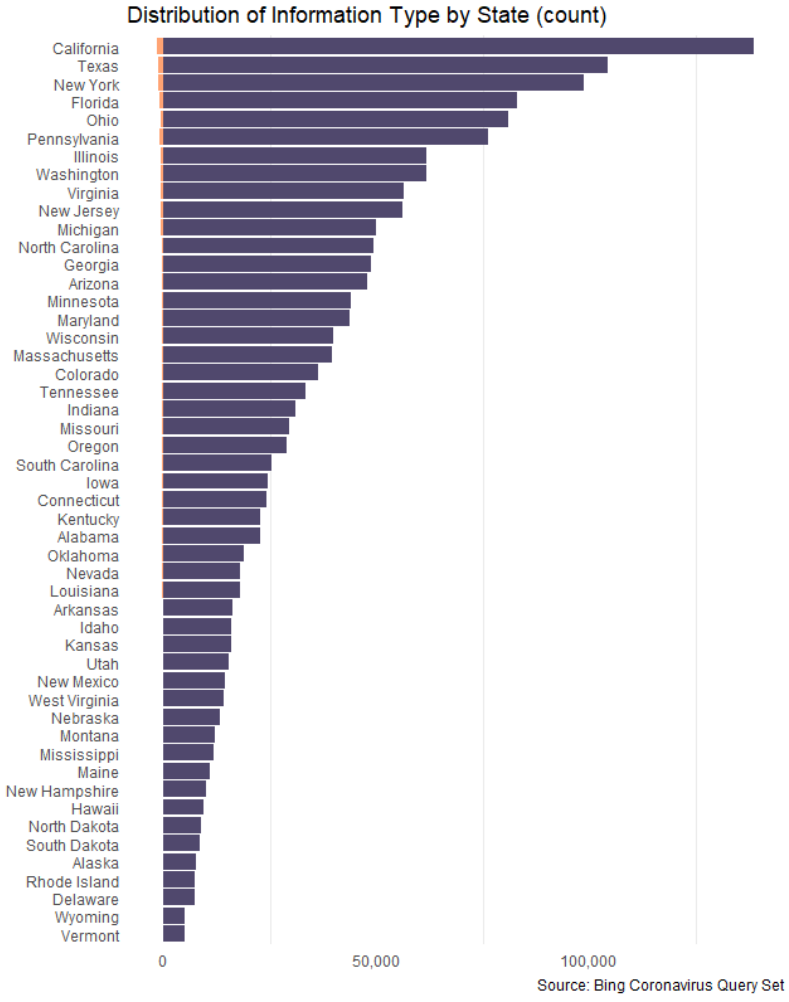


Figure 4.12 - Count of queries by state

investigated by importing data from the United States Census Bureau[48]. The US Census population data was imported using the R package *tidycensus* which allow direct access by using an API key from the US Census. The relationship between state population and the total number of queries is visualized in a scatterplot (Figure 4.13). An additional loess curve was added to visualize the overarching trend across all states. The curve remains linear

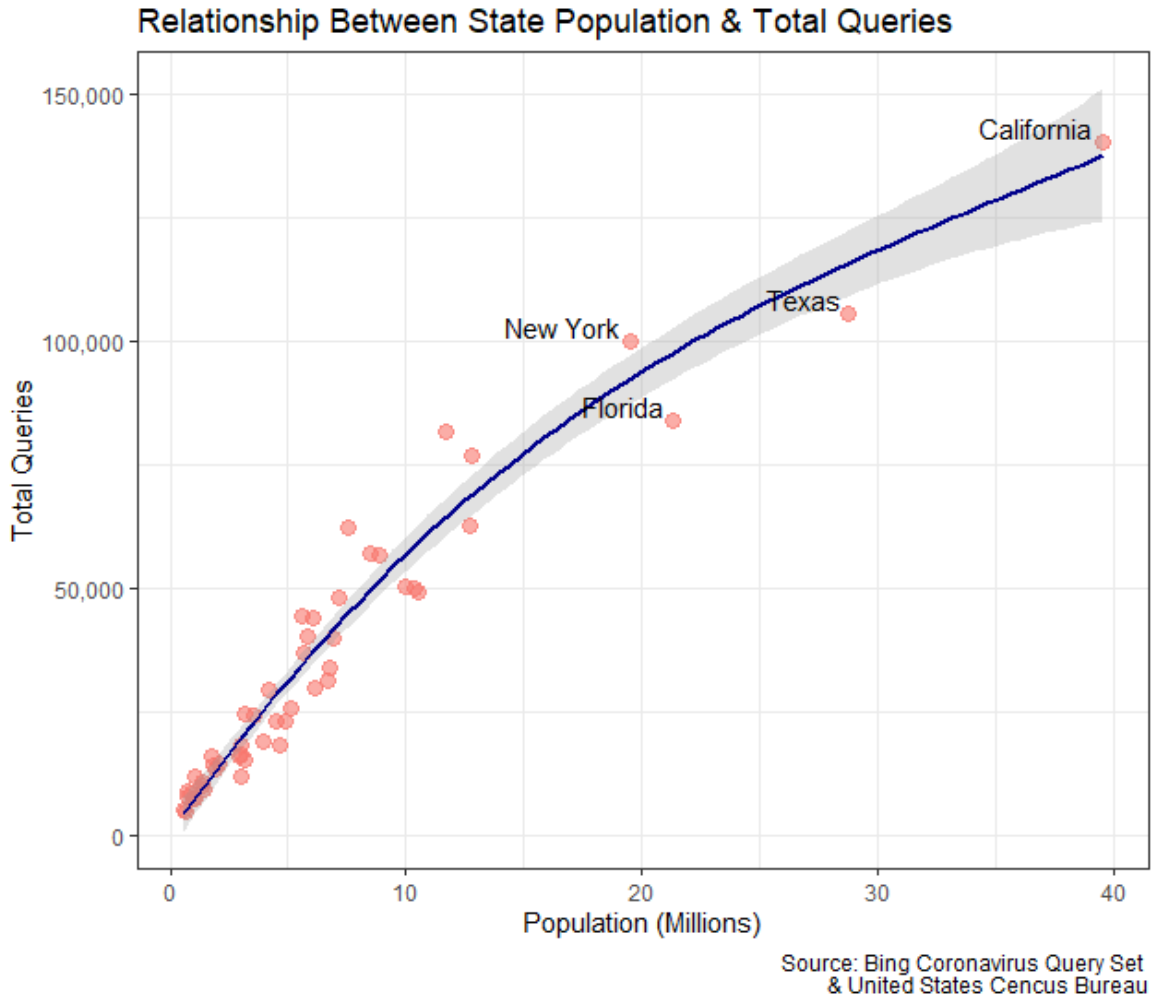


Figure 4.13 - Population vs total queries

up to a population of around 10 million and 55-60.000 total queries. After that, the curve begins to flatten as it moves towards the max population in California (~40 million). As the population grows the relative proportion of total search queries gets smaller. The same trend is evident for both regular- and misinformation queries (Appendix A, 4.6.1). While the correlation between information types was already explored by month in Figure 4.11, a similar approach can be pursued to investigate the relationship on a state level. By visualizing the information in a scatterplot with linear smoothing, it is possible to spot potential state outliers which either has

[48] https://www.census.gov/data.html

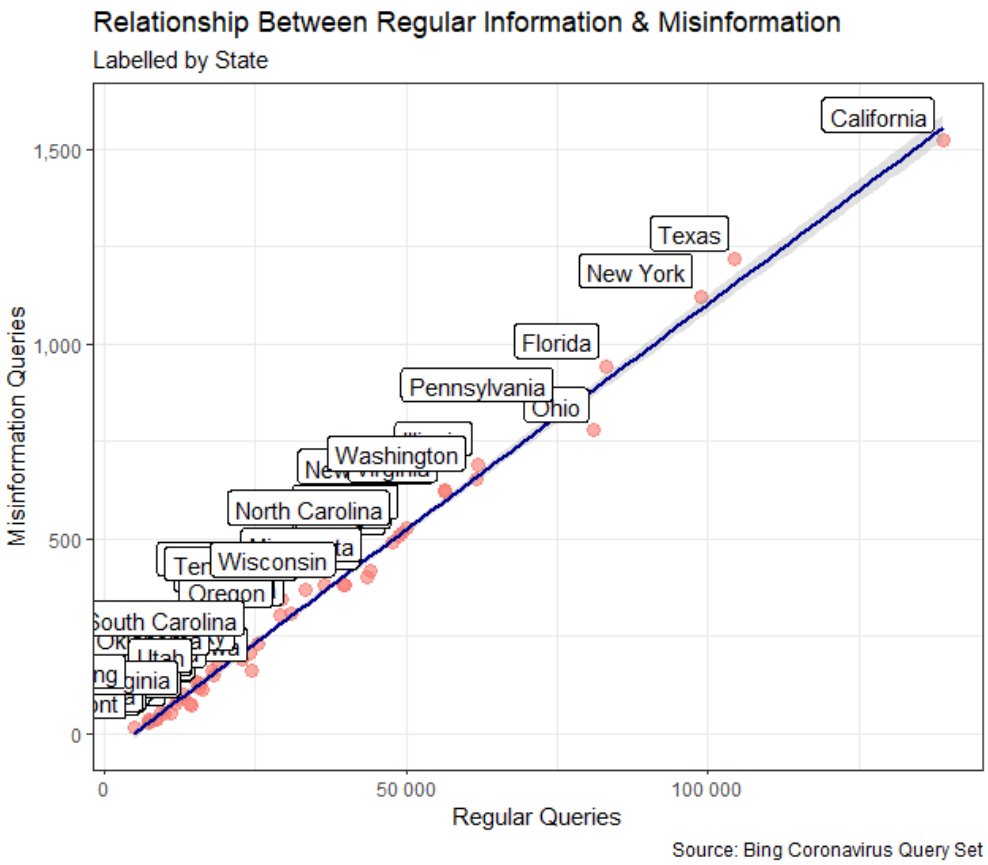


Figure 4.14 - Regular queries vs misinformation queries

proportionally more or less misinformation (Figure 4.14). This shows that the misinformation by state generally increases linearly with the number of regular search queries. There are some outliers such as Ohio; however, the differences are minor and can be challenging to spot in this representation, especially among the states with lower values. Rather than investigating differences in a scatterplot, the data can be plotted in geographic visualizations.

4.3.2.1 Visualizing Misinformation on a Map of The United States

While visualizing geographic data on a map is a good and popular approach to represent data (especially spatial data), there are several important considerations included in the process. Simply plotting the total amount of misinformation by state creates a beautiful plot with apparent differences between the states (Appendix A, 4.6.2). Like earlier plots, it shows that the states with the highest amount of misinformation are California, Texas, New York, and Florida. However, as shown in Figure 4.12 and Figure 4.13, these states are also the ones with the highest population as well as the highest number of total queries. This creates a misinformative map, as it is essentially just visualizing state misinformation counts without considering other factors such as misinformation counts relative to total queries or population. While some information could potentially be extracted from such a map, it is better represented in the previous bar- and scatterplots. As a lot of the states are closely clumped together with relatively low values (Figure 4.14), the colour scale was changed to use misinformation log-odds rather than the common probability fractions. This makes the differences between states more apparent, and more evenly distributed across

the colour scale. Figure 4.15 shows misinformation queries relative to total queries from each state, using misinformation log odds as the fill colour. This representation uses the entire dataset ranging from January to August 2020. The log odds scale ranges from -5.5 to -4.0, which corresponds to a range of 0.3% to 1.9%

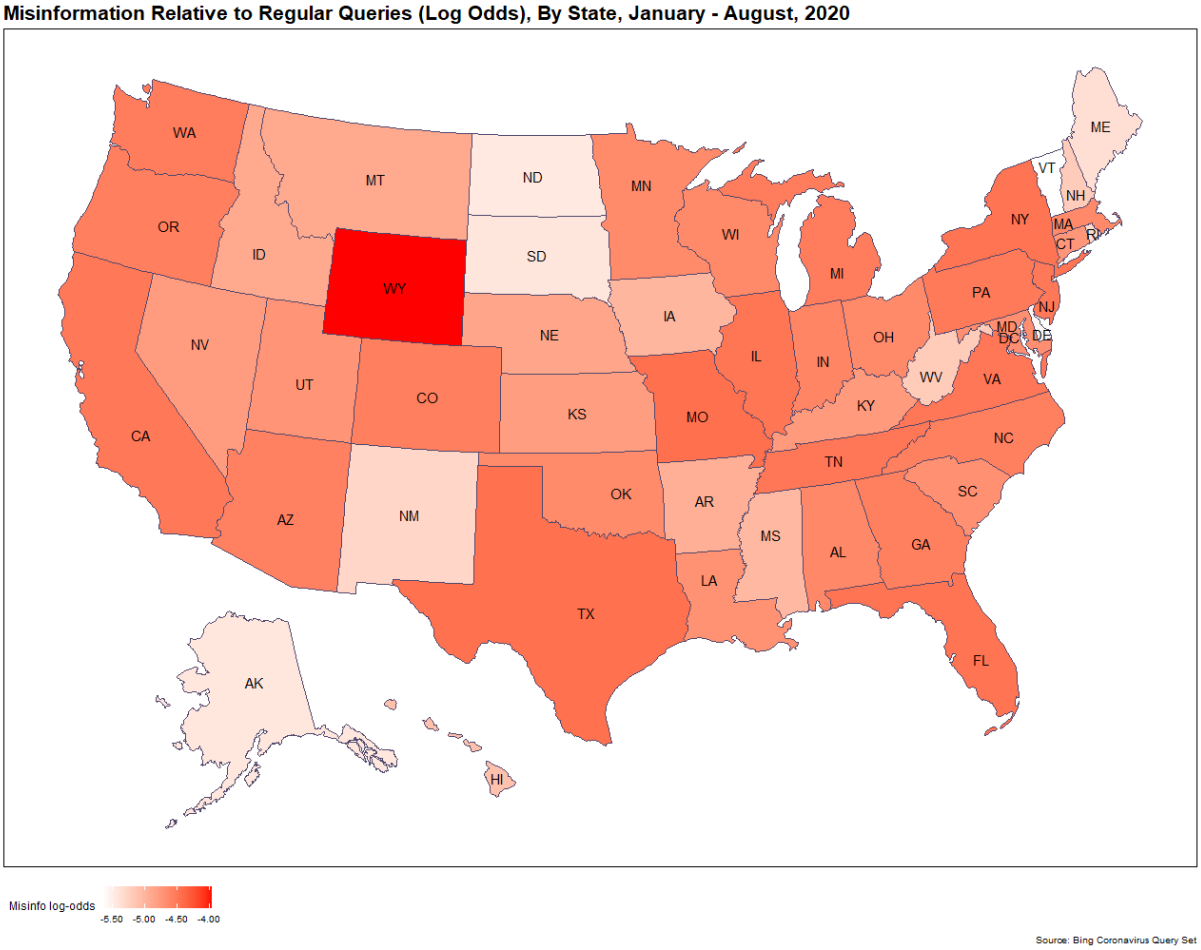


Figure 4.15 - Proportion of misinformation per state (log odds)

proportional misinformation. While differences might seem very large when looking at the map, it is essential to remember the relatively small percentage variation between the states. The state with the least amount of misinformation proportional to total search queries is Vermont in the North East (0.3%). New Hampshire and Maine also had a low proportion of misinformation during the eight months, where the rest of the northeast was slightly above medium level with log odds-scores ranging from -4.6 to -4.4 (~0.5%). The South, going from Virginia in the east to Texas in the west, was generally at a medium level with minor differences. Notably, West Virginia was at a low level of misinformation along with Arkansas and Mississippi in the central south. The Midwest was generally at a low to medium level of misinformation. North- and South Dakota was at low levels (~-5.45 / 0.4%), and increased levels can be noticed in states located more to the east. Missouri, Illinois, and Michigan had the highest levels of misinformation in the Midwest at slightly above medium levels (~-4.5 / 1.1%). In the West, the states down the coastline ranging from Washington to California and inland to Arizona were all slightly above medium levels. The states in the centre and north of the West (Nevada, Utah, Idaho, and Massachusetts) were all below the average (~0.7%). The two significant outliers in the West is New Mexico and Wyoming. New Mexico

has the least amount of misinformation at 0.5%, where Wyoming has the highest proportion of misinformation in the entire United States, indicated by the deep red colour (1.9%). Alaska and Hawaii both had low levels of misinformation (0.4% and 0.6%). While this approach provides an excellent overall picture of the proportion of misinformation over time, it does not show any changes that occurred across the months. To visualize these changes; an additional plot was made, including misinformation per month (Figure 4.16). This plot displays misinformation proportional to total queries for the given month. This means that it does not show overall changes (relative to total queries for all months), but rather how the information search behaviour has been for the individual months. A state coloured in grey means that there were no misinformation queries from that state in that month.

The North East started with a relatively high proportion of misinformation with only Vermont having low proportional level. The relative levels remained high across the following months, with the most significant changes observed in Maine. In May the overall level dropped while Vermont, New Hampshire, and Vermont had no misinformation queries. The level gradually dropped from May to July, and in August the misinformation levels were at the lowest. Interestingly, in August, the states change behaviour with Ohio, Pennsylvania, and New York being very low on the scale, and Vermont, New Hampshire and Maine moving to medium levels. In the South, the overall relative levels of misinformation were medium to high in the first four months. Mississippi starts out with very low values in January, spiking to a high level in February, before moving down to medium levels. A general trend can be observed of a decreasing relative misinformation level, especially noticeable in Texas which has high proportions in the early months of COVID-19 and decreasing to become one of the states with the lowest level in August. The same trend can be observed in Georgia and Florida. The Midwest generally had higher levels of proportional misinformation in the states to the east, and lower levels in the west. Like in the South, this behaviour changes in August where the eastern states have very low levels and the west have higher levels. North Dakota had no misinformation queries in January, and South Dakota had none in May. Illinois changes from having the highest level in January to March, dropping to medium level in the following months and having the lowest level in August. The West follows a similar trend to the other areas with states along the coast having high levels in the early months and dropping to low levels in the last months. The states further inland have low to medium levels in most months, before having higher levels of misinformation than the coastal states in August. The states with the largest cities, California and Arizona, have high levels of misinformation in the early months and lowest levels in August. The most significant outlier in the overall plot (Figure 4.15) is Wyoming, is also the state with the largest relative changes by month. In January, Wyoming had no misinformation queries, moving to medium levels in February, back down to very low levels in March, before increasing to very high levels from April to August. In August, Wyoming has a much higher relative level than the rest of the country. Alaska and Hawaii have their highest level

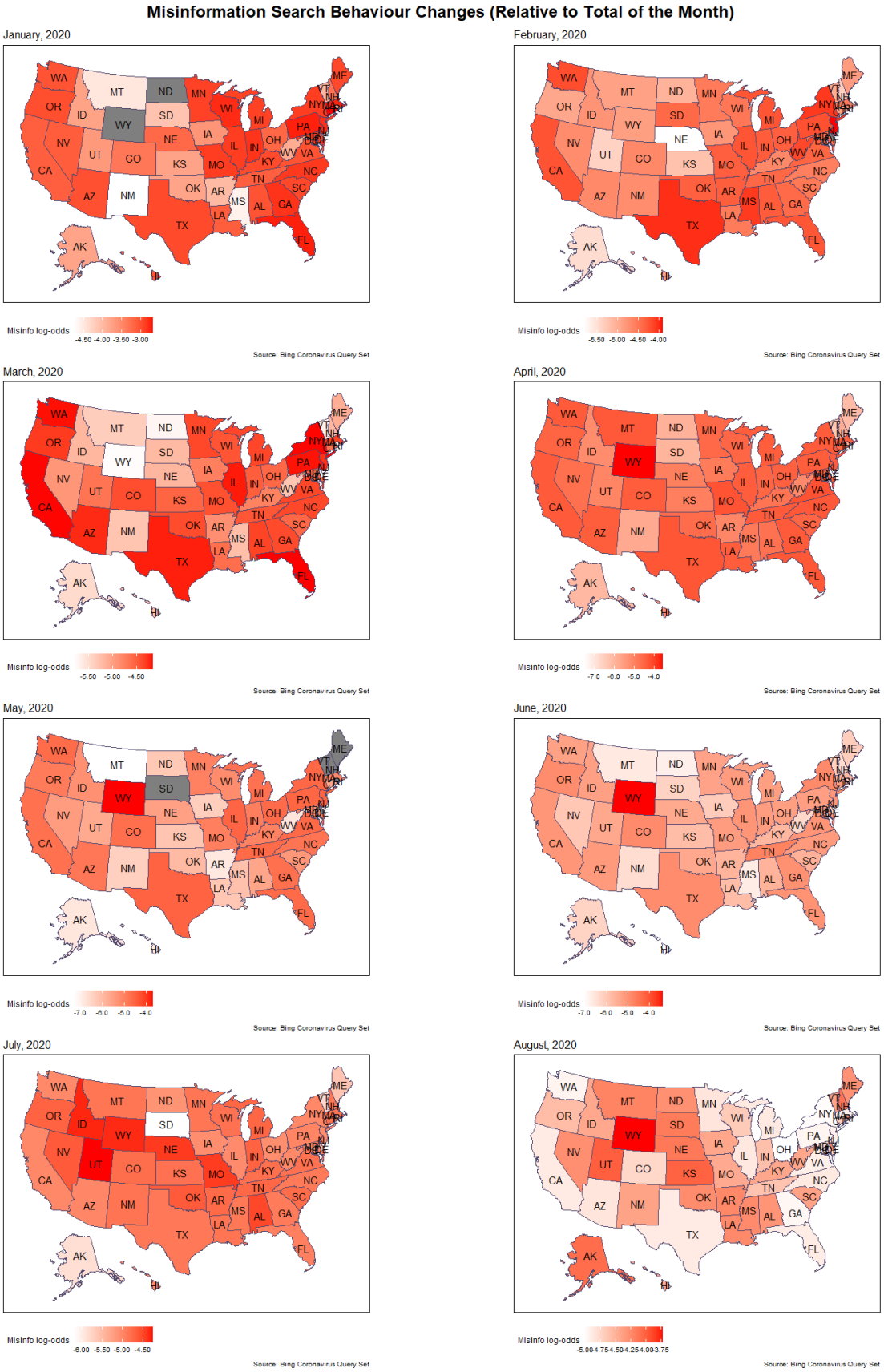


Figure 4.16 - Proportion of misinformation per state. Facetted by month

of misinformation in August and have otherwise remained at low levels. An exciting trend in the monthly misinformation distributions is that states with the largest cities in the United States (New York NY, Los Angeles CA, Chicago IL, Houston TX, Phoenix AZ) has relatively high levels from January to April, drops to medium levels from May to July, and has misinformation levels among the lowest in August. This could indicate that people in larger cities have gained knowledge during the COVID-19 process and have more experience now than in the beginning. However, as all the plots in Figure 4.16, are relative to the given month this could also just be an indication of some smaller states, like Wyoming or Kansas, having significantly higher proportion in August. In order to further explore this trend, additional plots were made with a fixed scale across all months. This was not possible with the log-odds scale, as some states had months with missing values. Instead, a fixed scale was created based on misinformation proportion (misinformation count / total count). Figure 4.17 shows changes from January to August, and it clearly shows an overall lower percentage of misinformation across all the United States. Furthermore, it confirms the previous trend of states with more populated cities having high initial misinformation percentage, which changes to a very low rate in August. The central states are generally at around medium levels, which

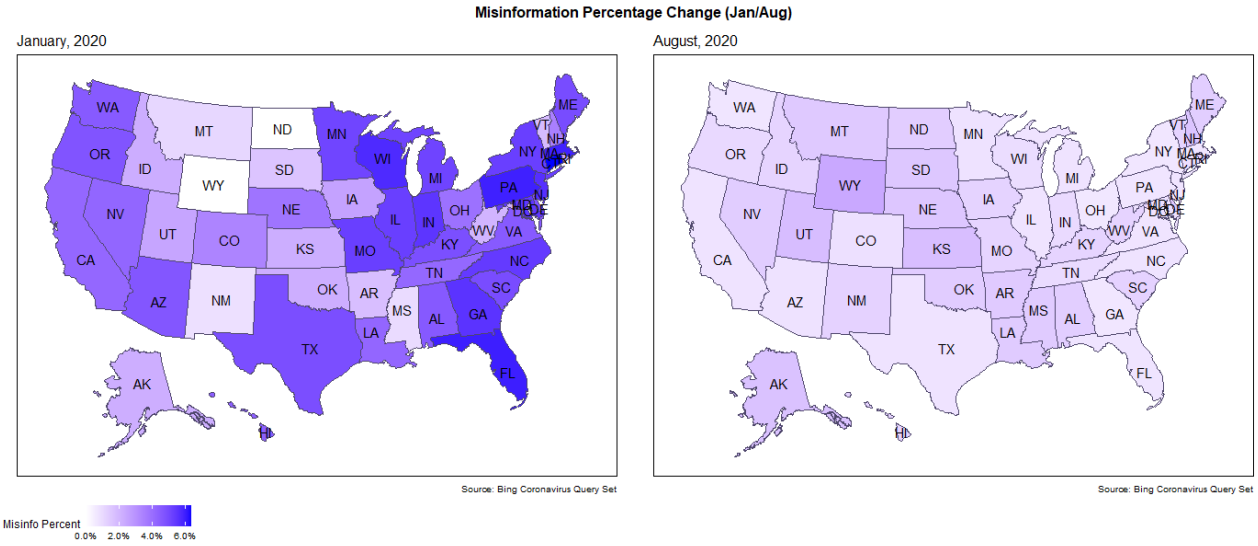


Figure 4.17 - January vs August. Percent misinformation

ends up being the highest in August. Finally, the highest level of misinformation is in the eastern part of the US in January. However, this representation comes with some limitations. Namely the fact that the overall misinformation percentage was much higher in January (4.8%) compared to ranges between 1.3% in February to 0.9% in August. This means that the colour scale is heavily impacted by the high percentages in January. Dropping January from the plotting scale provides a better understanding of the overall changes in the subsequent months.

This was done in Figure 4.18, showing the broad differences between February, May, and August[49]. In February, a general trend of lower misinformation in central states can be observed. This is also apparent in May, where the

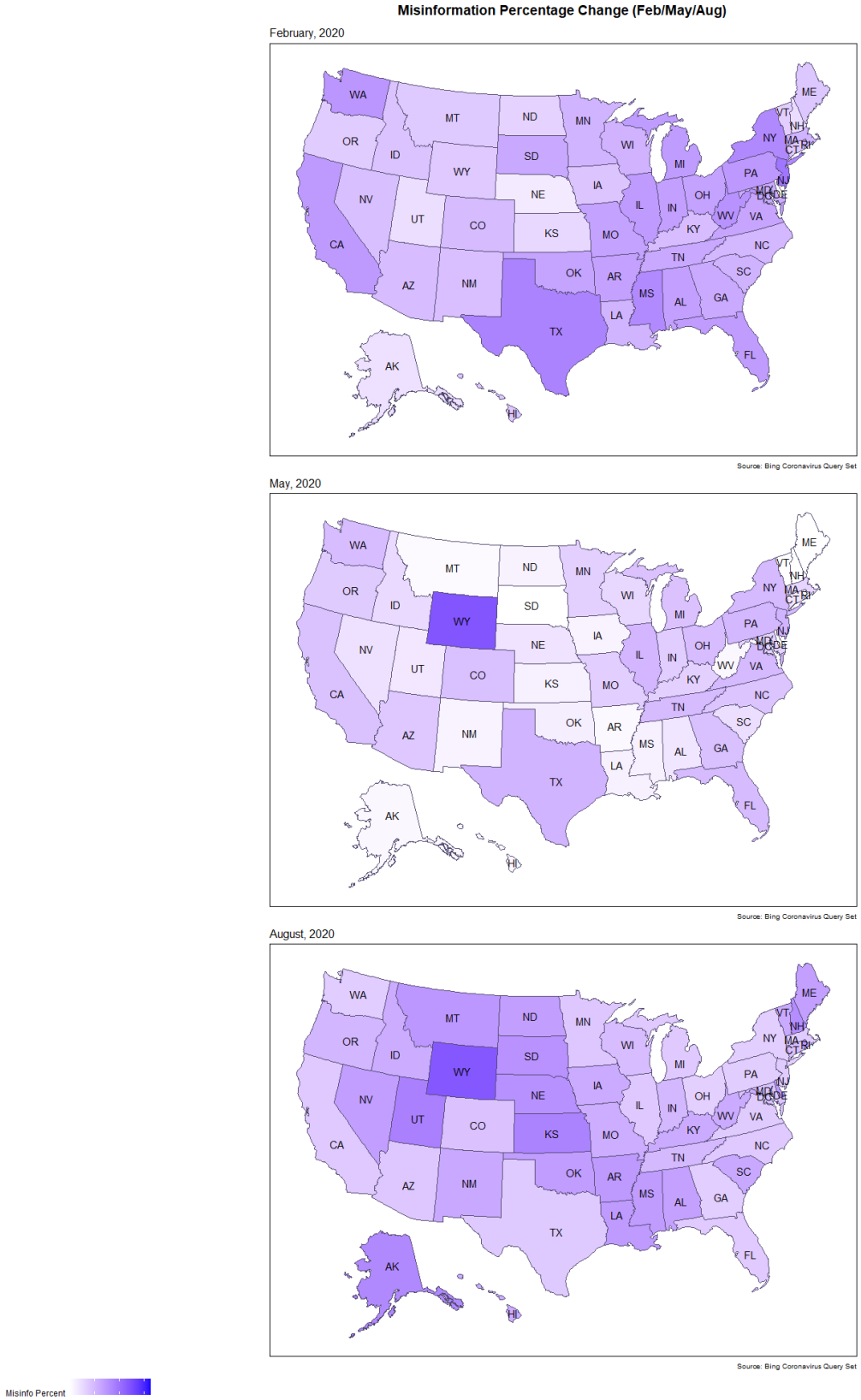


Figure 4.18 - February, May, and August. Misinformation percent

overall misinformation level is lower, except Wyoming, which is now at the highest relative levels across all states and months. Finally, in August, the central states have the highest levels of misinformation, and Alaska and Hawaii have much more relative misinformation than previously. The trend of misinformation being centred around the largest cities in the early months and moving to a low level in August can also be confirmed. Finally, the overall changes across the different regions were summarized in a line plot (Figure 4.19). This confirms previous observations, presented in an alternate way. Here it is clearly seen how the North East has the highest percentage of misinformation in January and ending up with the lowest rate in August. The opposite is true for the West, including Wyoming and adjacent states, which changed a lot during the eight months. It also clearly shows the early spike in January, with more considerable percentual differences, dropping significantly in February and settling on smaller relative differences between the regions for the rest of the months.

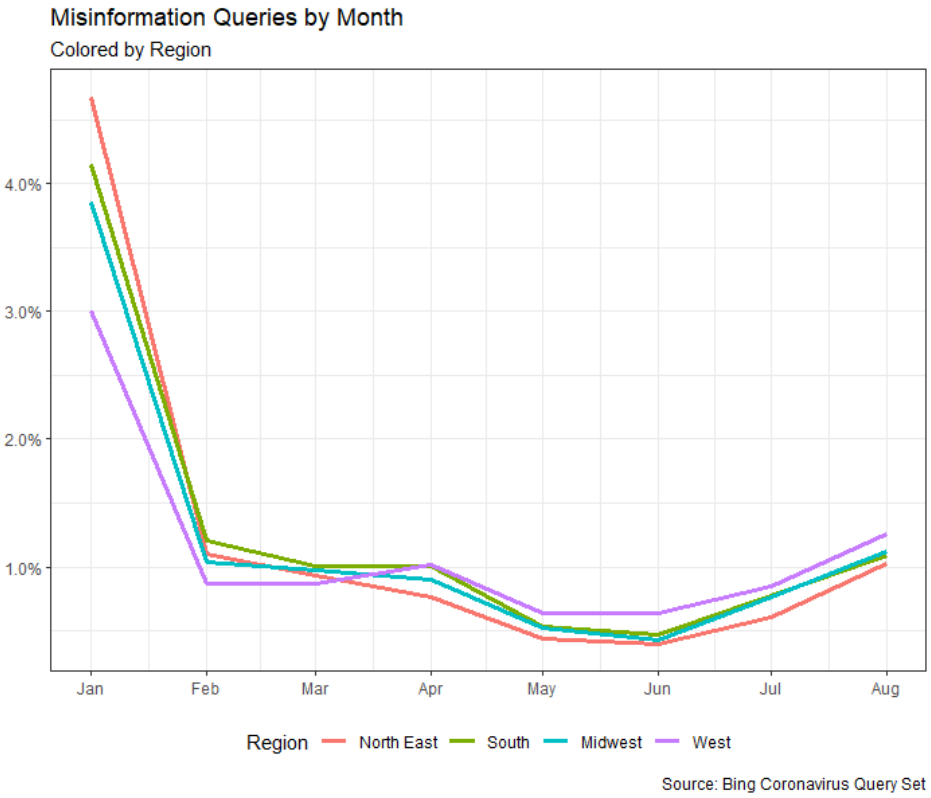


Figure 4.19 - Misinformation by region

4.3.3 Top Search Queries

The previous sections have explored overall distributions and differences between the different states and regions. Search queries related to misinformation is observed to start high, gradually decreasing before seeing an increase

[49] The scale was calculated based on all months between January and August.

again in July and August. This section will explore the defining search queries of the eight months, both as a whole and on an individual month level.

Figure 4.20 shows the top 10 queries from January to August, based on frequency. Regular queries are all general in nature, and includes "*coronavirus*", "*coronavirus update*", "*coronavirus map*", and "*covid 19 update*". They either express interest in the coronavirus as a whole or status reports in the form of map representations, statistics or dashboards. All of them are correctly classified as regular information, as none of them relates to any current misinformation claims. As for misinformation, "*QAnon"* is by far the leading term followed by "*herd immunity"*, "*hydroxychloroquine"*, *"Bill Gates coronavirus"*, and "*malaria drug for coronavirus*". All the misinformation queries relate directly to misinformation claims identified in the earlier work (section 3.4.2), except for "*carona virus*". The latter is obviously a misspelling, suggesting that coronavirus misspelled is classified as being misinformation, which is unintended. This problem will be discussed in a later section.

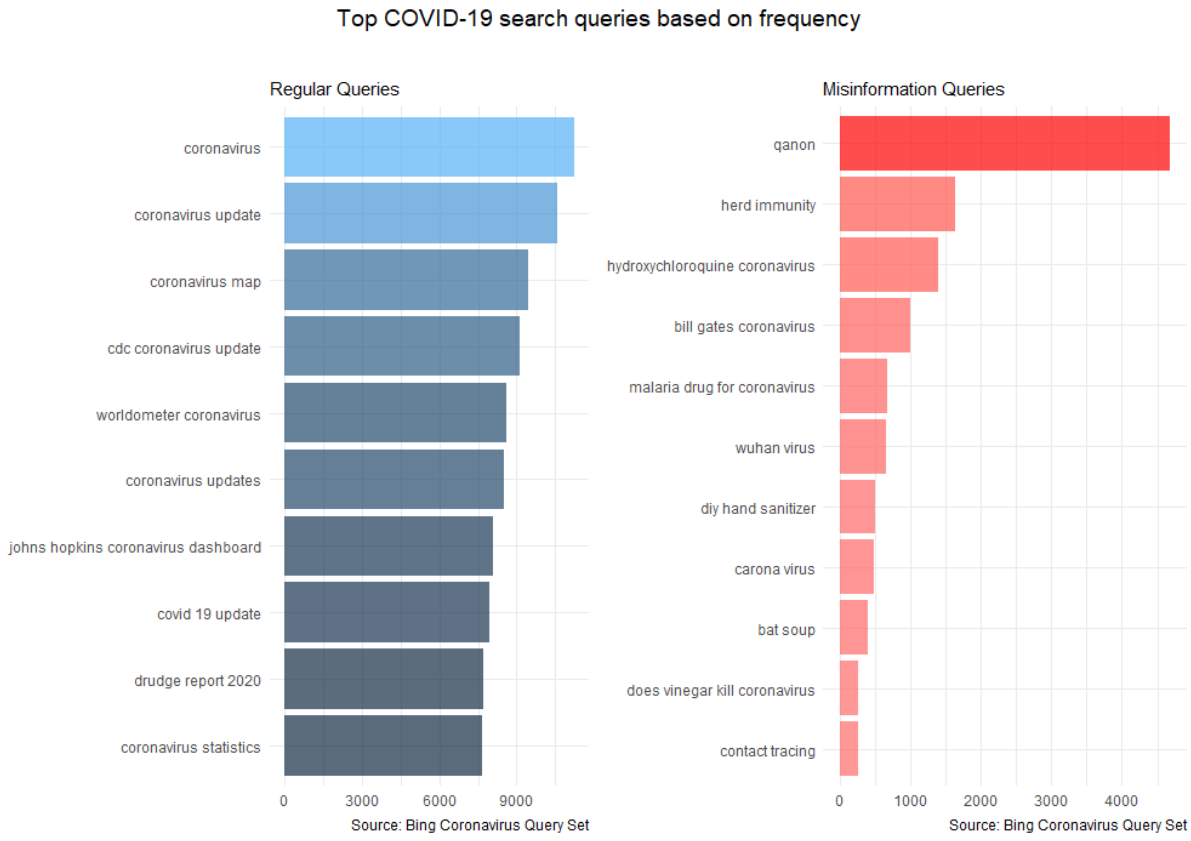


Figure 4.20 - Top regular/misinformation queries

To investigate how the top queries of all eight months have changed during the outbreak; additional line plots were made (**Error! Reference source not found.** and Figure 4.22). This includes the top five queries for each of the i nformation types and can be used to determine whether the top queries of the entire period have been relevant

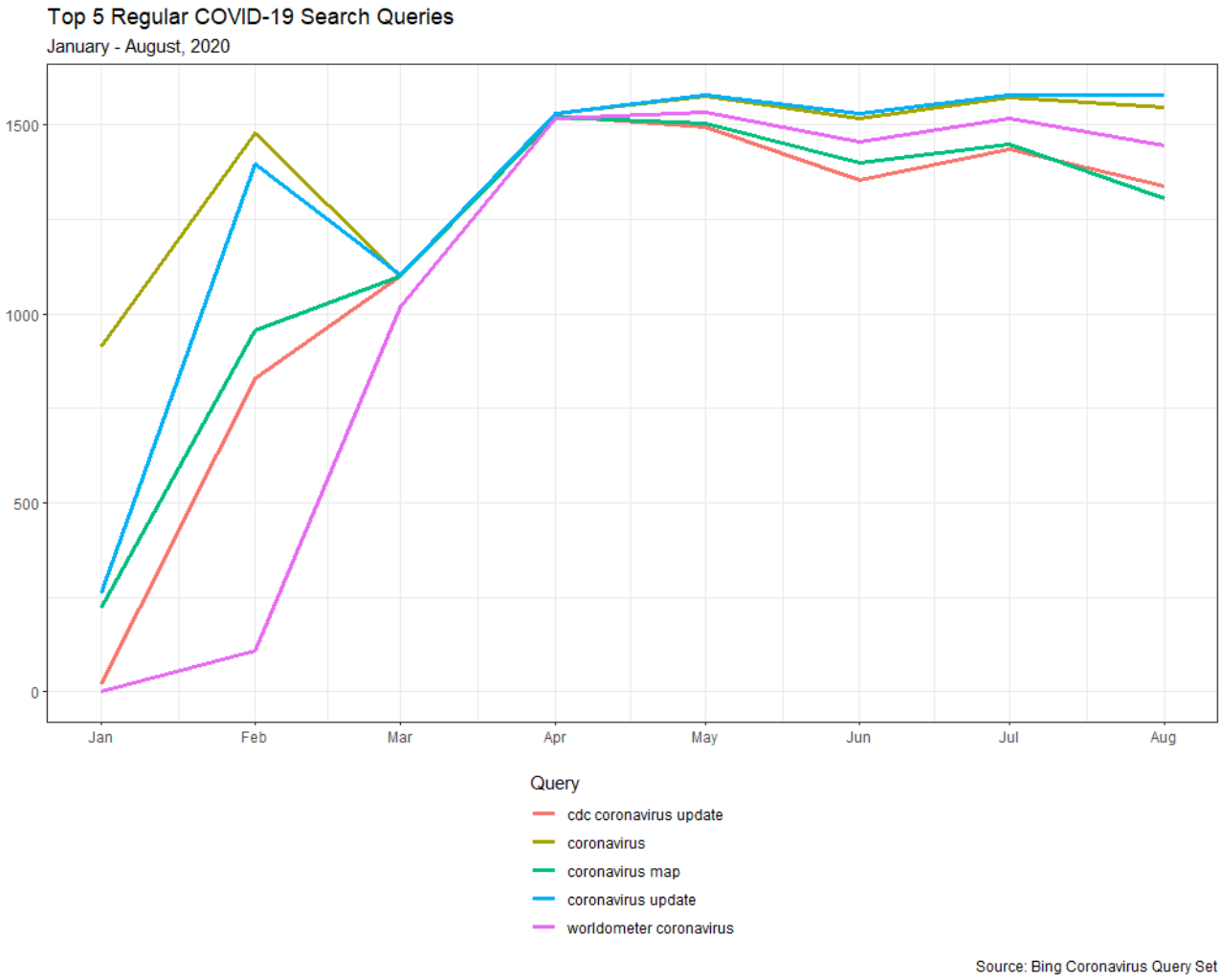


Figure 4.21 - Top 5 regular queries. Changes over time

across the individual months or if their high placement is caused by extreme spikes at specific times. The plots are based on total counts and not relative values. Furthermore, it is based on unique search queries and not search topics, which would include all queries related to a relevant word like, for example, hydroxychloroquine. The top 5 regular queries all follow a similar trajectory across the months with exceptions in January and February. Surprisingly, from March and forward, the lowest counts are found in March. This is the month with the overall highest number of queries, so it could be expected that the queries would have had high counts. A possible explanation could be more query variation within the monthly queries. From April to August the queries were all stable at high levels, the most popular being "*coronavirus*" and "*coronavirus update*". All queries, except "*coronavirus update*", are decreasing towards August, which could suggest a lower overall recent interest.

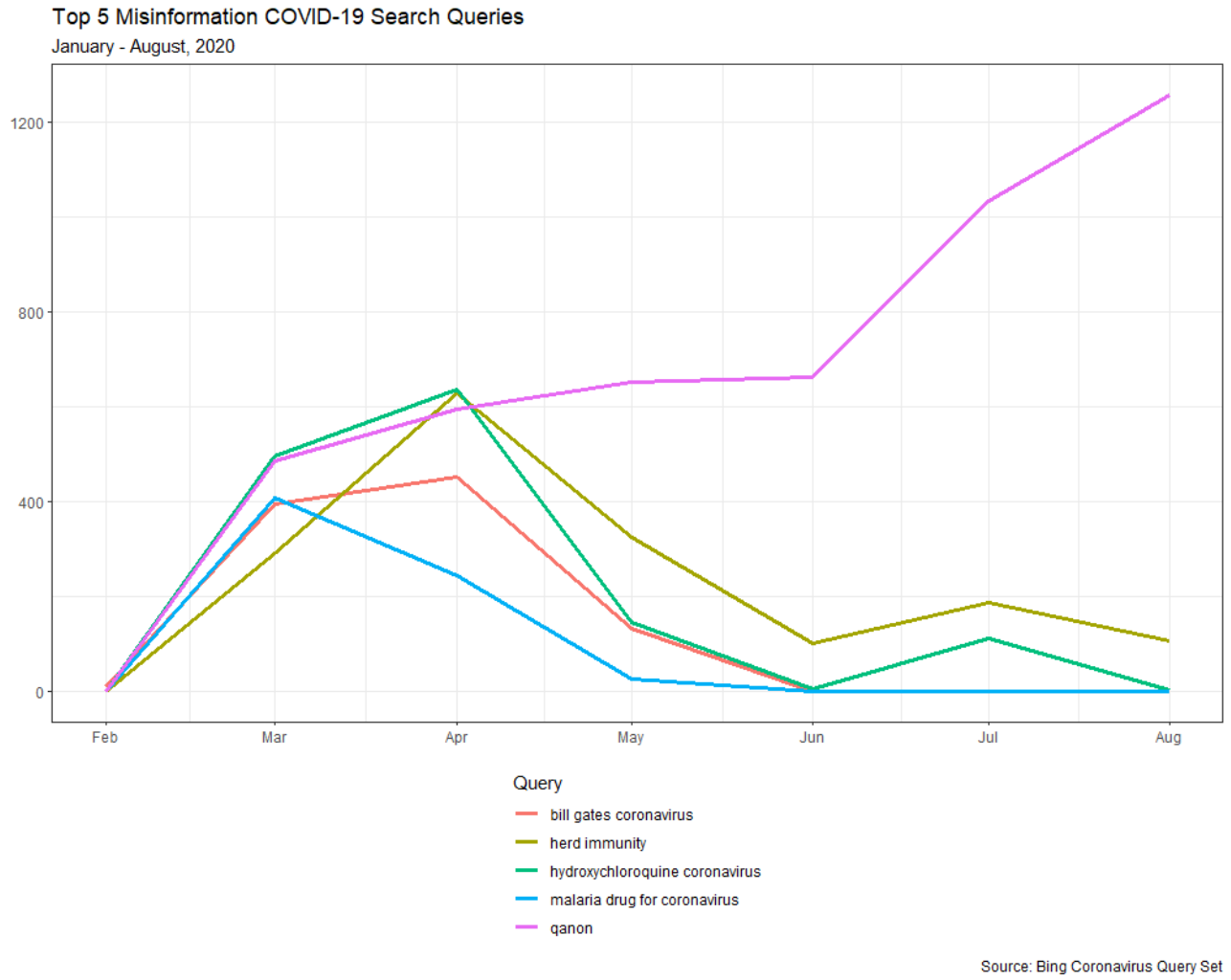


Figure 4.22 - Top 5 misinformation queries. Changes over time

As for search queries related to misinformation, "*bill gates coronavirus*", "*herd immunity*", "*hydroxychloroquine coronavirus*", and "*malaria drugs for coronavirus*" all saw major increases in March and April. The overall trajectory of these four queries corresponds to the changes in general information/misinformation (Figure 4.11). They have an early spike, drop heavily in May and fluctuates on lower levels for the remaining months. "*Bill gates coronavirus*" and "*malaria drug for coronavirus*" are not represented in June to August and "*hydroxychloroquine coronavirus"* has no hits in June. The significant outlier is "*qanon*" which has the same initial spike in March and April, slowly increasing until June before seeing a significant increase through July and August. Exploring the top 5 queries by state also shows Qanon as the clearly most dominant misinformation query (Figure 4.23). It has the smallest counts in the states with more total queries and generally increases in states with lower total queries. Across all states, it remains at high levels, the highest being in Wyoming (~90%) and the lowest in California (~30%). Queries related to herd immunity and Bill Gates are generally well represented in states with more total queries and gets gradually smaller as the total query counts get smaller. Overall, the queries are more balanced in states with more total counts, suggesting more query variation as the population/counts increases.

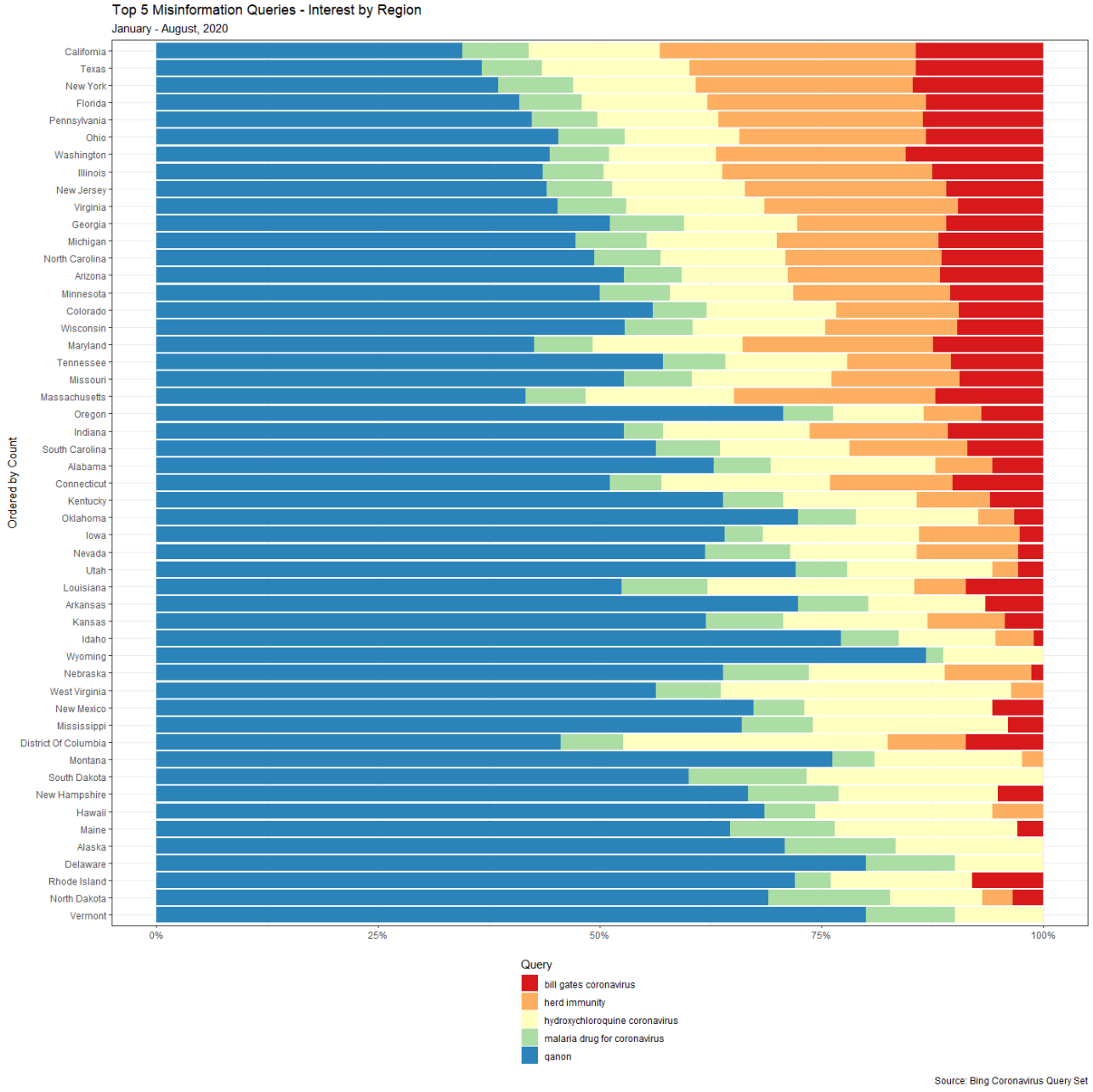


Figure 4.23 - Top 5 misinformation popularity by state

The above graphs were all based on overall search query frequency, which might not accurately represent the broad corpus of search queries. They are useful to identify the most used unique queries across the entire dataset; however, additional variables can be helpful in this context. The Bing dataset includes information on query popularity score, i.e. how popular a query was at a given time and place, as well as whether its intent was implicit or explicit. As well as providing insight into the overall changes across months, it is useful in evaluating the results of the classification process as it might reveal queries that could be considered wrongly classified as misinformation. Figure 4.24 shows the top queries per month based on the popularity score coloured by implicitness of the query. The purple colour is explicitly related to COVID-19, and all include coronavirus or COVID in the query. The

orange colour expresses implicit intent and includes queries like "*dr. Charles Lieber*[50]", "*hydroxytoluene*", and "*qanon*". Generally, all months have low popularity scores (range 0-100), suggesting that misinformation queries do not reach widespread popularity when compared to other queries. January, which has high overall misinformation (Figure 4.9), is mostly made up of various misspellings of coronavirus such as "carona virus" or "corono virus". Due to the novelty of the virus, it makes sense that misspellings would occur here. The most popular queries were "Wuhan virus" and "bat soup" which both relates to the origin of the virus. In February, the most popular queries were concerned with drug treatments administered in Asia. This relates to the myth that various types of HIV, flu, and malaria drugs can cure the coronavirus. In the time of writing, no drugs have been proven to cure COVID-19. This trend continues in March, April, and May, which also introduces new queries related to various measures

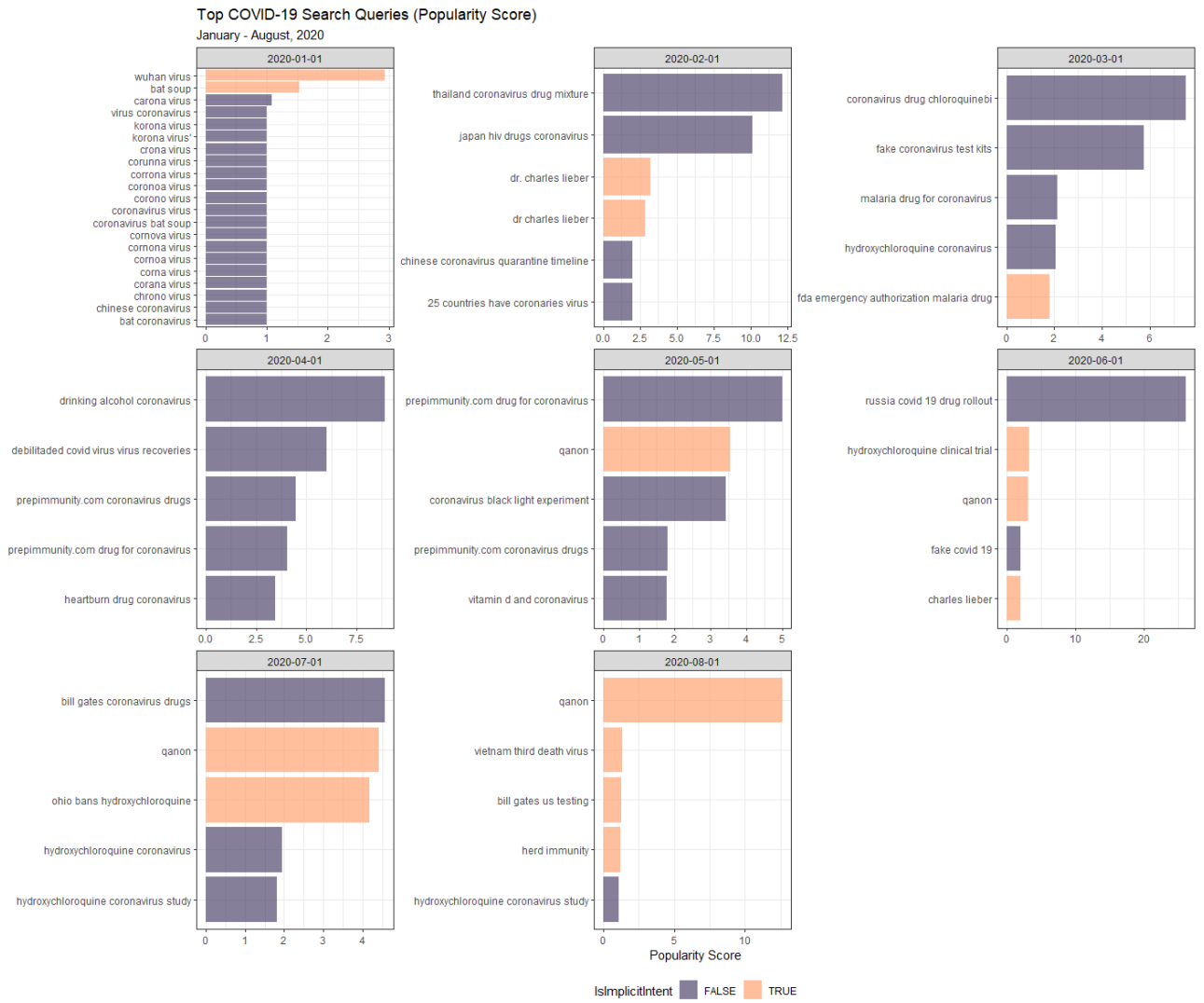


Figure 4.24 - Top queries per month by popularity score

[50] Charles Lieber, a Harvard professor, was the center of a myth claiming he was responsible for creating the coronavirus in collaboration with China which led to his arrest, which turned out not to be related to COVID-19.

http://mythdetector.ge/en/myth/disinformation-fake-bbc-cnn-abc-about-harvard-professor-allegedly-arrested-creating-covid-19

of self-medication or cures, including hydroxychloroquine, alcohol, vitamins and UV light. From May, various conspiracy theory related queries start to increase in popularity, including QAnon, Charles Lieber (also represented in February) and Bill Gates related queries. Generally, a trend can be spotted of queries in the earlier months being mostly related to various rumours surrounding origin and cure of COVID-19. These are also represented in the later months, but here more alternative conspiracy theories are also gaining popularity.

In order to provide further insight into the defining search terms for each month, rather than merely relying on the frequency or overall popularity scores by Bing, the queries can be investigated using other metrics such as TF-IDF values or weighted log odds with the R package *tidylo*[51] (Silge, 2019). Essentially these approaches lower the weight of highly used words and increase the weight of less used words (Silge & Robinson, 2017). This can reveal queries not represented in the previous frequency-based plots. A plot was made using weighted log-odds (Figure 4.25), which can work better than TF-IDF scores in some contexts (Schnoebelen, 2019. Schnoebelen & Silge.,

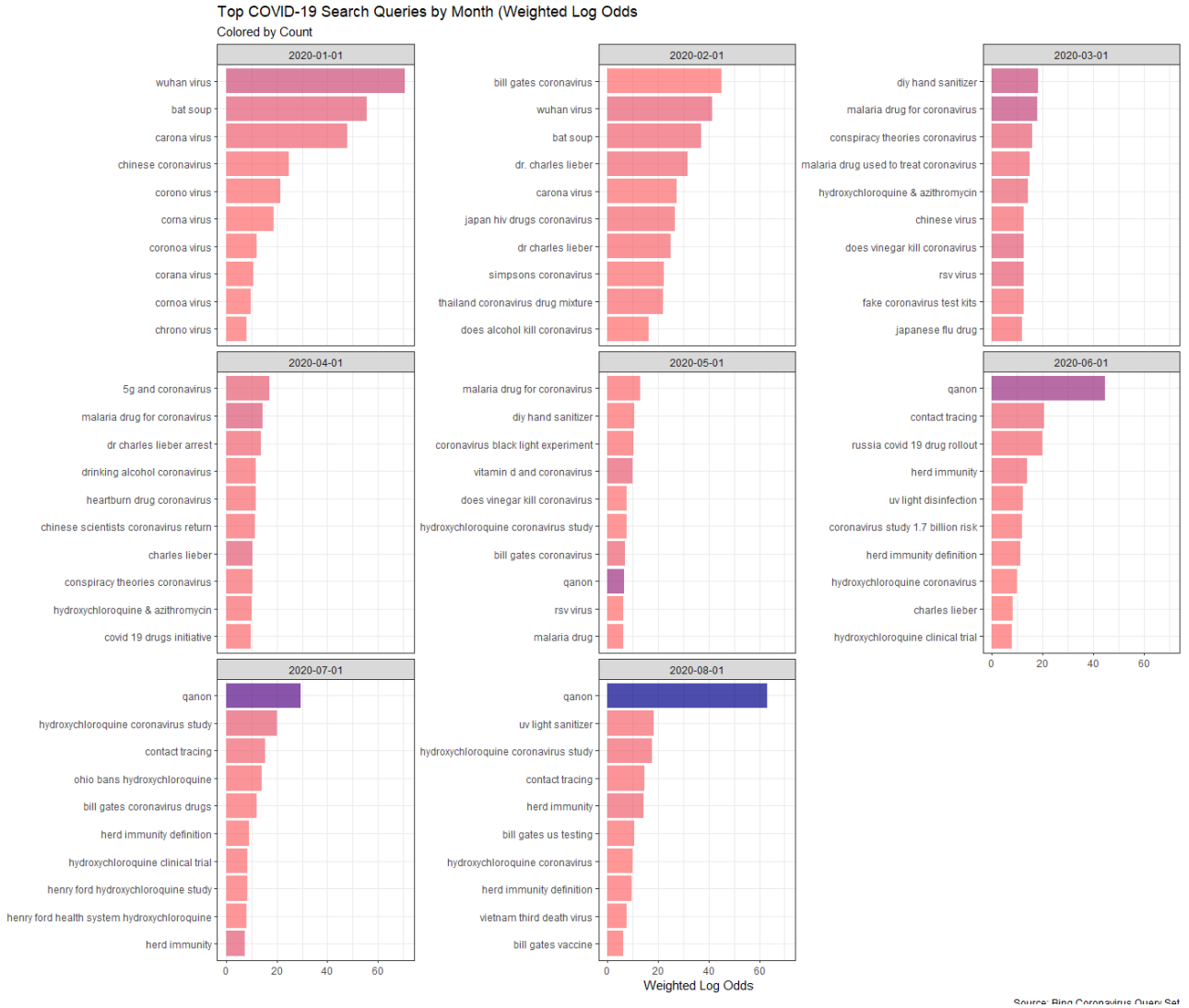


Figure 4.25 - Top queries per month by weighted log odds (colored by count)

[51] https://cran.r-project.org/web/packages/tidylo/index.html

2020). An additional plot was made for regular queries which can be seen in Appendix A, 4.7.5. To illustrate how weighted log-odds provides different results than frequencies; the bars were coloured by count. In January, the overall picture is the same, and most of the queries are misspellings of coronavirus except for the top queries "*Wuhan virus*" and "*bat soup*". These two carries over into February which was not shown in the previous Figure 4.24 and the following months introduces other conspiracy related queries like "*bill gates coronavirus*", "*5g and coronavirus*" and *"Simpsons coronavirus*" related to a myth of Simpsons predicting COVID-19 in 1993 (Lee, Forbes, 2020)[52]. The graph introduces several new queries that were not present in the frequency-based plots, but overall, the themes are similar. Most queries are related to myths about the origin and cure of the coronavirus, some are related to popular conspiracy theories, and few queries could be interpreted as assigning blame or stigma towards Asian groups ("Wuhan virus", "Chinese virus", "Chinese scientists coronavirus return"). The trend previously observed trend is less pronounced here but is still present. Bat soup, one of the early myths about COVID-19 origin, is exclusively searched for in the first months[53]. March and April had far more queries than any other months, and those are dominated by queries relating to myths of alternative virus cures and prevention (alcohol, vinegar, malaria drugs, DIY hand sanitizer). UV light started appearing as a top search term in May, following the claims made by Donald Trump in late April. From June to August, *QAnon* starts to appear as a top query. It was also featured in May with a higher count, but lower weighted log-odds score. In the last three months, it scores very high on both scales, making it a very defining query in recent times, culminating in August where it is far ahead of the other queries. Herd immunity starts to appear as a top term from June, which could suggest an increased interest in knowing if, and when, COVID-19 will end.

One of the primary findings of the above sections is that queries are generally gathered in clusters relating to similar topics. These topics change slightly over time, suggesting an overall change in search intent. These topics can be further explored in a network plot, here based on bigrams extracted from misinformation queries (Figure 4.26). The alpha of the lines between points was set according to total counts – a higher counts draw a more solid line. The network only includes query combinations of a minimum of two words, meaning single term queries like QAnon, were not included. Some significant clusters can be immediately observed when inspecting the network plot. The largest groups surround the term *coronavirus*, and most of the other clusters are connected to it through different numbers of links. At the bottom, a collection surrounding the term virus can be identified. Several of the terms here are various misspellings of *corona* which was very frequent in January. The bottom left all relates to drugs and includes popular misinformation claims surrounding COVID-19 drug treatment. To the right of the centre is a minor cluster with alternative treatments (alcohol, UV light, vinegar, and disinfectants). Although hydroxychloroquine could be considered as part of this cluster, it is not used in the same contexts. Instead, it is commonly used next to coronavirus and sometimes used in specific contexts like "*Ohio bans hydroxychloroquine*", and

[52] https://www.forbes.com/sites/brucelee/2020/05/09/did-the-simpsons-episode-really-predict-covid-19-coronavirus-and-murder-hornets/
[53] https://nationalpost.com/life/covid-19-bat-soup

"… clinical trials". Along the top-right edge are queries used alone without any connection to the rest of the network. Henry Ford likely refers to the debunking of various mask myths by Henry Ford Health Systems[54], but the relationship is not clear as the terms are not connected to any other queries. The same is relevant for other queries like *Charles Lieber* and *bat soup*.

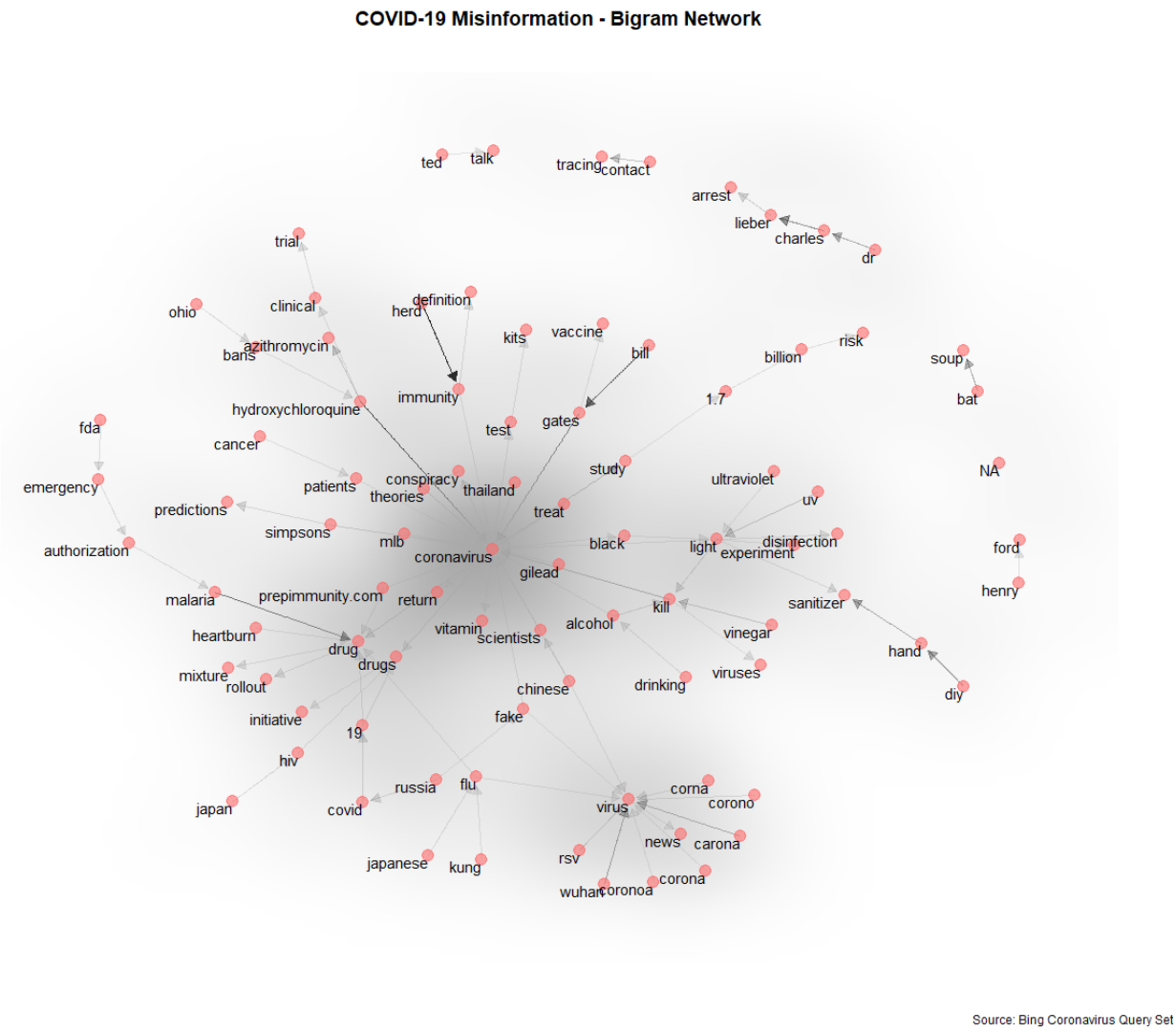


Figure 4.26 - Network plot of misinformation queries

4.3.4 Exploring Other Data Sources (CoronaNet and Election)

In order to explore the findings of the previous sections, additional data can be used. The CoronaNet database (Cheng et al., 2020) provides further information on multiple different factors that could be interesting to investigate in the context of misinformation. It has information on every political intervention implemented on different state levels. While exploring all of these would be beyond the scope of this thesis, they can be used to get

[54] https://www.henryford.com/blog/2020/07/debunking-covid19-mask-myths

a general idea of whether political decisions impact the level of misinformation. This is a complex area involving several different causal connections, and this section will not attempt to reveal causality between the factors, but instead briefly explore potential associations between the different data sets.

Initially, the overall count of implemented political interventions was investigated (Figure 4.27). The data was initially filtered to only use observations that had a compliance level of *mandatory* within the relevant states (Appendix A, 4.8). The hypothesis is that many mandatory policies were more likely to prompt misinformation queries than voluntary measures. However, this was not found to be the case when observing the figure. No clear

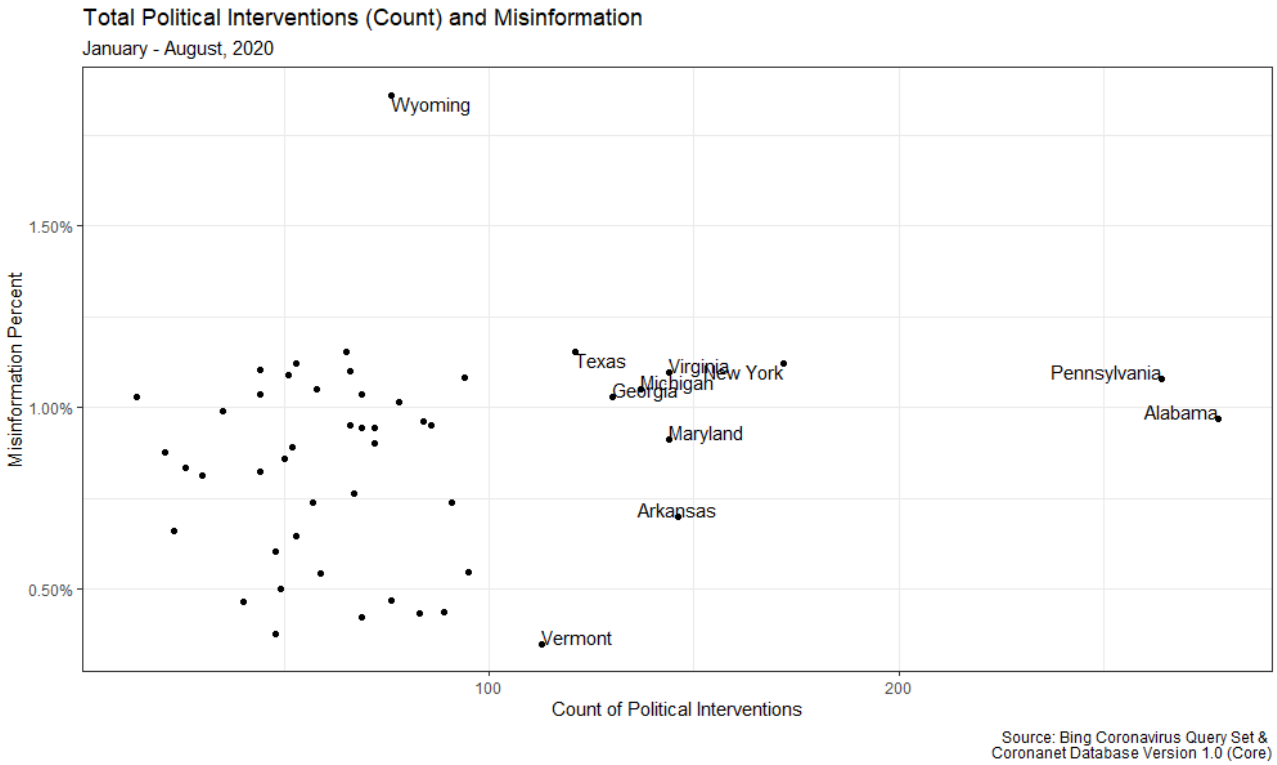


Figure 4.27 - Relationship between political interventions and misinformation level

pattern between increased implemented policies and overall misinformation percentage can be observed. The states with higher count of political interventions such as Alabama, Pennsylvania, or the cluster between New York and Texas do not have higher misinformation than several of the other states with lower counts. Some states even had many implemented policies while remaining on a very low proportion of misinformation (Vermont, Arkansas). Wyoming had the highest percentage of misinformation, but this is not directly associated with overall political interventions which are below the medium level. The associations can be further explored by looking at the individual types of implemented policies. An excerpt of these can be seen in Figure 4.28, and the full range of policies can be found in Appendix A, 4.8). It should be noted that each type of policy has several sub-types. An example is *Quarantine* which both includes measures of self-quarantine at home, quarantine after contact with the virus or quarantine in external locations such as hotels enforced by the government. This section only covers the top level. The x-axis was calculated by adding the active days per sub-type together, returning a total number of

days per type. While these results are obviously higher than the actual days gone since the initial outbreak, it still serves the purpose of providing insight into the strength of measures taken by a state within the relevant type. However, this representation does not reveal any clear patterns between misinformation percent and degree of implemented policies within a particular type. Generally, the focus on policies is similar for all states with some significant outliers. Pennsylvania and New York have had a higher level of measures related to quarantines, but other states are at a similar misinformation level despite the lower active days of quarantine. Alabama has had unusually many interventions of the social distancing type, but several states are still at a higher level of misinformation. Wyoming has proven to have the highest level of misinformation but generally had a low number

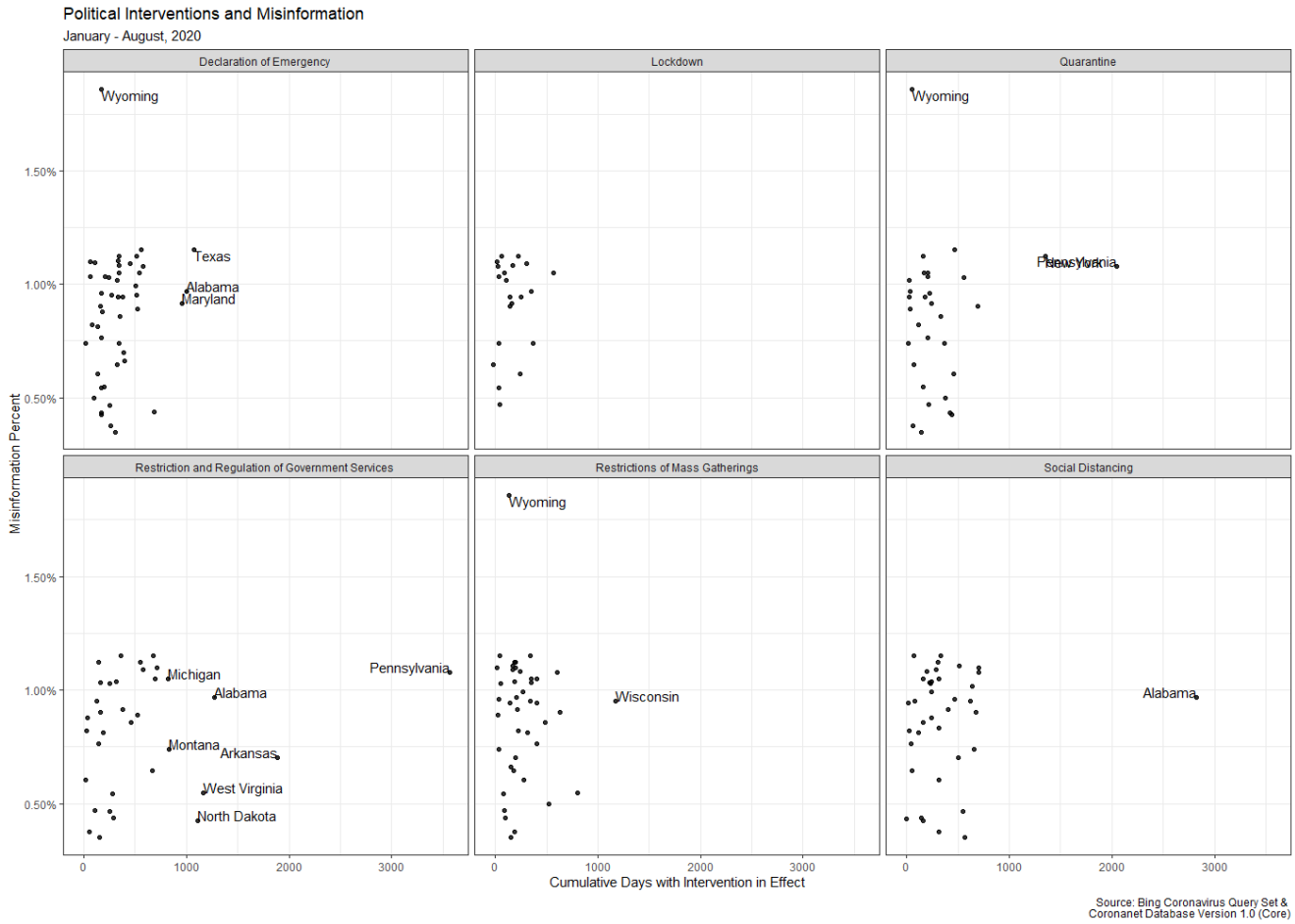


Figure 4.28 - Relationship between specific types of policies and misinformation level

of days impacted by new policies – it is even missing in some categories as no measures of that type has been implemented. Exploring the new COVID-19 related approaches in relation to misinformation provides no additional explanations to what can be causing increased levels of misinformation. Some states, like Pennsylvania, have had a lot of new policies of the *quarantine* and *restriction and regulation of government serves* types, and scores high on misinformation. This means that they could potentially be associated, but due to the large number of states that have equal levels of misinformation and fewer new policies, the association is weak at best.

Another possible hypothesis is that political orientation could have an impact on the level of misinformation per state. In order to investigate this further, a map of states based on results from the 2016 election extracted from the election dataset was created and coloured by the winning party of the state (republican = red, democratic = blue). It should be noted that it is challenging to show all the relevant dimensions in one visualization. For instance, the winning margin could be pertinent to explore whether profoundly red or intensely blue states could have an overall impact on the findings. In this case, the colour is simply an expression of the winning party, and the misinformation proportion was implemented in the alpha setting. This means that states with lower alpha have a lower degree of misinformation, and more saturated colours indicate a higher level of misinformation. Figure 4.29 shows a map visualization with misinformation calculated as a percentage of total queries from that state. The data covers the entire time period from January to August. The republican states which are highly saturated compared to the rest are Wisconsin, North Carolina, Alabama, and Florida. The high-ranking democratic states are Washington, Nevada, Virginia, Maryland, New York, New Jersey, Massachusetts, and Maine. While there are more democratic states with high scores of misinformation, several democratic states remain at very low scores. Only four Republican states are at high scores, and the rest are either at medium or low levels. In this representation, no clear connection between political orientation and misinformation level can be observed. It should also be noted that due to the complexity of the voting system in the United States including the relationship between population and electoral votes as well as the *winner takes all* rule (*The Electoral College*, 2020)[55], it can be difficult to make any conclusions on a state being more misinformative without considering the population and whether the state was a

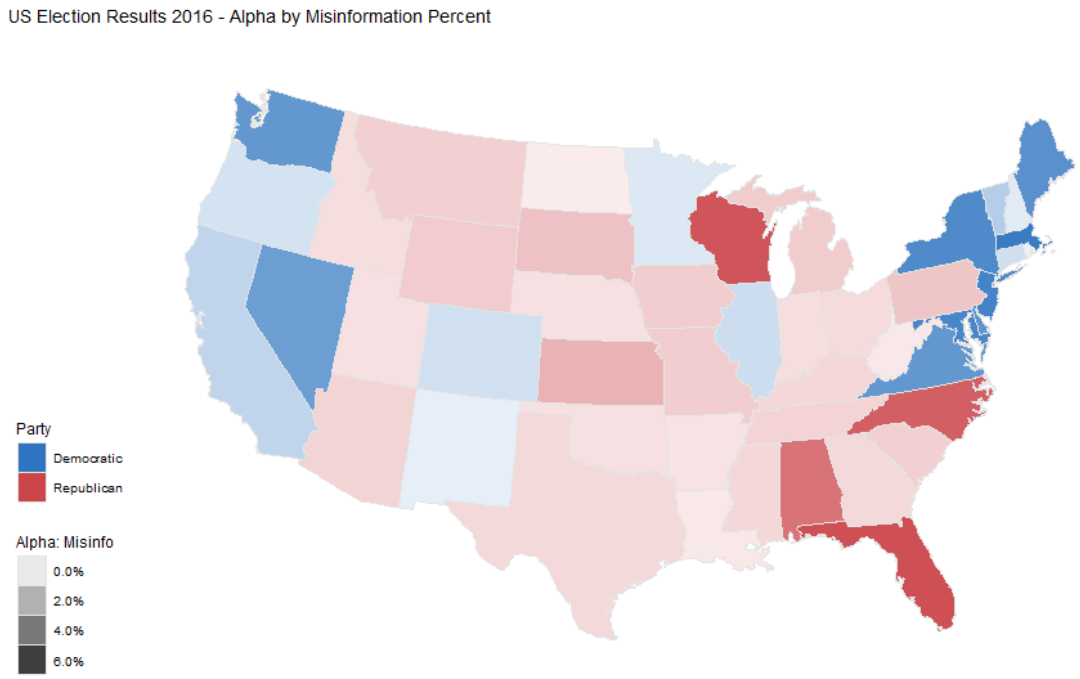


Figure 4.29 - Relationship between political orientation and misinformation level

[55] thttps://usafacts.org/visualizations/electoral-college-states-representation

swing state or not in the relevant election. If California, which accounts for 11.94% of the total US population, is democratic, it could still have more individual republican voters than Wyoming, which only accounts for 0.17% of the population[56]. This would make it difficult to assume that political orientation impacts misinformation levels without considering population. Figure 4.30 implements population in the alpha determined by the proportion of misinformation divided by state population and converted to log-odds. A change can be observed when compared to the above Figure 4.29. However, no significant differences can be observed between the political parties. Both have states with high and low levels, so assuming political orientation impacts misinformation based on this data would be a stretch. Michigan has a high level of misinformation, but it was also the closest swing state in the 2020 election with only a 0.23% winning margin for the republicans[57]. This means that even if a high proportion of the population expresses interest in misinformation, they are not necessarily republican as almost 50% of the voters were democratic. Wyoming has remained an outlier throughout the exploratory analysis, and here it can be observed that it also has a high level of misinformation proportional to the relatively small population. Wyoming is also among the most deeply republican states in the US with a distribution of republican to democratic of 70.1% to 18.2%. This could suggest that the level of misinformation is associated with their political orientation. However, comparing Wyoming to the second most Republican state Utah, which also has a high proportion of republicans compared to democrats (56% to 28%), the same trend is not observed. Utah, on the other hand, has very low levels of misinformation.

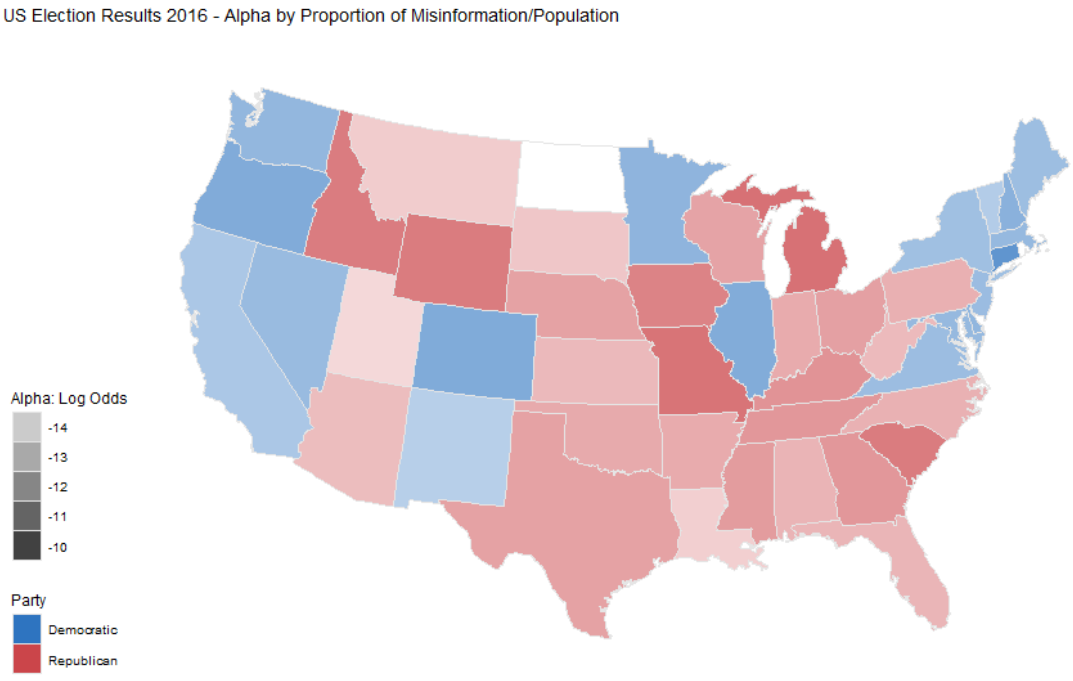


Figure 4.30 - Relationship between political orientation, population, and misinformation level

[56] https://worldpopulationreview.com/states
[57] https://en.wikipedia.org/wiki/Swing_state

Overall, the data does not suggest an association between political orientation and misinformation search queries. The same was observed between implemented mandatory policies and misinformation level.

4.4 Is The Data Representative?

One of the primary limitations of this study is the overall generalizability of the findings. All the queries come from the Microsoft Bing search engine, which in 2020 has 2.44% of the global market share. Meanwhile, Google far exceeds any other search engine and holds a market share of 91.54%[58]. This section will briefly present various data collected from Google Trends and compare results to those from the previous sections. Google data is aggregated and imported in formats that make direct comparisons difficult, but it can be used to investigate top queries from the Bing dataset as well as overall COVID-19 interest. Exploring Google Trends is helpful to this thesis in order to strengthen external validity. A few important limitations of the Google Trend platform should be noted: It only allows for comparison of max five search queries, the results provided are relative and scaled within their own environment meaning scaled according to the top query of the imported set. The primary use of the platform, as the name suggests, is to spot trends over time.

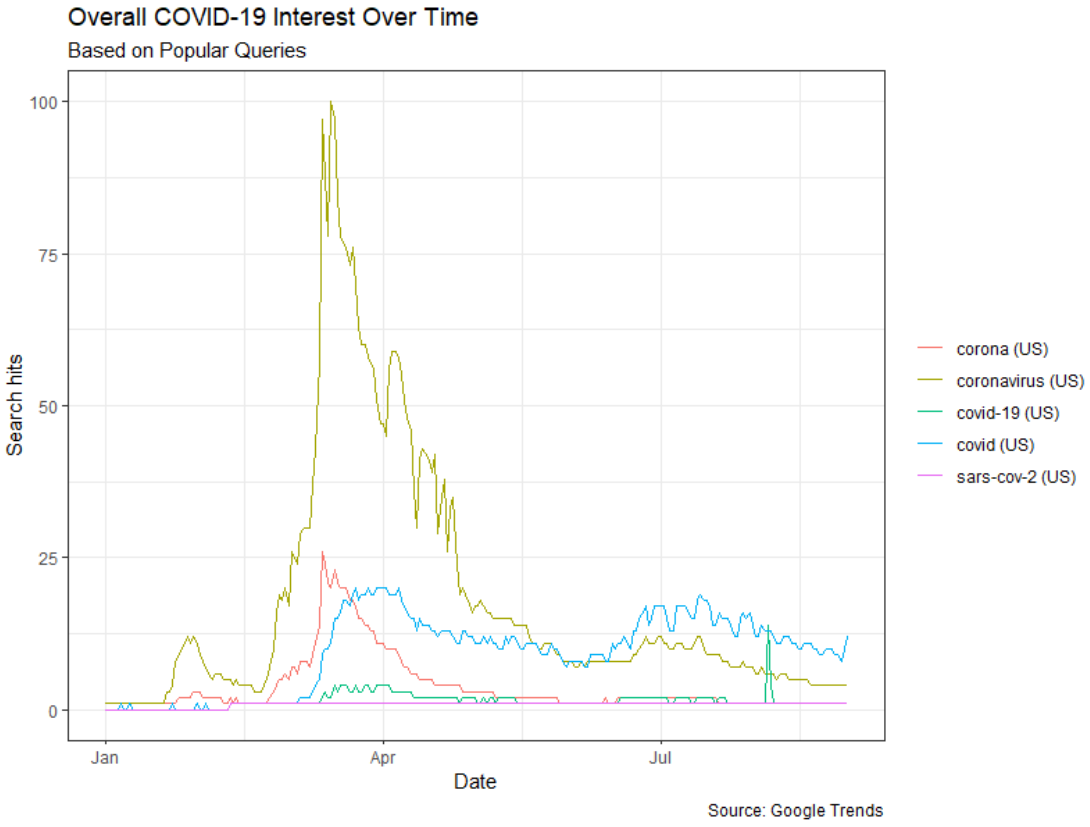


Figure 4.31 - Google Trends interest over time

[58] https://www.webfx.com/blog/seo/2019-search-market-share/

The initial step was to gauge the overall interest in COVID-19 during the period. This was done using top search queries directly related to the coronavirus. These included the queries "*coronavirus*", "corona", "*covid*", "covid-19", and "SARS-CoV-2" (Rovetta & Bhagavathula, 2020a), and the data is visualized in Figure 4.31. This shows similar overall interest as the findings from the Microsoft Bing dataset. The interest is relatively low in January and February, increasing significantly in March and April, moving to lower levels in the following months before seeing another increase in July and finally a drop towards August. Looking at this graph; the data from Microsoft Bing can be considered representative of overall COVID-19 interest. In order to explore the generalizability further, two additional line plots were made corresponding to the top 5 search queries from the Bing data over the entire period.

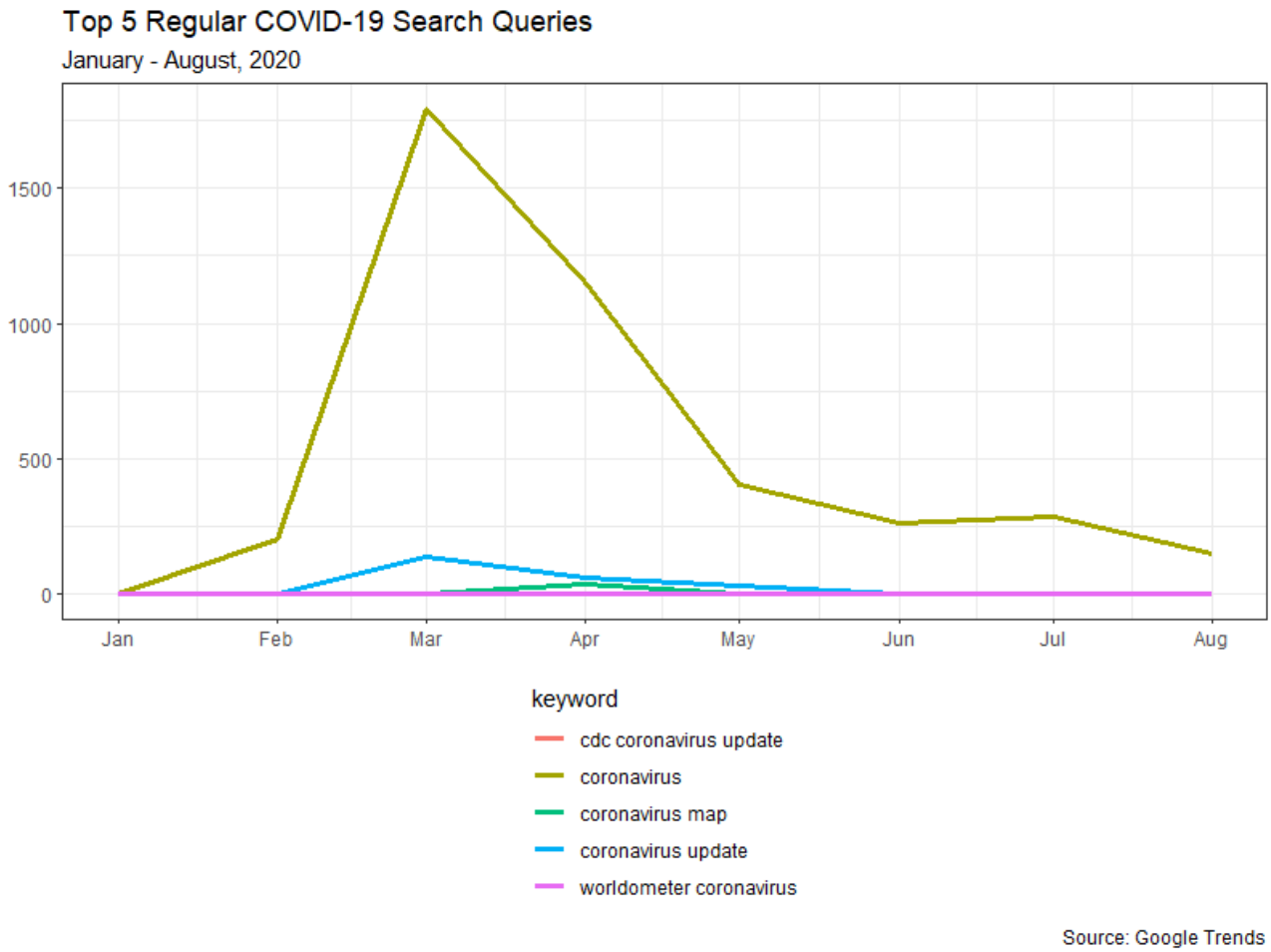


Figure 4.32 - Top 5 regular queries using Google Trends data

Figure 4.32 shows Google popularity of the top-5 queries from the Bing dataset. This is substantially different than Figure 4.21, and only "*coronavirus*" seems to have been a general term during the eight months. "*Coronavirus update*" has a small spike in March and April, but generally most of the top 5 queries are not popular at Google. This suggests that other search terms might be more popular on the Google platform. Looking at the previous Figure 4.31of overall COVID-19 Google queries also suggests that coronavirus is by far the most popular query, relative

to related terms. Looking at the top 5 misinformation queries from the Bing data (Figure 4.22), extracted from Google Trends, the popularity is more similar (Figure 4.33). Early increases can be observed on both platforms in March and April. The interest drops towards June increases again in July before decreasing in August. QAnon was found to be the significant outlier in the Bing data, starting to appear in March, seeing minor increases towards June, before increasing significantly from June to August. A similar trend can be spotted in the Google Trends data, although it is comparatively lower than other queries in the early months, which was not the case at the Bing platform. Bill gates queries were more prevalent at Google, but the spike surrounding him happened at the same time at both platforms. *Hydroxychloroquine, herd immunity,* and *malaria* follow the same general trend in both datasets. Generally, the two search platforms show similar interests, increasing and decreasing at the same points in time.

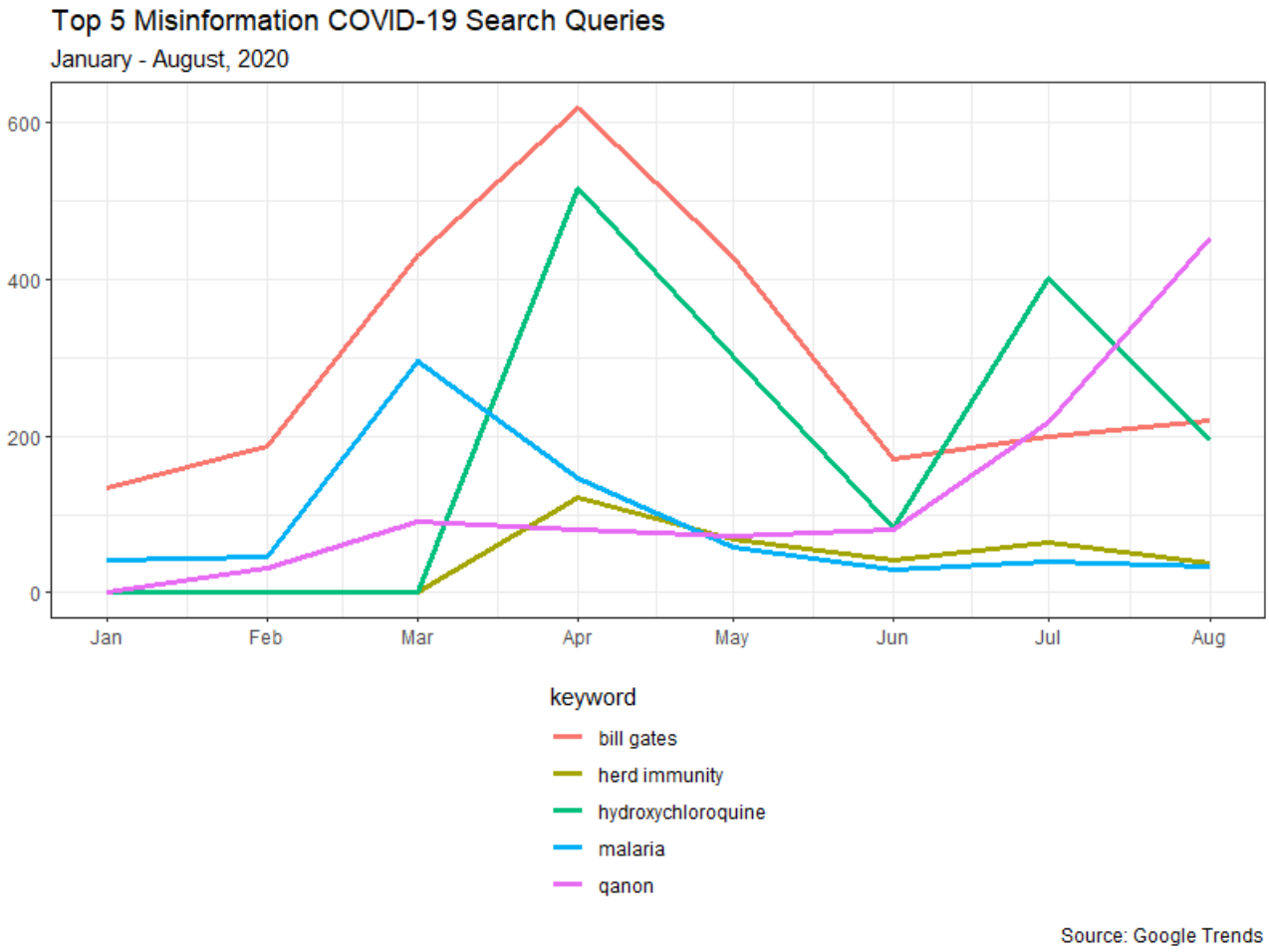


Figure 4.33 - Top 5 misinformation queries using Google Trends data

Finally, a comparison of the widespread misinformation per state was made (Figure 4.34). Note that this comparison is purely based on the top 5 previously identified misinformation queries. This is due to the API restrictions set by Google (limited to 5 queries), so the comparison is obviously skewed as the Bing data uses much more distinct queries to determine the extent of misinformation. Generally, when looking at previous misinformation distributions by state (Figure 4.23), the results are not showing the same trends. Several queries are high on the meter according to one platform, while being lower on the other. Furthermore, the query distribution within the states is much more evenly distributed. The Bing data showed QAnon vastly more represented across all states, and both bill gates and herd immunity had higher percentages as well. In order to get a more accurate comparison of the two, additional queries would have to be collected from Google, or the Bing data would have to be limited to only consider the top 5 misinformation queries.

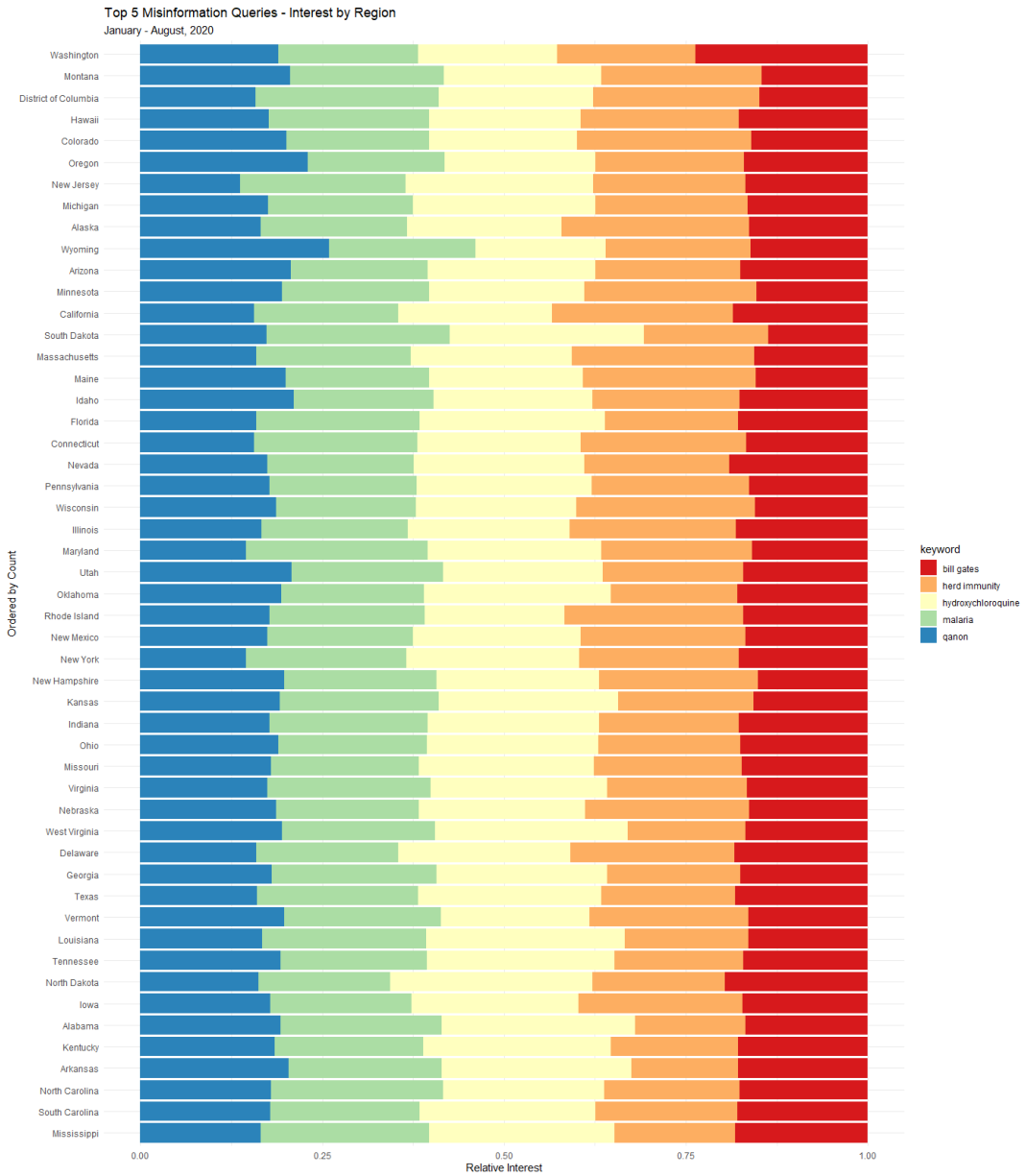


Figure 4.34 -Top 5 misinformation query popularity by state. Google Trends data

5 Discussion & Conclusion

This thesis explored Microsoft Bing search queries from the United States. The main objectives were to identify misinformation queries, determine the extent of these across the US states, identify the most frequent queries, and explore potential explanations. This final section will answer each of the research questions and discuss the main findings as well as relevant limitations and other reflections. Suggestions for further research will be discussed when applicable.

RQ1. How can search queries related to COVID-19 misinformation be identified?

This research question was fundamental, at it provided the foundation for any subsequent analysis and findings. The process of identifying misinformation queries was divided into several steps: creating a keyword index, applying the index to the data and extracting purposive samples for manual coding, and finally using the output of the coding to perform supervised machine learning. If the study had adhered strictly to computational grounded theory, as suggested by Nelson (2020), this initial step would have used topic modelling and unsupervised learning to identify meaningful patterns. It was decided not to do so, as the primary objective was to determine the two major groups of search queries; regular and misinformation.

The first major challenge was to identify essential keywords related to COVID-19 misinformation claims and myths. Care was taken in the selection process, both to ensure that information was gathered from credible sources, but also to try and cover a wide range of misinformation types. A total of 100 different misinformation claims were identified, and each was assigned a minimum of 1 keyword describing the content of the claim. Any search query with one of the defined keywords was labelled as misinformation and used to extract samples for manual coding. It was decided to sample with a 50/50 split between regular- and misinformation queries. Due to misinformation only being represented by ~1% of all queries, it was not possible to do a simple random sample, as this would likely have returned very few misinformation queries, making supervised learning difficult. This also meant that not some queries related to specific keywords were probably left out, as only 1000 of the ~17.500 misinformation queries were used for manual coding. Ideally, all the misinformation queries, or at the very least a more extensive selection, would be coded. This would ensure that all the different claims were represented when training the models. This was not possible within the scope of this thesis but should be considered in further research. Another reflection on this process is that COVID-19 misinformation claims have been rapidly changing during the outbreak. Most of the work on the keyword index was done during the summer of 2020, but new theories and claims are continually being introduced, making a frequent iteration of the process necessary. Even in the last few weeks, new ideas have emerged related to the ongoing election in the United States, which would all be relevant to implement. A significant question when working with machine learning is whether the findings are still relevant. This could be addressed by frequently updating the keyword index and repeating the following analysis with the new data, which I suggest doing in further research. While my findings are relevant to the concept

of misinformation, it can't be considered exhaustive as it was only possible to cover a certain number of claims within the time available.

Finally, some reflections on the automatic text classification using supervised learning. The model used to classify the entire dataset was a generalized linear model (logistic regression) set to classification. Very similar performance was achieved with random forest and SVM models, with XGBoost a fair bit behind. The lack of performance by the boosted tree model was most likely caused my limited knowledge of how to tune its hyperparameters accurately. Despite random forest performing slightly better when looking at the metrics, the glm model was selected due to its simplicity and very similar performance. Furthermore, the random forest classified some observations related to Wuhan as misinformation, which should be classified as regular information. The text analysis revealed that all misspellings of *corona* were classified as misinformation, especially noticeable in January, which should be corrected in similar work going forward. Various natural language processing methods such as lemmatization or stemming was attempted at the pre-processing stage but found to provide to return a lot of regular queries as misinformation. On the other hand, the selected method might have proven to be too restrictive, which resulted in a potentially narrow understanding of misinformation. This could be explored in further research while also implementing synonyms and alternate spellings.

RQ2. What is the overall extent of expressed interest in COVID-19 misinformation in the United States?

How does this differ across states?

The overall extent of misinformation, as represented in search queries across the United States, was observed to be approximately 1%. The level of misinformation was observed highest in January (4.8%), but several top queries of this month included misspellings of corona, and the result should be revisited after removing those. The remaining months saw gradually decreasing levels from 1.3% in February to 0.9% in August. The average number of queries across all months was 216.810, with January being the lowest at 16.542 and March being the highest at 430.810. The total number of queries across the different states were found to be closely associated with the population of states. The state-levels of misinformation was found to be correlated with the total number of queries from that state. Wyoming was found to be a significant outlier compared to the other states. It started at no or relatively low levels of misinformation before spiking significantly in April and staying at high levels for the remaining months. Furthermore, it was observed at very high misinformation levels both compared to the total number of queries and population size. A trend was observed of misinformation, in the early months and in the states with the largest cities, being at high levels. This gradually changed across the months, and in August were at very low levels, where states further inland saw increased levels of misinformation. The data at hand suggest that misinformation is moving towards the rural areas of the United States rather than primarily coming from areas in and around larger cities, but further research is required to confirm this.

RQ3. What are the top search queries related to misinformation, and how has this changed over time?

The top search queries were identified based on frequencies and weighted log odds. The top 5 regular search queries were "*coronavirus*", "*coronavirus update*", "*coronavirus map*", "*cdc coronavirus update*", and "*worldometer coronavirus*". These were mostly generic and primarily focused on COVID-19 updates, statistics, and reports. The top 5 misinformation queries, when measured by frequency, were "*qanon*", "*herd immunity*", "*hydroxychloroquine coronavirus*", "*bill gates coronavirus*", and "*malaria drug for coronavirus*". The popularity of these queries saw large popularity increases in March, April, and July, except for *QAnon*, which continued to increase significantly during the entire period. This was especially relevant from June to August, where QAnon reached twice the hits than any other misinformation query had reached at any other point during the eight months. QAnon was also found to be very popular across all states, while bill gates and herd immunity queries were predominantly coming from states with higher total query counts. Top misinformation queries were compared to numbers from the Google Trends platform and were generally found to follow a similar trajectory across the months. Comparing state-level data to Google Trends, very different trends were observed, as Google data was much more balanced. A possible explanation of this difference is that Google has much more data available, resulting in a more even number of queries from each state.

Three overall clusters of misinformation types were observed throughout the months. Initially, people were searching for debunked claims on the origin of the coronavirus, later this changed to alternative cures and self-medicating, and the final months saw large increases of misinformation queries related to various conspiracy theories, dominated by QAnon.

RQ4. What are some possible explanations for variations between states?

While the primary objective of this thesis was to explore the overall distributions and extent of misinformation; additional data sources were briefly investigated to find possible explanations. The CoronaNet dataset was used to examine whether political interventions could be associated with increases in misinformation queries. The CoronaNet data is extensive and could easily be covered on its own in a separate project. The data was filtered to only include policies that were implemented with a compliance level of mandatory. The total number of implemented policies were not found to have any impact on the level of misinformation. The same was relevant when investigating the different types of implemented political measures. While some states did shows signs of associations between COVID-19 policies and misinformation level, this was countered by a larger number of states that did not show any impact. Additionally, political orientation was explored as a potential source of increased misinformation interest. No connection was found between political orientation and misinformation levels. Although the most problematic state of Wyoming is also the most deeply red state of the United States, this was countered by other republican states having low overall levels.

Further research should continue to investigate this by implementing new and different data sources. It should be noted that while the population was covered to an extent in this paper, it would be challenging to investigate

causality based on search query logs alone. Even if a state is observed as a significant outlier in the dataset, there are many other factors that should be considered (Abay et al., 2020). Examples of these include unique users, search engine popularity, overall internet usage of the state, and scorings in freedom of speech (Mavragani & Ochoa, 2019), all of which would be relevant areas in future work.

PS. How has COVID-19 prompted misinformation seeking behaviour changed during the recent pandemic in the United States?

The above research questions were all instrumental in answering the main problem statement. The general interest in COVID-19 related topics was found to be significantly higher in March and April following the announcement of the virus as a global pandemic (WHO, 2020). Misinformation was represented in ~1% of all search queries, with the highest level in January. Wyoming was observed having significantly more misinformation queries than any other state, both relative to total queries and population. A trend was observed of misinformation levels, in the early months, being high in states with the largest cities, before dropping to lower levels in recent months. The opposite was observed for several rural states which saw increasing levels from January to August. *QAnon* was the most popular search query related to misinformation. Other top misinformation topics were herd immunity, bill gates, hydroxychloroquine, and malaria drugs. These spiked in popularity in March and April before gradually declining in popularity, while Qanon kept increasing during the entire period. Three significant misinformation categories were identified. In the early months, the most popular queries related to the origin of the coronavirus, which later changed to be more concerned with miracle cures and self-medication measures. From May, conspiracy theories started to be frequently represented in the queries, and they were much more popular than other misinformation types in the final months. Similar findings were observed in a very recent study by Cornell University (Evanega et al., 2020). The thesis found no association between implemented policy measures, political orientation and misinformation levels, but this should be investigated more in further research. Another possible extension of this study is to use a similar approach to explore infodemics in other countries and compare results to the United States.

This study has provided insight into the COVID-19 infodemic in the United States. At the time of writing, very few similar studies were identified, making the study highly relevant. The research was motivated by misinformation countermeasures suggested by the World Health Organization. The United States has been impacted severely by COVID-19 - I hope this, and further research, can assist in finally limiting the rapid spread.

6 REFERENCES

1st WHO Infodemiology Conference, WHO Infodemic Management. Retrieved 20 August 2020, from https://www.who.int/teams/risk-communication/infodemic-management/1st-who-infodemiology-conference

Coronavirus disease 2019 (COVID-19). Situation report – 82. Retrieved 19 August 2020, from https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200411-sitrep-82-covid-19.pdf?sfvrsn=74a5d15_2

A Common API to Modeling and Analysis Functions • parsnip. (n.d.). Retrieved 11 September 2020, from https://parsnip.tidymodels.org/

A Grammar of Data Manipulation • dplyr. (n.d.). Retrieved 11 September 2020, from https://dplyr.tidyverse.org/

Abay, K. A., Tafere, K., & Woldemichael, A. (2020). *Winners and Losers from COVID-19: Global Evidence from Google Search* (SSRN Scholarly Paper ID 3617347). Social Science Research Network. https://papers.ssrn.com/abstract=3617347

About RStudio. (n.d.). Retrieved 11 September 2020, from https://rstudio.com/about/

Aggarwal, C. C. (2018). *Machine Learning for Text*. Springer International Publishing. https://doi.org/10.1007/978-3-319-73531-3

Ahmed, W., Vidal-Alaball, J., Downing, J., & López Seguí, F. (2020). COVID-19 and the 5G Conspiracy Theory: Social Network Analysis of Twitter Data. *Journal of Medical Internet Research*, *22*(5), e19458. https://doi.org/10.2196/19458

Allington, D., Duffy, B., Wessely, S., Dhavan, N., & Rubin, J. (2020). Health-protective behaviour, social media usage and conspiracy belief during the COVID-19 public health emergency. *Psychological Medicine*, 1–7. https://doi.org/10.1017/S003329172000224X

Angelo Basteris, Mansourvar, M., & Will, U. K. (n.d.). *Google Trends and Seasonal Effects in Infodemiology: A Use Case About Obesity*.

Basch, C. H., Hillyer, G. C., Meleo-Erwin, Z. C., Jaime, C., Mohlman, J., & Basch, C. E. (2020). Preventive Behaviors Conveyed on YouTube to Mitigate Transmission of COVID-19: Cross-Sectional Study. *JMIR Public Health and Surveillance*, *6*(2), e18807. https://doi.org/10.2196/18807

Bode, L., & Vraga, E. K. (2018). See Something, Say Something: Correction of Global Health Misinformation on Social Media. *Health Communication*, *33*(9), 1131–1140. https://doi.org/10.1080/10410236.2017.1331312

Bradbury-Jones, C., & Isham, L. (2020). The pandemic paradox: The consequences of COVID-19 on domestic violence. *Journal of Clinical Nursing*, *29*(13–14), 2047–2049. https://doi.org/10.1111/jocn.15296

Bragazzi, N. L., Alicino, C., Trucchi, C., Paganino, C., Barberis, I., Martini, M., Sticchi, L., Trinka, E., Brigo, F., Ansaldi, F., Icardi, G., & Orsi, A. (2017). Global reaction to the recent outbreaks of Zika virus: Insights from a Big Data analysis. *PLOS ONE*, *12*(9), e0185263. https://doi.org/10.1371/journal.pone.0185263

Bryman, A. (2012). *Social research methods* (4th ed). Oxford University Press.

CDC. (2020, February 11). *Coronavirus Disease 2019 (COVID-19)*. Centers for Disease Control and Prevention. https://www.cdc.gov/coronavirus/2019-ncov/covid-data/covidview/08282020/percent-ili-visits.html

Cerbara, L., Ciancimino, G., Crescimbene, M., Parsi, M. R., Tintori, A., & Palomba, R. (n.d.). *A nation-wide survey on emotional and psychological impacts of COVID-19 social distancing*. 9.

Chen, X., & Sin, S.-C. J. (2013). 'Misinformation? What of it?' Motivations and individual differences in misinformation sharing on social media. *Proceedings of the American Society for Information Science and Technology*, *50*(1), 1–4. https://doi.org/10.1002/meet.14505001102

Cheng, C., Barceló, J., Hartnett, A. S., Kubinec, R., & Messerschmidt, L. (2020). COVID-19 Government Response Event Dataset (CoronaNet v.1.0). *Nature Human Behaviour*, *4*(7), 756–768. https://doi.org/10.1038/s41562-020-0909-7

Christopher Yee—Exploratory data analysis on COVID-19 search queries. Retrieved 15 July 2020, from https://www.christopheryee.org/blog/exploratory-data-analysis-on-covid-19-search-queries/

Cinelli, M., Quattrociocchi, W., Galeazzi, A., Valensise, C. M., Brugnoli, E., Schmidt, A. L., Zola, P., Zollo, F., & Scala, A. (2020). The COVID-19 Social Media Infodemic. *ArXiv:2003.05004 [Nlin, Physics:Physics]*. http://arxiv.org/abs/2003.05004

Compare Trends search terms—Trends Help. Retrieved 31 August 2020, from https://support.google.com/trends/answer/4359550?hl=en

Cook, S., Conrad, C., Mohebbi, A., & Matthew, H. *Assessing Google Flu Trends Performance in the United States during the 2009 Influenza Virus A (H1N1) Pandemic—ProQuest*. Retrieved 27 August 2020, from https://www-proquest-com.zorac.aub.aau.dk/docview/1308204512?accountid=8144

Coronavirus Search Trends. Google Trends. Retrieved 29 August 2020, from https://trends.google.com/trends/story/US_cu_4Rjdh3ABAABMHM_en

COSMO. (n.d.). Retrieved 15 August 2020, from http://copsy.dk/cosmo/

COVID: Top 10 current conspiracy theories. (n.d.). Alliance for Science. Retrieved 12 September 2020, from https://allianceforscience.cornell.edu/blog/2020/04/covid-top-10-current-conspiracy-theories/

COVID-19 deaths per capita by country. (n.d.). Statista. Retrieved 6 October 2020, from https://www.statista.com/statistics/1104709/coronavirus-deaths-worldwide-per-million-inhabitants/

COVID-19 Misinformation Types -. (n.d.). Retrieved 24 August 2020, from https://covid19misinfo.org/covid-19-claim-types/

COVID-19 Mythbusters – World Health Organization. (n.d.). Retrieved 31 August 2020, from https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public/myth-busters

Craswell, N., & Szummer, M. (2007). Random walks on the click graph. *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '07*, 239. https://doi.org/10.1145/1277741.1277784

Create Elegant Data Visualisations Using the Grammar of Graphics • ggplot2. (n.d.). Retrieved 11 September 2020, from https://ggplot2.tidyverse.org/

Creswell, J., & Clark, V. (2020, October 1). *Designing and Conducting Mixed Methods Research*. SAGE Publications Inc. https://us.sagepub.com/en-us/nam/designing-and-conducting-mixed-methods-research/book241842

Cronin, P., Ryan, F., & Coughlan, M. (2008). *Undertaking A Literature Review: A Step-By-Step Approach. British Journal of Nursing*.

Danish Health Authority (2020). *Questions and answers on novel coronavirus*. (n.d.). Retrieved 6 October 2020, from https://www.sst.dk/en/english/corona-eng/faq

Dastin, J. (2018, October 10). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

Disinformation by Fake BBC, CNN, ABC about Harvard Professor Allegedly Arrested for Creating COVID-19 | Drupal. (n.d.). Retrieved 30 September 2020, from /en/myth/disinformation-fake-bbc-cnn-abc-about-harvard-professor-allegedly-arrested-creating-covid-19

Downey, D., Dumais, S., Liebling, D., & Horvitz, E. (2008). Understanding the relationship between searchers' queries and information goals. *Proceeding of the 17th ACM Conference on Information and Knowledge Mining - CIKM '08*, 449. https://doi.org/10.1145/1458082.1458143

Dr. Fauci Says There Is No Truth At All to This Common Mask Myth. (n.d.). MSN. Retrieved 16 September 2020, from https://www.msn.com/en-us/Health/medical/dr-fauci-says-there-is-no-truth-at-all-to-this-common-mask-myth/ar-BB16SbRz

Evanega, S., Lynas, M., Adams, J., & Smolenyak, K. (n.d.). *Quantifying sources and themes in the COVID-19 'infodemic'*. 8.

Eysenbach, G. (2002). Infodemiology: The epidemiology of (mis)information. *The American Journal of Medicine*, *113*(9), 763–765. https://doi.org/10.1016/S0002-9343(02)01473-0

Eysenbach, G. (2006). *Infodemiology: Tracking Flu-Related Searches on the Web for Syndromic Surveillance*. 5.

Eysenbach, G. (2009). Infodemiology and Infoveillance: Framework for an Emerging Set of Public Health Informatics Methods to Analyze Search, Communication and Publication Behavior on the Internet. *Journal of Medical Internet Research*, *11*(1). https://doi.org/10.2196/jmir.1157

Eysenbach, G. (2011). Infodemiology and Infoveillance: Tracking Online Health Information and Cyberbehavior for Public Health. *American Journal of Preventive Medicine*, *40*(5, Supplement 2), S154–S158. https://doi.org/10.1016/j.amepre.2011.02.006

Factsheet-infodemic_eng.pdf. (n.d.). Retrieved 19 August 2020, from https://iris.paho.org/bitstream/handle/10665.2/52052/Factsheet-infodemic_eng.pdf?sequence=14

Fisher, M. (2020, April 23). R0, the Messy Metric That May Soon Shape Our Lives, Explained. *The New York Times*. https://www.nytimes.com/2020/04/23/world/europe/coronavirus-R0-explainer.html

Florkowski, C. M. (2008). Sensitivity, Specificity, Receiver-Operating Characteristic (ROC) Curves and Likelihood Ratios: Communicating the Performance of Diagnostic Tests. *The Clinical Biochemist Reviews*, *29*(Suppl 1), S83–S87.

Frutos, R., Serra-Cobo, J., Chen, T., & Devaux, C. A. (2020). COVID-19: Time to exonerate the pangolin from the transmission of SARS-CoV-2 to humans. *Infection, Genetics and Evolution*, *84*, 104493. https://doi.org/10.1016/j.meegid.2020.104493

Gencoglu, O., & Gruber, M. (2020). Causal Modeling of Twitter Activity During COVID-19. *ArXiv:2005.07952 [Cs]*. http://arxiv.org/abs/2005.07952

General Resampling Infrastructure. (n.d.). Retrieved 11 September 2020, from https://rsample.tidymodels.org/

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, *457*(7232), 1012–1014. https://doi.org/10.1038/nature07634

Google Trends. (n.d.). Google Trends. Retrieved 29 August 2020, from https://trends.google.com/trends/?geo=US

Greenhalgh, T., & Peacock, R. (2005). Effectiveness and efficiency of search methods in systematic reviews of complex evidence: Audit of primary sources. *BMJ*, *331*(7524), 1064–1065. https://doi.org/10.1136/bmj.38636.593461.68

Greenwell, B. & Boehmke, B. (2019). *Hands-On Machine Learning with R*. https://bradleyboehmke.github.io/HOML/

Grolemund, G., & Wickham, H. (2020). *R for Data Science*. https://r4ds.had.co.nz/

Hammersley, M. (2020). Reflections on the Methodological Approach of Systematic Reviews. In O. Zawacki-Richter, M. Kerres, S. Bedenlier, M. Bond, & K. Buntins (Eds.), *Systematic Reviews in Educational Research: Methodology, Perspectives and Application* (pp. 23–39). Springer Fachmedien. https://doi.org/10.1007/978-3-658-27602-7_2

Hiemstra, D. (2020). Reducing Misinformation in Query Autocompletions. *ArXiv:2007.02620 [Cs]*. http://arxiv.org/abs/2007.02620

How the myths surrounding bat soup came to represent our collective fear and confusion over COVID-19. (n.d.). National Post. Retrieved 30 September 2020, from https://nationalpost.com/life/covid-19-bat-soup

Hua, J., & Shaw, R. (2020). Corona Virus (COVID-19) "Infodemic" and Emerging Issues through a Data Lens: The Case of China. *International Journal of Environmental Research and Public Health*, *17*(7), 2309. https://doi.org/10.3390/ijerph17072309

IFLA -- How To Spot Fake News. (n.d.). Retrieved 28 August 2020, from https://www.ifla.org/publications/node/11174

Islam, M. S., Sarkar, T., Khan, S. H., Mostofa Kamal, A.-H., Hasan, S. M. M., Kabir, A., Yeasmin, D., Islam, M. A., Amin Chowdhury, K. I., Anwar, K. S., Chughtai, A. A., & Seale, H. (2020). COVID-19–Related Infodemic and Its Impact on Public Health: A Global Social Media Analysis. *The American Journal of Tropical Medicine and Hygiene*. https://doi.org/10.4269/ajtmh.20-0812

Jansen, B. J., Spink, A., Taksa, I., & College, B. (n.d.). *Handbook of Research on Web Log Analysis*. 628.

January 1, S. B. on, & 2020. (2020, January 1). 2020 Search Market Share: 5 Hard Truths About Today's Market. *WebFX Blog*. https://www.webfx.com/blog/seo/2019-search-market-share/

Jolley, D., & Douglas, K. M. (2017). Prevention is better than cure: Addressing anti-vaccine conspiracy theories. *Journal of Applied Social Psychology*, *47*(8), 459–469. https://doi.org/10.1111/jasp.12453

Jun, S.-P., Yoo, H. S., & Choi, S. (2018). Ten years of research change using Google Trends: From the perspective of big data utilizations and applications. *Technological Forecasting and Social Change*, *130*, 69–87. https://doi.org/10.1016/j.techfore.2017.11.009

Jurczyk, T. (n.d.). *Gains vs ROC curves. Do you understand the difference? | TIBCO Community*. Retrieved 22 September 2020, from https://community.tibco.com/wiki/gains-vs-roc-curves-do-you-understand-difference

Koehrsen, W. (2018, March 10). *Beyond Accuracy: Precision and Recall*. Medium. https://towardsdatascience.com/beyond-accuracy-precision-and-recall-3da06bea9f6c

Kouzy, R., Abi Jaoude, J., Kraitem, A., El Alam, M. B., Karam, B., Adib, E., Zarka, J., Traboulsi, C., Akl, E. W., & Baddour, K. (n.d.). Coronavirus Goes Viral: Quantifying the COVID-19 Misinformation Epidemic on Twitter. *Cureus*, *12*(3). https://doi.org/10.7759/cureus.7255

Kuhn, M. (2008). Building Predictive Models in *R* Using the **caret** Package. *Journal of Statistical Software*, *28*(5). https://doi.org/10.18637/jss.v028.i05

Lazar, J. (2017). *Research methods in human computer interaction* (2nd edition). Elsevier.

Lee, B. Y. (n.d.). *Did 'The Simpsons' Episode Really Predict COVID-19 Coronavirus And Murder Hornets?* Forbes. Retrieved 30 September 2020, from https://www.forbes.com/sites/brucelee/2020/05/09/did-the-simpsons-episode-really-predict-covid-19-coronavirus-and-murder-hornets/

Leitner, S., Gula, B., Jannach, D., Krieg-Holz, U., & Wall, F. (2020). Infodemics: A call to action for interdisciplinary research. *ArXiv:2007.12226 [Physics, q-Fin]*. http://arxiv.org/abs/2007.12226

Li, H. O.-Y., Bailey, A., Huynh, D., & Chan, J. (2020). YouTube as a source of information on COVID-19: A pandemic of misinformation? *BMJ Global Health*, *5*(5), e002604. https://doi.org/10.1136/bmjgh-2020-002604

Lu, D. (2019). Google's hate speech AI may be racially biased. *New Scientist*, *243*(3243), 7–7. https://doi.org/10.1016/s0262-4079(19)31505-2

Maslow, A. H. (1943). A theory of human motivation. *Psychological Review*, *50*(4), 370–396. https://doi.org/10.1037/h0054346

Mavragani, A., & Ochoa, G. (2019). Google Trends in Infodemiology and Infoveillance: Methodology Framework. *JMIR Public Health and Surveillance*, *5*(2), e13439. https://doi.org/10.2196/13439

Microsoft Research – Emerging Technology, Computer, and Software Research. (n.d.). Microsoft Research. Retrieved 1 September 2020, from https://www.microsoft.com/en-us/research/

Microsoft/Bing-COVID-19-Data. (2020). Microsoft. https://github.com/microsoft/Bing-COVID-19-Data (Original work published 2020)

Microsoft/BingCoronavirusQuerySet. (2020). Microsoft. https://github.com/microsoft/BingCoronavirusQuerySet (Original work published 2020)

Misinformation Watch -. (n.d.). Retrieved 13 September 2020, from https://covid19misinfo.org/misinfowatch/

Mitchell, M. (2019, November 8). *Programming Languages For Data Scientists*. Medium. https://towardsdatascience.com/programming-languages-for-data-scientists-afde2eaf5cc5

Nations, U. (n.d.-b). *UN supporting 'trapped' domestic violence victims during COVID-19 pandemic*. United Nations; United Nations. Retrieved 1 September 2020, from https://www.un.org/en/coronavirus/un-supporting-%E2%80%98trapped%E2%80%99-domestic-violence-victims-during-covid-19-pandemic

Nelson, L. K. (2020). Computational Grounded Theory: A Methodological Framework. *Sociological Methods & Research*, *49*(1), 3–42. https://doi.org/10.1177/0049124117729703

Nguyen, H., Richards, R., Chan, C.-C., & Liszka, K. J. (2016). RedTweet: Recommendation engine for reddit. *Journal of Intelligent Information Systems*, *47*(2), 247–265. https://doi.org/10.1007/s10844-016-0410-y

Person, B., Sy, F., Holton, K., Govert, B., Liang, A., Garza, B., Gould, D., Hickson, M., McDonald, M., Meijer, C., Smith, J., Veto, L., Williams, W., & Zauderer, L. (2004). Fear and Stigma: The Epidemic within the SARS Outbreak. *Emerging Infectious Diseases*, *10*(2), 358–363. https://doi.org/10.3201/eid1002.030750

Porter, M. F. (2001). *Snowball: A language for stemming algorithms*. http://snowball.tartarus.org/texts/introduction.html

Preprocessing Tools to Create Design Matrices. (n.d.). Retrieved 11 September 2020, from https://recipes.tidymodels.org/

Principles of Epidemiology | Lesson 1—Section 1. (2020, May 11). https://www.cdc.gov/csels/dsepd/ss1978/lesson1/section1.html

Qin, L., Sun, Q., Wang, Y., Wu, K.-F., Chen, M., Shia, B.-C., & Wu, S.-Y. (2020). Prediction of Number of Cases of 2019 Novel Coronavirus (COVID-19) Using Social Media Search Index. *International Journal of Environmental Research and Public Health*, *17*(7), 2365. https://doi.org/10.3390/ijerph17072365

R Markdown. (n.d.). Retrieved 11 September 2020, from https://rmarkdown.rstudio.com/

R: What is R? (n.d.). Retrieved 11 September 2020, from https://www.r-project.org/about.html

Randolph, J. (n.d.). *A Guide to Writing the Dissertation Literature Review*. https://doi.org/10.7275/B0AZ-8T74

Rao, C. O., Sujata. (2020, July 3). Wall Street shifts bets to big pharma as COVID-19 vaccine race progresses. *Reuters*. https://www.reuters.com/article/us-health-coronavirus-stocks-idUSKBN24333M

Robinson, J. D & Silge J. (n.d.). *Text Mining with R*. Retrieved 2 September 2020, from https://www.tidytextmining.com/

Rocher, L., Hendrickx, J. M., & de Montjoye, Y.-A. (2019). Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications*, *10*(1), 3069. https://doi.org/10.1038/s41467-019-10933-3

Rosenberg, H., Syed, S., & Rezaie, S. (2020). The Twitter pandemic: The critical role of Twitter in the dissemination of medical information and misinformation during the COVID-19 pandemic. *CJEM*, *22*(4), 418–421. https://doi.org/10.1017/cem.2020.361

Rovetta, A., & Bhagavathula, A. S. (2020a). Global Infodemiology of COVID-19: Analysis of Google Web Searches and Instagram Hashtags. *Journal of Medical Internet Research*, *22*(8), e20673. https://doi.org/10.2196/20673

Rovetta, A., & Bhagavathula, A. S. (2020b). COVID-19-Related Web Search Behaviors and Infodemic Attitudes in Italy: Infodemiological Study. *JMIR Public Health and Surveillance*, *6*(2), e19374. https://doi.org/10.2196/19374

Rowley, J., & Slack, F. (2004). *Conducting a Literature Review*. https://www-emerald-com.zorac.aub.aau.dk/insight/content/doi/10.1108/01409170410784185/full/pdf?title=conducting-a-literature-review

RStudio | Open source & professional software for data science teams. (n.d.). Retrieved 11 September 2020, from https://rstudio.com/

Scharkow, M., & Vogelgesang, J. (2011). Measuring the Public Agenda using Search Engine Queries. *International Journal of Public Opinion Research*, *23*(1), 104–113. https://doi.org/10.1093/ijpor/edq048

Schnoebelen, T. (2019, April 11). *I dare say you will never use tf-idf again*. Medium. https://medium.com/@TSchnoebelen/i-dare-say-you-will-never-use-tf-idf-again-4918408b2310

Schnoebelen, T., Silge [aut, J., cre, cph, & Hayes, A. (2020). *tidylo: Weighted Tidy Log Odds Ratio* (0.1.0) [Computer software]. https://CRAN.R-project.org/package=tidylo

Serrano, J. C. M., Papakyriakopoulos, O., & Hegelich, S. (n.d.). *NLP-based Feature Extraction for the Detection of COVID-19 Misinformation Videos on YouTube*. 7.

Shahi, G. K., Dirkson, A., & Majchrzak, T. A. (2020). An Exploratory Study of COVID-19 Misinformation on Twitter. *ArXiv:2005.05710 [Cs]*. http://arxiv.org/abs/2005.05710

Silge, J & Hvitfeldt, E. (2020). *Supervised Machine Learning for Text Analysis in R*. Retrieved 17 September 2020, from https://smltar.com/

Silge, J. (2018, December 24). *Text classification with tidy data principles*. Julia Silge. https://juliasilge.com/blog/tidy-text-classification/

Silge, J. (2019, July 8). *Introducing tidylo*. Julia Silge. https://juliasilge.com/blog/introducing-tidylo/

Silge, J & Kuhn, M. (2020). *Tidy Modeling with R*. Retrieved 17 September 2020, from https://www.tmwr.org/

Simple, Consistent Wrappers for Common String Operations. (n.d.). Retrieved 11 September 2020, from https://stringr.tidyverse.org/

Soreni, N., Cameron, D. H., Streiner, D. L., Rowa, K., & McCabe, R. E. (2019). Seasonality Patterns of Internet Searches on Mental Health: Exploratory Infodemiology Study. *JMIR Mental Health*, *6*(4), e12974. https://doi.org/10.2196/12974

Special Report: COVID-19 Myths – NewsGuard. (n.d.). Retrieved 12 September 2020, from https://www.newsguardtech.com/covid-19-myths/

Strauss, A., & Corbin, J. (1990). Grounded Theory Research: Procedues, Canons and Evaluative Criteria. *Zeitschrift Für Soziologie*, *19*.

Suh, J., Horvitz, E., White, R. W., & Althoff, T. (2020). Population-Scale Study of Human Needs During the COVID-19 Pandemic: Analysis and Implications. *ArXiv:2008.07045 [Cs]*. http://arxiv.org/abs/2008.07045

Sundhedsstyrelsen. *Nyt samarbejde mellem Statens Serum Institut og Forskerservice om COVID-19 data til forskning*. Retrieved 28 August 2020, from https://www.ssi.dk/aktuelt/nyheder/2020/nyt-samarbejde-mellem-statens-serum-institut-og-forskerservice-om-covid-19-data-til-forskning

Swing state. (2020). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Swing_state&oldid=980084721

Tangcharoensathien, V., Calleja, N., Nguyen, T., Purnat, T., D'Agostino, M., Garcia-Saiso, S., Landry, M., Rashidian, A., Hamilton, C., AbdAllah, A., Ghiga, I., Hill, A., Hougendobler, D., van Andel, J., Nunn, M., Brooks, I., Sacco, P. L., De Domenico, M., Mai, P., … Briand, S. (2020). Framework for Managing the COVID-19 Infodemic: Methods and Results of an Online, Crowdsourced WHO Technical Consultation. *Journal of Medical Internet Research*, *22*(6), e19659. https://doi.org/10.2196/19659

Tasnim, S., Hossain, M. M., & Mazumder, H. (2020). Impact of Rumors and Misinformation on COVID-19 in Social Media. *Journal of Preventive Medicine and Public Health*, *53*(3), 171–174. https://doi.org/10.3961/jpmph.20.094

Terefe, B., Rovetta, A., Rajan, A. K., & Awoke, M. (2020). Coronavirus-related online web search desire amidst the rising novel coronavirus incidence in Ethiopia: Google Trends-based infodemiology. *MedRxiv*, 2020.07.23.20158592. https://doi.org/10.1101/2020.07.23.20158592

Textrecipes package | R Documentation. (n.d.). Retrieved 17 September 2020, from https://www.rdocumentation.org/packages/textrecipes/versions/0.3.0

The Electoral College: How it works & data on representation by state. (n.d.). USAFacts. Retrieved 2 October 2020, from https://usafacts.org/visualizations/electoral-college-states-representation/

Tidy Characterizations of Model Performance. (n.d.). Retrieved 11 September 2020, from https://yardstick.tidymodels.org/

Tidymodels. (n.d.). Retrieved 11 September 2020, from https://www.tidymodels.org/

Tidyverse. (n.d.). Retrieved 11 September 2020, from https://www.tidyverse.org/

Ting, K. M. (2010). Precision and Recall. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of Machine Learning* (pp. 781–781). Springer US. https://doi.org/10.1007/978-0-387-30164-8_652

US States—Ranked by Population 2020. (n.d.). Retrieved 2 October 2020, from https://worldpopulationreview.com/states

Usher, K., Bhullar, N., Durkin, J., Gyamfi, N., & Jackson, D. (2020). Family violence and COVID-19: Increased vulnerability and reduced options for support. *International Journal of Mental Health Nursing*, *29*(4), 549–552. https://doi.org/10.1111/inm.12735

Vazquez, M. (n.d.). *Calling COVID-19 the "Wuhan Virus" or "China Virus" is inaccurate and xenophobic*. Retrieved 24 August 2020, from https://medicine.yale.edu/ysm/news-article/23074/

Webster, J., & Watson, R. (2002). *Analyzing the Past to Prepare for the Future_ Writing a Literature Review—4132319.pdf*. Management Information Systems Research Center. Uni Minnesota.

WHO. (n.d.). *Weekly Epidemiological Update—28 September, 2020*. Retrieved 6 October 2020, from https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200928-weekly-epi-update.pdf?sfvrsn=9e354665_6

WHO welcomes preliminary results about dexamethasone use in treating critically ill COVID-19 patients. (n.d.). Retrieved 16 September 2020, from https://www.who.int/news-room/detail/16-06-2020-who-welcomes-preliminary-results-about-dexamethasone-use-in-treating-critically-ill-covid-19-patients

Wickham, H. (2014). Tidy Data. *Journal of Statistical Software*, *59*(1), 1–23. https://doi.org/10.18637/jss.v059.i10

Windfeld, A. (2019). *Automatic Classification of Relevance Aspects in Complex Game Requests*. 11.

Wright, M. N., & Ziegler, A. (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, *77*(1). https://doi.org/10.18637/jss.v077.i01

Yong, S. by E. (n.d.). How the Pandemic Defeated America. *The Atlantic*. Retrieved 6 October 2020, from https://www.theatlantic.com/magazine/archive/2020/09/coronavirus-american-failure/614191/

Zarocostas, J. (2020). How to fight an infodemic. *The Lancet*, *395*(10225), 676. https://doi.org/10.1016/S0140-6736(20)30461-X

Zeraatkar, K., & Ahmadi, M. (2018). Trends of infodemiology studies: A scoping review. *Health Information and Libraries Journal*, *35*(2), 91–120. https://doi.org/10.1111/hir.12216

Zika virus. (n.d.). Retrieved 29 August 2020, from https://www.who.int/news-room/fact-sheets/detail/zika-virus

Zweig, M., & DeVoto, E. (2015). *Observational studies: Does the language fit the evidence? Association vs. causation*. HealthNewsReview.Org. https://www.healthnewsreview.org/toolkit/tips-for-understanding-studies/does-the-language-fit-the-evidence-association-versus-causation/

7 APPENDICES

The main code and R data are submitted but can also be downloaded from my GitLab. Contact me if there are problems with access: awindfeld@gmail.com / awindf18@student.aau.dk

https://gitlab.com/awindfeld/thesis

A. Markdown PDF containing all R code and visualizations
B. Spreadsheet with misinformation claims and keywords
C. Manually coded sample
D. Literature foundation submitted for approval
E. Email with literature approval by supervisor
F. Executive summary
G. R code
H. R data-file