CSCI 5502 Project Progress Report*

The Pandemic within COVID-19: Assessing Misinformation Susceptibility

Son Pham
CU Boulder
Boulder, Colorado
soph3006@colorado.edu

Kyle Rogers
CU Boulder
Boulder, Colorado
kyro3301@colorado.edu

Reiko Matsuda-Dunn
CU Boulder
Boulder, Colorado
rema8973@colorado.edu

Ryan Karasopoulos
CU Boulder
Boulder, Colorado
ryka6853@colorado.edu

ABSTRACT

This paper focuses on how COVID-19 information was communicated within and between different countries, reactions of governments to the pandemic, and attitudes and risk perceptions people had towards the virus. The major questions to answer are how digital communications influenced people's interpretation of the news, what their responses were to the new laws and mandates, their beliefs and concerns regarding the pandemic versus other world issues, and the similarities and trends among the different countries.

CCS CONCEPTS

• **Information systems** → **Presentation of retrieval results**; **Data management systems**; **Information integration**; **Database design and models**; • **Mathematics of computing** → *Probability and statistics.*

KEYWORDS

Data Mining, COVID-19, Risk Perception, Communication, Misinformation

ACM Reference Format:
Son Pham, Kyle Rogers, Reiko Matsuda-Dunn, and Ryan Karasopoulos. 2021. CSCI 5502 Project Progress Report: The Pandemic within COVID-19: Assessing Misinformation Susceptibility. In *Proceedings of CSCI-5502 Data Mining Summer 2021 (CSCI-5502 '21).* ACM, New York, NY, USA, 7 pages. https://doi.org/10.475/123_4

1 INTRODUCTION

The worldwide pandemic known as the corona virus disease 2019 (COVID-19) placed the world under lockdown and greatly changed the global perception of the veracity of trustworthy sources and how information is presented. Various influences ranging from social media and local news depictions of the virus to a lack of trust by individual and political actors have cultured an environment that facilitates the frequent spread of misinformation. In this study, we will examine survey data from 12 different countries across the globe conducted by Roozenbeek et. al. [3]. The responses to the provided questions are coupled to form predictions on the level of individual comprehension relating to COVID-19. Elements such as trust in various entities, questions pertaining to personal feelings of risk, and basic informal personal questions allowed for an assessment of collective misunderstanding across a diverse range of surveyed countries.

*This updated proposal is Part 3 of the CSCI-5502 Data Mining Project

CSCI-5502 '21, July 2021, Boulder, CO, USA

ACM ISBN 123-4567-24-567/08/06.
https://doi.org/10.475/123_4

1.1 Motivation

The general spread of misinformation in modern society is an interesting and important topic. Acceptance of misinformation as truth can influence critical outcomes such as election results, public trust, and public health. This data set not only contains information that may help explain the mechanism behind the spread of misinformation regarding COVID-19, but the intersection of these fields.

As the COVID-19 pandemic has continued to affect people's lives, understanding the effects of misinformation are key to assessing trends and potential outcomes moving forward. After conducting preliminary analyses, the team found that many questions were more difficult to assess direct misinformation metrics to. The survey data sampled worked to maintain a non-biased and non-intrusive questioning scheme, so the team was driven to look deeper into relationships between specific question results. This allowed for more accurate characterization and assessment of potential misinformation.

2 LITERATURE SURVEY

A vigorous literature survey was conducted to ensure the authors' analyses remain illustrating original and interesting results. This was broken down into two parts: studies outside of the data set that assessed mental health and collaboration amidst COVID-19, and studies done on parts of the data set.

2.1 External studies

Given the urgent and drastic nature of COVID-19, there are many publications and corresponding data sets available. This study wanted to focus on data outside of the more common death rate, positive test rate, and overall case rate data sets. As such, data regarding mental health and perceptions was a focus of this study. One article of interest involved collaboration analysis between countries as a result of COVID-19 [3]. A strengthened bilateral research relationship was evident between multiple countries, and the study identified the US, UK, and China as the largest contributors to COVID-19 research. Another article of interest observed psychological effects from COVID-19 in Spain [2]. Survey data was collected during the initial stages of the pandemic and was given a rating on the Impact of Event scale, which assesses psychological distress caused by a traumatic life event.

2.2 Studies based on this data set

Prior work concluded that a belief in COVID misinformation was rare, however this type of misinformation is often perceived as highly reliable by a small but persistent cohort. Higher trust in science and math abilities were associated with lower susceptibility to misinformation. Researchers used basic statistical methods such as ANOVAs and ordinary least squares (OLS) linear regression to evaluate survey results [3]. Significant differences were found in the perceived reliability of misinformation between countries. Linear regression found that predictors of susceptibility to misinformation included identifying as politically right-wing, self-identifying as a member of a minority group, and using social media as a source of information. Predictors of lower susceptibility to misinformation include trust in scientists and higher numeracy (math) scores.

Another publication for the data set focuses on risk perceptions of the surveyed individuals [1]. This study involves 10 of the 12 countries from the original study. Based on van der Linden's (2015, 2017) risk perception model, the questions from the survey are grouped based on mental and emotional experiences, the social-cultural perception of risk, and relevant individual differences. This publication highlights some of the main factors contributing to risk perception, as well focus on the most apparent risk-prone countries involved in the study.

3 PROPOSED WORK

One published paper for this data set reviews survey results from five different countries: The UK, Ireland, the United States, Mexico, and Spain. However, data on twelve different countries was collected, the seven additional countries being China, Sweden, Japan, Korea, Italy, Australia, and Germany. We will examine the complete data set as well as correlations of different survey results via clustering and other classification techniques. While previous literature has provided correlations of the survey data, there is much room for additional study.

This study would also like to implement techniques similar to those used by Dryhurst et. al. to cluster the data presented and potentially identify a numeric metric for misinformation related to COVID-19 [1]. This numerical value would be utilized in a similar nature to the ranking of individuals on the Impact of Event scale. This could allow relation of traumatic events and misinformation.

Initial use of clustering has confirmed correlation results found by [3]. By iterating over large amounts of attributes, questions that alone may not reflect significant aspects pertaining to misinformation have been coupled together. This is done k-means clustering methods, with initial cluster locations being chosen by the team. These clusters will be assigned a value as described above to reflect the level of misinformation potential.

Density-based clustering has shown interesting results that reveal diversity among individuals who are susceptible to misinformation regarding COVID-19. So far, it has been critical that a metric measuring "susceptibility to misinformation" is included as an attribute. The development of a predictive model does not seem promising. However, clustering with DBSCAN has already yielded results that provide more nuanced insight into the data than OLS.

3.1 Data Cleaning and Pre-processing

Given the data will be processed using Python, each data set is imported from the available comma-separated values files (csv's) into Pandas DataFrames. String data and floats are used. Filtering of participants who did not accurately report their results (i.e. someone submitting all 7's or all 1's) is accounted for as well. There will also be additional filtering of null values, which will be assessed as valid or invalid to the survey data. The treatment of null values will be evaluated on a case by case basis, as the team is experimenting with several different methods and different sets of attributes.

3.2 Data Integration

The first steps of the data integration involve combining the different data sets into one complete set. This is accomplished more easily via consistent attribute naming, which is done in the data pre-processing phase. As some attributes are the same survey questions but named differently in the .csv files, a standard set of attribute names have been chosen.

After appending all of the various data sources into one large DataFrame, remaining header information from each source was removed. Integrated data is stored as multiple file types of the extension type .csv and .feather. The latter allowed for more rapid loading and manipulation of the data. These files converted each attribute to a floating point value or string, and eliminated invalid answers (i.e. anything outside of the normal polling range).

3.3 Statistical Analysis and Decision Trees

The first analysis will be to explore statistical information of the dataset. Although some of the attributes are either nominal or ordinal, statistical analysis may be useful for the available numerical attributes. Parameters such as mean, median, mode, range, and quantiles will help summarize the dataset where logical do to so.

Another data mining technique is to determine the similarities and dissimilarities between object groups relative to the attributes of the dataset. This process can be performed on any type of attribute and, for the dataset in focus, it may be possible to be performed on a mixed set of attributes. These techniques may be effective at showing how alike or unalike people surveyed in certain countries are in comparison to others. Measuring similarity and dissimilarity will lead to clustering technique which will be discussed in the next section.

Correlation techniques have been employed thus far to determine the relationships between all attributes. It was shown that most of the subsections of attributes (i.e. questions that dealt with the same topic) illustrated high correlation. However, minor relationships between attributes previously thought to be distant were present as well. This provides more knowledge to apply to clustering techniques, such that correlated attributes can be reduced.

A third method of analyzing the data will be to create a decision tree model that could predict categorical labels among the different attributes. Information gain selection will be performed to determine the best attributes to split on and help make the decision trees as effective as possible. This technique may help determine the potential attributes that may be behind scenarios such as the mechanisms that influence the spread of misinformation and the probabilities of one perception affecting another.

3.4 Data Clustering and Presentation

For the COVID-19 survey data analyzed we will be performing clustering on various sets of attributes. We will target differences in perceptions of COVID-19 risks by country. Clustering results can be mapped on to residency and differences in trends can be evaluated. Survey data such as education level and numeracy scores will be interesting attributes to cluster on. It will be important to first identify if there is a correlation between these attributes and select one or the other accordingly. Preliminary correlation analysis has been performed, allowing for a more focused approach to locating valid attributes. Trust in scientists, journalists, and government responses will be other important attributes to include in the model. It is expected that many iterations of the clustering model will be performed, and not all attributes will lead to interesting results, so some may ultimately be omitted. This has been illustrated with some initial correlation analysis.

Both k-means clustering and density-based clustering methods have been identified by the team as essential for accurate manipulation of the data. Frequent iterations will be necessary to allow for implementation of the many attributes present in the data set.

Results will be visualized with histograms for each clustered category and scatter plots to map clusters. An interactive visualization using Plotly will be a potential stretch goal. Possibilities include a dashboard that incorporates maps to demonstrate trends across countries. Based on initial implementation of Plotly graphics and tools, the team feels that this goal is attainable.

4 DATA SET

The dataset is available from https://osf.io/vhnk7/, which includes survey answers from ten thousand participants across eleven different countries. These survey questions include demographic data (seven attributes), 90 questions regarding perceptions of COVID-19 risks, preparedness, information sources, trust in society, political views, and four probability math questions. The final analysis results of the original researchers is also provided. This includes ordinal scores for several categories such as "Trust in government" or "Trust in journalists" summarized from survey results.

The primary data (i.e., original survey answers) is available in fifteen separate csv files (separated by country). There are several different formats among these csv files, so some integration was required. After the initial integration, 12,820 objects were assembled in the data set. As previously described, this integrated data was available for processing in multiple formats. Additionally, the survey questions were separately documented for ease of attribute selection.

5 EVALUATION METHODS

Evaluation methods will include a brief review of summary statistics, clustering, and visualization. The former will consist of a correlation analysis to compare individual categories in survey results. The original study relied on multiple linear regression and an ANOVA to produce results for five countries. While this project will examine all 12 countries surveyed, an important component of this process will be to avoid repeating this work, but it is expected that new results should support those of the original study. These correlation results will inform the selection of attributes used for clustering.

Clustering analysis will be performed to discover trends based on country, education level, and survey answers. This will likely provide the bulk of the project results. Clustering results could be evaluated by retaining a portion of the data set as a test set. After some initial clustering, this method will be utilized to further focus on specific attributes and create classification parameters for more accurate clustering. Standard metrics such as accuracy, sensitivity, specificity, and precision can also be used to quantify this evaluation.

Because we will also be looking into decision tree analysis, the two models can be compared and possibly merged.

An additional stretch goal is to analyze longitudinal results. A small (just under 6280 objects), longitudinal dataset is provided from the same source. This contains survey data collected two months after the first survey and could be used to supplement this work. Additionally, known population responses by country could be investigated. This would require finding a dataset from an additional source, and assuming that our samples are equally representative of the population as a whole. This final step could determine if misinformation susceptibility, as determined by our model, is a predictor of a measurable response such as vaccination rate.

6 TOOLS

As mentioned previously, Python will be the primary resource utilized by the team. This will involve the use of various IDE's across the team, as well as many libraries constructed for Python. These libraries include NumPy (for computations and array manipulation), Pandas (for data manipulation and DataFrame creation), Scikit-Learn (for clustering), Plotly (for plotting and visualization), and potential use of Tkinter (for user interface options). The Feather file type has also been utilized to reduce load times and computational expenses for the large data set analysis. Git and GitHub are used in conjunction for a version-controlled communal code repository, allowing team members to record and review respective version histories and commits. Additional tools include Overleaf for an online interface for group participation in report generation, and Google Documents/Sheets/Presentations for group interaction and further collaborative efforts.

7 PRELIMINARY RESULTS

Results from 15 different surveys were successfully combined, producing a cleaned dataset of 12,802 objects with 104 attributes. After initial integration of the data was achieved, preliminary statistics of the set in its entirety were desired. For nominal, ordinal, and categorical data, this included the total count of tuples in the dataset, the unique values for each attribute, and the top value and their frequencies for each attribute. For numerical data, the count, mean, min, max, standard deviation, and the 4-quartile ranges were computed. In this dataset, the majority were of the discrete numerical data type, with the remainder being nominal and binary types.

7.1 Statistical Analysis

Shown in Figure 1 is a box plot of a numeric attribute, "Num2b", which contains a mathematical probability question for the participants to test their numerical skills. The box plot shows the results of participants, normalized from 0-1, with the red bar being the correct answer normalized. As will be described later, there is a visible correlation between numerical aptitude of participants and their susceptibility to misinformation.

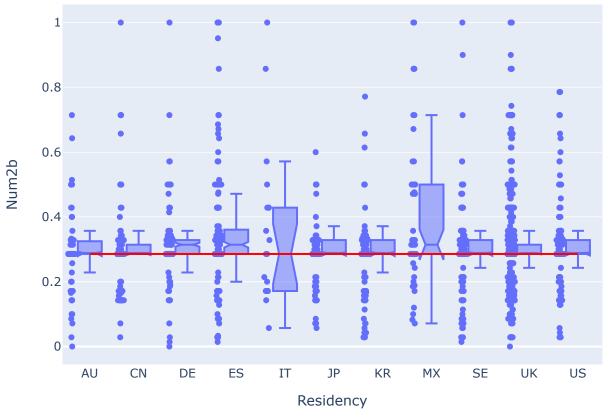
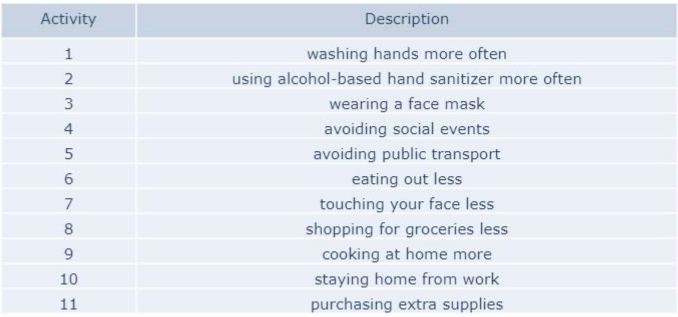


Figure 1: Box Plot of Num2b Attribute

7.2 Apriori Analysis

Apriori analysis was performed on the attribute "Prep", which included tuples of itemsets for the participants. Figure 3 shows the frequent-1 itemsets for each individual country. The goal was to visualize the level of support each country has for each activity shown in Figure 2, so a support of 0.1 was chosen. This analysis was performed to see how participants were following recommendations from their respective government and non-profit organizations. The figure shows that most countries follow recommendations to "wash hands more often" and "avoiding social events", but are dispersed for "wearing a face mask", with East Asia countries such as Japan and Korea with high participation in mask wearing, while western countries such as the United Kingdom had low participation.

Figure 4 shows the same apriori analysis of the preparation attribute, but at a global scale. Preliminary analysis show that there are high supports (near 0.6) for activities such as wearing masks and avoiding social activities, but activities such as mask wearing, staying home from home, and purchasing extra supplies were near or below 0.3. Figure 5 shows the frequent-3 itemsets that exceed the support of 0.4. The plot shows that, globally, participants are more likely to wash hands more often and avoid social events and restaurants,



Activity	Description
1	washing hands more often
2	using alcohol-based hand sanitizer more often
3	wearing a face mask
4	avoiding social events
5	avoiding public transport
6	eating out less
7	touching your face less
8	shopping for groceries less
9	cooking at home more
10	staying home from work
11	purchasing extra supplies

Figure 2: Preparation Attribute Table

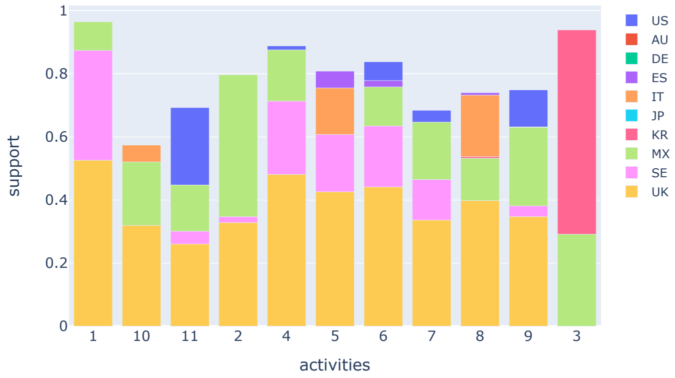


Figure 3: Individual Country Frequent Itemset at Support=0.1

but very unlikely to perform recommendations such as wearing a face mask (which one can argue as being an important activity to follow).

7.3 Clustering Analysis

Some early density-based clustering also yielded interesting results. By utilizing the DBSCAN method, the team constructed clusters with the attributes related to sources of information that participants found, how much they trusted those sources of information, and a question that served as a measure of susceptibility to misinformation, Canada_Q1. The Canada_Q1 attribute polled each participant on "How much do you agree or disagree with the following statements? - Getting sick with the coronavirus/COVID-19 can be serious." This attribute was identified early on by the team as being a focal point for misinformation. The DBSCAN results (shown in Figure 6) for this set of attributes generated six clusters, with clusters 5 and 6 being susceptible to misinformation (cluster 6 was the most susceptible). Interestingly, the most susceptible group had a low trust in social media, while cluster 5 (susceptible but slightly less so) had a higher trust in social media. Of those less susceptible, there was a large diversity in degrees of trust in social media. Susceptible clusters had a low trust in the World Health Organization.

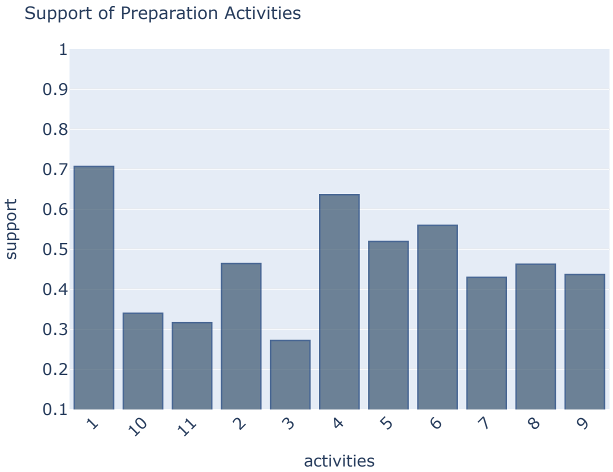


Figure 4: Global Frequent-1 Itemset at Support=0.4

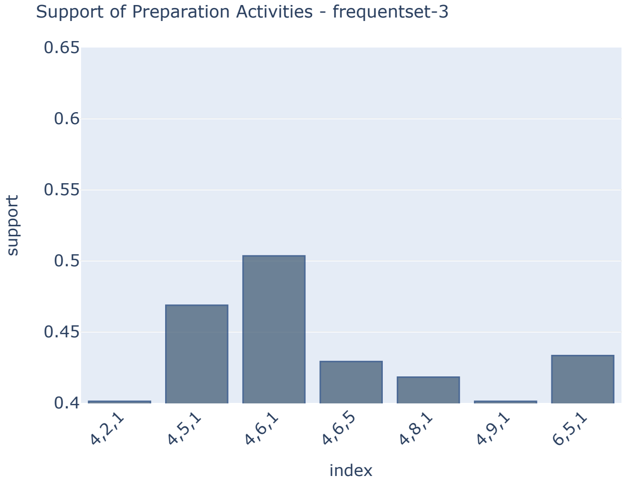


Figure 5: Global Frequent-3 Itemset at Support=0.4

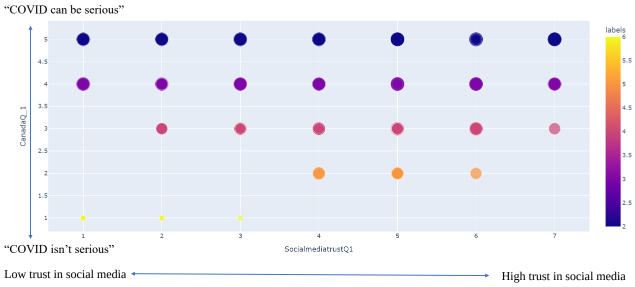


Figure 6: Preliminary DBSCAN results. Clusters are represented by the color scale at the right, and the size of each point represents the degree of trust in the WHO.

8 MILESTONES

Each group member will be involved in completing parts of the study. A detailed milestone assessment has been developed to ensure proper communication and completion throughout the study. The dates laid out below are targeted completion dates.As the team has progressed, various milestones have been achieved, which has been documented in the following sections.

8.1 Milestones Completed

8.1.1 Week of June 21st - Project Part 1.

Reiko: COVID-19 data sets, presentation slides
Son: GitHub repository, presentation slides
Ryan: MongoDB server, presentation slides
Kyle: Overleaf project, presentation slides

8.1.2 Week of July 5th - Project Part 2.

Reiko: Report generation - Data set, literature survey (prior work), evaluation methods, Proposed work (Data Clustering and Presentation)
Son: Report generation - Introduction, Proposed work (Statistical Analysis and Decision Trees), Tools, Milestones
Ryan: Report generation - Tools, Formatting, Editing
Kyle: Report generation - Literature survey section

8.1.3 Week of July 12th - Data integration/processing.

Reiko: Data integration and cleaning
Son and Ryan: Initial exploratory statistics collected
Kyle: Initial k-means exploration

8.1.4 Week of July 19th - Data Processing.

Reiko: DBSCAN clustering
Son: Data processing and visualization
Ryan: Initial Correlation analysis
Kyle: K-means implementation

8.1.5 Week of July 26th - Project Part 3.

Reiko: Other density clustering, data visualization, report updates
Son: Statistical Analysis, Apriori Analysis, report updates
Ryan: Correlation analysis, report updates
Kyle: Classification implementation, report updates

8.2 Milestones To Complete

8.2.1 Week of August 2nd - Final Results.

Reiko: Continued density-based clustering investigation, visualization
Son: Decision tree analysis, visualization
Ryan: Other k-means implementation, visualization
Kyle: Finalization of classifications, visualization

8.2.2 Week of August 9th - Project Parts 4-7.

Reiko: Final report generation, final visualizations
Son: Final report generation, final visualizations
Ryan: Final report generation, final visualizations
Kyle: Final report generation, final visualizations

REFERENCES

[1] Sarah Dryhurst, Claudia Schneider, John Kerr, Alexandra Freeman, and Gabriel Recchia. 2020. Risk perceptions of COVID-19 around the world. (2020).

[2] Rocio Rodriguez-Rey and Helena Garrido-Hernansaiz. 2020. Psychological Impact of COVID-19 in Spain: Early Data Report. (2020).

[3] Jon Roozenbeek, Claudia Schneider, Sarah Dryhurst, John Kerr, Alexandra L J Freeman, Gabriel Recchia, Anne Marthe van der Bles, and Sander van der Linden. 2020. Susceptibility to misinformation about COVID-19 around the world. (2020).