CSCI 5502 Project 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, if there are trends in their political beliefs and in susceptibility to misinformation, peoples beliefs and concerns regarding the pandemic versus other world issues, and the similarities and trends among the different countries. The results of the study illustrated that attributes pertaining to trust had the highest information gain and were thus the main focus for clustering and other analysis. The attribute Canada_Q1 and multiple worry and affected attributes were also prioritized. Conclusions stated that social media trust somewhat indicated higher susceptibility, while general low trust in groups and government indicated higher susceptibility as well. Only 3.8 percent of the dataset was assessed as susceptible to misinformation, and that small group was very diverse which limited classification and model generation. Political affiliation was also diverse in this set, preventing strong assessment of susceptibility. Each country also was assessed for support for various COVID preparation activities, and the more predominant mask-wearing technique was found to be lower-supported in the US, UK, and Sweden.

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

*This report is the final part 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

ACM Reference Format:
Son Pham, Kyle Rogers, Reiko Matsuda-Dunn, and Ryan Karasopoulos. 2021. CSCI 5502 Project 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, 15 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. [4]. 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.

The team was concerned first and foremost with aspects that impacted peoples views significantly and thus increased the potential for susceptibility. These included observing political affiliation and social media habits, as the team felt these had great impact after preliminary studies. Also, impact globally and how responses varied on a per country basis were important to the team as differences in preventative actions and perceptions were of interest. Finally, comparison of perception of significance of COVID relatively to other international issues was use to tie multiple trust and global aspects together, giving the team a more in-depth understanding.

Visualizations for this study can be viewed in the referenced link created by the group [2]. Some description of the visualizations is described in the presentation video present on the team's GitHub page.

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. A more thorough analysis of the data for the final results confirmed initial hypothesis involving specific correlations, while also reducing significance of other attributes based on their reduced correlation.

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 [4]. 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 [3]. 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 [4]. 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 [4]. 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 RESULTS

Results from 15 different surveys were successfully combined, producing a cleaned dataset of 12,744 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. In addition, clustering techniques were applied to the data to allow for advanced understanding of trends and posed questions for the data. The clustering methods focused on by the team were k-means clustering and density-based clustering, and while other methods were considered they were not heavily used. Various classification techniques were also performed to provide insight for attributes to cluster on or perform statistical analyses on. Frequent itemsets were determined using the Apriori algorithm, and Bayesian classification was executed to search for potential relationships between independent attributes.

One question on the survey 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 an indicator of susceptibility to misinformation - specifically those who responded "Not Severe" to this question would be more susceptible to misinformation.

7.1 Statistical Analysis

Before any analysis took place, it was helpful to understand the dataset and any potential biases or skews it may contain. In terms of demographic information, the United Kingdom was vastly over-represented in terms of the number of respondents; the UK had 5,099 data points whereas Spain, the next largest country category, had only 1,381 data points, with the rest of the countries averaging only about 750 participants. As such, the results of all analyses involving summarized data will skew towards a UK perspective. Politically the data is rather diverse, following an approximate normal curve centered around a Neutral political perspective (4 on a scale of 1-7), though leaning slightly Liberal.

Though this particular dataset contains only a few thousand data points, there are over 100 distinct attributes to parse through in the search for interesting patterns. Manually trying all combinations of attributes in the hopes of finding an enticing correlation is logistically infeasible, so the primary first concern is to gather preliminary data to try to ascertain related or meaningful connections between attributes before more complicated clustering or categorization analysis is to be performed. To start, the correlation coefficient can be easily computed on all numeric columns. After dropping columns with strings or date stamps (Date, Residency, etc.), the basic correlations were computed, and made into a reference table. The results from this initial simplified attempt were unsurprisingly circumspect, as many questions in the survey were almost identical to one another, yielding the highest correlations between questions of the same class (e.g. one question asking participants for their exact age, and another asking them a multiple-choice question on their age range). These classes included trust in various social groups, trust in government or formal entities like the WHO, and feelings towards the government response to Coronavirus over varying time periods. To find non-obvious correlations between questions, when one cluster of questions' maximum correlation was between another question in the same class, one of those attributes was removed, and the correlation table recomputed. This process was repeated a number of times until the maximum correlation between attributes was mostly in other attribute classes. From this, the largest correlation coefficient values, both positive and negative, were extracted into a simplified reference table, the first few columns of which can be seen in Figure 1. The full, unsimplified correlation chart was also generated and used as a reference guide.

Attribute	Max	MaxVal	Min	MinVal
DemGen	CultCog_1	0.12	FinitePool_2	-0.09
quota_age	DemAge	0.97	Personal_2	-0.09
GenSocTrust	Trustingroups_6	0.31	COVIDeffect_1	-0.11
Trustingroups_1	FriendstrustQ1	0.23	Personal_1	-0.09
Trustingroups_6	GenSocTrust	0.31	FinitePool_4	-0.29
Trustingroups_11	Govresponse_8	0.49	CultCog_1	-0.22
COVIDexp	SARS	0.17	COVIDeffect_4	-0.22
COVIDeffect_1	Personal_2	0.54	GenSocTrust	-0.11
COVIDeffect_4	COVIDeffect_1	0.24	COVIDexp	-0.22
SARS	DemHealthcare	0.2	COVIDeffect_4	-0.21
CultCog_1	Govrestrict_1	0.22	govtrustQ1	-0.24
prosocial	WHOtrustQ1	0.27	CultCog_1	-0.11
CanadaQ_1	FinitePool_2	0.45	CanadaQ_3	-0.2
CanadaQ_2	COVIDeffect_4	0.21	COVIDexp	-0.17

Figure 1: First Rows of Simplified Correlation Reference Chart

Though this analysis was done by eye with little regard to formal process, this new reference table nevertheless contains information relating to potential interactions between classes of questions, and proved useful as a starting point in the future analysis. From this process, we can see that the maximum correlation between what we decided would be the benchmark for assessing misinformation ("Coronavirus Infection can be Serious [1-7]") is most highly correlated with the column addressing individuals' personal worry about Coronavirus, with a correlation coefficient of about 0.45, implying a connection between the objective severity of viral infection and how the respondents felt about their own safety surrounding the pandemic. A breakdown of how the participants' worry over Coronavirus correlated with their perceived danger can be found in Figure 2. Notably, most participants thought of Coronavirus infection as being a serious condition, despite their worry about becoming infected. Perhaps more interesting is the correlation between susceptibility to

misinformation and whether or not participants believed that their country would be badly affected by the pandemic. Figure 2 depicts the slight negative correlation between respondents' acknowledgement of the severity of COVID-19 to their belief in the level at which the pandemic will infect those in their home country.

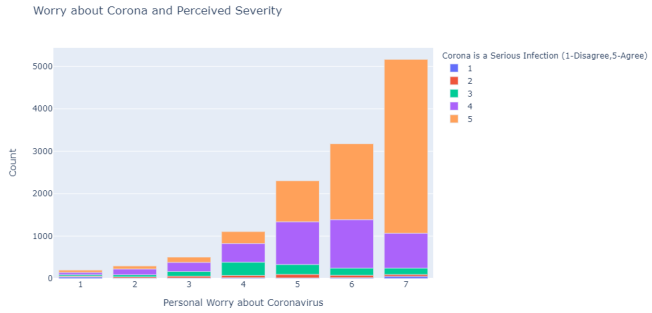

Figure 2: Bar Chart Comparing Worry to Perceived Severity

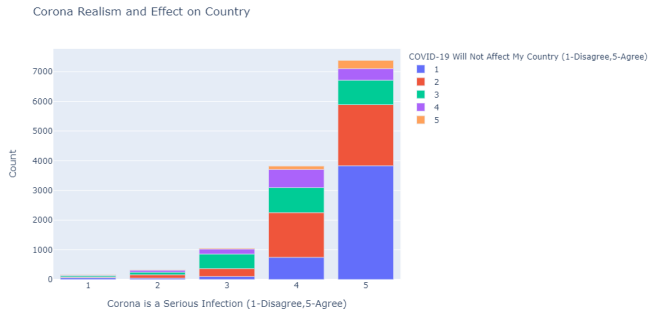

Figure 3: Bar Chart Comparing Perceived Severity to Perceived Effect on Home Country

Shown in Figure 4 is a box plot of a numeric attribute, "Num2b", which contains a mathematical probability question for the participants to test their numerical skill set. The box plot shows the results from participants, normalized from 0-1, with the red bar being the correct answer. As will be described later, there is a visible correlation between numerical aptitude of participants and their susceptibility to misinformation. Figure 4 shows IT and MX having a wide 25-75 percent quartile range, though the median results were at or near the correct answer. UK shows more data points due to the dataset containing much more participants residing in the country compared to others. Figure 5 shows a more interesting pattern among the different countries. Similar trends is shown with IT and MX, where participants are more likely to provide the incorrect answers. Additionally, there is a deviation with AU from the correct answer. There were three more numerical attributes, two (Num1 and Num2a) of which does not show any new information, and one (NumeracyQ1) which has not been analyzed. Four of the numerical attributes can be viewed in the visualization website [2].

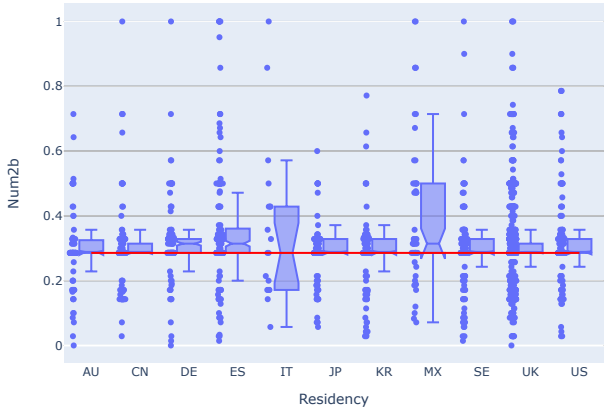

Figure 4: Box Plot of Num2b Attribute

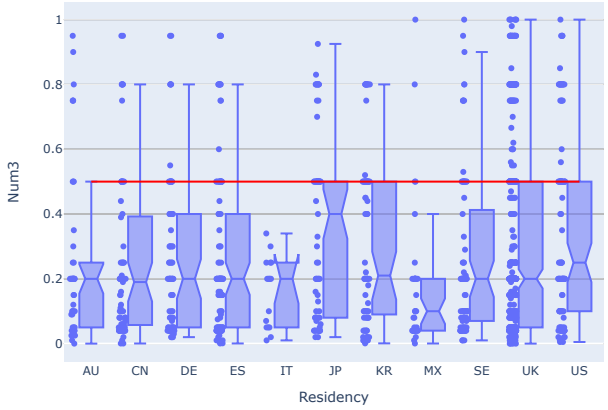

Figure 5: Box Plot of Num3 Attribute

7.2 Apriori Analysis

Apriori analysis was performed on the attribute "Prep", which included tuples of itemsets from the participants which can be mined for frequent itemsets. Figure 7 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 6, so a support of 0.01 was chosen. This analysis was performed to see how participants were following recommendations from their respective government and global agencies, possibly correlating to their trust in these recommendations. The figure shows that most countries follow recommendations to "wash hands more often" and "avoiding social events", but may not trust that "wearing a face mask" may help prevent spreading and contacting the virus. The figure shows that East Asia countries such as Japan and Korea have high participation in mask wearing, while western countries such as the United Kingdom, Sweden, Mexico, and the United States had low participation. The US and Sweden, in particularly, have only mask-wearing supports of 0.0704 and 0.0744, respectively. Currently, the US, SE, and the UK remain among some of top the countries with the highest COVID-19 cases per 1 million population. This exploratory analysis supports the theory that misinformation, in this case the idea that masks will or will not help protect against the virus, affects the progression of the pandemic. Further analysis using decision trees and clustering will help answer the questions of how susceptibility to misinformation can be predicted.

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 6: Preparation Attribute Table

Figure 8 shows the same apriori analysis of the preparation attribute, but at a global scale. Preliminary analysis shows that there are high global supports (near 0.8) for activities such as washing hands more often and

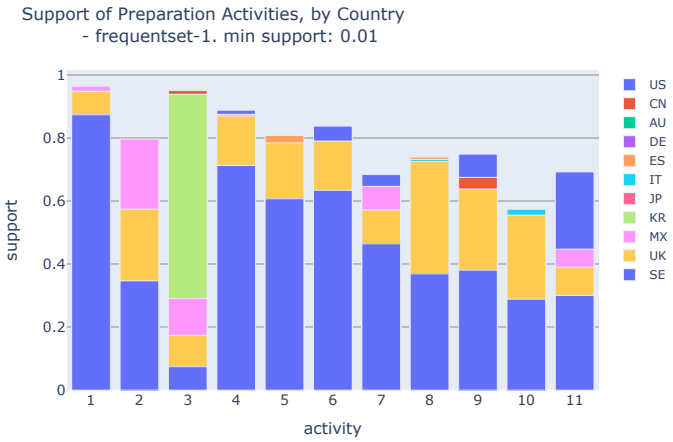

Figure 7: Individual Country Frequent Itemset at Support=0.1

avoiding social activities, but activities such as mask wearing, staying home from work, and purchasing extra supplies were near or below 0.4. Figure 9 shows the frequent-2 and 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, but very unlikely to perform recommendations such as wearing a face mask. The visualization webpage will allow the user to visually perform apriori data mining [2].

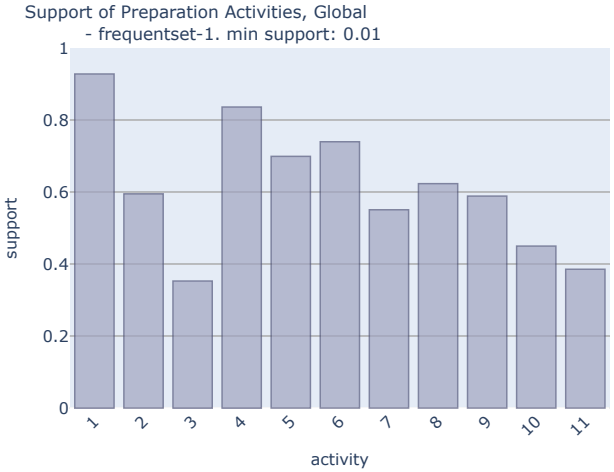

Figure 8: Global Frequent-1 Itemset at Support=0.01

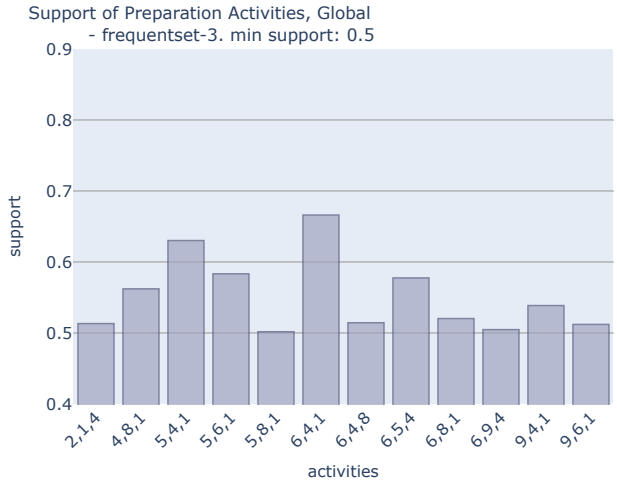

Figure 9: Global Frequent-3 Itemset at Support=0.4

attribute	gain
PosterstrustQ1	0.0535
WHOtrustQ1	0.0454
workplacetrustQ1	0.0367
SocialmediatrustQ1	0.0326
FinitePool_2	0.0236
FriendstrustQ1	0.0221
Govrestrict_3	-0.0072
Trustingroups_9	-0.0073
Politics	-0.0076
FinitePool_3	-0.0081
Personal_6	-0.0082
Friends_6	-0.0094

Figure 10: Information Gain Table of Six highest and Six lowest

7.3 Bayesian Analysis

Decision tree and Bayesian analysis was performed to determine whether a classification model can be built to predict which attribute can contribute to the susceptibility to misinformation. The table on Figure 10 shows the information gain values for the highest six attributes and lowest six attributes. It is interesting to note that the attributes with the highest information to be gained from branching are ones relating to trust. Because PostertrustQ1 has the highest information gain among the attribute, it would be selected as the splitting attribute.

Current analysis shown in Figures 11, 12, 13 used Vaccine1, Vaccine2, and CanadaQ1 as possible classes to classify to. Figure 11 shows the probability that a person is willing to take the vaccine if one was available, willing to recommend to others to take the vaccine, or neither. A note to point out here is that the majority of the population is willing to accept that the vaccine will help cure the virus, and only a minority disagrees.

Figure 12 shows the probability relative to the age group attribute. If a person were to accept the vaccine, the probability that the age groups whom would be accepting would be age 45-54, or more broadly, from 25-64. There seem to be a lower probability of teenagers and young adults accepting the vaccine. Figure 13 is analysis done on the CanadaQ1 attribute, separated by age group. Similarly, the age groups who would be most

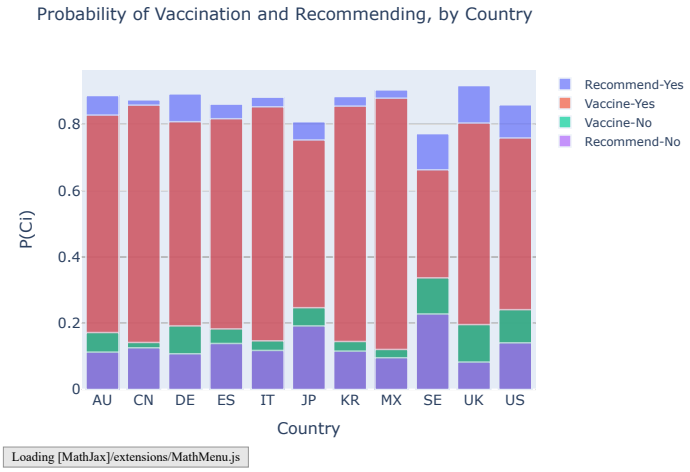

Figure 11: Probability of a Person Accepting a vaccine and Recommend to Others vs. Not Accepting or Recommending, By Country

likely to believe that the virus is serious are groups 45-64, and the lowest probability is with the 18-24 age group.

7.4 Clustering Analysis

7.4.1 *Density-based clustering.*

Density-based clustering (DBSCAN) was applied to find patterns in susceptibility to COVID misinformation

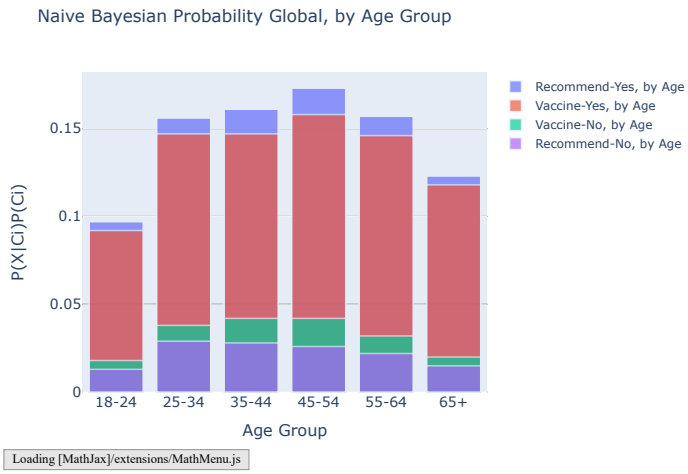

Figure 12: Naive Bayesian Probability of VaccineQ1, by Age Group

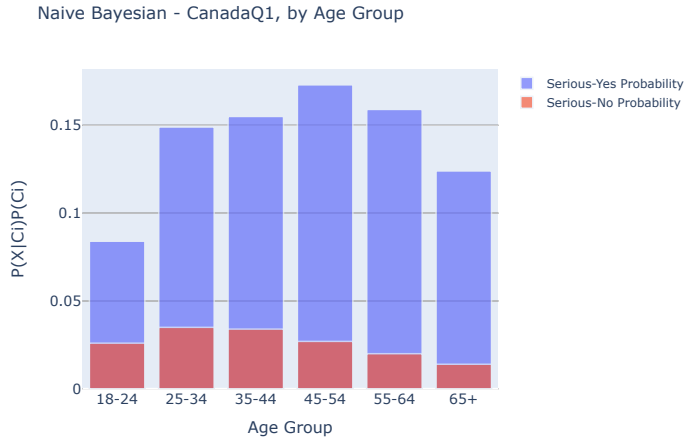

Figure 13: Naive Bayesian Probability of CanadaQ1, by Age Group

and trust in sources of information and political viewpoints. Two separate clustering analyses were made: one focusing information sources and another focusing on politics. The former used attributes pertaining to information sources, how much they trusted those sources of information, and Canada_Q1, the metric of misinformation susceptibility. The latter political analysis used attributes representing opinions on the appropriate degree of government control in individual lives, self-identified political leanings, balancing personal cost with social benefits, and Canada_Q1. The initial set of attributes used was then reduced based on correlations found between attributes (see **??** in 10). Some with attributes were retained if they had a weak correlation to others (r<0.4) in the set and contributed to a greater separation of clusters. Various values of epsilon between 0.5 and 2 were experimented with. In both analyses, an epsilon of 1 created the best separation of the susceptible groups (answered 1 and 2 for Canada_Q1).

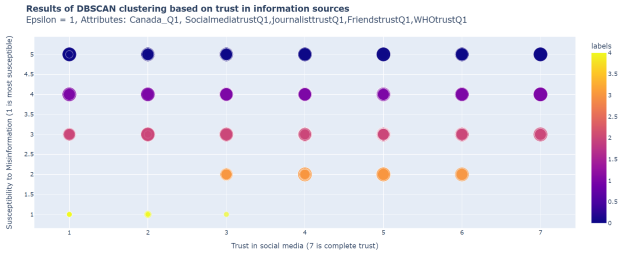

Figure 14: DBSCAN results on information source attributes. Clusters are represented by the color scale at the right, and the size of each point represents the degree of trust in the WHO.

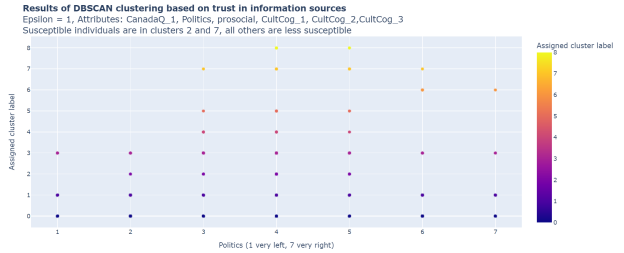

Figure 15: DBSCAN results on political attributes.

The DBSCAN results (shown in Figure 14) for attributes related to information sources generated five clusters, with clusters 1 and 2 being susceptible to misinformation (cluster 1 was the most susceptible). Distributions of each cluster were examined, and those of clusters 1 and 2 can be seen in 16 and 17, respectively. All others are available in the appendix. Interestingly, the most susceptible group had a somewhat low trust in social media, while cluster 2 (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 all other information sources examined.

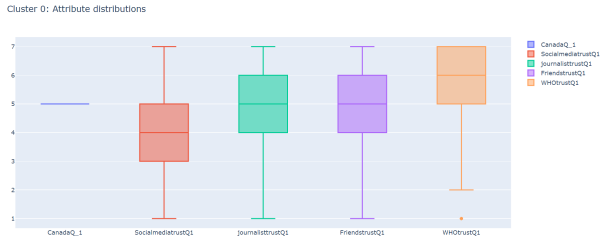

Figure 16: Distribution of trust in information sources for cluster 1, the most susceptible to misinformation.

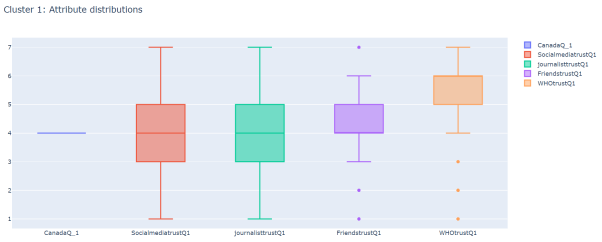

Figure 17: Distribution of trust in information sources for cluster 2, which is somewhat susceptible to misinformation.

The DBSCAN results for attributes related to political views (shown in Figure 15) generated eight clusters, with clusters 2 and 7 being susceptible to misinformation. Distributions of each cluster were examined, and those of clusters 2 and 7 can be seen in 18 and 19, respectively. All others are available in the appendix. Cluster 2 had relatively centrist views, and in these attributes this group was not significantly different from the entire dataset. Cluster 7 participants identified as a bit further right-wing, but interestingly also indicated that they believed in greater government intervention in everyday life.

7.4.2 K-means clustering.

Additional clustering results were obtained. A k-means algorithm was implemented and many low-dimensional analyses were performed. These revealed a relationships between worry around COVID and social media use, trust in the government and the WHO, and consistent faith in reported estimates of COVID metrics.

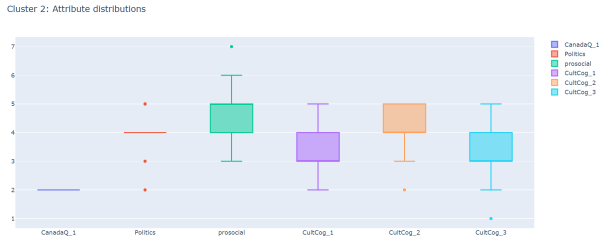

Figure 18: Distribution of political attributes for cluster 2. This group is susceptible to misinformation and has relatively centered political views.

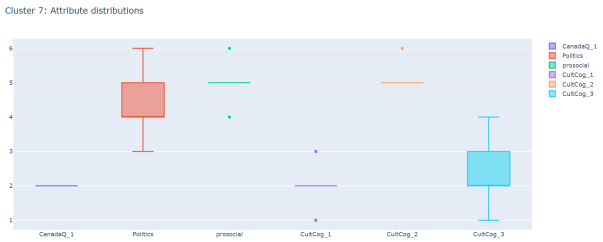

Figure 19: Distribution of political attributes for cluster 7. This group is susceptible to misinformation and has self-identified right-leaning political views, but a belief in more government intervention in everyday life.

8 APPLICATIONS

After obtaining detailed results via classification, clustering, and statistical analyses, the team assessed the information gained from the study. From the basic statistical information, it was determined that a proper understanding of the severity of Coronavirus infection was most positively correlated with participants' personal worry about the pandemic, and most negatively correlated with their perception of infection rates in their home countries. Analyzing the data for individual countries and political affiliations revealed that misinformation was equivalently uncorrelated with both nationality and political affiliation across the board, indicating that misinformation is a global phenomenon, regardless of party stance or state allegiance. The lack of large correlations, positive or negative, between attributes indicates that a more nuanced approach to determining susceptibility to misinformation is in order, and current prejudices that attribute one's belief in misinformation to political standing or national origin are unjustified for those represented by this data set.

The Bayesian classification performed yielded results in information gain that allowed for more accurate determination of splitting metrics for clustering. The gain values illustrated that attributes related to trust illustrated the highest gain, and were thus one of the focal points for the k-means clustering approach.

Density-based clustering showed that the groups that are susceptible to misinformation regarding COVID can be quite diverse in their political views. Some can be defined as identifying as more conservative, but the distribution in political leanings is quite large. This indicates that a more nuanced understanding of the factors driving this susceptibility should be employed. Interestingly, those that were most susceptible to misinformation had very low trust levels in all sources of misinformation, even their friends. These individuals may be particularly resistant to any form of information sharing.

9 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.

9.1 Milestones Completed

9.1.1 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

9.1.2 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

9.1.3 Week of July 12th - Data integration/processing.
Reiko: Data integration and cleaning
Son and Ryan: Initial exploratory statistics collected
Kyle: Initial k-means exploration

9.1.4 Week of July 19th - Data Processing.
Reiko: DBSCAN clustering
Son: Data processing and visualization
Ryan: Initial Correlation analysis
Kyle: K-means implementation

9.1.5 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

9.1.6 Week of August 2nd - Final Results.
Reiko: Continued density-based clustering investigation, visualization
Son: Bayesian analysis, visualization
Ryan: Other k-means implementation, visualization
Kyle: Finalization of classifications, visualization

9.1.7 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] Son Pham, Reiko Matsuda-Dunn, Kyle Rogers, and Ryan Karasopoulos. [n.d.]. COVID-19 Misinformation Project Visualizations. Retrieved Aug 11, 2021 from https://pharsalus.herokuapp.com/
[3] Rocio Rodriguez-Rey and Helena Garrido-Hernansaiz. 2020. Psychological Impact of COVID-19 in Spain: Early Data Report. (2020).
[4] 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).

10 APPENDIX

Figure 20: Table of attributes used in DBSCAN analyses with descriptions and whether or not they were included after assessing correlation coefficients.

Attribute Name	Description	Selected for which Analysis	Used in Final Clustering
CanadaQ_1	How much do you agree or disagree with the following statements? - Getting sick with the coronavirus/COVID-19 can be serious.	Both	Yes
MediaExp_2	Have you come across information about coronavirus/COVID-19 from: - Social media or online blogs from individuals	Information sources	No
MediaExp_3	Have you come across information about coronavirus/COVID-19 from: - Journalists and commentators in the media (TV, radio, newspapers)	Information sources	No
MediaExp_4	Have you come across information about coronavirus/COVID-19 from: - Government or official sources such as websites or public speeches/broadcasts within the country you are living in	Information sources	No
MediaExp_5	Have you come across information about coronavirus/COVID-19 from: - Official messages from your place of work or education	Information sources	No
MediaExp_6	Have you come across information about coronavirus/COVID-19 from: - Friends and family	Information sources	No
MediaExp_7	Have you come across information about coronavirus/COVID-19 from: - World Health Organisation	Information sources	No
SocialmediatrustQ1	Thinking about the coronavirus/COVID-19 information that you saw on social media, how much did you trust the information?	Information sources	Yes
JournalisttrustQ1	Thinking about the coronavirus/COVID-19 information that you saw from journalists and commentators in the media, how much did you trust the information?	Information sources	Yes
govtrustQ1	Thinking about the coronavirus/COVID-19 information that you saw from the government or official sources in the country you are living in, how much did you trust the information?	Information sources	No
FriendstrustQ1	Thinking about the coronavirus/COVID-19 information that you saw from friends and family, how much did you trust the information?	Information sources	Yes
WHOtrustQ1	Thinking about the coronavirus/COVID-19 information that you saw from the World Health Organisation, how much did you trust the information?	Information sources	Yes
Politics	Where do you feel your political views lie on a spectrum of left wing (or liberal) to right wing (or conservative)?	Political views	Yes
prosocial	To what extent do you think it is important to do things for the benefit of others and society even if they have some costs to you personally?	Political views	Yes
CultCog_1	How strongly do you agree or disagree with each of these statements? - The government interferes far too much in our everyday lives.	Political views	Yes
CultCog_2	How strongly do you agree or disagree with each of these statements? - Sometimes government needs to make laws that keep people from hurting themselves.	Political views	Yes
CultCog_3	How strongly do you agree or disagree with each of these statements? - It's not the government's business to try to protect people from themselves.	Political views	Yes
CultCog_4	How strongly do you agree or disagree with each of these statements? - The government should stop telling people how to live their lives.	Political views	No
CultCog_5	How strongly do you agree or disagree with each of these statements? - The government should do more to advance society's goals, even if that means limiting the freedom and choices of individuals.	Political views	No
CultCog_6	How strongly do you agree or disagree with each of these statements? - Government should put limits on the choices individuals can make so they don't get in the way of what's good for society.	Political views	No

The following figures include attribute distributions of clusters resulting from applying the DBSCAN algorithm to attributes related to trust in information sources:

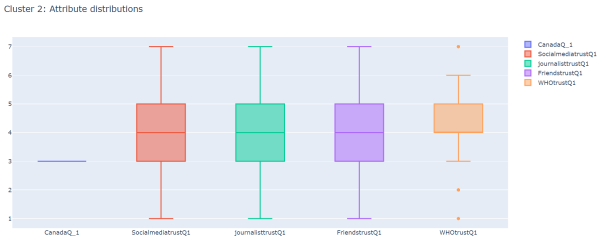

Figure 21: Cluster 3

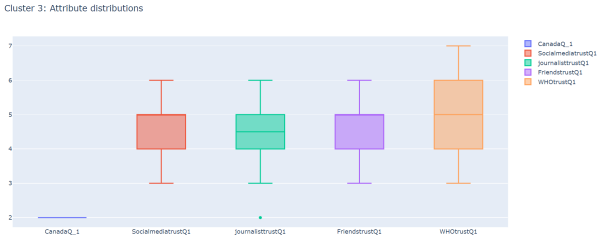

Figure 22: Cluster 4

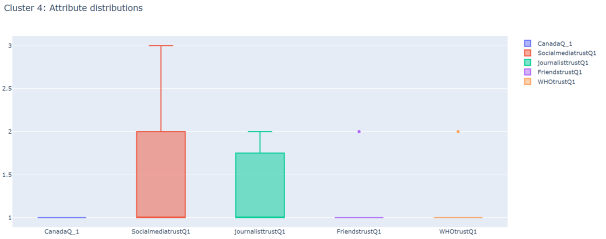

Figure 23: Cluster 5

The following figures include attribute distributions of clusters resulting from applying the DBSCAN algorithm to attributes related to political views:

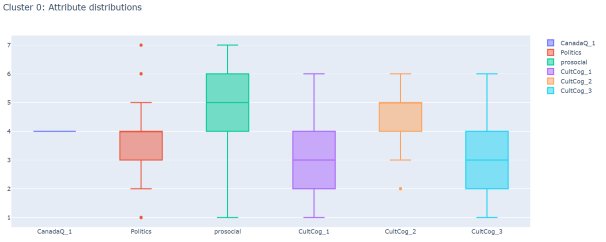

Figure 24: Cluster 0

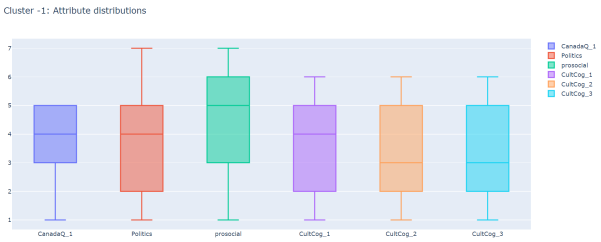

Figure 25: Cluster 1

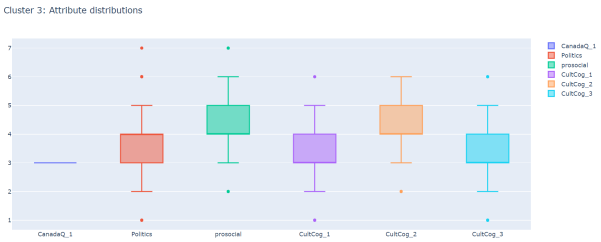

Figure 26: Cluster 3

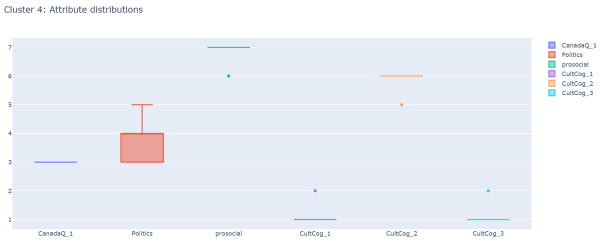

Figure 27: Cluster 4

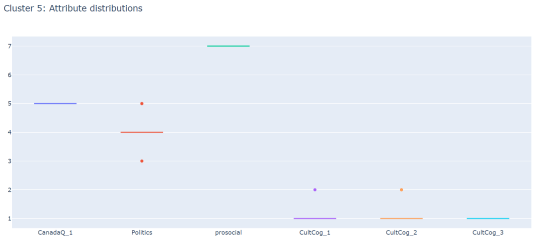

Figure 28: Cluster 5

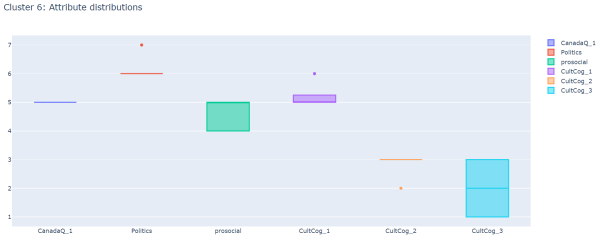

Figure 29: Cluster 6

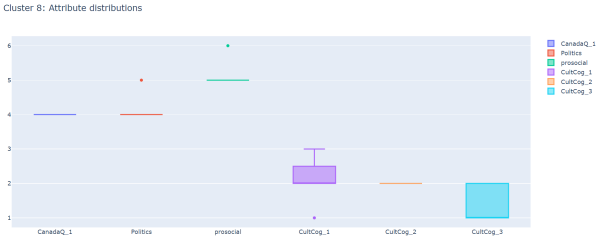

Figure 30: Cluster 8