Causal Analysis and the impact of COVID-19 on Reddit Posts

Drishika Dey
Supriya Chaudhary
Madhumitha Rajarajan
Harsh Mishra
University of Illinois at Chicago
USA

1 ABSTRACT

Demystifying the characteristics of social media users' sentiments is a crucial requirement to setting up immediate responses to healthcare events. With adverse healthcare situations, such as the COVID-19 pandemic, public sentiment can determine crisis management steps - such as setting up a lockdown restriction. Social media platforms such as Reddit often have consolidated bulletin pages (subReddits) for specific topics, such as the COVID posts related to a particular country (for e.g., r/coronavirusUK). Our current work focuses on modeling Reddit users' posts, to find causal relations with COVID-19 infections and deaths. Our work is preceded by a literature survey in similar fields of causality, social media attention, COVID-19 and epidemiology. To do this, we begin with exploratory data analysis and cleaning of SocialGrep's data. This is followed by modeling a pipeline similar to [9]. Such a pipeline would require an algorithm to find a causal graph such as NO TEARS [28], followed by utilising Bayesian Networks to find the conditional probabilities. Further comparisons are made with similar algorithms NoCurl [27] and No Fears [23] . We utilise Area Under the Receiver Operating Characteristic as the metric to evaluate the same. Our results demonstrate that causal methods can effectively be used to identify factors that influence public sentiment. The complete implementation of our pipeline can be found on Github: https://github.com/harshm16/causal_inference_reddit_data

2 INTRODUCTION

COVID-19 is an ongoing global pandemic and it has affected more than 639 Million people and caused more than 6.54 Million deaths as of November 2, 2022, making it one of the deadliest in history. It was deemed an international public health emergency on January 30, 2020, and a pandemic on March 11, 2020, by the World Health Organization (WHO) [16].

Social media channels have become the most common resource for the public to share their experience and emotion during the pandemic. Many researchers ([9], [14], [20]) have used causal models in order to identify the spread and the impact of COVID-19 using social media datasets. Social media has played a huge role in risk communication as well as risk mitigation by delivering up-to-date information on the pandemic and its flow throughout the world.

For effective crisis management especially during adverse situation such as COVID-19, it is crucial to comprehend the characteristics of public opinion and their sentiment reflected through their posts on social media, especially in areas which had a higher amount of cases and which were more susceptible to negative sentiment. Predicting causal sentiment can move the focus from warning people to calming people and improve catastrophe preparedness. Although numerous studies have exploited Twitter data in predictive context during COVID-19 pandemic [9] [21] [22] [10]; Causal analysis of public sentiment has not yet been explored in other social media platforms. Such a research could potentially help with crisis management , while viewing the user's activity online. According to Statista, Reddit is ranked as one of the popular social networks globally and has over 430 million active users.[7]. Reddit visitors are mainly from the USA, UK, Canada, Australia, and Germany. In this study, we aim to propose a causal method to analyze the causal relationships between the pandemic characteristics such as number of deaths and Reddit posts in terms of the public sentiment.

We are motivated to find the causal connections between statistics such as the increase in infections or deaths and Reddit sentiment or activity. Further, we are interested in learning multiple causal methods of causal discovery and applying them to our use-case. Additionally, we proceed to compare our proposed solutions to find the optimal causal model.

3 RELATED WORK

A few studies conducted prior to our analysis were mainly used to address social media and its impact on COVID-19. A study on sentiment analysis on social media was conducted by Neri and Aliprandi [3]. This study correlates findings from a public broadcasting service and a private company by analysis at both the theoretical and empirical levels, examining political communication in the media. Sentiment Analysis of text data is often done using DistilBERT [25], a distilled version of the BERT model. It uses knowledge distillation, a process wherein a smaller, more condensed model is trained to replicate the behavior of a larger model to extract the sentiment. We also notice that negative tweets have a high polarity, and users from public organisations that tend to tweet negatively during Covid-19 gain a larger following, according to [26]. Pomama, Cessa and Ciaccio [17] defined the association between adverse events following AEFI immunization and COVID-19 vaccination by creating a workable methodology, following the basic structure of the World Health Organization (WHO). They considered potential causal association between the administration of the COVID-19 vaccination and AEFI based on their suggested model.

Ma et al.[15] and Guo et al.[11] talk about the causal impact of the different types of policies set in place during the pandemic. They investigated the topic of determining the causal relationships between various COVID-19 related policies and the dynamics of the epidemic in various counties at any given moment. In order to achieve this, they combined and analyzed data on various COVID-19-related policies (treatments) and epidemic dynamics (outcomes) for various United States counties across time. Bhaskaran and Bacon [6] conducted a study by fitting age and sex-adjusted logistic models for the two outcomes namely COVID-19 and non-COVID deaths, associations between individual-level characteristics were also estimated. These deaths were categorized based on if underlying cause of death was listed as a COVID-19 code on the death certificate.

On this same subject of covid related deaths, an article by Bajaj, Gadi, Spihlman, Wu, Choi and Moulton [4] shows that every system in the body experiences ongoing biological changes as we age, and the immune system is no exception. Some of these changes cause the immune system to perform less effectively, as shown by an increased vulnerability to respiratory illnesses like the flu and new coronaviruses. However, younger people have a larger variety of naive immune cells that give them the capacity to fight off new infections and successfully respond to foreign antigens, leading to milder disease states or even asymptomatic infection, as is seen in the majority of younger people who test positive for the novel coronavirus. Hence, age plays a huge part in the response of the host body to covid infections, something that is reflected in our project as well.

Sarker and Lakamana [18] investigated people who tested positive for COVID-19 self-reporting symptoms on Twitter. Our main paper Gencoglu et. al., [9] monitored Twitter activity during COVID-19. This study successfully captures domain knowledge and identifies features that affects public attention and sentiment of COVID-19. Their results show generalizability of causal inference for their chosen countries with high accuracy. They discuss the factors affecting Twitter activity and sentiment in the early stages of the pandemic. Similar to this work, we distinguished causal relations between Reddit posts, COVID infections and deaths.

The algorithm we referred to was DAGs with NO TEARS [28] affecting Twitter activity. In this, the structure learning issue is formulated as a purely continuous optimization problem over real matrices, which completely escapes this combinatorial limitation. Zheng, Aragam, Pradeep Ravikumar and Xing accomplished this by using a unique, smooth, and accurate characterization of acyclicity and the graph is not subjected to any structural constraints, such as bounded treewidth or in-degree.

Kaiser and Sipos [12] further argued that these techniques are ineffective for deducing true causality from empirical data. The NOTEARS algorithm does not perform any invariance checks with regard to the data parameterization. For instance, how the procedure responds when the units or scale of some of the variables are changed. The edge orientation will be skewed toward explaining variables with more variance if the variables are on distinct scales (with different variances) and the resulting edge orientation in the related DAG will be random if the variables have the same variance.

To address this, Yu, Gao, Yin, and Ji [1] introduced the DAG-NoCurl method, a new technique that effectively solves the optimization problem in two steps: The Hodge decomposition of graphs is used to learn an acyclic graph by projecting the cyclic graph to the gradient of a potential function and then the Hodge decomposition of graphs is first used to obtain an initial cyclic solution to the optimization issue.

NOFEAR, a different solution, extends current algebraic characterizations of acyclicity to a class of matrix polynomials. The Karush-Kuhn-Tucker (KKT) optimality criteria [5] for the NO TEARS formulation are then demonstrated to be unsatisfiable, save in a simple situation, which explains a behavior of the related algorithm. This is done by focusing on a one-parameter-per-edge configuration. The KKT conditions for an identical reformulation are then derived by Wei, Gao, and Yu [24]. They demonstrate their need and tie them to explicit constraints that specific edges must not exist in the graph.

4 FORMAL PROBLEM DESCRIPTION

Reddit being one of the most popular social media platforms, and a source of news during the pandemic, causal analysis of the pandemic characteristics with respect to the public sentiment through social media activity is still an area that is remained to be explored.

While traditional machine learning techniques are an effective way to model predictive patterns and for hypothesis discovery through correlation, clustering or exploratory analysis, causal inference of related phenomena is not achievable without causal computational modeling. The main aim of this project is to establish a causal relationship in the context of social media and pandemic characteristics to help in maximizing the spread of accurate information to the public as well as to enhance disaster preparedness for similar future scenarios. From the inference of causal impact on Reddit activity can aid in unraveling correlations between the pandemic variables.

We propose a structural causal modeling of Reddit activity in order to understand the underlying causes affecting the public decision making through social media during a pandemic. To evaluate our idea, we used SocialGrep [19] to collect data containing Reddit posts spanning over 57 days in 7 countries selected and categorize several attributes of COVID-19 pandemic that might have an effect on Reddit activity.This includes important attributes like Sentiment and Reddit usage. Previous causal modeling of Reddit posts has not been done. Further the problem also must cover external attributes such as Single Households and Lockdown Announcements which may affect the Causal Model. All the data is collected for the countries which utilise Reddit the most.

Further our problem is two-fold. Firstly, a Causal Discovery problem where we find the Causal Structural Model via a method to discover Directed Acyclic Graph. To model the DAG we use the following notation , Let V denote a set of d numbers of random variables, $X = (X_1, ..., X_d) \in \mathbb{R}^d$ be an observation on V, and $\mathbb{D}$ denotes the space of DAGs $\mathcal{G} = (V, E)$ on V, we aim to learn a DAG $\mathcal{G} \in \mathbb{D}$.

Secondly, we also propose a Causal Bench marking problem by comparing multiple Causal Structural learning algorithms such as NO TEARS [28], NoCurl [27], and No Fears [23]. Finally we also wish to compare the Causal Models via evaluation metrics and find interesting Causal Effects between Reddit Activity or Sentiment and the other variables in our Bayesian Network.

5 OUR SOLUTION

We propose a solution for our problem with the three causal structure discovery algorithms mentioned below. Our implementations for these models were taken from the the authors' official repositories [28] [27] [23] . Our results include the comparisons of the three models as well.

NO TEARS DAG: We hypothesize that daily Reddit activity and sentiment during COVID-19 has a causal relationship with the characteristics of the pandemic (our features 1-6). In order to validate our approach, we first use a structure learning algorithm (NO TEARS [28]), to get the causal diagram G that describes the conditional dependencies between variables in our dataset. So a typical formulation for generalized linear model is given by the Structural equation model where a weighted adjacency matrix defines the graph. This basically formulates the operation on the continuous spaces instead of discrete space of DAGs. That is,

$$\begin{aligned} \min_{W \in \mathbb{R}^{n \times n}} \; & L(W) \\ \text{s.t. } & G(W) \in \mathrm{DAGs} \end{aligned} \tag{1}$$

where *G(W)* is the *n*-node graph produced by the weighted adjacency matrix *W*, and the score or loss function that needs to be minimized *L*. Solving the above equation is still a combinatorial optimization problem as the acyclicity constraint is difficult to enforce as it is discrete, whereas the loss function is continuous one. The approach used by NOTEARS is by reformulating the learning problem as a continuous optimization one, which is achieved by introducing *h(W)*, a continuous measure of "DAG-ness" which basically quantifies the severity of violations from acyclicity as W changes as given by Eqn. 2.

$$\begin{aligned} \min_{W \in \mathbb{R}^{n \times n}} \; & L(W) \\ \text{s.t. } & h(W) = 0 \end{aligned} \tag{2}$$

We assume that RA and S cannot be parent nodes, i.e., they cannot be the cause of other nodes. Also, Po65, RU, SH, and LA cannot be affected by any other variable, i.e., they cannot be child nodes. Once we get the casual structure, we use that alongside the training data to learn the parameters of a Bayesian network using Causalnex [2].

The NOTEARS algorithm is a new hybrid algorithm, which is considered to be better than the classical structure discovery algorithms as it makes lesser assumptions about the data. These algorithms though may require larger datasets ,and may take higher processing times.

No Curl DAG: There have been studies which show that data chosen for NO TEARS must be scaled well and show experimental results of how it may not be able to model real world data well. The studies also show that if the variables used in NO TEARS have different variances the causal graph produced by the algorithm has been shown to be biased towards the features with larger variances[13]. No Curl removes the continuous constraint optimization present in NO TEARS. This is done by reforming the problem's DAG space.

$$\{\mathcal{G}_{W \circ ReLU(grad(p))}\} = \mathbb{D} \tag{3}$$

The DAG is represented as a product of a skew symmetric matrix and a function p (which is a measure of the topological ordering of DAGs). This Hadamard product is then minimized instead, as shown below

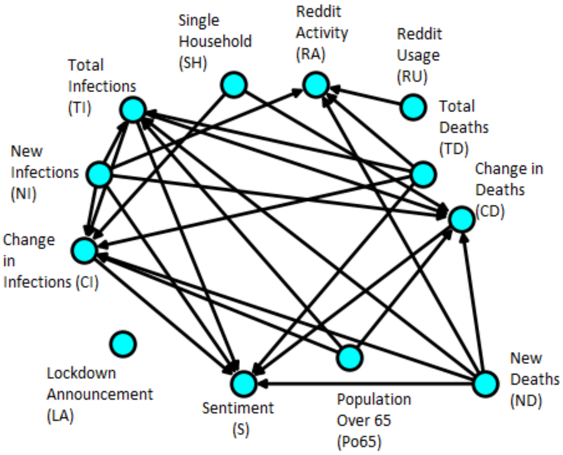

Figure 1: Predicted Causal Diagram using NOTEARS alogirthm with Continuous Lockdown Announcement feature. - Baseline Model

$$(W^*, p^*) = \arg\min_{W,p} F(\gamma(W, p), \mathbf{X}) \tag{4}$$

This will result in the final graph being generated in a more efficient manner. The No Curl algorithm matches NO TEARS in terms of accuracy and even performs better in cases where a DAG is created from dense matrices.

No Fears DAG: Building on NO TEARS, a more accurate algorithm called No Fears was explored in this paper, which uses a score-based structure learning in order to infer causal connections that exist between the Reddit Activity and pandemic characteristics. By revisiting the NO TEARS formulation given in Eqn(2), the weighted adjacency matrix is obtained by squaring the element wise squaring of the parameter matrix. For this constrained optimization, the Karush-Kuhn-Tucker (KKT) optimality conditions is not satisfied except in a trivial case. Due to this reason, the NO TEARS algorithm takes a long time to converge to an exact acyclic solution even with the addition of the smoothing function *h(W)*.

An equivalent reformulation in this approach is obtained by taking the adjacency matrix as the absolute value of the parameter matrix in order to accomodate for the failure to satisfy the KKT conditions from above, given a local minimum. Hence the formulation becomes,

$$\begin{aligned} \min_{W \in \mathbb{R}^{n \times n}} \; & L(W) \\ \text{s.t. } & h(|W|) = 0 \end{aligned} \tag{5}$$

Furthermore, if the score function *L(W)* is convex, then the KKT conditions are also sufficient for local minimality. The KKT conditions can thus be understood through edge absences: together these must be sufficient to ensure acyclicity, but each absence must also be necessary in preventing the completion of a cycle. Based on this understanding, a local search algorithm called KKT-search (KKTS) is proposed to satisfy the conditions. KKTS proves to work really well as post processing when applied to the output of other

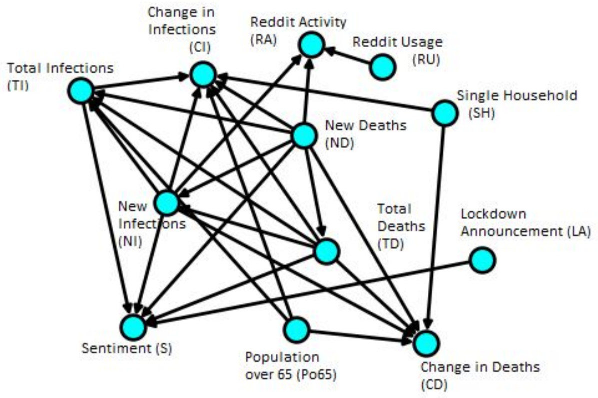

Figure 2: Predicted Causal Diagram using NO TEARS algorithm with binary Lockdown Announcement feature.

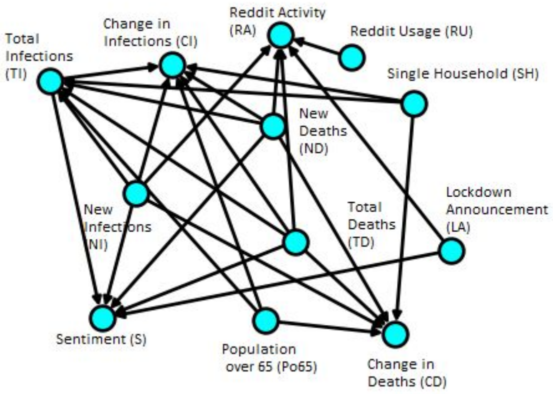

Figure 3: Predicted Causal Diagram using NO CURL algorithm.

algorithms by reducing the structural Hamming distance (SHD) substantially.

6 DATA DESCRIPTION AND CLEANING

The dataset for this project is taken from Kaggle procured using SocialGrep [19]. It contains a comprehensive collection of about 4.5M posts mentioning COVID-19, extracted from Reddit up until October 25th, 2021. This data had columns including the title of the post, the timestamp it was created as UTC, followed by details like a permalink, subReddit name and ID. We cleaned this data by dropping rows with NaN values and extracting the date and time from the UTC timestamp. Utilising this date value we filter the dataset for the 57 days considered (from January 22nd to March 18th 2020). This would allow us to compare inferences with the original model by Gencoglu et. al. [9] Further we extract a column for the location from the tweets. These locations consist of the 7 countries we have selected to evaluate the model consisting of America, Australia, Brazil, Canada, Japan, New Zealand, United Kingdom. We also extract sentiment for the Reddit posts via a fine-tuned distilBERT [28] model. This generates a sentiment score for each post in the range of [-1,1].

Another data source we will be referring to in this research is from "COVID-19 repository" which contains the daily number of officially reported COVID infections and death by the Center for Systems and Engineering at Johns Hopkins University [8]. This data has a time series for each country that we consider in the model.

From the above mentioned data sources, we extract 12 features to help us corresponding to the variables mentioned above. Firstly from the COVID-19 data we get the daily count of total and new infections , followed by the percent change in infections. Similarly the deaths were also calculated. For the 7 countries we also gather the data concerning the people over 65, the percentage of single households, the date of first restriction imposed and the Reddit usage. From the Reddit dataset we also find the daily Reddit activity (as the number of posts) and the average sentiment of all posts , for each country.

Based on the distribution of each feature we convert our dataset into a conditional probability table , with each feature measured as HIGH or LOW depending on the mean value of the feature. For e.g. if the value for the feature is greater than its mean it is classified as HIGH. Further sentiment is classified as NEGATIVE if the score is lesser than 0 , otherwise it is classified as POSITIVE. This helps us create our conditional probability table of all features which is utilised further.

To characterize the pandemic straightforwardly, we assign 12 attributes to their corresponding variables – (1) *Total Infections* as TI, (2) *New Infections* as NI, (3) *Change in Infections* as CI, (4) *Total Deaths* as TD, (5) *New Deaths* as ND, (6) *Change in Deaths* as CD, (7) *Sentiment* as S, (8) *Reddit Activity* as RA, (9) *Reddit Usage* as RU, (10) *Population over* 65 as Po65, (11) *Single Household* as SH, (12) *Lockdown Announcement* as LA. These form our set of final variables , which will be mentioned in the generated Causal Diagrams below.

7 EXPERIMENTAL SETUP AND RESULTS

We run our experiments by utilizing NoTears [28], NoCurl [27], and No Fears [23] to first obtain the causal structure from our features. We used the constraints mentioned in **Our Solution** section and train the algorithms with a edge threshold value of 0.3. We use Bayesian Networks to then find the conditional probabilities from the DAGs. In this section we explain and compare in detail our findings from the three algorithms.

7.1 Structure Learning

NoTears DAG Baseline: As a baseline for our experiments we first used NoTears [28] to predict the casual structure, which can be found in Figure 1. We notice our target variable Reddit Activity (RA) is affected by Reddit Usage (RU), New Deaths (ND) and

Country	AUC-NOTEARS	AUC-NoCurl	AUC-No Fear
Australia	0.845	0.857	0.887
Brazil	0.860	0.756	0.756
Canada	0.809	0.837	0.837
Japan	0.858	0.861	0.867
New Zealand	0.789	0.934	0.934
United Kingdom	0.930	0.951	0.954
United States	1	0.99	1
Mean AUC	0.870	0.885	0.891

Table 1: AUC result for each country

New Infections (NI). We also notice how other variables such as Sentiment(S) are affected by Infections and Deaths.

Reddit Usage affecting Reddit Activity is an expected association, as popularity of Reddit in a country (Reddit usage) would affect the Reddit activity. Interestingly we also noted that the Lockdown Announcement (LA) did not causal connections which was not expected. This is as we expected online sentiment change in accordance to negative sentiment research such as work by [26].

NoTears DAG Optimised: We optimise our previous work which encoded the Lockdown Announcement (LA) as a continuous variable (measured as the number of days from the Lockdown Announcement) to a categorical value. Hence the Lockdown Announcement was encoded as Yes if a government restriction had been announced and No otherwise. This allowed us to predict the model's causal structure, which can be found in Figure 2. We see that the variable for Lockdown Announcements (LA) has Causal Relationships in the graph. For example, our secondary target variable Sentiments (S) is affected by Lockdown Announcement (LA), Total Deaths (TD), New Deaths (ND) and New Infections (NI) and Total Infections (TI).

NoCurl DAG: The figure 3 shows the DAG predicted by the NoCurl algorithm [27]. We see that Reddit Activity has two new variables affecting it compared to the graph in figure 2. Total Death (TD) and Lockdown Announcement are the additional variables that influence Reddit Activity. On the other hand, we don't see any change in the variables that affect Sentiment.

NoFears DAG: Finally we used the NoFears [23] algorithm, the predicted DAG for which can be seen in figure 4. Similar to NoCurl, NoFears also predicts a dependency between Lockdown Announcement and Reddit Activity. But predicts no effect of Deaths on Reddit Activity.

7.2 Evaluation

We used the graph structures from Figure 2, 3, 4 to then learn the parameters of a Bayesian Network. The features were first discretized into HIGH or LOW depending on the mean value of the feature. These "new" set of features are then used to predict the Reddit activity for each day, as High or Low using the rest of the variables.

We then validate the output by a k-fold cross validation. We used Leave-One-Out Cross Validation to train the model on k-1 countries and validated on the last country. We used AUC (Area under the ROC curve) as a metric to evaluate our model's prediction of **Reddit activity** given other variables. This can be seen in Table 1.

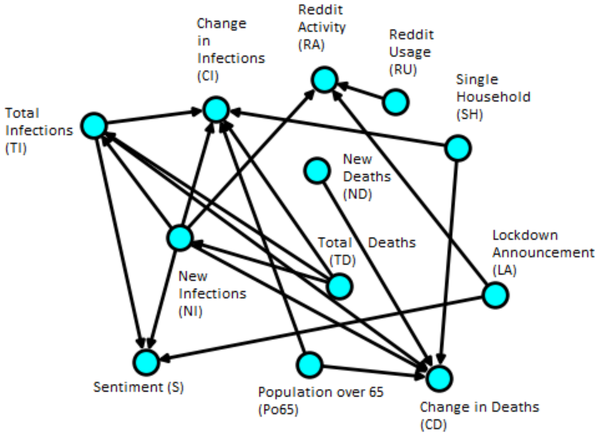

Figure 4: Predicted Causal Diagram using NO FEAR algorithm.

Our results show that all of the models were able to predict the Reddit activity with a good probability for all the countries selected in our experiments. From Table 1 we can observe that NoFear just about out performed NoCurl and NoTears in predicting Reddit Activity. We hypothesize that the increase in AUC score may be as a result of Lockdown Announcement affecting Reddit Activity as this causal effect is only measured in NoCurl and NoFear.

Conditional Probability - We further used the Bayesian Network to determine the marginal probabilities. The method enables simple model querying with a range of observations. We show a few of the marginal probabilities across the 3 models in Table 2. We set our target variables, which reflect the Public attention and sentiment to High Reddit Activity and Negative Sentiment respectively.

First, in order to see if the domain knowledge and the Bayesian network's likelihood calculations are concurrent, we check the effect of % of single-person households and % of 65+ people affecting total infections. The NoTears model satisfies this notion, where the probability of Total Infections being High is higher in the case when % of single-person households was Low and % of 65+ people

Query	Variable	Prob. NO TEARS	Prob. NoCurl	Prob. No Fear
SH % = H & Po65 % = L	TI = H	0.311	0.5	0.532
SH % = L & Po65 % = H	TI = H	0.51	0.49	0.532
NI = H & ND = H	RA = H	0.6	0.5	0.516
NI = L & ND = L	RA = H	0.12	0.15	0.13
LA = Y	RA = H	0.326	0.25	0.38
LA = N	RA = H	0.326	0.11	0.20
LA = Y	S = Neg.	0.719	0.721	0.82
LA = N	S = Neg.	0.427	0.405	0.58
ND = H	S = Neg.	0.684	0.564	0.791
ND = L	S = Neg.	0.620	0.620	0.739

Table 2: Marginal Probability Results

were high. On the other hand NoCurl and NoFear models suggest that the Total Infection rate was probabilistically high in both the scenarios. We then move to our target variable, Reddit Activity, and see how Infections, Deaths and Lockdown Announcement affect it. All the models indicate that higher infection and death rates have a positive causal relationship with Reddit Activity. As previously mentioned the NoTears model did not predict any casual relationship between Lockdown Announcement and Reddit Activity, thus we don't see any change in conditional probability in that case. NoCurl and NoFear models although do suggest that Reddit Activity was during the period when Lockdown was announced, compared to when it was not. Finally, we also find conditional probabilities for the Sentiment being negative, while conditioning on Lockdown Announcement and then on New Deaths. All our models unanimously predict Sentiment to be negative with a high probability in both the scenarios. From table 2 we can infer that Lockdown Announcement and rise in Covid cases/deaths had a negative impact on the general sentiment of the public.

8 DISCUSSION

In this project, we strove to discover causal associations between COVID-19 and Reddit activity in a span of almost three months. From our experiments, we attempted to discover the causal associations between COVID-19 patterns and Reddit activity as well as public sentiment during the early stages of the pandemic. Our results highlight the findings of some of the expected causal relationships, like Reddit Activity in a given country being affected by the popularity of Reddit in that country (Reddit Usage). Our results also show alignment with the scientific literature on COVID-19. We see that % Single Household and % of Population over 65 affect the Total Infections. The higher the % Population over age of 65, the higher the probability of Total Infections, and vice versa in the case of % Single Household. These findings align with the literature on the transmission of COVID-19, as researched in [15] and [11]. Coming onto our target variables. We observe that when new infections and new deaths are high, Reddit activity is roughly 5 times more likely than when the situation is opposite, as seen in Table 2. We see similar trends in the case when an announcement of lockdown is made. The probability of Reddit Activity being high was almost doubled during the period when Lockdown was announced.

We also see variations in the predicted DAG with the change in the algorithm used. In the case of NoTears we saw that Reddit Activity was affected by Reddit Usage, New Deaths and New Infections. Whereas, with the addition of using NoCurl, two new variables affected the target variable, Total Deaths and Lockdown Announcement. On the other hand, NoFears showed no casual effect of Deaths on Reddit Activity.

9 FUTURE SCOPE AND CONCLUSION

Future scope: Additional data like including more countries that uses Reddit as a major social media platform to share COVID news or extending the range of COVID-19 dates to more than three months, can be incorporated in order to discover interesting causal relationships. Also, we have not included the upvotes, downvotes and comments which would make for a larger Reddit dataset. Bayesian Networks also enable interventional computations, like do-calculus, n addition to the simple observational computations that we performed, like the calculation of marginal probabilities. Conducting such interventional studies which allow simulations of different what-if situations, could further help understand the structure in social media data.

We also saw the predicted graph structure changed from figure 1 to figure 2, after changing the Lockdown Announcement variable from continuous to a binary variable. This poses a question of finding out how different features affect the structure predicting algorithms and impact the finding of the eventual DAG. Feature and model selection is something that may very well influence the casual relationships discovered. Finding ways to better select a model without any domain knowledge bias is something that is potentially an open problem.

Furthermore, in the context of this study, ground truth causal associations do not exist even for a few variables, preventing the direct measurement of performance of causal discovery methods. We would like to clarify that we are aware of these and any other pertinent restrictions on our study.

Conclusion: We believe that understanding the factors affecting public attention and sentiment during any pandemic is crucial for making public policies. Inference of these patterns in a causal model from social media can help us in the pursuit of timely decisions and suitable policymaking, and consequently, high public engagement. Computational methods such as causal inference and causal reasoning seem to be the perfect tool to solve such problems as they help us disentangle correlations and causation between the observed variables of the adverse phenomenon.

REFERENCES

[1] [n. d.]. https://arxiv.org/pdf/2106.07197v1.pdf

[2] [n. d.]. Causalnex.network.BayesianNetwork¶. https://causalnex.readthedocs.io/en/latest/causalnex.network.BayesianNetwork.html

[3] 2020. (2020). https://doi.org/10.3390/books978-3-03928-573-0

[4] Varnica Bajaj, Nirupa Gadi, Allison P. Spihlman, Samantha C. Wu, Christopher H. Choi, and Vaishali R. Moulton. 2020. Aging, immunity, and covid-19: How age influences the host immune response to coronavirus infections? https://www.frontiersin.org/articles/10.3389/fphys.2020.571416/full

[5] Ronny Bergmann and Roland Herzog. 2019. Intrinsic formulation of KKT conditions and constraint qualifications on smooth manifolds. https://arxiv.org/abs/1804.06214

[6] K Bhaskaran, SCJ Bacon, SJW Evans, CJ Bates, CT Rentsch, B MacKenna, L Tomlinson, AJ Walker, A Schultze, CE Morton, and et al. 2021. Factors associated with deaths due to covid-19 versus other causes: Population-based cohort analysis of UK Primary Care Data and linked national death registrations within the OpenSAFELY platform. (2021). https://doi.org/10.1101/2021.01.15.21249756

[7] S. Dixon and Jul 26. 2022. Biggest social media platforms 2022. https://www.statista.com/statistics/272014/global-socialnetworks-ranked-by-number-of-users/

[8] Ensheng Dong, Hongru Du, and Lauren Gardner. 2020. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases* 20, 5 (2020), 533–534. https://doi.org/10.1016/S1473-3099(20)30120-1

[9] Oguzhan Gencoglu and Mathias Gruber. 2020. Causal Modeling of Twitter Activity during COVID-19. *Computation* 8, 4 (2020), 85. https://doi.org/10.3390/computation8040085

[10] Erfaneh Gharavi, Neda Nazemi, and Faraz Dadgostari. 2020. Early Outbreak Detection for Proactive Crisis Management Using Twitter Data: COVID-19 a Case Study in the US. *CoRR* abs/2005.00475 (2020). arXiv:2005.00475 https://arxiv.org/abs/2005.00475

[11] Yan-Rong Guo, Qing-Dong Cao, Zhong-Si Hong, Yuan-Yang Tan, Shou-Deng Chen, Hong-Jun Jin, Kai Sen Tan, De-Yun Wang, and Yan Yan. 2020. The origin, transmission and clinical therapies on coronavirus disease 2019 (COVID-19) outbreak- A n update on the status. *Military Medical Research* 7 (12 2020). https://doi.org/10.1186/s40779-020-00240-0

[12] Marcus Kaiser and Maksim Sipos. 2021. Unsuitability of notears for causal graph discovery. https://arxiv.org/abs/2104.05441v1

[13] Marcus Kaiser and Maksim Sipos. 2022. Unsuitability of notears for causal graph discovery when dealing with dimensional quantities. *Neural Processing Letters* 54, 3 (2022), 1587–1595. https://doi.org/10.1007/s11063-021-10694-5

[14] Jing Ma, Yushun Dong, Zheng Huang, Daniel Mietchen, and Jundong Li. 2021. Assessing the Causal Impact of COVID-19 Related Policies on Outbreak Dynamics: A Case Study in the US. *CoRR* abs/2106.01315 (2021). arXiv:2106.01315 https://arxiv.org/abs/2106.01315

[15] Jing Ma, Yushun Dong, Zheng Huang, Daniel Mietchen, and Jundong Li. 2022. Assessing the causal impact of COVID-19 related policies on outbreak dynamics: A case study in the US. *Proceedings of the ACM Web Conference 2022* (2022). https://doi.org/10.1145/3485447.3512139

[16] Elisabeth Mahase. 2020. Covid-19: Who declares pandemic because of "alarming levels" of spread, severity, and inaction. *BMJ* (2020), m1036. https://doi.org/10.1136/bmj.m1036

[17] Cristoforo Pomara, Francesco Sessa, Marcello Ciaccio, Francesco Dieli, Massimiliano Esposito, Giovanni Maurizio Giammanco, Sebastiano Fabio Garozzo, Antonino Giarratano, Daniele Prati, Francesca Rappa, and et al. 2021. Covid-19 vaccine and death: Causality algorithm according to the who eligibility diagnosis. *Diagnostics* 11, 6 (2021), 955. https://doi.org/10.3390/diagnostics11060955

[18] Abeed Sarker, Sahithi Lakamana, Whitney Hogg-Bremer, Angel Xie, Mohammed Ali Al-Garadi, and Yuan-Chi Yang. 2020. Self-reported covid-19 symptoms on Twitter: An analysis and a research resource. (2020). https://doi.org/10.1101/2020.04.16.20067421

[19] SocialGrep. 2021. *The Reddit COVID dataset.* https://www.kaggle.com/datasets/pavellexyr/the-reddit-covid-dataset

[20] Edgar Steiger, Tobias Mußgnug, and Lars Eric Kroll. 2020. Causal analysis of COVID-19 observational data in German districts reveals effects of mobility, awareness, and temperature. *medRxiv* (2020). https://doi.org/10.1101/2020.07.15.20154476 arXiv:https://www.medrxiv.org/content/early/2020/07/23/2020.07.15.20154476.full.pdf

[21] Mike Thelwall and Saheeda Thelwall. 2020. A thematic analysis of highly retweeted early COVID-19 tweets: Consensus, information, Dissent and Lockdown Life. *Aslib Journal of Information Management* 72, 6 (2020), 945–962. https://doi.org/10.1108/ajim-05-2020-0134

[22] Jeremy Turiel, Delmiro Fernandez-Reyes, and Tomaso Aste. 2021. Wisdom of crowds detects covid-19 severity ahead of officially available data. *Scientific Reports* 11, 1 (2021). https://doi.org/10.1038/s41598-021-93042-w

[23] Dennis Wei, Tian Gao, and Yue Yu. 2020. DAGs with No Fears: A Closer Look at Continuous Optimization for Learning Bayesian Networks. *CoRR* abs/2010.09133 (2020). arXiv:2010.09133 https://arxiv.org/abs/2010.09133

[24] Dennis Wei, Tian Gao, and Yue Yu. 2020. Dags with no fears: A closer look at continuous optimization for learning Bayesian Networks. https://arxiv.org/abs/2010.09133v1

[25] Fan Yu, Jiawei Guo, Wei Xi, Zhao Yang, Rui Jiang, and Chao Zhang. 2021. Audio distilbert: A distilled audio Bert for speech representation learning. *2021 International Joint Conference on Neural Networks (IJCNN)* (2021). https://doi.org/10.1109/ijcnn52387.2021.9533328

[26] Haiyan Yu, Ching-Chi Yang, Ping Yu, and Ke Liu. 2022. Emotion diffusion effect: Negative sentiment COVID-19 tweets of public organizations attract more responses from followers. *PLOS ONE* 17, 3 (2022). https://doi.org/10.1371/journal.pone.0264794

[27] Naiyu Yin Yue Yu, Tian Gao and Qiang Ji. 2021. DAGs with No Curl: An Efficient DAG Structure Learning Approach. In *Proceedings of the 38th International Conference on Machine Learning.*

[28] Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. 2018. Dags with no tears: Continuous optimization for structure learning. https://arxiv.org/abs/1803.01422

‘