Spring 6-13-2021

Examining Polarized COVID-19 Twitter Discussion Using Inverse Reinforcement Learning

Sydney Lister
Sydney.A.Lister.21@Dartmouth.edu

Examining Polarized COVID-19 Twitter Discussion Using Inverse Reinforcement Learning

Sydney Lister

Undergraduate Thesis

Computer Science Department

Advisor: Prof. Soroush Vosoughi

Dartmouth College

June 1, 2021

Abstract

In this work, we model users' behavior on Twitter in discussion of the COVID-19 outbreak using inverse reinforcement learning to better understand the underlying forces that drive the observed pattern of polarization. In doing so, we address the largely untapped potential of inverse reinforcement learning to model users' behavior on social media, and contribute to the body of sociology, psychology, and communication research seeking to elucidate the causes of socio-cultural polarization. We hypothesize that structural characteristics of each week's retweet network as well as COVID-19 data on cases, hospitalizations, and outcomes are related to the Twitter users' reward function which leads to polarized discussion of COVID-19 on the platform. To derive the state space of our inverse reinforcement learning model, we compute the relative modularity of retweet networks formed from retweets about COVID-19. The action space is determined by the distribution of mask-wearing sentiment in tweets about COVID-19. We build a fine-tune a BERT text classifier to determine mask-wearing sentiment in tweet. We design state features which reflect both structural characteristics of the retweet networks and COVID-19 data on cases, hospitalizations, and outcomes. Our results indicate that polarized Twitter discussion about COVID-19 weighs more heavily on features relating to the severity of the COVID-19 outbreak and less heavily on features relating to the structure of retweet networks. Overall, our results demonstrate the aptitude of inverse reinforcement learning in helping understand user behavior on social media.

Acknowledgements

I would first like to thank Professor Vosoughi for his guidance and support throughout the course of my time at Dartmouth, and especially during my time doing research and writing my thesis. It has been an honor to be your student and thesis advisee. I would also like to thank Professor Preum and Professor Mehnaz for their time and consideration in serving on my Honors Thesis committee. I am grateful to the entire Dartmouth Computer Science Department - it is a privilege to have learned from dedicated professors in exciting curricula over the past four years. I would not have been able to write this thesis without the love and support of my family and friends.

Contents

Chapter 1

Introduction

Section 1.1

Polarization and COVID-19

The recent rise of social media allows individuals to access information from around the world on an unprecedented scale. From 2005 to 2015, use of social media skyrocketed - 65% of American adults now use social media [46]. Many of these users are young people (90% of Americans aged 18-29 use social media platforms) but usage amongst Americans 65 and older has also tripled from 2005-2015 [46]. Given its prominence, social media have become invaluable tools in spreading information and opinions. Twitter in particular has captured the attention of the American public in recent years. On Twitter, users can interact with and post short messages known as "tweets". As of October 2020, 22% of American adults use Twitter [2]. Roughly 42% of Twitter users are active on the platform daily [2]. A staggering 500 million tweets are sent every day [2].

Coupled with the rise in social media usage is another phenomenon: the rise of partisan news and polarized discussion. Americans are more ideologically polarized now than in recent decades. The 2020 U.S. presidential election exemplifies the deep divides amongst the American public. Before the election, roughly 80% of all American voters believed core American values set them apart from the other side [1]. Roughly 90% felt that a victory by the hand of the opposition would yield "lasting harm" to the country [1].

Representatives in the House of Representatives and the Senate are acutely polarized as well. Political scientists have shown that Democratic and Republican legislators began pulling apart in the 1970s reaching a degree of polarization not seen since the post-Civil War era [11]. Polarization amongst the American public has been exacerbated by the rise of partisan news media. Researchers blame both the increasingly partisan nature of news media and the human tendency to seek out information that aligns with their political ideologies for the rise of polarization [26]. Partisan news outlets put viewers in an "echo chamber" of their beliefs, which might in turn polarize them [54]. These programs are often designed to persuade and bolster the opinions of viewers, who often already share similar beliefs. Given that 71% of Twitter users use the platform as their primary news source [2], Twitter and related social media are now used as tools to proliferate partisan news . Social media as a whole, including Twitter, have made the production, distribution, and discovery of news cheaper and easier.

This is due in part because, even as access to information increases, human attention limits the amount of information individuals can consume. Limits on human attention require content to be filtered such as to only provide an individual with a digestible amount of information. On social media, this information is filtered according to consumer biases. As a result, individuals now find themselves in “echo chambers,” where they only consume information that aligns with their opinions and beliefs. The outcomes and implications of the polarization induced by social media can be dangerous.

The 2020 outbreak of the COVID-19 pandemic exemplifies the pervasiveness of polarization in the United States. When asked whether they thought the United States did a good job dealing with the outbreak of COVID-19, 78% of Republicans said yes compared to 29% of non-Republicans [40]. This discrepancy in support for the governing coalition was the largest amongst the 14 nations surveyed [40]. The divide extends across many issues related to the pandemic, including contact tracing, the response of public health officials, mask wearing, vaccinations, and the return to “normal”. Coverage of such topics on social media has been riddled with misinformation and unfounded information.

There have already been serious consequences to this trend. Echo chambers on social media reinforced this misinformation and individuals' misguided beliefs, leading to racially motivated assaults and discriminatory events against specific racial groups in the United States and beyond [55]. Moreover, the spread of misinformation and unfounded information in the face of a public health threat interferes with the nation's recovering from such a crisis. These dangerous effects demanded researchers to take a closer look at polarization over social media surrounding the COVID-19 pandemic. Preceding the outbreak of the pandemic, scientists studied polarization through the lens of sociological, psychological, and communication theories. Recent advances in the computer science field, however, could offer new perspectives on how polarization on social media platforms evolves.

Section 1.2

Inverse Reinforcement Learning

The problem of inverse reinforcement learning was primarily motivated by its potential to inform computational models of animal and human learning. Computational models that use reinforcement learning rely on a fixed and known reward function. For example, scientists observed evidence of reinforcement learning occurring in bee foraging [39] and songbird vocalization [50]. Yet in other instances animal and human learning, specifically those involving multi-attribute reward functions, the reward function must be discovered through empirical observation. One example of such behaviors is human economic decision making, where different considerations must be taken into account. Computer scientists have also found this problem applicable to the task of driving. When adults learn to drive, they learn by demonstration rather than a specified reward function. They learn by a process called apprenticeship learning, meaning learning from an expert. Inverse reinforcement learning critically solves the problem of extracting a reward function given observed, optimal behavior [44]. The inverse reinforcement learning problem is informally characterized as follows:

Given

1. Measurements of an agent's behavior over time under varied conditions,
2. Measurements of the agent's sensory inputs,
3. A model of the environment

Determine The reward function being optimized.

The three givens along with the reward function form a Markov Decision Process (MDP). More formally, a MDP is a tuple (S, A, T, R) where S is a finite set of states; A is a finite set of actions, T is a set of state transition probabilities; and R is the reward function, which we assume to be bounded in absolute value by 1. We let MDP/R denote an MDP without a reward function. For the purpose of this work, we choose a probabilistic approach to inverse reinforcement learning because of its resolution of the ambiguity other inverse reinforcement learning techniques face. Namely, strategies such as linear programming, structured maximum margin prediction and matching feature expectations result in a large set of reward functions (including those of that are all zeros) for which the observed policy is optimal. Instead, we take a probabilistic approach known as maximum entropy inverse reinforcement learning which results in a single stochastic policy.

Section 1.3

Our Goal

With COVID-19 continuing to spread across the world and trends of polarization continuing on social media, it becomes increasingly important to understand the formation and evolution of echo chambers. Inverse reinforcement learning opens the doors to gaining a different perspective on how polarization on social media occurs. More specifically, we seek to learn the reward function underpinning the human behavior that results in polarization on Twitter. In doing so, we hope to better elucidate the learning process of human

ideological polarization. Polarization on social media helps propagate perceptions, even when corrective measures are taken [20]. During a time where the United States faces a grave public health and economic threat, proliferating only grounded truths is of utmost importance. To achieve this, we present an maximum entropy inverse reinforcement learning model to learn the reward function underlying the trend of polarization surrounding COVID-19 on Twitter. To create the model, we collected tweets related to COVID-19 from February, 2020 to September, 2020. Using this data, we derive both a state space and an action space. We also incorporate data detailing number of cases, number of hospitalizations, and number of deaths due to COVID-19 between February, 2020 and September, 2020 as features of the state space.

We derive the state space by constructing a retweet network using the collected tweets. A retweet denotes that a user x has shared another user's tweet. We construct directed, weighted graphs based on users retweeting other users in a given week, wherein a vertices u, v in graph G represent Twitter users and there is an edge e from u to v if user v retweeted user u. We also consider how many times a user retweets another user by adding a weight w to edge e denoting the number of times user v retweeted user u. We then find the modularity of such retweet networks to classify the relative polarization occurring in each network as "high", "medium", or "low". Each week can thus be associated with a state by classifying the relative modularity of the week's retweet network. The classifications of each retweet network form the state space for our inverse reinforcement learning model.

To derive the action space, we creating a text classification model which classifies tweets as revealing pro- or anti-mask-wearing sentiment. We use the pretrained Bidirectional Encoder Representations from Transformers (BERT) model and add a layer which classifies tweets as pro- or anti- mask-wearing. Using these classifications, we can then find the distribution of pro- and anti- mask tweets amongst all of the collected tweets. We use entropy to measure the level of agreement amongst mask-wearing sentiment classifications. The distribution of mask-wearing sentiment classifications for a given

week can be classified as "high", "medium", or "low". Each week can thus be associated with an action by classifying the relative entropy of the distribution of mask-wearing sentiment in tweets from that week. The classifications of the distribution of tweet sentiments form the action space for our inverse reinforcement learning model.

Using structural characteristics of the constructed retweet networks as well as COVID-19 case data, we define features of each state of the agent. For each week and its associated state, we calculate the density, the number of connected components, and the density of the largest connected component in the week's retweet network. These basic structural characteristics will help constitute the network-related state features. We also calculate the number of new cases of COVID-19, number of new COVID-19 hospitalizations, and number of new deaths due to COVID-19. These measurements will help constitute the COVID-19-related state features. Then, for each state of the agent, we calculate the average of both the network-related measurements and the COVID-19 related measurements across all weeks associated with the given state.

A defining feature of our work is the use of inverse reinforcement learning models to examine human socio-cultural trends on social media. Our hypothesis is three-fold:

- Pro- and anti-mask-wearing sentiment can be predicted based on users' tweets alone.
- Retweet networks constructed will show modularity that increases over time. Increased modularity over time would indicate the formation of "echo chambers" within conversations related to COVID-19 on Twitter.
- Structural characteristics of each week's retweet network as well as COVID-19 data on cases, hospitalizations, and outcomes are related to the Twitter users' reward function which leads to polarized discussion of COVID-19 on the platform.

This work expands on previous applications of inverse reinforcement learning models which, to the best of our knowledge, are limited predominately to modelling route choice,

football players' strategies, and robot navigation. Significantly, we present an approach that offers a new computational technique to examining user behavior on social media.

Section 1.4

Overview of Results

We constructed a text sequence classifier using a pretrained Bidirectional Encoder Representations from Transformers (BERT) model and added a layer which classifies tweets as pro- or anti- mask-wearing. The classifier did not perform significantly better than a dummy classifier, however.

We also constructed a retweet network based on our collected data. For a given week, we create an edge between two users if one retweeted the other. The weight of this edge denotes the number of times one of the users retweeted the other. We examine the modularity of these networks, which reveals the strength of division of a network into communities. Contrary to our hypothesis, the modularity of retweet networks did not increase over time. However, our measurements of modularity remain relatively high across all weeks in the collection period.

We were indeed able to extract a reward function using the state space and action space generated from both text classifications of tweets and structures of the retweet networks. Twitter users' rely more heavily on COVID-19-related features than retweet network-related features when demonstrating polarized COVID-19 discussion on Twitter. This might signal that event-related characteristics might be more influential on Twitter polarization related to COVID-19 than structural characteristics of retweet networks. Further, this result indicates that, given a well formed state and action space, inverse reinforcement learning can be used to discover the reward function of human behavior on social media. Also, our results demonstrate that inverse reinforcement learning can be an effective approach to test whether certain factors influence human behavior on social media.

Chapter 2

Related Work

Section 2.1

Polarization

Americans are more ideologically polarized now than in recent decades. The 2020 U.S. presidential election exemplifies the deep divides amongst the American public. Before the election, roughly 80% of all American voters believed core American values set them apart from the other side [1]. Roughly 90% felt that a victory by the hand of the opposition would yield "lasting harm" to the country [1]. Representatives in the House of Representatives and the Senate are acutely polarized as well. Political scientists have shown that Democratic and Republican legislators began pulling apart in the 1970s reaching a degree of polarization not seen since the post-Civil War era [11]. However, the trend of polarization is not limited to the United States. Across the globe, social media users form polarized opinions based on information (as well as misinformation) regarding socio-cultural activities, products, and services. Consider the following examples of polarizing events taking place outside of the United States:

1. In Nigeria, fake news of a infant's bloodied corpse induced polarizing opinions, and the ensuing violence led to the killing of 10 people [34].

2. In Chile, secondary school students organized a campaign against metro fare evasion that resulted in protests, instability, violence, and deaths as well as damage to cultural heritage sites and museums [57].

3. In India, efforts to suppress the release of a controversial film backfired, resulting in widespread distribution of the film over social media [17].

These events reveal the potential of polarized opinions to cause widespread human suffering and destruction of property [45]. With the recent rise of social media, users across the globe are given the ability to access and share information on a massive scale. This widespread consumption of information via social media might induce polarized opinions that can impact behavior. Sociology, psychology, and communication research have long acknowledged the importance of understanding polarization.

In the sociological domain, game theory, muted group theory, social identity theory, as well as theories of social distance and protracted social conflict have been used to explain polarization. Woon employs behavioral game theory to understand the connection between voter behavior and candidate positioning that leads to polarization amongst political candidates [60]. Korn defines muted group theory as inequity in language yielding marginalized groups unable to express themselves, in turn creating a loss and distortion of information that produces polarization [29]. Other sociologists have argued that strong social identity, or social attachments to one's group memberships, create attitudes that are both partisan and uncontested ideologically [25, 59]. Social distance in the sociological domain refers to the level of trust felt towards another individual or group. Akerlof suggests that social distance might induce ideological behavior that results in polarization [3]. Protracted social conflict based in ethnic, racial, cultural, and religious hatreds can also result in polarization [7].

The psychological domain offers further theoretical explanations for polarization. Cognitive dissonance, the phenomenon of positive feelings experienced when confronting opinion-reinforcing information, might extend towards consumption on social media, influencing polarization [41]. Similarly, confirmation bias is the human tendency

towards information that confirms preexisting opinions and can result in polarization [58]. On the other hand, motivated skepticism is the human tendency of being more skeptical towards information that does not align with one's preexisting opinions and can also lead to polarization [22, 38]. Further psychological explanations of polarization include theories of motivated reasoning [53] and anti-reflexivity [37].

Polarization can also be explained through the lens of communication theories. For one, the media plays an influential role dictating the importance of issues, thus polarizing the audience [24]. Similarly, cultivation theory explains the alignment between reality portrayed through the media and society's lived reality, which can lead to polarization [51]. Elaboration Likelihood Models [4, 23], flaming [13], and spiral of silence theories [14, 31] further elucidate the causes of polarization.

While various disciplines attempt to explain the roots of socio-cultural polarization, the rise of social media and polarized discussions on social media offer new data sources for investigating this phenomenon. Echo chambers, wherein users consume information which aligns with their held opinions [10, 19], and their potential to have polarizing effects demand further analysis of polarization induced by social media.

One timely issue that has been the source of polarizing debate is the outbreak of COVID-19. Potential for polarized discussion extends across many issues related to the pandemic, including contact tracing, the response of public health officials, mask-wearing, vaccinations, and the return to "normal". The present work offers a computational approach for looking at polarization. By using a probabilistic learning approach to investigate user behavior in COVID-19 Twitter discussion, we demonstrate a technique for testing hypotheses of the causes of socio-cultural polarization on social media that differs from traditional sociological, psychological, and communication approaches.

Section 2.2

Analysis of COVID-19 Discussion on Twitter

Since the outbreak of COVID-19, studies have leveraged Twitter data to monitor and analyze discussion related to the virus.

Studies employ different approaches to investigate the content and sentiment of COVID-19 Twitter discussion. Manguri et al. use the TextBlob library in python to perform sentiment analysis on tweets collected across seven days in April 2020 containing the keywords "COVID-19" and "coronavirus" [36]. Similarly, Boon-Itt and Skunkan explore public perception of COVID-19 on Twitter by finding frequent keywords and performing both sentiment analysis and topic modelling on roughly 100,000 collected tweets [9]. Specifically relating to masks, Sanders et al. use clustering and sentiment analysis techniques to organize tweets in the first 5 months of 2020 relating to mask-wearing into high-level themes and then use automatic text summarization to extract key words and topics for individual clusters [49]. In addition,[42] pre-train a transformer-based model on a large corpus of tweets from January, 2020 to April, 2020 related to COVID-19 to create COVID-Twitter-BERT [42]. Rodriguez et al. use communicative content analysis to compare the spread misinformation on Twitter and Sina Weibo in early February, 2020 [48]. Jiang et al. seek to quantify partisan and geographic differences in online conversations about COVID-19 in the United States from January, 2020 to April, 2020 [28].

While previous work investigates the content and sentiment of discussion about COVID-19 on Twitter, many studies leverage data from short time frames towards the beginning of the outbreak. As the COVID-19 outbreak continues into 2021, more data has become available for analysis and Twitter users' opinions on various issues are continually shaped. As more information is disseminated regarding the efficacy of wearing masks in curtailing the spread of COVID-19, further analysis is necessary to capture the scope of Twitter discussion on the topic. By leveraging data from across seven

months of the COVID-19 outbreak, the present work considers the evolution of Twitter discussion of COVID-19 and mask wearing. Specifically, we present a mask-wearing sentiment classifier trained on tweets about COVID-19 spanning 28 weeks. Thus, our analysis of COVID-19 and mask-wearing discussion on Twitter might offer insight that takes into account a wider trajectory of the outbreak.

Section 2.3

Twitter Retweet Networks

Previous work uses various aspects of user interactions on Twitter to draw conclusions about how information is shared on the platform. However, Bi & Cho demonstrate that retweet networks provide a clearer signal to identify users' topical interests [8]. Retweet networks are conceptualized through the use of graphs, wherein nodes are users and edges denote one user retweeted another. The aptitude of retweet networks in investigating patterns of information sharing can be seen in many studies. For example, in investigating political polarization on Twitter, it is demonstrated that retweet networks have a partisan structure while user-to-user mention networks do not [15]. Stewart et al. use retweet networks to elucidate the contributions of Twitter trolls in politically divided conversations on Twitter [52]. Retweet networks have also been used to find which users have retweeted or have been retweeted in the past [61]. In working with retweet networks, many studies calculate the modularity of the network. Modularity can be defined as how well a network can be divided into smaller clusters [43]. Thus, in the retweet network of a polarized discussion, one would expect the graph to have a higher measure of modularity.

Since retweet networks demonstrate promising results in analyzing polarized discussions on Twitter, the present work constructs retweet networks for each week in the collection period. We use modularity to examine communities formed within each retweet network.

Section 2.4

Inverse Reinforcement Learning

The problem of inverse reinforcement learning was primarily motivated by its potential to inform computational models of animal and human learning.

Reinforcement learning seeks to find an optimal policy which maximizes the rewards of an agent [56]. Computational models that use reinforcement learning rely on a fixed and known reward function. For example, scientists observed evidence of reinforcement learning occurring in bee foraging [39] and songbird vocalization [50]. Yet in other instances animal and human learning, specifically those involving multi-attribute reward functions, the reward function must be discovered through empirical observation. In these cases inverse reinforcement learning is used as a tool to estimate the reward function and learn a demonstrator's policy of action [44]. A real-world example where the reward function unknown is human economic decision making, where different considerations must be taken into account. Computer scientists have also found this problem applicable to the task of driving. When adults learn to drive, they learn by demonstration rather than a specified reward function. They learn by a process called apprenticeship learning, meaning learning from an expert. Inverse reinforcement learning critically solves the problem of extracting a reward function given observed, optimal behavior [44]. The inverse reinforcement learning problem is informally characterized as follows:

Given

1. Measurements of an agent's behavior over time under varied conditions,
2. Measurements of the agent's sensory inputs,
3. A model of the environment

Determine The reward function being optimized.

The three givens along with the reward function form a Markov Decision Process (MDP). More formally, a MDP is a tuple (S, A, T, R) where S is a finite set of states; A

is a finite set of actions, T is a set of state transition probabilities; and R is the reward function, which we assume to be bounded in absolute value by 1. We let MDP/R denote an MDP without a reward function.

As inverse reinforcement learning was first proposed, the learner receives an optimal policy as input, as well as the state space, action space, and fully known transition probabilities. A critical challenge of this form of inverse reinforcement learning is that many reward functions (e.g. ones in which all reward values are zero) can be considered optimal for the given policy [44]. This creates ambiguity in the solution to an inverse reinforcement learning problem. Thus, many approaches have been used for inverse learning, including margin-based optimization and entropy-based optimization.

The goal of maximum margin prediction is to learn a reward function that better explains the demonstrated policy by some margin as compared to alternative policies [5]. This approach eliminates ambiguity of inverse reinforcement learning solutions by converging to a solution that maximizes some margin. Different approaches for defining and calculating the margin can be employed. Across maximum margin prediction techniques, a hypothesis for the policy is created each time the reward weights are updated. Thus, when the expert displays sub-optimal behavior, a linear reward function cannot be found [62].

To remedy this, the principle of maximum entropy can be applied under the assumption that the distribution over behaviors or trajectories that maximizes entropy minimizes commitments to any particular path beyond the constraints of matching feature expectations [27]. This approach is most prevalent in resolving the ambiguity of other inverse reinforcement learning solutions.

Inverse reinforcement learning has been implemented to estimate the reward function in a variety of problems. For example, maximum entropy inverse reinforcement has been used to model route choice of taxi drivers [62]. Similar inverse reinforcement learning techniques hav also been employed to understand football players' strategies [32] and

robot navigation [30]. Liu et al. model human risk decision making using maximum entropy inverse reinforcement learning [33].

Despite its promise for understanding human behaviors, few studies examine social media users' behavior using inverse reinforcement learning. One study investigates the effects of social-psychological feedback on human behavior on social media using inverse reinforcement learning [16]. In another study, inverse reinforcement learning is used to understand differences in behavior between troll and non-troll accounts on Twitter [35]. Still, the potential of inverse reinforcement learning to help understand human behavior on social media remains largely untapped. Thus, in the present work, we examine polarization on Twitter using a maximum entropy inverse reinforcement learning framework to better understand users' behavior.

Chapter 3

Data Collection and Preprocessing

Section 3.1

Data Collection

We rely on the publicly available COVID-19 Twitter data set which has been collected on an ongoing basis since January 28, 2020. Chen et al. used Twitter's streaming API and Tweepy to track 76 specific English keywords relating to COVID-19 [12]. As of May 3, 2021, the authors collected 1,443,871,621 tweets. For this purpose of this work, we use tweets collected between January, 2020 and September, 2020. We randomly sample 100,000 unique tweets per day between this time frame.

We also collect data from The COVID-19 Tracking Project at *The Atlantic*. This includes cross-checked data from 56 states and territories in the United States regarding testing, hospitalizations, and patient outcomes [47]. The data set spans January, 2020 up to March, 2021. For the purpose of this work, we store the number of new cases, new deaths, and new hospitalizations. New cases of COVID-19 include both confirmed and probable cases. New hospitalizations include patients hospitalized with confirmed or suspected COVID-19 cases. New deaths include fatalities with confirmed or probable COVID-19 diagnosis. We only use data collected between January, 2020 and September, 2020.

Section 3.2

Preprocessing

Amongst the 100,000 sampled tweets per day, we only include tweets that are written in the English language. We store data containing: the hashtags included in the tweet; the user ID of the tweet's author; the Twitter username of the author; the name of the author; the text of the tweet; any linked URLs contained in the tweet; the author's Twitter bio; the user ID of the retweeted tweet's author; the text of the retweeted tweet; and the time of the tweet. If the tweet is an original tweet, i.e. not a retweet, we store "None" under the retweeted user's ID and the retweeted text. We extract the date the tweet was written from the time of the tweet. For some tweets, the time of the tweet is stored as "None", and such tweets are discarded from the data set. Using the time of the tweet, we find the Gregorian week number in which the tweet was written. We use the Gregorian week number as an index to collect all tweets written in the same week for examining both mask-wearing sentiment and retweet network structure. The corresponding 2020 dates for each Gregorian calendar week number can be found in Table A.1.

Chapter 4

Methods

Section 4.1

Mask-Wearing Sentiment Classification

4.1.1. Bidirectional Encoder Representations from Transformers (BERT)

The foundation of our mask-wearing sentiment classifier is a pretrained Bidirectional Encoder Representations from Transformers (BERT) classifier. BERT is a language representation model designed to pre-train deep representations of unlabeled text [18]. The model is pre-trained using a large English corpus using a masked language modeling (MLM) objective. BERT can be fine-tuned with only one additional output layer and is suitable for a wide range of natural language processing tasks without requiring significant task-specific architecture modifications.

For the purpose of this work, we use the 'bert-base-uncased' pretrained model. This model is uncased, so it does not make a difference between 'english' and 'English' when processing textual data.

4.1.2. Mask-Wearing Sentiment Classifier

We create a model which classifies the text of a tweet as either pro-mask-wearing or anti-mask-wearing. We examine tweets spanning the collection period and parse the

hashtags included in the tweet. We compile a list of common hashtags that reveal an anti-mask-wearing sentiment:

- #NoMask
- #BurnYourMasks
- #AntiMask
- #AntiMaskers
- #NoLockdowns
- #NoNewNormal
- #WeWillNotComply

For any tweet in the data set which contains at least one of the above hashtags, we label the tweet with a 0 to denote anti-mask-wearing sentiment. We found 5 tweets containing at least one of the above hashtags. Similarly, we compile a list of common hashtags that reveal a pro-mask-wearing sentiment:

- #WearAMask
- #MasksSaveLives
- #WearADamnMask
- #WearAMaskSaveALife
- #WearYourMask

For any tweet in the data set which contains at least one of the above hashtags, we label the tweet with a 1 to denote pro-mask-wearing sentiment. We found 23 tweets containing at least one of the above hashtags.

We collect all tweets with the label 0 or 1 to form a labeled data set to train and validate a supervised learning model. Thus, 28 tweets are included in this data set.

We then transform the text of each tweet in the data set into a list of strings called tokens, where each token is one word or punctuation mark. We accomplish this using the tokenizer from the *Transformers* library. We then map each token to an ID, which is an integer representation of the token. After tokenizing the data, we split this data set using 90% for training and 10% for validation.

We use a pretrained Bidirectional Encoder Representations from Transformers (BERT) model (bert-base-uncased). We fine-tune the model for pro-/anti-mask-wearing sentiment prediction using the data set defined above. We add a linear classification layer to the BERT model so that the fine-tuned model outputs either a 0 (anti-mask) or 1 (pro-mask).

After validation, the fine-tuned BERT model performs with 67% accuracy. We compare this with a "dummy" classifier which uses a stratified strategy to make predictions. The dummy classifier performs with a 66% accuracy.

4.1.3. Discussion

The fine-tuned BERT classifier does not perform significantly better than the baseline dummy classifier. One possible explanation for this result is the lack of labeled training and validation data. Consequentially, it is possible that our classifier over-fits the training data, and performs poorly when generalizing to other examples. One possible solution to this problem is to get more labeled data. We could expand the list of hashtags such that more tweets could be included in the labeled data set. This approach runs the risk of mislabeling tweets as pro- or anti-mask-wearing in the training data. Another approach would be to derive the relevant hashtags using clustering, as in [49]. In this study, k-means clustering is used to cluster tweets in their embedding space, and k-means is applied again to each cluster to form a two-level cluster hierarchy [49]. The authors then find key words to describe and differentiate between each of the clusters and sub-clusters. For the present work, performance might be improved by deriving key words and hashtags using a clustering approach to represent pro- and anti-mask-wearing sentiment.

Section 4.2

Retweet Networks

4.2.1. Creation of Retweet Networks

To examine the structure of retweets amongst Twitter discussion related to COVID-19, we construct a set of undirected, weighted graphs. For each week in the collection period, we find all of the tweets in which one user is retweeting another user. We identify tweets which are retweets by checking whether a retweeted user ID is stored for the given tweet. We filter the retweets for each week to include only users who are included in every week's collection of retweets. Figure 4.1 displays the number of retweets for each week after filtering users.

For each week, we construct a weighted, un-directed graph where each there is an edge between a user u and a user v if either user retweeted the other. The weight of the edge is equal to the number of retweets between the two users. Thus, the weight of the undirected edge from u to v is given by:

$$w_{u,v} = w_{v,u} = r_{u,v} + r_{v,u} \tag{4.1}$$

In Equation 4.1, $r_{u,v}$ is the number of times user u retweeted user v. $r_{v,u}$ is the number of times user v retweeted user u.

Figure 4.2 shows the graphs constructed for the weeks of February 24 to March 1, March 2 to March 8, March 16 to March 22, and March 23 to March 29, 2020.

4.2.2. Modularity

We use a measurement of network modularity to investigate polarization in discussions on Twitter surrounding COVID-19. The modularity of a network measures the strength of division of the network into different communities. Thus, in the retweet network of a polarized discussion, one would expect the graph to have a higher measure of modularity. Using our constructed retweet networks, we calculate the modularity of the graph for

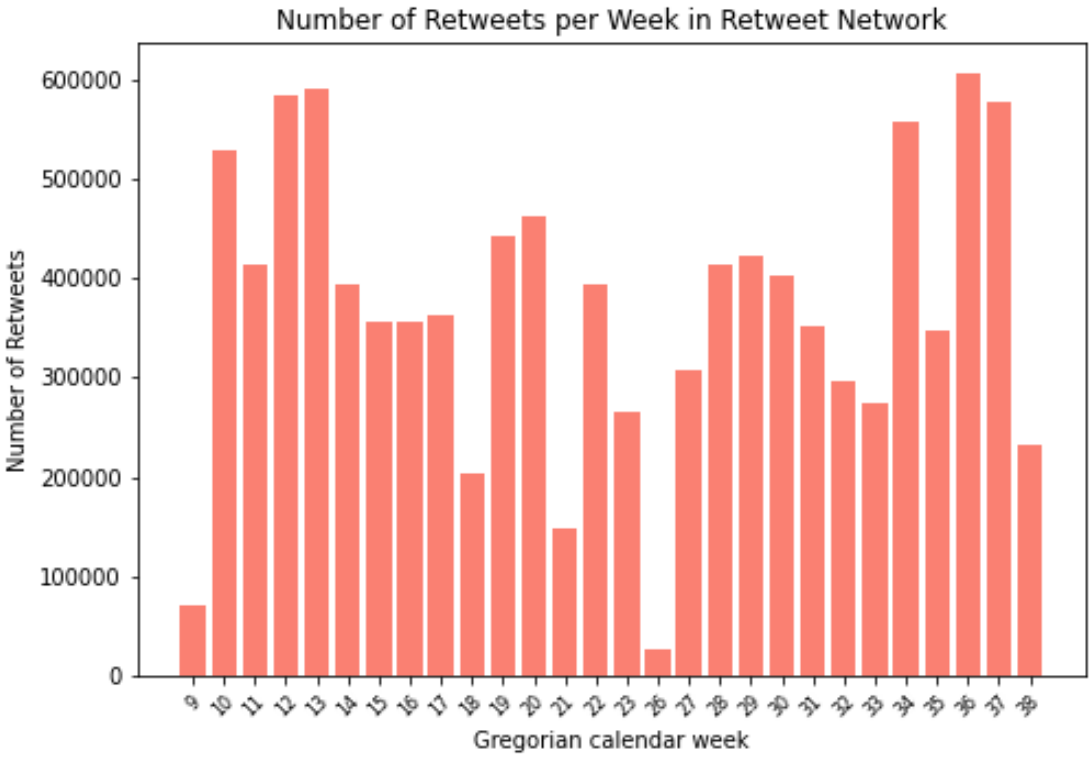

Figure 4.1: The number of retweets included in each week's retweet network. We identify a week by it's Gregorian calendar number. The corresponding 2020 dates for each week can be found in Table A.1. We did not create retweet networks for weeks with less than 10,000 cleaned retweets.

Figure 4.2: Retweet networks for four weeks in 2020. Each week contains the same set of Twitter users. Each node in the network is a Twitter user, and an edge between two users denotes one user retweeted the other. The graph for week 9 (upper left) has 87 nodes and 63 edges. The graph for week 10 (lower left) has 262 nodes and 258 edges. The graph for week 12 (upper right) has 163 nodes and 139 edges. The graph for week 13 (lower right) has 188 nodes and 168 edges.

Week	# Nodes	# Edges	# CCs	Modularity
9.0	87	63	24	0.914792
10.0	262	258	10	0.802715
11.0	190	169	22	0.900658
12.0	163	139	24	0.910247
13.0	188	168	20	0.905614
14.0	189	168	21	0.907649
15.0	220	204	17	0.877533
16.0	234	220	17	0.882972
17.0	226	210	17	0.869142
18.0	182	164	19	0.908149
19.0	191	175	16	0.887520
20.0	213	197	17	0.871320
21.0	132	109	23	0.891405
22.0	123	102	21	0.918800
23.0	164	143	22	0.887070
26.0	65	42	23	0.932396
27.0	191	173	19	0.902213
28.0	241	232	11	0.871118
29.0	252	241	13	0.898735
30.0	245	236	14	0.879035
31.0	209	203	12	0.847282
32.0	91	69	22	0.886917
33.0	91	67	24	0.807574
34.0	146	131	16	0.868075
35.0	127	105	22	0.898015
36.0	129	108	21	0.897098
37.0	179	166	16	0.877060
38.0	149	133	17	0.897547

Table 4.1: Summary statistics of retweet networks for each week. CCs indicates connected components of a network. The calculations of modularity are given by Equation 4.3.

each week. For the purpose of this work, we use the Louvain algorithm to help calculate modularity. The Louvain algorithm is robust towards weighted graphs, thus making it suitable for this application [6].

In computing modularity we first find the partition of the graph nodes which maximizes modularity based on the Louvain heuristics [6]. This is called the modularity optimization step. The Louvain algorithm for community detection has two phases that are repeated iteratively. During the first phase, each node in the weighted graph containing N nodes is assigned to its own community. Then, for each node i we examine each

of its neighbors j and calculate the gain of modularity achieved by removing i from its community and placing it in j's community. The change in modularity given by moving i into community C is given by:

$$\Delta Q = \left[\frac{\sum_{in}+2k_{i,in}}{2m} - \left(\frac{c+k_i}{2m}\right)^2\right] - \left[\frac{\sum_{in}}{2m} - \left(\frac{\sum_{tot}}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)^2\right] \tag{4.2}$$

Where $\sum_{in}$ is the sum of the weights of the edges inside C, $\sum_{tot}$ is the sum of the weights of the edges incident to nodes in C, k_i is the sum of the weights of the edges incident to node i, $k_{i,in}$ is the sum of the weights of the edges from i to nodes in C, and m is the sum of the weights of all of the edges in the network. We place i in the community where the gain, ΔQ, is both positive and maximum, or else it stays in its original community. This phase is repeated until no node can be moved to improve modularity.

In the second phase of the Louvain algorithm, we build a new network wherein the nodes are the communities found in the first phase. This is called the community aggregation step. The weight of the edges between two new nodes are equal to the sum of the weight of the edges between nodes in the corresponding two communities. Edges between nodes of the same community lead to self-loops in the new graph. Once this step is complete, the modularity optimization step is reapplied to the resulting weighted network followed by another iteration of the community aggregation step. Iteration of these two phases continues until there are no more changes and maximum modularity is attained.

To find the partition that maximizes the modularity as defined above, we use the Python-Louvain API [6]. We then compute the modularity of the partition using the same library as follows:

$$Q = \frac{1}{2m}\sum_{i,j}[A_{ij} - \frac{k_i k_j}{2m}]\delta(c_i, c_j) \tag{4.3}$$

Where A_{ij} is the edge weight between node i and node j; $k_i k_j$ are the sum of the edge weights attached to nodes i and j, respectively; m is the sum of all edge weights; c_i, c_j

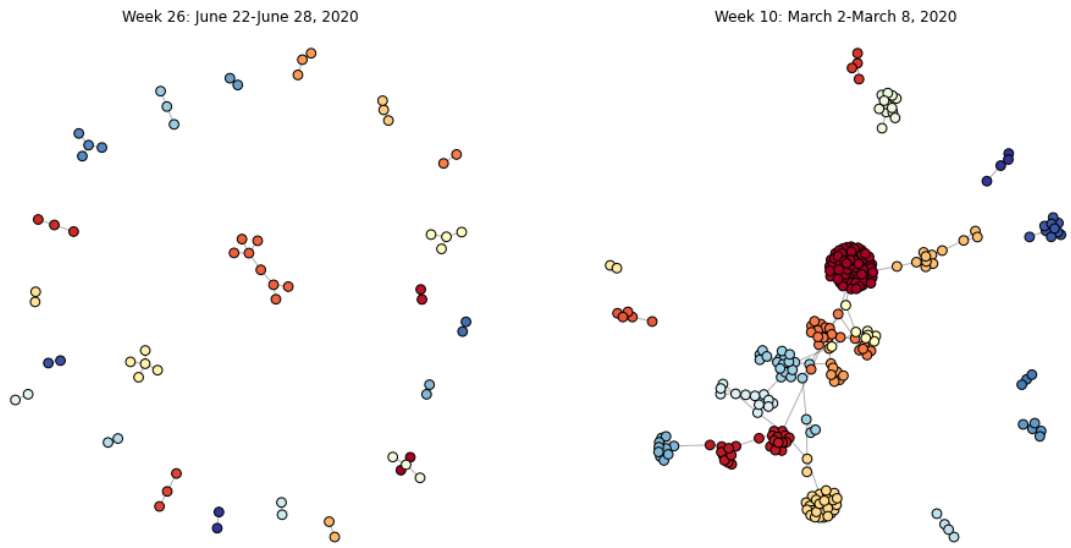

Figure 4.3: Examples of two retweet networks in which different color nodes represent different communities in the network. Communities are computed by finding the partition of nodes which maximizes the modularity of the network. The network on the left is the network with the highest modularity across the collection period. The network on the right is the network with the lowest modularity across the collection period.

are the communities of nodes i and j, respectively; and δ is the Kronecker delta function. Figure 4.3 shows two retweet networks wherein different colors of nodes represent different communities in the retweet networks. We store the maximized modularity for each retweet network.

The modularity, number of nodes, and number of edges in the retweet networks over the course of 28 weeks can be shown in Table 4.1.

4.2.3. Discussion

We hypothesized that the networks would display increasing modularity over time. This result would indicate the formation of "echo chambers" within conversations related to COVID-19 on Twitter. The resulting networks did not demonstrate increased modularity over time. There are several possible explanations for such a result. For one, it is possible that formation of echo chambers did not occur within conversations related to COVID-19 on Twitter. However, the measurements of modularity stayed relatively high (i.e. close to 1) and consistent as can be seen in Table 4.1. This could indicate that polarization levels stayed relatively high and consistent across the observed 28 weeks.

There are also other possible ways to construct the graph of retweets. For this study we observe the structure of a weighted, un-directed graph. However, it is possible to construct a directed graph using this data, wherein a directed edge from user u to user v denotes that user v retweeted user u. The weight of such an edge would denote the number of times user v retweeted user u. If the network were to be constructed as such, it would be necessary to devise an implementation of community detection and modularity computation which takes into account directed and weighted edges.

Section 4.3

Inverse Reinforcement Learning Model

We model the Twitter community behavior of polarization surrounding COVID-19 using Markov Decision Process. We argue that the Markovian property of transition and action holds for polarization with certain assumptions when setting up the state and action spaces. Specifically, we set states as the relative level of modularity of a given week's retweet network as compared to the modularity of other networks across the collection period. The state can be represented by 0 (low modularity), 1 (medium modularity), or 2 (high modularity). The next state of the retweet network we assume depends on the previous week's retweet network state and current action. Our action space is comprised of the relative entropy level of mask-wearing sentiment classification for a given week as compared to other weeks across the collection period. Lower entropy indicates a higher level of agreement regarding mask-wearing. The action can be represented by 0 (low entropy), 1 (medium entropy), or 2 (high entropy). We consider a trial to be a span of time or trajectory of states whose lengths range from 2 weeks to 28 weeks.

On Twitter, the group of users discussing COVID-19 resemble an agent in a reinforcement learning task. The goal of the group of users is to collect the largest reward in a single trial from forming communities or "echo chambers" within the group of users without knowing the structure or modularity of communities. We can regard this process of a group of users learning an unknown reward function. Because of the irrational nature

of groups of users, the true reward function (modularity of communities over time) is never learned. Instead, the group of users' sentiment regarding mask-wearing provides data that demonstrate potential underlying influence of historical trials to current ones. This means the data is not independent with each other between different trials.

4.3.1. State Space

We use the constructed retweet networks to derive the state space for our inverse reinforcement learning model. First, we sort the stored modularity values for each week's retweet network in increasing order. We classify each week as 0 (low modularity), 1 (medium modularity), or 2 (high modularity). The state of the agent is then the modularity level of the given week's retweet network.

4.3.2. Action Space

We use classifications of tweets based on their sentiment regarding mask-wearing to generate the action space for our inverse reinforcement learning model. We classify the mask-wearing sentiment of all of the tweets in our data set as pro- or anti- mask-wearing. For each week in the collection period, we calculate the entropy of the distribution of classifications from that week. We sort the entropy values in increasing order and classify each week as 0 (low entropy), 1 (medium entropy), or 2 (high entropy). Lower entropy indicates a higher level of agreement regarding mask-wearing. In other words, we use entropy level as a proxy for the observed level of agreement amongst Twitter users sentiment on mask-wearing. In designating the mask-wearing sentiment of users' tweets as the action space, we also make the assumption that the next state of the agent, i.e. the modularity of the following week's retweet network, relies on both the current state and the entropy level of mask-wearing sentiment of users' tweets for the current week. The action taken by the agent for a given week is then the entropy level of mask-wearing sentiment of the week's tweets.

4.3.3. State Features

We construct the state features in a way that reflects information from historical trials. Table 4.2 shows our state features, calculated for each week's retweet network.

In constructing the state features, we assume potential influences on the state of a given week's retweet network, apart from users' sentiment regarding mask-wearing, include features related to COVID-19 and features related to the structure of retweet networks.

To derive features related to COVID-19, we use data from The COVID-19 Tracking Project at *The Atlantic*. This includes cross-checked data from 56 states and territories in the United States regarding testing, hospitalization, and patient outcomes. For the purpose of this work, we store the number of new cases, new deaths, and new hospitalizations. New cases of COVID-19 include both confirmed and probable cases. New hospitalizations include patients hospitalized with confirmed or suspected COVID-19 cases. New deaths include fatalities with confirmed or probable COVID-19 diagnosis. We aggregate the number of new cases, the number of new deaths, and the number of new hospitalizations for each of the 28 weeks in our collection period for retweet networks. Then, for each low, medium, or high modularity state, we calculate the average number of new cases, average new hospitalizations, and average new deaths for weeks in this state. We use the average number of new cases, the average number of new hospitalizations, and average number of new deaths as features of a given state.

We also include features related to the retweet networks in our state features. We include three basic characteristics of retweet networks: density of edges, number of connected components, and density of the largest connected component. The density of edges is calculated by:

$$d = \frac{2m}{n(n-1)} \tag{4.4}$$

For a retweet network G, n is the number of nodes and m is the number of edges in G. The number of connected components in retweet network G is the number of subgraphs

Feature	Value	Meaning
1	Float	Avg. number of new deaths for this state
2	Float	Avg. number of new hospitalizations for this state
3	Float	Avg. number of new cases for this state
4	Float	Avg. density of retweet network for this state
5	Float	Avg. number of CCs in retweet network for this state
6	Float	Avg. density of largest CC in retweet network for this state

Table 4.2: State features with value and meaning. CC denotes connected component.

in G where each pair of nodes in the subgraph is connected. The largest connected component is the largest subgraph of G in which each pair of nodes is connected, and the density of the largest connected component is calculated by applying Equation 4.4 to the subgraph with the largest number of nodes. For each low, medium, or high modularity state, we calculate the average density, average number of connected components, and average density of the largest connected component in the retweet networks. We use the average density, average number of connected components, and average density of the largest connected component as features of a given state.

Overall, we derive 6 state features for each state. Features and their meanings are listed in Table 4.2

4.3.4. Maximum Entropy Inverse Reinforcement Learning

In a traditional Markov Decision Process (MDP), the aim is to find a policy for an agent. The policy specifies which action an agent will choose when in a given state. A MDP assumes that the reward function of agents is known. In real world scenarios, humans often do not explicitly know the reward function of their actions and behaviors. In imitation learning, an agent observes the states and actions of an expert over the course of different trajectories. A trajectory can also be known as a path, or a sequence of states that an agent will take in one example demonstration. Inverse reinforcement learning takes a different approach, reducing the problem to recovering the reward function that makes behavior of an agent closely mimic the demonstrated behavior. Inverse reinforcement learning assumes that there is a linear relationship between the reward value of a state R and the specified features associated with states, $\mathbf{f}_{s_j}$. This

linear function is parameterized by some reward weights, θ. Thus, the reward value of a trajectory ζ, is the sum of state rewards along that trajectory. This is equal to the reward weight applied to the path feature counts, which are the sum of the state features along the path:

$$reward(\mathbf{f}_\zeta) = \theta^T \mathbf{f}_\zeta = \sum_{s_j \in \zeta} \theta^T \mathbf{f}_{s_j} \tag{4.5}$$

The most prevalent inverse reinforcement learning model is the maximum entropy inverse reinforcement learning model, as proposed by [62]. This model finds the reward weights which maximizes the entropy of trajectories:

$$\theta = argmax_\theta L(\theta) = argmax_\theta \sum_{examples} \log P(\tilde{\zeta} \mid \theta, T) \tag{4.6}$$

We can use gradient descent to optimize reward weights. The gradient is the difference between the expected empirical feature counts and the learning agent's expected feature counts. The learning agent's feature counts can be expressed in terms of expected state visitation frequencies, D_{s_i}.

$$\nabla L(\theta) = \tilde{\mathbf{f}} - \sum_\zeta P(\zeta \mid \theta, T)\mathbf{f}_\zeta = \tilde{\mathbf{f}} - \sum_{s_i} D_{s_i}\mathbf{f}_{s_i} \tag{4.7}$$

D_{s_i} is derived using Algorithm 1 in [62] and $\tilde{\mathbf{f}}$ is calculated by:

$$\tilde{\mathbf{f}} = \sum_{Path\zeta_i} P(\zeta_i)\mathbf{f}_{\zeta_i} \tag{4.8}$$

Now, we can find a policy distribution based on the reward weights that maximizes the entropy of trajectories. The policy distribution takes the form:

$$P(a_t \mid s_t) \propto exp(\theta^T \mathbf{f}(s_t)) \tag{4.9}$$

Using the maximum entropy inverse reinforcement learning algorithm as defined above, we find reward weights which make the Twitter community behavior appear

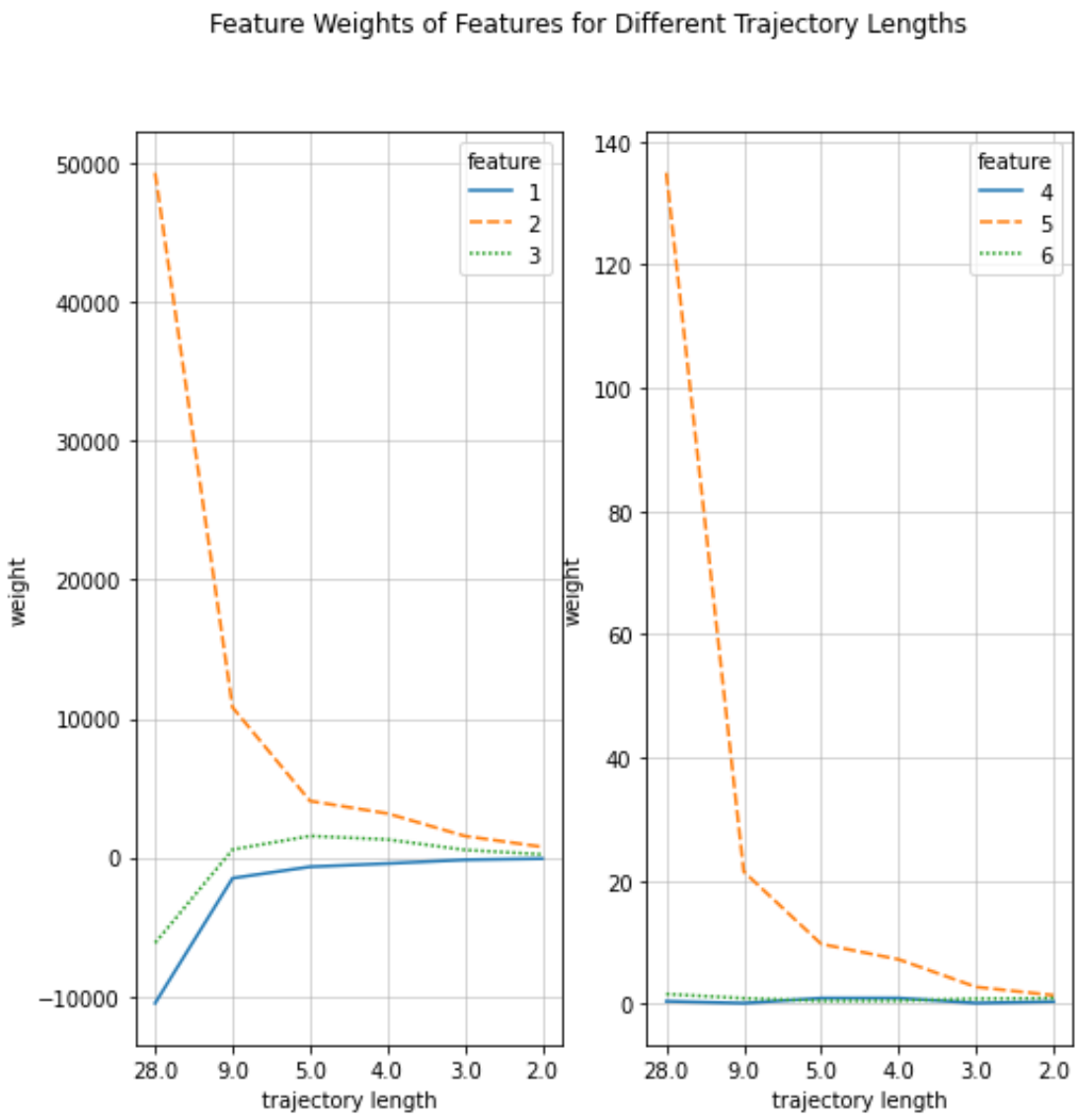

Figure 4.4: Weights of features for inverse reinforcement learning models trained using different trajectory (i.e. length n for path $\zeta = \{s_1, ..., s_n\}$).

near optimal under the learned reward function. Using our 28 weeks of collected data as expert behavior, we find calculate optimal reward weights for different trajectory lengths. To generate different trajectories, we partition the 28 week sequence of states into sub-sequences ranging from 2 state (week) long trajectories to a single 28 week long trajectory. The reward weights for different trajectory lengths can be seen in Figure 4.4. The reward weights from the maximum entropy inverse reinforcement learning model trained on trajectories of length 2 can be seen in Figure 4.5.

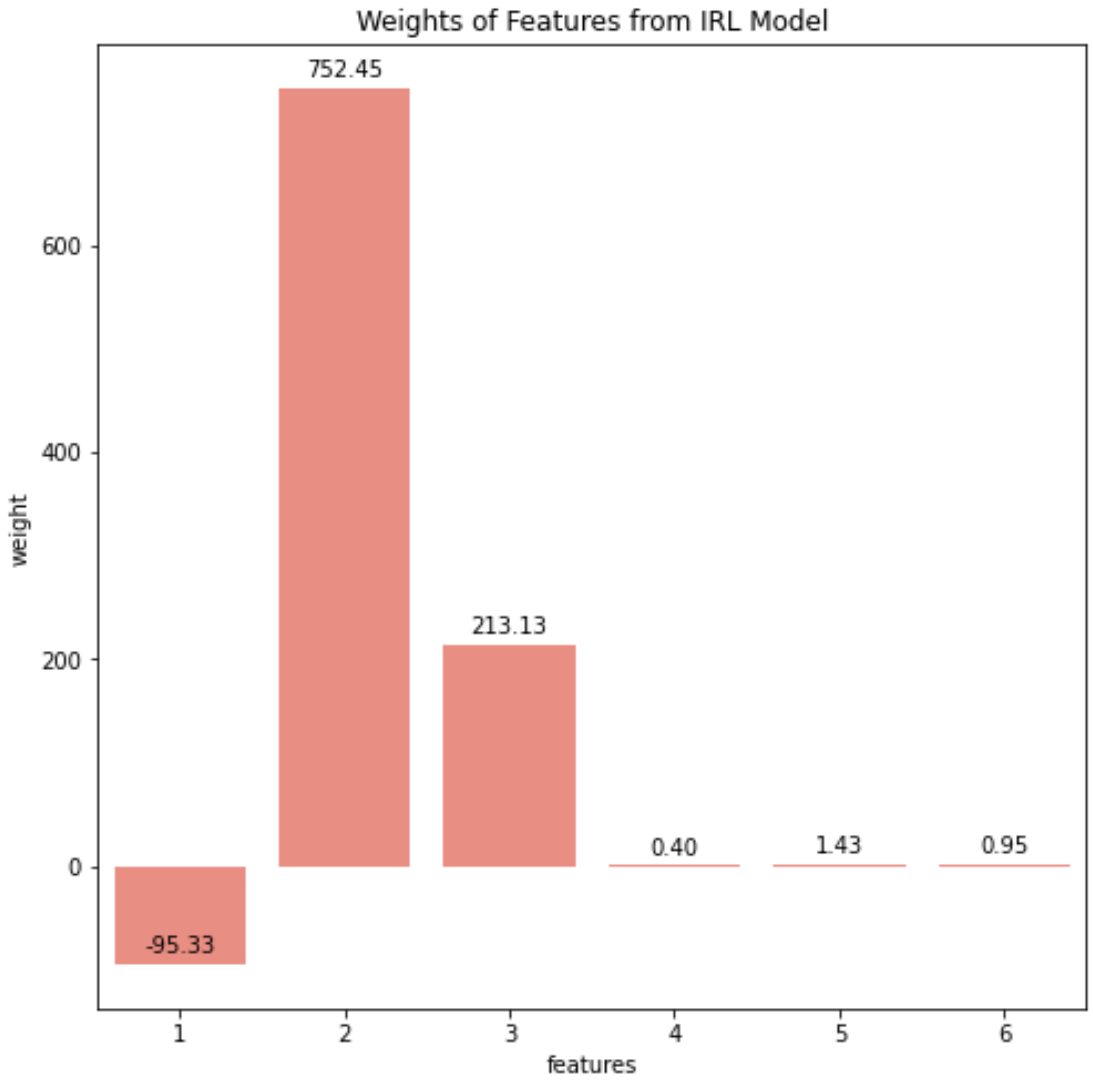

Figure 4.5: Weights of features from the inverse reinforcement learning model with trajectory length of 2 and 14 trajectories. The meaning of the 6 features are listed in Table 4.2.

4.3.5. Discussion

The weights of features for trajectories of length 2 are in Figure 4.5, and weights of features across all generated trajectory lengths are in Figure 4.4. Interestingly, overall, Twitter users show higher weight of "Average number of new hospitalizations for this state" (feature 2), possibly indicating a general effect of COVID-19 resource availability on Twitter polarization in COVID-19 discussion. Moreover, Twitter users weigh significantly more on COVID-19-related features and weigh less on characteristics of the retweet networks. The "Average number of new deaths for this state" (feature 1) and "Average number of new cases for this state" (feature 3) still show higher weight than the network-related features. This might indicate that polarization of Twitter discussion of COVID-19 is more heavily impacted by the severity of the outbreak at a given time, rather than the structure of Twitter discussions of the topic. Amongst the network-related features, however, "Average number connected components in retweet network for this state" (feature 5) weighed highest. Thus, the number of sub-networks in which Twitter users all retweet one another might be more influential on polarization than other characteristics of the retweet network.

The learned weights of the model demonstrated consistent relative magnitudes across different trajectory lengths. Mainly, Twitter users weight more on "Average number of new hospitalizations for this state" (feature 2) and less on "Average density of retweet network for this state" (feature 4) and "Average density of largest connected component in retweet network for this state" (feature 6).

Polarization of opinions has occurred surrounding many events across the world. Recently, examples of polarizing issues have included climate change in the United States [21], fake news of a baby's bloodied corpse [34], metro fare evasion in Chile [57], and a the suppression of a controversial film in India [17]. Although polarization can arise due to a wide variety of events, our results indicate that polarized COVID-19 discussion on Twitter might be more heavily influenced by the severity of the COVID-19 outbreak at a given time. Users tend to weigh more heavily on features specific to COVID-19

rather than those that could be shared amongst a variety of events. Thus, the driving forces of polarized Twitter discussion of COVID-19 might be related to COVID-19, rather than a common influence across all Twitter discussions.

This result contributes to the various theories and processes used to explain polarization, including sociological theories, psychological theories, and selection processes. While different instances of polarization share commonalities, this work indicates that event-specific features might also be heavily influential on ideological polarization.

These results demonstrate the ability of inverse reinforcement learning to provide insights into the underlying influences of Twitter users' behavior. However, our interpretation of Twitter users' behavior introduces bias to our work. Specifically, our choice of state space, action space, and state features derived from both Twitter data and national COVID-19 data heavily influence the findings of this research. These results should only be viewed as a preliminary analysis of polarized Twitter discussion of COVID-19.

Chapter 5

Limitations and Discussion

Section 5.1

Limitations

One limitation to our work is the lack of available twitter data labeled as pro- or anti-mask-wearing. As a result, it is possible that our mask-wearing sentiment classifier over-fits the training data, and is less robust towards new data samples.

Our method of labeling the available mask-wearing-sentiment data also leaves room for error in our classification. For example, a tweet that includes the hashtag '#antimask' does not necessarily indicate that the author holds an anti-mask-wearing stance. Moreover, based on our labeling approach, it is possible that our training data is not limited to unique tweets. For example, an original tweet that contains the hashtag '#wearamask' could be retweeted by another user, and both tweets would be included in our data set despite having the same text. Thus, it is possible that the number of unique tweets included in the training data is smaller than the total number of tweets.

In the analysis of retweet networks, the algorithm used to calculate the modularity of our network accounts for weighted but not directed edges. With this in mind, we created retweet networks containing undirected edges between users. As a result, information is lost as to which user retweeted the other. Calculation of modularity then also does not account for which users are disseminating information.

Our method of filtering users to be included in the retweet network does not take into account whether retweeted users are present in all weeks of the collection period. Thus, each retweet network has a different number of nodes, which could impact our comparisons of modularity between retweet networks.

In defining the state and action spaces of the inverse reinforcement learning model, the interpretation of Twitter users' behavior introduces bias to our work. Social media provides ample examples of human data as potential sources for investigation using inverse reinforcement learning models. The decision of which data constitutes the states and actions of the agent, as well as who or what is considered the agent, is entirely up to the investigator. By choosing the modularity of the retweet network to be the state of the agent, the entropy of mask-wearing-sentiment classification to be the action of the agent, our definition of state features as seen in Table 4.2, and the Twitter conversation surrounding COVID-19 to be the agent itself, some degree of bias is introduced to this work. Thus, these results should only be viewed as a preliminary analysis of polarized Twitter discussion of COVID-19.

In regards to the COVID-19 related state features, the data set used to calculate the features 1, 2, and 3 in Table 4.2 consists of national data from the United States, whereas the collected Twitter data features all tweets written in English. Thus, the data used to derive those state features does reflect the severity of the COVID-19 outbreak for Twitter users outside of the United State who are present in our Twitter data set.

Section 5.2

Discussion

The fine-tuned BERT classifier does not perform significantly better than the baseline dummy classifier. One possible explanation for this result is the lack of labeled training and validation data. Consequentially, it is possible that our classifier over-fits the training data, and performs poorly when generalizing to other examples. One possible solution to this problem is to get more labeled data. We could expand the list of hashtags such

that more tweets could be included in the labeled data set. This approach runs the risk of mislabeling tweets as pro- or anti-mask-wearing in the training data. Another approach would be to derive the relevant hashtags using clustering, as [49]. In this study, k-means clustering is used to cluster tweets in their embedding space, and k-means is applied again to each cluster to form a two-level cluster hierarchy. The authors then find key words to describe and differentiate between each of the clusters and sub-clusters. For the present work, performance might be improved by deriving key words and hashtags using a clustering approach to represent pro- and anti-mask-wearing sentiment.

We hypothesized that the networks would display increasing modularity over time. This result would indicate the formation of "echo chambers" within conversations related to COVID-19 on Twitter. The resulting networks did not demonstrate increased modularity over time. There are several possible explanations for such a result. For one, it is possible that formation of echo chambers did not occur within conversations related to COVID-19 on Twitter. However, since modularity levels remained relatively high (i.e. close to 1) across the observed 28 weeks, it is possible that there was a high level of polarization in Twitter discussions from the beginning of our collection period.

There are also other possible ways to construct the graph of retweets. For this study we observe the structure of a weighted, un-directed graph. However, it is possible to construct a directed graph using this data, wherein a directed edge from user u to user v denotes that user v retweeted user u. The weight of such an edge would denote the number of times user v retweeted user u. If the network were to be constructed as such, it would be necessary to devise an implementation of community detection and modularity computation which takes into account directed and weighted edges.

The weights of features for trajectories of length 2 are in Figure 4.5, and weights of features across all generated trajectory lengths are in Figure 4.4. Interestingly, overall, Twitter users show higher weight of "Average number of new hospitalizations for this state" (feature 2), possibly indicating a general effect of COVID-19 resource availability on Twitter polarization in COVID-19 discussion. Moreover, Twitter users weigh sig-

nificantly more on COVID-19-related features and weigh less on characteristics of the retweet networks. The "Average number of new deaths for this state" (feature 1) and "Average number of new cases for this state" (feature 3) still show higher weight than the network-related features. This might indicate that polarization of Twitter discussion of COVID-19 is more heavily impacted by the severity of the outbreak at a given time, rather than the structure of Twitter discussions of the topic. Amongst the network-related features, however, "Average number connected components in retweet network for this state" (feature 5) weighed highest. Thus, the number of sub-networks in which Twitter users all retweet one another might be more influential on polarization than other characteristics of the retweet network.

The learned weights of the model demonstrated consistent relative magnitudes across different trajectory lengths. Mainly, Twitter users weight more on "Average number of new hospitalizations for this state" (feature 2) and less on "Average density of retweet network for this state" (feature 4) and "Average density of largest connected component in retweet network for this state" (feature 6).

Polarization of opinions has occurred surrounding many events across the world. Recently, examples of polarizing issues have included climate change in the United States [21], fake news of a baby's bloodied corpse [34], metro fare evasion in Chile [57], and a the suppression of a controversial film in India [17]. Although polarization can arise due to a wide variety of events, our results indicate that polarized COVID-19 discussion on Twitter might be more heavily influenced by the severity of the COVID-19 outbreak at a given time. Users tend to weigh more heavily on features specific to COVID-19 rather than those that could be shared amongst a variety of events. Thus, the driving forces of polarized Twitter discussion of COVID-19 might be related to COVID-19, rather than a common influence across all Twitter discussions.

This result contributes to the various theories and processes used to explain polarization, including sociological theories, psychological theories, and selection processes.

While different instances of polarization share commonalities, this work indicates that event-specific features might also be heavily influential on ideological polarization.

These results demonstrate the ability of inverse reinforcement learning to provide insights into the underlying influences of Twitter users' behavior. However, our interpretation of Twitter users' behavior introduces bias to our work. Specifically, our choice of state space, action space, and state features derived from both Twitter data and national COVID-19 data heavily influence the findings of this research. These results should only be viewed as a preliminary analysis of polarized Twitter discussion of COVID-19.

Chapter 6

Conclusions

Section 6.1

Our Findings

We hypothesized that pro- and anti- mask-wearing sentiment could be predicted based on users' tweets alone. The model created did not support this claim, achieving an accuracy score that was only 1% higher than that of a dummy classifier. These results could be due to lack of training data to fine-tune our BERT model.

We also hypothesized that our constructed retweet would show modularity that increases over time. We thought that increased modularity over time would indicate the formation of "echo chambers" within conversations related to COVID-19 on Twitter. Evidence did not support this claim. Table 4.1 shows the modularity of the retweet networks across the 28 weeks included in the data set. Modularity values fluctuated over time, staying above 0.80 for all 28 weeks. This might indicate that, rather than polarization increasing over time, discourse surrounding COVID-19 on Twitter remained polarized since the beginning of our collection period.

Finally, we hypothesized that structural characteristics of each week's retweet network as well as COVID-19 data on cases, hospitalizations, and outcomes are related to the Twitter users' reward function which leads to polarized discussion of COVID-19 on the platform. The weights of features for trajectories of length 2 are in Figure 4.5,

and weights of features across all generated trajectory lengths are in Figure 4.4. Overall, Twitter users show higher weight of "Average number of new hospitalizations for this state" (feature 2), possibly indicating a general effect of COVID-19 resource availability on Twitter polarization in COVID-19 discussion. Moreover, Twitter users weigh significantly more on COVID-19-related features and weigh less on characteristics of the retweet networks. The "Average number of new deaths for this state" (feature 1) and "Average number of new cases for this state" (feature 3) still show higher weight than the network-related features. Amongst the network-related features, however, "Average number connected components in retweet network for this state" (feature 5) weighed highest.

This result indicates that COVID-19 data on cases, hospitalizations, and outcomes are more influential in Twitter users' reward function which leads to polarized discussion of COVID-19 on the platform. While this result is promising, it relies on the derivation of the state and action space from our modularity calculations and mask-wearing sentiment classification.

Section 6.2

Future Work

Future work on mask-wearing sentiment classification using tweets should focus on finding better ways to label tweets as pro- or anti-mask-wearing. This could be accomplished by using different approaches to identify pro- or anti-mask-wearing hashtags, which could help more accurately label tweets for training. One such approach could involve using k-means clustering to form a two-level cluster hierarchy, as in Sanders et al., and finding key words or hashtags which separate the clusters [49]. Those key words or hashtags could be used to label training data as pro- or anti-mask-wearing. Alternatively, unsupervised learning approaches can help identify pro- or anti-mask-wearing sentiment in tweets without requiring labeling. Unsupervised learning techniques would eliminate

the bias introduced by the assumption that the use of a certain hashtag or key word definitively indicates mask-wearing sentiment.

In the analysis of retweet networks, implementing algorithms to calculate modularity of a weighted, directed graph would offer more insight into the dynamics of such networks. To the best of our knowledge, there lacks readily available python implementations for community detection on weighted, directed graphs, leading us to represent the retweet networks using un-directed graphs instead. Thus, details as to which users serve as prominent sources of information are lost. Moreover, information on individual users' retweet patterns are lost in the calculation of the weights of the edges in the network. By exploring different ways to calculate modularity in a weighted, directed network, more information can be extracted from the analysis of retweet networks. Further, in examining polarization on Twitter relating to COVID-19, it would be interesting to investigate the temporal dynamics of hashtag networks (i.e. an un-directed graph in which two users in the graph share an edge if their tweet contained the same hashtag) and user-to-user mention networks (i.e. a directed graph in which two users share an edge if one user mentioned the other).

Based on the bias introduced in defining the state space, action space, and state features of the inverse reinforcement learning model, future work on the subject should be wary of making assumptions about human actions in data. The decision of which aspects of data constitute an action versus a state of an agent can potentially introduce bias into an inverse reinforcement learning model. Due to the bias introduced, using inverse reinforcement learning to examine human behavioral data means that the extracted reward function might mis-specify human goals. Thus, evaluating the performance of inverse reinforcement learning models in extracting reward function from observed users' behavior should be a key component of future work with this approach.

Future work should also further verify these findings by performing inverse reinforcement learning using different state spaces, action spaces, and state features. For example, one possible state feature could be derived from the number of in-edges versus

the number of out-edges in a retweet network. By using different state features on the same state and action spaces, future work can test hypothesis of different explanations for the same observed behavior.

Appendix A

Dates of Retweet Networks

Week	Dates
9	2/24-3/1
10	3/2-3/8
11	3/9-3/15
12	3/16-3/22
13	3/23-3/29
14	3/30-4/5
15	4/6-4/12
16	4/13-4/19
17	4/20-4/26
18	4/27-5/3
19	5/4-5/10
20	5/11-5/17
21	5/18-5/24
22	5/25-5/31
23	6/1-6/7
26	6/22-6/28
27	6/29-7/5
28	7/6-7/12
29	7/13-7/19
30	7/20-7/26
31	7/27-8/2
32	8/3-8/9
33	8/10-8/16
34	8/17-8/23
35	8/24-8/30
36	8/31-9/6
37	9/7-9/13
38	9/14-9/20

Table A.1: Gregorian calendar week number and 2020 dates

References

[1] *Amid campaign turmoil, biden holds wide leads on coronavirus, unifying the country*, 2020.

[2] *Twitter by the numbers: Stats, demographics & fun facts*, 2021.

[3] George A Akerlof, *Social distance and social decisions*, Econometrica: Journal of the Econometric Society (1997), 1005–1027.

[4] Kevin Arceneaux, Martin Johnson, and John Cryderman, *Communication, persuasion, and the conditioning value of selective exposure: Like minds may unite and divide but they mostly tune out*, Political communication **30** (2013), no. 2, 213–231.

[5] Saurabh Arora and Prashant Doshi, *A survey of inverse reinforcement learning: Challenges, methods and progress*, Artificial Intelligence (2021), 103500.

[6] Thomas Aynaud, *Community detection for networkx's documentation*, 2018.

[7] Edward E Azar and Chung In Moon, *Managing protracted social conflicts in the third world: facilitation and development diplomacy*, Millennium **15** (1986), no. 3, 393–406.

[8] Bin Bi and Junghoo Cho, *Modeling a retweet network via an adaptive bayesian approach*, Proceedings of the 25th international conference on world wide web, 2016, pp. 459–469.

[9] Sakun Boon-Itt and Yukolpat Skunkan, *Public perception of the covid-19 pandemic on twitter: sentiment analysis and topic modeling study*, JMIR Public Health and Surveillance **6** (2020), no. 4, e21978.

[10] Andrei Boutyline and Robb Willer, *The social structure of political echo chambers: Variation in ideological homophily in online networks*, Political psychology **38** (2017), no. 3, 551–569.

[11] Royce Carroll, Jeff Lewis, James Lo, Nolan McCarty, Keith Poole, and Howard Rosenthal, *Dw-nominate scores with bootstrapped standard errors*.

[12] Emily Chen, Kristina Lerman, and Emilio Ferrara, *Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set*, JMIR Public Health and Surveillance **6** (2020), no. 2, e19273.

[13] Daegon Cho and K Hazel Kwon, *The impacts of identity verification and disclosure of social cues on flaming in online user comments*, Computers in Human Behavior **51** (2015), 363–372.

[14] Marco Clemente and Thomas J Roulet, *Public opinion as a source of deinstitutionalization: A "spiral of silence" approach*, Academy of Management Review **40** (2015), no. 1, 96–114.

[15] Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini, *Political polarization on twitter*, Proceedings of the International AAAI Conference on Web and Social Media, vol. 5, 2011.

[16] Sanmay Das and Allen Lavoie, *The effects of feedback on human behavior in social media: An inverse reinforcement learning model*, Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems, Citeseer, 2014, pp. 653–660.

[17] Andrea DenHoed, *Silencing "india's daughter"*, The New Yorker.

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, *Bert: Pre-training of deep bidirectional transformers for language understanding*, arXiv preprint arXiv:1810.04805 (2018).

[19] Seth Flaxman, Sharad Goel, and Justin M Rao, *Filter bubbles, echo chambers, and online news consumption*, Public opinion quarterly **80** (2016), no. S1, 298–320.

[20] DJ Flynn, Brendan Nyhan, and Jason Reifler, *The nature and origins of misperceptions: Understanding false and unsupported beliefs about politics*, Political Psychology **38** (2017), 127–150.

[21] Abel Gustafson, Seth A Rosenthal, Matthew T Ballew, Matthew H Goldberg, Parrish Bergquist, John E Kotcher, Edward W Maibach, and Anthony Leiserowitz, *The development of partisan polarization over the green new deal*, Nature Climate Change **9** (2019), no. 12, 940–944.

[22] Jiyoung Han and Christopher M Federico, *The polarizing effect of news framing: comparing the mediating roles of motivated reasoning, self-stereotyping, and intergroup animus*, Journal of Communication **68** (2018), no. 4, 685–711.

[23] Mary Lynn Miller Henningsen, David Dryden Henningsen, Michael G Cruz, and Joshua Morrill, *Social influence in groups: A comparative application of relational framing theory and the elaboration likelihood model of persuasion*, Communication Monographs **70** (2003), no. 3, 175–197.

[24] Ki Deuk Hyun and Soo Jung Moon, *Agenda setting in the partisan tv news context: Attribute agenda setting and polarized evaluation of presidential candidates among viewers of nbc, cnn, and fox news*, Journalism & Mass Communication Quarterly **93** (2016), no. 3, 509–529.

[25] Shanto Iyengar, Gaurav Sood, and Yphtach Lelkes, *Affect, not ideologya social identity perspective on polarization*, Public opinion quarterly **76** (2012), no. 3, 405–431.

[26] Kathleen Hall Jamieson and Joseph N Cappella, *Echo chamber: Rush limbaugh and the conservative media establishment*, Oxford University Press, 2008.

[27] Edwin T Jaynes, *Information theory and statistical mechanics*, Physical review **106** (1957), no. 4, 620.

[28] Julie Jiang, Emily Chen, Shen Yan, Kristina Lerman, and Emilio Ferrara, *Political polarization drives online conversations about covid-19 in the united states*, Human Behavior and Emerging Technologies **2** (2020), no. 3, 200–211.

[29] Jenny Ungbha Korn, *'genderless' online discourse in the 1970s: Muted group theory in early social computing*, Morgan & Claypool Publishers New York, NY, 2016.

[30] Henrik Kretzschmar, Markus Spies, Christoph Sprunk, and Wolfram Burgard, *Socially compliant mobile robot navigation via inverse reinforcement learning*, The International Journal of Robotics Research **35** (2016), no. 11, 1289–1307.

[31] Matthew J Kushin, Masahiro Yamamoto, and Francis Dalisay, *Societal majority, facebook, and the spiral of silence in the 2016 us presidential election*, Social Media+ Society **5** (2019), no. 2, 2056305119855139.

[32] Hoang M Le, Peter Carr, Yisong Yue, and Patrick Lucey, *Data-driven ghosting using deep imitation learning*, (2017).

[33] Quanying Liu, Haiyan Wu, and Anqi Liu, *Modeling and interpreting real-world human risk decision making with inverse reinforcement learning*, arXiv preprint arXiv:1906.05803 (2019).

[34] Eric B Logan, *'fake news' has killed nigerians. can a bill stop the violence?*, Los Angeles Times.

[35] Luca Luceri, Silvia Giordano, and Emilio Ferrara, *Detecting troll behavior via inverse reinforcement learning: A case study of russian trolls in the 2016 us*

election, Proceedings of the International AAAI Conference on Web and Social Media, vol. 14, 2020, pp. 417–427.

[36] Kamaran H Manguri, Rebaz N Ramadhan, and Pshko R Mohammed Amin, *Twitter sentiment analysis on worldwide covid-19 outbreaks*, Kurdistan Journal of Applied Research (2020), 54–65.

[37] Aaron M McCright, *Anti-reflexivity and climate change skepticism in the us general public*, Human Ecology Review **22** (2016), no. 2, 77–108.

[38] Aaron M McCright, Sandra T Marquart-Pyatt, Rachael L Shwom, Steven R Brechin, and Summer Allen, *Ideology, capitalism, and climate: Explaining public views about climate change in the united states*, Energy Research & Social Science **21** (2016), 180–189.

[39] P Read Montague, Peter Dayan, Christophe Person, and Terrence J Sejnowski, *Bee foraging in uncertain environments using predictive hebbian learning*, Nature **377** (1995), no. 6551, 725–728.

[40] Mara Mordecai and Aidan Connaughton, *Public opinion about coronavirus is more politically divided in u.s. than in other advanced economies*, 2020.

[41] Sendhil Mullainathan and Ebonya Washington, *Sticking with your vote: Cognitive dissonance and political attitudes*, American Economic Journal: Applied Economics **1** (2009), no. 1, 86–111.

[42] Martin Müller, Marcel Salathé, and Per E Kummervold, *Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter*, arXiv preprint arXiv:2005.07503 (2020).

[43] Mark EJ Newman and Michelle Girvan, *Finding and evaluating community structure in networks*, Physical review E **69** (2004), no. 2, 026113.

[44] Andrew Y Ng, Stuart J Russell, et al., *Algorithms for inverse reinforcement learning.*, Icml, vol. 1, 2000, p. 2.

[45] Kieron O'Hara and David Stevens, *Echo chambers and online radicalism: Assessing the internet's complicity in violent extremism*, Policy & Internet **7** (2015), no. 4, 401–422.

[46] Andrew Perrin, *Social media usage: 2005-2015*, 2015.

[47] The COVID Tracking Project, *National testing and outcomes data*.

[48] Cristina Pulido Rodríguez, Beatriz Villarejo Carballido, Gisela Redondo-Sama, Mengna Guo, Mimar Ramis, and Ramon Flecha, *False news around covid-19 circulated less on sina weibo than on twitter. how to overcome false information?*, International and Multidisciplinary Journal of Social Sciences **9** (2020), no. 2, 107–128.

[49] Abraham C Sanders, Rachael C White, Lauren S Severson, Rufeng Ma, Richard McQueen, Haniel C Alcântara Paulo, Yucheng Zhang, John S Erickson, and Kristin P Bennett, *Unmasking the conversation on masks: Natural language processing for topical sentiment analysis of covid-19 twitter discourse*, medRxiv (2021), 2020–08.

[50] Kenji Sejnowski and Terrence J Doya, *Birdsong vocalization learning*, Advances in Neural Information Processing Systems 7 **7** (1995), 101.

[51] Jim Shanahan, James Shanahan, Shanahan James, and Michael Morgan, *Television and its viewers: Cultivation theory and research*, Cambridge university press, 1999.

[52] Leo G Stewart, Ahmer Arif, and Kate Starbird, *Examining trolls and polarization with a retweet network*, Proc. ACM WSDM, workshop on misinformation and misbehavior mining on the web, 2018.

[53] April A Strickland, Charles S Taber, and Milton Lodge, *Motivated reasoning and public opinion*, Journal of health politics, policy and law **36** (2011), no. 6, 935–944.

[54] Cass R Sunstein, *Going to extremes: How like minds unite and divide*, Oxford University Press, 2009.

[55] Ella Torres, *Backlash against asians could hinder efforts to contain coronavirus, expert says*, ABC News.

[56] M Waltz and K Fu, *A heuristic approach to reinforcement learning control systems*, IEEE Transactions on Automatic Control **10** (1965), no. 4, 390–398.

[57] David A Wemer, *Bolivia reflects the deep polarization crisis in latin america*, Atlantic Council.

[58] Axel Westerwick, Benjamin K Johnson, and Silvia Knobloch-Westerwick, *Confirmation biases in selective exposure to political online information: Source bias vs. content bias*, Communication Monographs **84** (2017), no. 3, 343–364.

[59] Magdalena Wojcieszak and R Kelly Garrett, *Social identity, selective exposure, and affective polarization: How priming national identity shapes attitudes toward immigrants via news selection*, Human communication research **44** (2018), no. 3, 247–273.

[60] Jonathan Woon, *Primaries and candidate polarization: Behavioral theory and experimental evidence*, American Political Science Review **112** (2018), no. 4, 826–843.

[61] Tauhid R Zaman, Ralf Herbrich, Jurgen Van Gael, and David Stern, *Predicting information spreading in twitter*, Workshop on computational social science and the wisdom of crowds, nips, vol. 104, Citeseer, 2010, pp. 17599–17601.

[62] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, and Anind K Dey, *Maximum entropy inverse reinforcement learning.*, Aaai, vol. 8, Chicago, IL, USA, 2008, pp. 1433–1438.