Modeling Considerations for Quantitative Social Science Research Using Social Media Data

Ceren Budak[1], Stuart Soroka[1], Lisa Singh[2], Michael Bailey[2], Leticia Bode[2], Nitesh V. Chawla[3], Pamela E. Davis-Kean[1], Munmun De Choudhury[4], Richard De Veaux[5], Ulrike Hahn[6], Brad Jensen[2], Jonathan Ladd[2], Zeina Mneimneh[1], Joshua Pasek[1], Trivellore Raghunathan[1], Rebecca Ryan[2], Noah A. Smith[7], Karen Stohr[2], and Michael Traugott[1]

May 17, 2021

[1] University of Michigan
[2] Georgetown University
[3] University of Notre Dame
[4] Georgia Institute of Technology
[5] Williams College
[6] University of London
[7] University of Washington & Allen Institute for AI

This series of white papers is funded by National Science Foundation awards #1934925 and #1934494 as a collaboration between Georgetown University and University of Michigan. To learn more about The Future of Quantitative Research in Social Science research project, visit www.smrconverge.org.

1. Project Overview

In 2019, a group of computer and social scientists began a project to 'converge' the two disciplines, with the aim of harnessing data from social media to improve our understanding of human behavior. People all over the world use social media, search engines, smart devices, and other technologies that record their moment-to-moment behaviors (often called "digital traces"). Social media, in particular, provide a massive amount of information on the everyday activities, opinions, thoughts, emotions, and behaviors of individuals, groups, and organizations in near real-time. Today, most adults in the US use some form of social media (Perrin & Anderson, 2019) to share and discuss topics as wide-ranging as politics, employment, parenthood, leisure activities, travel, sports, and health. As such, these platforms provide new ways of gathering information on constructs relevant to all social science fields.

Expanding the availability and utility of this extremely rich but still underutilized set of data sources in the social sciences requires attention to the unique features of these data. Unlike many forms of standard social science data, social media data do not have a structure that is the product of a designed process initiated by the researcher to answer specific hypotheses or questions. Instead, such observational data are provided "as is," which often means they are raw, complex, and highly sparse in nature. Moreover, these data introduce unique bias concerns not typically at issue in traditional social science methods, including a good deal of uncertainty about who generated the data, or what population those data represent. This organic nature of the data, along with their magnitude and complexity, require methods for managing, structuring, and understanding them in order to create useful measures for social scientific inquiry. Appropriate methods for doing so are most commonly found within the toolbox of computer scientists, making a convergence of computer science and social science methods potentially very fruitful.

While employing methods from computer science to wrangle digital trace data in order to answer social science questions has enormous potential, it also presents a number of challenges. First, neither computer scientists nor social scientists are especially well versed in the others' methods. So before social scientists can begin using ideas and algorithms from computer science, they need to learn how to work with large-scale unstructured organic data and understand the general principles, tools, and methods used by computer scientists. Likewise, computer scientists can reach inaccurate conclusions if they fail to understand key considerations and objectives within social science research that may not traditionally apply in computer science. Second, it is often unclear who or what, exactly, are behind the accounts that produce the data appearing in social media data sets, and how the entities creating trace data might relate to the larger groups of people that social science researchers would like to understand. For many questions, social media data contain information about the presence of a behavior but not about why or under what circumstances that behavior may have occurred (or may not have occurred) – which is a key area of focus for many social scientists. Third, ethical questions around the use of digital trace data in research contexts require collaboration across these disciplines. Understanding what we need to know about the data that are gathered, what other data are needed to supplement them

to answer central research questions, and how to do so responsibly is critical for digital trace data to live up to their full potential. These are just some of the challenges involved when social media data are used for social science research.

The advantages to social science in effectively harnessing social media data are clear. But for computer scientists too, this convergence holds great opportunity. Designing algorithms with a new set of constraints and optimizing existing algorithms for this large-scale, real-time domain, while addressing privacy, bias, misinformation, and algorithmic fairness concerns will also advance computer science research.

To initiate this convergence, our group planned a set of topical meetings bringing together social scientists from multiple disciplines, including economics, psychology, political science, communications, sociology, and survey methodology with data scientists and computer scientists with the goal of creating a common set of methodologies for how to study complex human behaviors using social media data in a scientifically rigorous manner. The topics of these meetings address each stage of the research process as we have defined it: study design; data acquisition, sampling, and preparation; measurement and feature engineering; model construction; analysis, and visual storytelling. At each meeting, we also discuss criteria for the responsible conduct of research with social media data.[8]

In this paper, the fourth in a series of white papers, we provide a summary of the discussions and future directions that came from the topical meeting that focused on model construction with social media data. A particularly interesting aspect of this meeting was, in our view, discussion of the different disciplines' requirements and approaches to modeling and the different considerations that are used to assess model fit.

2. General Modeling Challenges

There are a number of characteristics that make modeling social media data difficult. First, the stimulus for conversation is unclear. Social media users are a self-selected group. Language and behavior are affected by many factors including age, cultural context, and socioeconomics. Next, social media conversations do not always have the same meaning over time. There is temporal variation which can differ across subgroups on the platform. Third, platform design and algorithms influence user behavior, and the ways in which behaviors are influenced by platforms can be difficult to theorize about or detect. Also, social media contains natural language, which is inherently challenging to model. Mapping the text of a tweet to some notion of what the author "means," even taking context into account, forces us to make a huge number of simplifying assumptions. Finally, many statistical and machine learning models assume relatively clean data, but social media has repetitive data, missing data, a vast, ever-changing vocabulary, and high levels of sparsity, requiring us to rethink how we can use existing models and what conditions are the norm for new ones.

[8] We note that this introduction section is similar for all the white papers.

While all of these challenges need to be addressed, there are a few worth highlighting because of their impact on high-quality social science research. We begin with the sparsity and data quality challenges social media data pose. Because the properties of social media data are not well understood (described later), new types of error (possibly non-random and biased) are accumulating at every stage from sampling to measurement construction to model design. It is unclear how one should measure this compounding error since no standards have been developed to understand and account for it. (We discuss issues of error propagation in Section 5.) Second, different social media platforms have different levels of data quality and access. Part of the variability stems from the different purposes and designs of the platforms. Part of it results because some companies work harder at removing poor quality information, e.g. spam, and content derived from machine-driven accounts like bots, while others allow poor quality information to permeate through platform conversations. Therefore, the reliability of the data and the data producers need to be better understood. Further, when generating features (variables or measurable properties) for data mining or machine learning models, the social media feature space is sparse, meaning it is a high-dimensional variable space with a lot of missing values or zero values. For many machine learning algorithms, this sparsity reduces their ability to build an accurate and/or a robust model.

The next challenge we highlight is the variability around how and why computer scientists and social scientists use models. Computer scientists generally build models for prediction. Their goal is to develop mathematical and computational models that can be used to describe data, e.g. clustering and topic modeling, or to make predictions using different types of data, e.g. support vector machines and multi-arm bandit learning. In contrast, social scientists generally use models to test issues of causality or estimate the nature and strength of associations that are typically hypothesis driven. Their goal is to develop a theoretical model that can be tested using data. These ideas are discussed in greater detail in the next section. However, we want to highlight that the challenge of model design with social media is even greater because the modeling goals of these two disciplines differ.

Finally, the hallmark of high-quality research is replication. However, because platforms remove accounts and adjust account information regularly, it is complicated to replicate studies that use social media data. In many cases, exact replication is impossible. Therefore, as a community, we need to think through the value of data snapshots for research and the viability of them given platform terms of service.

3. Modeling Taxonomy and Background

Each field has its own interpretation of what a ‘model’ is. Scientists use conceptual, mathematical, theoretical, analytical, physical, analogical models -- to name just a few. The degree to which these different classes of models are used varies across disciplines and modeling purposes. This creates an inherent difficulty for holding discussions of modeling in interdisciplinary spaces.

In order to establish a common language, in this paper we use a simple taxonomy that we believe captures a broad range of analytics relevant to social media research as well as the broad range of disciplines contributing to that research. We acknowledge that this taxonomy, much like any such scheme, is a simplification and not reflective of all research in this area; indeed, some scholars voice skepticism that models belong to distinctive categories and claim that anything can be a model (Frigg & Hartmann, 2020). We hold the position that taxonomies of models, much like models, are all wrong, but some are useful; and while we structure the rest of our paper around the simple taxonomy we are proposing, we also discuss the limitations of our proposed structure in later sections.

Our simple taxonomy categorizes models under four categories: *descriptive* (Section 3.1), *diagnostic/explanatory* (Section 3.2)*, predictive* (Section 3.3), and *prescriptive* (Section 3.4). Below, we provide a short description for each of these categories and present examples of different statistical and computational models that fit into these categories as they relate to social media data. We note that the models presented in each category may be useful for other categories. However, we have placed them in the category that is most applicable given the suggested taxonomy. We can of course provide only brief introductions to statistical and computational approaches here; but we provide citations to both textbook and applied examples for different approaches. Readers interested in more detail will, we hope, find these sources valuable.

Naturally, there are a number of other useful categorizations that are not as deeply explored here. For instance, models can be generative or discriminative. Discriminative models attempt to learn by identifying decision boundaries between groups. This tends to map to learning conditional probability distributions that distinguish groups. Generative models focus on determining the generation process of the data and learn the joint probability distribution. Generative models make modeling assumptions which allow them to use regularities or patterns in input data to generate new examples that plausibly could have been drawn from the original data set. This is another reasonable way to analyze different models and when useful, we refer to this taxonomy as well.

3.1 Descriptive Models

Descriptive models aim to describe what happened in the past. These models are most useful when exploring data to understand their structure or to determine the characteristics of the individuals or groups of individuals generating the data. The input to such models commonly includes a set of variables (X) that can be used to describe the cases or units that comprise the data. Descriptive models commonly create a subset or new group of either variables or cases that serve to simplify the data or to describe some interesting characteristic(s) about the data. Examples of models used for descriptive analysis include clustering, topic modeling, and

anomaly/outlier[9] detection. We select these examples because they are particularly useful for social media data analysis.

Clustering: The goal of clustering is to identify similar cases or units (typically individuals or objects) to create clusters or groups based on shared features/attributes/variables (Tan et al., 2005). Each cluster contains cases that are more similar to others within the group than to those outside the cluster. Grouping cases by a single known feature, e.g. gender, is not viewed as clustering, but rather as data segmentation since only one feature is used. Instead, clustering algorithms focus on efficiently grouping individuals using multiple features.

Many different types of algorithms have been created for clustering. For example, hierarchical clustering models use different notions of distance to build a hierarchy of clusters, where each step involves identifying which clusters are closest to each other and merging them into larger clusters (Murtagh & Contreras, 2011; Murtagh & Contreras, 2017). One way these models have been used to organize social media data is through hierarchical agglomerative clustering, used to identify clusters of tags used by individuals in social networks (Shepitsen et al., 2008) or as a way to cluster misinformation memes on Twitter (Ferrara et al., 2013). K-Means clustering also uses distance, but instead partitions the units into k clusters, where each individual is added to the cluster with the closest cluster centroid, thereby minimizing the within cluster variance (Lloyd, 1982). Variants of k-means have been used to cluster social media users with respect to textual similarity (e.g. Miller et al., 2014). Both hierarchical and k-means clustering are considered classic clustering methods that are used across disciplines. Another clustering approach is distribution-based clustering based on a statistical distribution, typically a Gaussian mixture model (Hastie et al., 2009). Each cluster contains points that follow a Gaussian distribution that is part of the Gaussian mixture. This model has been used for clustering words from posts on social media and in newspapers to understand core conversation and article themes (Yin & Wang, 2014).

Density-based clustering models search for dense regions in a noisy data space by labeling points based on their closeness to other points. Those points that can be easily reached are put in the cluster and those that cannot be easily reached by any group are considered outliers and are not added to any clusters (Tan et al., 2005). A variant of DBSCAN, a popular density-based clustering algorithm, has been used to cluster posts based on geotagging (Liu et al., 2018). These types of algorithms that remove noise can be important for social media domains. Finally, there are many clustering or community detection algorithms that create clusters based on network connectivity. Here, nodes are grouped together if they have more connections to nodes within the cluster than to nodes outside the cluster. A large number of models exist ranging from models created from local hierarchical clustering methods (Girvan & Newman, 2002) to ones that are based on more global optimizations that identify communities by finding groups of

[9] While anomalies and outliers are not synonymous, for the purposes of the methods discussed here, they can be used interchangeably.

nodes with more connections than would be expected if the network was random (Newman, 2006). All of these methods also have variants that identify time varying clusters and overlapping clusters (see Xu & Wunsch, 2005; Xu & Tian, 2015; Hruschka et al., 2009 surveys for more information).

Topic models: Topic modeling is a variant of basic clustering that focuses on finding clusters of words and phrases that can be grouped together to represent topics within a document or text collection. (Note that our unit of analysis has shifted here -- in the previous section we focus on clustering of individuals, while here we consider topic modeling of text.) There are a large number of different topic models that have been proposed (see Qiang et al., 2020 for recent survey). The most well-known generative model for topic modeling is Latent Dirichlet Allocation (LDA) (Blei et al., 2003). This algorithm assumes Dirichlet-multinomial distributions for both the words and the topics. While it works well on more traditional documents, it does not capture topics as well in noisy data domains with short texts, i.e. social media. A number of LDA variants have been proposed to address the challenges of noise and short documents (Wang et al., 2012; Quan et al., 2015; Nguyen et al., 2015; Moody, 2016; Hong & Davison, 2010). These models have been used extensively on social media. For example, Surian and colleagues uses both topic modeling (LDA) and community detection to cluster opinions about human papillomavirus (HPV) vaccines on Twitter (Surian et al., 2016). Graph-based topic models have also had success, particularly for temporal topic modeling (Churchill et al., 2018; Cataldi et al., 2010; Churchill & Singh, 2020). These models build networks or graphs, where each node represents a word or phrase in the document collection and edges exist between nodes that co-occur within the same post. Then topics are identified by finding ways to partition the graph into groups using variants of community detection algorithms. Here each group represents a different topic. In the context of social media data, for example, graphs have been used to build topics for the 2016 US presidential election (Churchill et al., 2018; Churchill & Singh 2020).

Anomaly detection: Anomaly detection or outlier detection focuses on identifying rare observations, objects, or events. Classic applications include fraud detection, fault detection, and health monitoring. They have also been applied in the analysis of social media on important topics such as detection of spam, fraud, and bullying (Wang et al., 2012). This problem is formulated as both an unsupervised and supervised learning problem. Unsupervised approaches do not rely on training data and assume that the majority of data define “normal” while deviations from this majority can be assumed to be anomalous. The classic anomaly detection techniques are density based (Tan et al., 2005). They identify neighborhoods for each observation and identify outliers by finding observations that have a smaller density than their neighbors or a smaller density than an expected density for the data set. Computational models vary in how they define a neighborhood and how they measure what is close and what is further. For event detection or other time series anomaly detection methods, algorithms will look for bursts in activity. For example, one can use bursts in conversation discussion about a particular topic or occurrence to identify events on Twitter (Wei & Singh, 2017), or use methods to aggregate hidden information during a crisis (Leavitt & Robinson, 2017). Supervised

approaches rely on training data that include known normal and anomalous examples. Because of the volume of data and the difficulty of finding anomalous examples (maintaining class balance), creating a training data set can be labor intensive. One example where supervised anomaly detection has been used is sarcasm detection (see Joshi et al., 2017 for a survey of semi-supervised and supervised methods). However, the class imbalance and the construction of the training data can have large impacts on performance results (Abercrombie & Hovy, 2016).

3.2 Predictive Models

These models are useful when aiming to predict what will happen in the future or what would have happened in a counterfactual past. The input to such models commonly includes two types of data: 1) variables (X) that describe the data, and 2) the outcomes (Y) observed for different predictors (X). Predictive models describe the relationship between X and Y; and this can in turn be used to predict Y ($\hat{Y}$) from X for future values.

Researchers typically need to make two choices when it comes to determining which predictive model to use: a) the set of variables that should be part of the model and b) the functional form of the relations considered. The first task is feature engineering, and social media data provides both an opportunity and a challenge. Given the vast amount of data available, the potential feature space is large. The second challenge is to pick the specific model. Example predictive models include regression, support vector machines (SVM), decision trees, and random forests. These models are supervised learning models, meaning that they need labeled training data to inform their predictions. Here we briefly describe models that have been used for different predictive tasks related to social media. We refer the reader to Hastie et al. (2009) for a more detailed discussion.

We pause to mention that when computer scientists write papers about predictive models, they typically compare results from multiple machine learning algorithms to gain insight about unique characteristics of the data set. In contrast, social science papers typically select a model and then use that as a way to support a particular theory of interest. This difference is important to note because the examples we present draw from both communities, and there are fewer examples of many of the machine learning techniques computer scientists are developing in social science research.

Linear regression and variants: Regression models are the most common models in the social sciences. The goal of regression is to model the relationship between a dependent variable (Y) and one or more independent variables (Xs). The outcome variable of interest dictates what type of regression needs to be performed. If it is a continuous variable, linear regression can be used. If it is a binary outcome, logistic regression is most commonly used. If the outcome variable is a value between 0 and 1, beta regression may be preferred. Regression is commonly fitted using the least squares approach but other loss functions are also possible. For instance, especially in a setting with a large number of independent variables, one might want to employ a penalized version of the least squares cost function, such as ridge regression (L2-norm penalty) or lasso regression (L1-norm penalty) (Hastie et al., 2009).

Regression models are also used for time series analysis. Most time-series econometrics is regression-based, and analyses of monthly, weekly, daily, or even hourly aggregates of social media content may be analyzed in this way. As another example, multilevel regression (hierarchical linear models) can be used to account for the data's hierarchical structure. These models can be particularly useful when the number of observations one has per unit is non-uniform. This is indeed the case for social media data where a small number of highly active users account for a large fraction of the content. As such, these models are frequently used in social media studies (e.g. Budak & Watts, 2015; Rajadesingan et al., 2020).

Linear and logistic regression have been used in a large number of social media studies, everything from personality trait prediction using user Facebook activity data as the predictors (Bachrach et al., 2012) to predicting political interactions associated with the 2016 US presidential election using posting behavior on Reddit (De Francisci Morales et al., 2021) to using photographs on Instagram to predict markers of depression (Reece & Danforth, 2017).

Support vector machine (SVM): SVMs are typically used for classification to predict a binary outcome. SVM is closely related to classical regression. The distinction is in their loss function. For instance, linear regression uses least squares, while linear SVM uses hinge loss. The objective of the model is to place points into the two classes in such a way that the "width" of the hyperplane (the decision boundary) between the two classes is maximized. The researcher can also choose from different kernel functions, depending on the task at hand. For instance, while a linear kernel can perform well in cases where there are no significant interactions between the independent variables, a different kernel, say a quadratic one, might be preferred when the interactions are significant. SVMs are generally more highly effective on non-linear high dimensional data. Different studies have used SVMs (and other methods) to predict health conditions like depression (Aldarwish & Ahmad, 2017; De Choudhury et al., 2013) and infer demographics (Chen et al., 2015).

Decision trees: Decision trees are another type of model that be used to perform classification (predict a class value) or predict a real number. The constructed model basically encodes a set of consecutive rules that can be used to predict an outcome. These rules can be viewed as a set of decisions related to subsets of variables that need to be made to reach an outcome. Algorithms for building decision trees choose variables that best split the data into subgroups containing similar features and outcomes. Different algorithms use different metrics for determining the quality of a "good" split, e.g. information gain. Decision trees are easily interpretable, handle missing data well, and are very fast. This presents an important advantage for social media data that can be rather large and has missing data. Interpretability is also important for social science research. However, such models can be prone to overfitting training data. There are strategies to combat that, e.g. setting a maximum tree depth, but they can introduce bias error. Because of the overfitting issue, these models are not used as extensively with social media data, when compared to other models, they do not always perform as well. For example, researchers use a number of classifiers including decision trees to predict information

credibility on Twitter (Castillo et al., 2011). When it performs well, it is useful for interpreting the most important features.

Random forest: Random forest models are ensemble methods for classification or regression. Intuitively speaking, these models use samples of the training data to construct a set of decision trees (weak learners). To make the final prediction, the random forest model outputs a majority vote (the mode class across all trees) for classification or the average of the predicted values for regression. These models are highly effective when the training data are imbalanced or contain missing values. They tend to work better than traditional decision trees on high dimensional data, but still may overfit data. Random forest models have been applied effectively to social media text data for spam detection (Chu et al., 2012; McCord & Chuah, 2011).

Naive Bayes: Naive Bayes is a generative model using Bayes theorem to model the posterior probability P(Y|X) using P(Y), P(X), and P(X|Y). It does this by assuming that the predictors are independent of each other, i.e. each feature in X is unrelated to other features in X. Even though this may not be true, in practice this assumption is reasonable for many data sets. It has performed well for document classification tasks and therefore, may be useful for some social media classification tasks. Compared to logistic regression, this approach has a higher bias but lower variance. Under what circumstances should a researcher prefer one over the other? Past work provides some insights (Ng & Jordan, 2002). Logistic regression performs better than Naive Bayes for large data sets. However, the generative model (Naive Bayes) reaches its asymptote faster. For instance, Ng and Jordan observed that some of the 15 data sets from the UCI ML repository did not have enough samples for logistic regression to perform better than Naive Bayes (Dua & Graff, 2019). As an example, Dilrukshi and colleagues use emoticons in a Naive Bayes model and an SVM model to predict sentiment (Dilrukshi et al., 2013).

Neural network models: While the previous mentioned models are examples of classic supervised learning models, neural network models have gained popularity recently, particularly in the form of deep learning models. Neural network models are inspired by the way a human brain works. These models are composed of layers of simple, connected nodes that each compute a function using different combinations of the input data, and ultimately learn to approximate an underlying mapping function from the inputs to the output. Each layer focuses on different patterns of the input data and learns rules based on these patterns. This is accomplished by learning weights and model parameters given a specific network structure (architecture). Typically, the optimization problem being solved is non-convex. Neural models have been particularly useful in modeling images and text. A more detailed treatment of the topic can be found in Aggrawal (2018).

There are several neural network architectures with different features. One's choice of the exact neural network model depends on the application, but a few commonly used ones are multilayer perceptrons (MLP), convolutional neural networks (CNNs), recurrent neural networks (RNNs). MLPs are feed-forward artificial neural networks. This means that information only flows in one direction through the network. They consist of three main layers: input layer, hidden

layer, and an output layer, where each node or neuron in one layer is connected to all the neurons in the next layer (fully connected). While different learning techniques can be employed on this network, one well known technique is backpropagation. During training, output values are compared to expected output values and information about the error is used to help adjust the weights of the nodes in order to obtain output values that are closer to the expected outputs. MLPs are used in classification and regression. Because MLPs are fully connected, they sometimes overfit data. CNNs attempt to combat this by having multiple hidden layers grouped into three blocks—convolutions, pooling, and fully-connected layers. The first two blocks focus on feature extraction and the last block maps the features to the output. CNNs are commonly used in image recognition, image classification, and object detection. RNNs allow information from prior inputs to influence current inputs and outputs, allowing for easier modeling of sequential data. This "memory" distinguishes it from the previous two models. RNNs are particularly useful for natural language processing and speech recognition applications. All of these models have been used successfully for image classification tasks (Wang et al., 2016; Guillaumin et al., 2009). As more social media image training data sets become available (Ulges et al., 2010; Deng et al., 2009), we may begin to see more of these methods used for social science related research.

A challenge with using many of these models for text tasks is that they need large amounts of training data and a large compute infrastructure to build. To help mitigate this challenge, computational linguists and computer scientists have been building general purpose language models. Use of these models has been shown to improve the prediction performance in NLP tasks over classic machine learning methods described earlier in this section. These language models are built using millions or billions of examples of raw text. The general idea is the following. While words alone give researchers insight into the text, understanding contexts surrounding words and identifying words having similar contexts can be important for many machine learning tasks. While this knowledge of word contexts can be generated by experts in dictionaries like WordNet (Fellbaum, 1998), or by using clustering algorithms that group similar words (Brown et al., 1992), another approach is to create word vectors where each dimension of the vector maps to the frequency with which a word occurs in a specific context (Smith, 2020). This allows algorithms to consider the distribution of contexts associated with any word. Word representations can be associated with words irrespective of contexts. Therefore, language models that use word vectors can be built considering different amounts of context. Non-contextual models generate a single word representation (word embedding) for each word even if a word can be used in multiple contexts. Word2Vec (Mikolov et al., 2013), Glove (Pennington et al., 2014), and FastText (Joulin et al., 2016) are examples of non-contextual word embeddings. Contextual models generate word representations based on other words in the sentence. BERT, the Bidirectional Encoder Representations for Transformers, is a contextual model. It is widely used and has been shown to provide state-of-the-art results on a wide variety of natural language processing tasks (Devlin et al., 2019). These language models can perform relatively well with some task-specific fine-tuning. Some relevant prediction tasks that have made use of these

language models with social media data sets include demographic inference (Liu et al., 2021), stance detection (Kawintiranon & Singh, 2021; Ghosh et al., 2019), sentiment (Tian et al., 2020), and bot identification (Kudugunta & Ferrara, 2018).

Finally, while we mention each predictive model in isolation above, various studies combine models through ensemble or stacking (Tang et al., 2014) approaches. A voting ensembler method uses votes (e.g. predicted as 0 or 1) from multiple classifiers and applies majority voting. A stacking approach combines multiple models through a second level meta-classifier where the prediction outputs by the level-1 classifiers are used as inputs to the meta-classifier. These techniques have been used in past social media work to collect and classify social media content (King et al., 2017) and identify social movement organizations (Bozarth & Budak, 2020), among others.

3.3 Diagnostic/Explanatory Models

These models aim to identify the relationship between variables -- i.e., relationship between X and Y or how X explains Y -- without any interest in the prediction of Y *per se*. Predictive models aim at estimating $\hat{Y}$ (predicted Y values) when a new observed X is collected while the explanatory models aim at estimating $\hat{\beta}$ (estimates of model coefficients), the effect of X on Y. In other words, while the input to the explanatory models is similar to predictive ones (X's and Y's), the goal is not to predict the value of the dependent variable Y ($\hat{Y}$) but to understand the relationship between X and Y ($\hat{\beta}$) in terms of the coefficient for any independent variable.

Researchers often use the set of methods listed under predictive, and to some extent descriptive, models also for explanatory goals. Standard regression techniques are the most common approach to explanatory analysis. Regression models do not, on their own, identify causality. In many instances, causal claims depend on the assumption that the Xs are 'causally prior' to the Ys. For instance, we might theorize that some demographics drive social media behavior because those demographics are determined prior to, and exogenously from, that behavior. Most cross-sectional survey-based work relies on this approach to causality, for instance (Hughes et al., 2012; Orchard et al., 2014), as does a good amount of social media analysis (see Boulianne, 2015). Some research designs or analytic approaches are better equipped to test such causal mechanisms, including quasi experimental designs, e.g. interrupted time series design (Budak et al., 2017), regression discontinuity analysis, propensity score matching (Aral et al., 2009; Zhang et al., 2017a), and/or the use of instrumental variables (Bollen, 2012). In each of these cases, the objective is to model Y as a function of some set of independent variables, where the latter are regarded as possible 'explanations' of Y but also correlated with the Xs. In other cases, researchers are not necessarily interested in a direct causal relationship between the independent and dependent variables, but instead interested in how the independent variable influences one or more mediator variables, which in turn influences the dependent variable. These are solved using mediation analysis (e.g. Zhang et al., 2017b), which is used to model complex systems.

3.4 Prescriptive Models

These models are helpful in prescribing optimal actions to achieve a particular preferred outcome. The input to such models commonly includes three types of data: 1) variables (X) that describe the data, 2) action (A) taken, and 3) the outcomes (Y) observed for different values of X and actions A. Action A can be thought of as an intervention. The goal of this family of models is to prescribe actions that optimize (maximize or minimize) Y given X. Examples include reinforcement learning, agent-based modeling, and network optimization.

Reinforcement learning: The premise of reinforcement learning differs from unsupervised and supervised learning. The learning problem is setup with an agent or set of agents that make decisions within an uncertain environment and a policy that serves as the rules of the learning task. The agent is an automated unit that receives observations and a reward from the environment and sends actions to the environment. Policy refers to the mapping that selects actions based on the observations from the environment (inputs). Through trial and error, the agent makes decisions. When a good decision is made, the agent is rewarded. When a bad decision is made, it is punished. In this way, the model learns to perform the task so as to maximize the expectation of a long-term reward. Successful reinforcement learning models are able to effectively balance exploring new parts of the search space (learning more) against exploiting the knowledge they have. Classic examples of using reinforcement learning include playing games (Mnih et al., 2013) and training autonomous cars (Sallab et al., 2017). Some newer examples that are particularly relevant to social media include using reinforcement learning to counter misinformation (Kaiser et al., 2020), to adaptively collect data online (Li et al., 2016), and to identify bots (Luceri et al., 2020). The reason it is effective in these spaces is because the model is adaptable, i.e. the computer learns from its own mistakes. In the context of a prescriptive analysis, once the agent optimizes for the conditions that exist, e.g. detecting bot behavior, the model can be adapted to learn how bots change their behavior over time, helping to reduce their influence on the platform. For a survey of foundational research on reinforcement learning, we refer the reader to Kaelbling et al. (1996).

Network optimization: Network optimization models attempt to use networks to solve different decision problems efficiently. A network or graph contains a set of nodes or vertices and a set of edges. These edges may have direction or weights depending on the specific network constraints. While there are many different network optimization problems, one that is useful in the context of social media is maximizing spread (of a behavior) or minimizing spread (of a virus or misinformation), i.e. a network diffusion process. In this problem, the edges or relationships can influence the behavior of a node. Each node that is "active" has a probability of being influenced by other active nodes or of influencing other active nodes. By simulating this diffusion process, researchers can model how information/opinion diffuses and through what paths. These models and optimization tasks have important implications for observing and affecting what societies care about (Agrawal, 2011). These approaches can help with the prescriptive task of attempting to maximize the spread of information (Kempe et al., 2003), place

sensors to monitor and detect diffusion (Leskovec et al., 2007), or limit/slow down the spread of diffusion (Budak et al., 2011) by understanding how the diffusion model must be changed to change the overall spread through the system (see the survey by Guille and colleagues for more information (Guille et al., 2013)).

Agents-based models: Agents-based models (ABM) consist of a group of autonomous decision-making entities or agents that model a system. Agents are simulated within the system to understand how the system will change when different conditions are specified. Similar to reinforcement learning, agents make decisions based on a set of rules. ABM models tend to include multiple types of agents that interact with each other in different ways, sometimes in competition. ABM models emulate real world systems, allowing agents to evolve to the point where unexpected behaviors begin to emerge. ABM models can describe systems in a flexible way, can use simple rules to model meaningful group behavior, and are useful for capturing emergent behavior. Foundational social science research used ABM to understand phenomena such as segregation (Schelling, 1971) and voting behavior (Kollman et al., 1992). We refer the reader to (De Marchi & Page, 2014) for a survey of relevant studies in political science and to (Heath et al., 2009) for a broader analysis of ABM related papers. More recent models will incorporate a range of learning techniques within the behavior of the agents, including neural networks and different types of reinforcement learning. Bayesian models can be used to represent the factors influencing an agent's decision-making process. Some applications where agents-based models have been used successfully include traffic flow management, operational risk models, and product adoption dynamics (Sun et al., 2006; Talukdar, 2002; Cowell et al., 2007). In the context of prescriptive models for social media, ABMs can be used to understand how to intervene to increase the spread of high-quality information or decrease the spread of poor-quality information. These models are commonly tested through simulation. The limitation of these models is that they are customized and therefore can be difficult to create quickly for different contexts.

System dynamics models: While ABMs can be used to model complex systems, other system dynamics models also exist. These simulation models focus on mathematically capturing the system dynamics, including different feedback loops that are the hallmark of these types of models. System dynamics models ignore the details of the system and instead focus on a higher level, general representation of the system (see Weisbuch for a more detailed discussion (2018)). One approach to this is to model the real world as stocks or entities, flows between stocks, and information to make adjustments to the flows. These types of systems dynamics models are used extensively within engineering fields and have been used to model health care capacity and delivery (Homer & Hirsch, 2011), movement during forced displacement (Anderson et al., 2007), and physical systems (Ford, 1999). One example uses a system model with different data sources including social media to understand forced displacement (Singh et al., 2019).

4. Cross-Disciplinary Distinctions in Model Construction

Section 2 described some general approaches to modeling social media content. This section focuses on cross-disciplinary issues that arose during the meeting. We view these issues as raising important issues for consideration by those interested in cross-disciplinary research and collaboration.

A first concern for any discussion of model construction is a definition of what exactly "modeling" is. There appears to be some general agreement within each discipline about what modeling is, but there quite clearly are differences across disciplines. These differences make a discussion of model construction complex. However, identifying and considering the differences is, in our view, a useful (and also a necessary) step in developing better approaches to model construction for social media data.

Below, we discuss four of the differences in the ways in which social scientists and computer scientists approach model construction: (1) differences in what falls into the 'modeling' category, (2) differences in emphasis on inductive versus deductive modeling, and (3) differences in models based on quantities for which there is an identifiable 'ground truth'.

4.1 What is Modeling?

For the typical social scientist, a model is used to generate and/or test an hypothesis. Data are collected. Measures are developed. And after some recoding and descriptive analyses, the analyst presents a model (a regression model, for instance) that explores the relationship between X and Y. This does not capture all the various ways in which social scientists think of modeling, of course. But the vast majority of social scientific data analyses proceed in this manner.

Computer scientists are inclined to interpret modeling more broadly. The decisions about what to measure are part of modeling. So too are the various mathematical transformations needed to produce the variables that eventually find their way into an analysis. For a computer scientist, all of the various decisions that turn raw material into measures and analyses are commonly thought of as being part of model construction. While these are parts of model construction, computer scientists also place a lot of emphasis on the development and optimization of computational models for different descriptive and predictive tasks.

There clearly are advantages to thinking about modeling using the broader framework that is more common in computer science. Social scientists have a strong tradition of examining the importance of measurement; and the consequences that result when a single measure of a construct is designed in several different ways. However, unlike computer science where measurement and modeling are intertwined, the social scientist often designs the measures as they relate to data collection and then incorporate these into the models during the analysis stage.

The importance of decisions related to measurement and prediction are especially clear for the analysis of social media data, where the transformation of raw data into usable measures requires so many different conceptual and mathematical decisions. Indeed, some measures are

quite clearly developed by modeling them, using machine learning for instance. Where social media data are concerned, and perhaps more generally as well, there are advantages to including the development of measures as part of the process of model construction. What may in fact be more important is to recognize that different forms of modeling are taking place throughout the social media research process from theoretical models that support different hypotheses to predictive models to support data preparation and measurement construction to computational models which use changing data to make decisions.

4.2 Inductive versus Deductive Modeling

Inductive modeling explores data as a way of developing theory; deductive modeling uses models to test existing hypotheses. Both approaches are of obvious importance for social scientists and computer scientists alike. But social scientists tend to focus primarily on deductive approaches whereas computer scientists are more comfortable using a combination of approaches.

This difference is linked to the differences in 'modeling' discussed in the preceding subsection – the inclination to view measurement as a component of modeling necessarily requires a view of modeling that is at times primarily inductive. Where model construction for social media is concerned, we see inductive approaches as being critical. Part of the task of researchers in a relatively new field is simply to get a sense of the data, after all. We are only beginning to understand what social media data look like and how they can be analyzed. Simple descriptive data can be a major contribution to the study of social media; and given the relative absence of theory (in comparison to more long-standing areas of social-scientific study, at least) the potential for deductive modeling is relatively limited.

Our argument here is not that there is no theory in analyses of social media; nor is it that social media analysis should be exclusively inductive. We nevertheless suspect that social scientists may sometimes benefit from an approach that takes more seriously the benefits of inductive modeling to better understand social media content. (Though we note that computer scientists may benefit from considering the importance of deductive, theory-oriented modeling as well; see, e.g., Lundberg et al., 2021.)

4.3 What is the 'Ground Truth'?

There are differences in approaches to model construction for quantities in which there is or is not a 'ground truth.' Consider first a machine-learning algorithm or dictionary designed to identify instances of sarcasm in social media. There is a 'ground truth' in this instance – whether or not the post was intended to be sarcastic. Modeling sarcasm can consequently proceed with an aim of reliably identifying that ground truth. Where the post was intentionally sarcastic, our model should identify that sarcasm. Where the post was not intended to be sarcastic, our model should not identify sarcasm. Any difference between our estimation and the 'ground truth' represents a failure in our model. (It is of course very hard to identify sarcasm in social media

using any automated system, at least partially due to the difficulty of gathering unbiased training data (Bamman & Smith, 2015).)

What if we want to identify whether a post is funny or not, or contains offensive language? This is more complex, because funny (or offensive) is not just about the intention of the person who posted, but also about audience members' interpretation of the post. And individuals will invariably disagree on which posts are funny and which are not. There may be an audience average for how funny a post is, but even that will vary across audiences. There is no 'ground truth.' And there is, as a consequence, no clear indication of whether the defined measure is correct.

Many researchers are used to working with data for which there is no ground truth – consider survey questions that capture attitudes that are evident only because we have asked the survey question, for instance. This kind of data makes for a very different approach to model construction, however. Where 'ground truths' are readily available, the task of model construction is to approximate that ground truth. Where ground truths are not available (such as, for instance, in sentiment coding, where the sentiment of text is often subjective), model construction must depend on strong theory and/or indications of concurrent validity, perhaps through a quantity or set of mathematical quantities that can be used to identify what should be expected.

5. Issues that Apply Across Models

Another way to look at challenges is to consider those that are applicable to many models in both social and computational science -- common difficulties, rather than cross-disciplinary differences. In this section, we highlight some of the common challenges that were prominent in our meetings.

5.1 Correctly Identifying the True Purpose of Modeling

It is common practice to map a particular research question to a category of an analysis/modeling family. A big part of the research process is to take a high-level question and formulate it into a set of measurable constructs. This process, when carried out carelessly—as it often is—can lead the researcher to misclassify the question at hand.

One of the most common misclassifications relate to predictive vs. explanatory. Predictive approaches are typically employed within computer science. For example, a computer scientist may build a model to predict the weather or the topic of discussion. Even if these models are set up to provide longitudinal estimates of an outcome, they do not inherently capture causal relationships. If the actual goal of the researcher is to identify causal relationships, such an approach is likely to fail since the goal is about the prediction (the what), not the explanation (the why). Yet, predictive models are used for explanation in various social science settings (Lundberg et al., 2021). Delineating the intended purposes of a forecasting model, which is understandably hard, can provide clues as to whether causal modeling needs to be a part of the

predictive approach. Approaches from the causal inference literature that allow capturing these relationships may be incorporated. It is important to note the following: not every question needs to involve an explanatory model. For instance, consider the policy of whether an individual leaving home should be handed an umbrella. A good policy here needs to simply reliably predict whether it will rain that day. This provides a simple categorization of problems into "β-problems" (the focus on the regression coefficients, more common in social science) and the "$\hat{Y}$ problems" (the focus on the actual prediction, more common in machine learning). The umbrella case is an example of a $\hat{Y}$ problem.

Another example of misclassification happens between predictive and prescriptive approaches. Take, for instance, the example of predicting leading indicators of toxicity on social media. While this can be seen as a predictive, or even an explanatory, analysis at the surface, the goal of the researcher/policymaker/platform is rarely just to predict/understand. In reality, the true goal is likely closer to determining the factors that can be used to develop an intervention strategy for reducing toxicity. In short, we too often are asking predictive questions while the true goal is prescriptive. A more bold version of this claim goes as follows: the goal of a model is almost *never* to simply predict. *Somebody* will take an action based on those predictions. This mind shift can lead the researchers to more carefully consider the quality of their predictions and how the accuracy of these predictions would impact the actions taken. This has ethical implications, as well as practical ones.

5.2 Time and Modeling

Social media data are not a monolith. Neither do they exist in a vacuum. As we appropriate social media data for all sorts of models, the question is -- what are we capturing and what are we missing? How does that change over time, as platform politics, norms, users, and the broader social contexts (e.g. real-world crisis events) evolve? Implied in all these questions is the importance of time. Time plays an important role in modeling. Researchers should 1) carefully construct their time scales, 2) assess IID (independent and identically distributed) assumptions, and 3) quantify and mitigate the impact of drift.

Carefully constructing time scales: There are also issues of detection of important phenomena that we want to predict without having paid any prior attention to preceding factors that might have explained or predicted it. An event occurs that requires (in an intellectual or scientific way) an explanation. How far back in the social media data file do we go to look for explanations? Should the model place a premium on as much advance notice as possible? Or should the emphasis be on the level of accuracy in prediction irrespective of the preceding time period, i.e. only short advance notice? This may be another way of describing the evaluation of models in terms of "fit for purpose."

Assessing IID assumptions: The assumption of independent and identically distributed (IID) random variables is not always reasonable in the analysis of social media data. Take for instance a text classifier for detecting misinformation based on hand-crafted features (e.g., word lists) or "neural" features based on distributed representations of words. Classical machine

learning models generally treat each data point as IID. However, in a world where social media users learn from each other and real-world events dictate their behavior, IID assumptions can be unrealistic. This makes time series analysis an important tool in the researcher toolbox.

Algorithmic drift: Trained models can get stale over time for various reasons. Algorithmic or concept drift occurs when the relationships in the underlying data change, thereby degrading the performance of the predictive model. For instance, social media norms continually change, making language detected by a language model obsolete. Important social media users might leave the platform, significantly changing the dynamics of influence modeled through network models. Or, the platform can take action to change its design or blacklist large groups of users. Algorithmic drift can also happen because of changes occurring in the real world. How can researchers handle these issues? First, it is crucial to continually test models to check for drift. Furthermore, it is desirable to build models that keep learning and have a feedback loop. Online learning techniques are commonly used in these circumstances (Hoi et al., 2018).

5.3 Error in Variables

All quantities derived from social media include error. Some of this error is *systematic*, a product of sampling bias (e.g., a sample that is too liberal or conservative) or variable specification (e.g., a positively-biased measure of sentiment). Dealing with systematic error typically requires a careful consideration of sampling and measurement issues. This is a common component of good social scientific research on social media.

Dealing with the random error in social media-derived measures and models is a much less frequent feature of social media analysis, particularly (although not exclusively) in the social sciences. Indeed, uncertainties in measures derived from social media are rarely well understood, and as a consequence many measures are treated as absolute values rather than stochastic quantities. This can lead to incorrect results from modeling, including attenuated bias (where a relationship is biased towards zero).

There are approaches to modeling that take this issue into account, including techniques for the 'propagation of error,' bootstrapping, and errors-in-variables models (Mooney, 1996; Van Huffel & Lemmerling, 2013). None are currently regular features of social-media-focused modeling exercises; all should be considered as ways of improving models of social media.

6. Case Studies - Putting Models to Work

How might all of this work in practice? We have thus far been writing about modeling social media data in very general terms. Here we make those generalities a little more concrete in a hypothetical case study.

The goal of our hypothetical study is to better understand the proliferation of both correct and false information about Covid-19 vaccinations. We begin with some general modeling challenges (from Section 2). We must get access to a sufficient body of content – enough to provide a credible account of the circulation of information about Covid-19; and we must be able

to identify Covid-19 content based on some relatively simple processing of natural language. The Covid-19 case has some distinct advantages: it is likely that the bulk of Covid-19 content can be identified with a very simple set of keywords; and we need only track that content from January 2020. The limited time frame reduces concerns about change in language over a long period of time; it also likely means that we have a manageable body of data as well. Our interest may be in social media generally, but we are faced with a dilemma: Facebook is used by a larger proportion of the public than Twitter, but Facebook content is typically more difficult to gather than Twitter content. Again, in this unique case, researchers can get access to these data either through different companies or through open collaborations that exist related to the pandemic (Banda et al., 2020; Facebook, 2021). Both data sources likely provide some measures of engagement (likes, re-posts) that may be central to our analysis. Our platform selection thus hinges on our research question – not just whether it is focused on one platform or another, but whether it requires data that may be more or less accessible in one platform or another, or whether it is important to look at data on multiple platforms simultaneously.

Note that we are already engaging in modeling, broadly construed. The moment we begin to make choices about how to identify which posts are relevant to our study, we begin to build a 'model' of the proliferation of Covid-19 vaccination information. What is the goal of that model? As Section 3 has outlined, we can categorize models into four different types. A descriptive model could seek to answer the question, what is 'virality' of correct versus false information about Covid-19 vaccinations? Even this relatively prosaic task requires some complex models. We must identify relevant content, and then build a model that identifies correct versus false information. We might accomplish this through clustering, or topic modeling, or anomaly detection. We might even rely on a dictionary search, which itself is another type of (very simple) model. Having captured the frequency (number of posts) and proliferation (number of likes and/or re-posts) of correct versus false posts, we are in a position to provide descriptive information about the virality of correct and false information about Covid-19 vaccinations.

A predictive model might seek to answer a more ambitious question, can we predict the virality of a given post about Covid-19 vaccinations? In this instance we need many of the same quantities as above – we need models that identify relevant content, and that identify both correct and false information about vaccinations. But rather than simply describing the frequency of correct and false information, we need to try to use other features of social posts (or social media users) that can predict the number of likes or re-posts that a given post might receive. The popularity of the user surely matters. So too might other words, pictures or videos in the post. We might try to use words to predict the number of likes with a regression model. The outcome in this instance is a model that can predict, based on the content of a given post alongside other factors, the likely virality that post will achieve. Or we might consider the task of labeling whether or not a post about a vaccine has accurate or false information in it. This could be done using classic machine learning models like random forest or support vector machines. However, given the subtlety of this task and the important role of language, a better approach might be to

use state-of-the-art language models (e.g. BERT) to increase accuracy, especially if we have limited labeled data. We might alternatively use some posts to predict the popularity of other posts through classification by decision trees, logistical regression, etc.

A diagnostic/explanatory approach is not fundamentally different from a prescriptive approach, as explained in Section 3. The difference here lies mainly in the target of our analysis: do we aim to make good predictions of virality (predictive), or do we aim to understand the various drivers of virality (explanatory)? Again, regression and classifiers are the modeling tools of choice. Rather than looking at the outcomes, however, we focus primarily on the drivers. To what extent does user popularity matter, versus the inclusion of photos, versus the use of inflammatory language, and so on. This is the main objective of an explanatory model.

Prescriptive models have their sights set on a more distant target. Again, we need to model the identification and categorization of relevant content. We also likely need to do much of what is necessary for an explanatory or predictive task. But our question is now something like: What can we do to increase the virality of correct information about vaccinations? So we now take the information learned from an explanatory model and attempt to extend that outwards, through, for instance, reinforcement learning broadly, or agent-based models or systems dynamics models more specifically. An explanatory model can help us see the factors that seem to increase virality; these prescriptive models offer a more complete account of what social media content about vaccinations might look like under different conditions (premised, presumably, on what we have learned in an explanatory model).

The kind of model(s) we require is thus fundamentally linked to the kind of questions we want to ask. Our approach will be further refined by some of the considerations discussed in Section 4. Is there already a good body of theory that predicts the virality of correct and false information about vaccinations? To the extent that theories have already been developed, our approach may test those theories through primarily deductive modeling exercises. We might test whether virality is driven by emotion-laden language, for instance. If existing theory is lacking, then taking an inductive approach will be necessary – and may be preferable even if good theory exists insofar as inductive models may turn up other not yet understood drivers of virality.

Is there a 'ground truth' where our models and measures are concerned? In this hypothetical example, some information is factually correct, and some information is not. Fact checkers can be helpful in this context. But there will be gray areas, posts in which information is partly true and partly false, and these instances may produce some complexities for the modeling exercise. Indeed, there seems to be limited agreement on which sources produce false information (Bozarth et al., 2020). Concerns raised in Section 5, including algorithmic drift and errors in variables will also be relevant in this context. The selection of models must thus always be concerned not just with the availability of data and type of research question (descriptive or otherwise), but with the ways in which measurement validity can be maximized and error – in many different forms – can be minimized.

7. Modeling Ethics and Algorithmic Bias

The previous sections highlight different possibilities for error and bias to arise. In the end, one central question is whether the results of an analysis or the inferences drawn from it are biased. There was clear agreement that all models have some form of bias. Where there was disagreement was the type of bias that different researchers focused on in their research. Is the bias a result of a social construct of interest? Is the bias a result of the population under examination not being representative of the population to which the results will be generalized? Is the bias a form of inductive bias that results from a particular statistical or machine learning model being used? Bias can influence every aspect of an inferential task. The formulation of the research question can be biased, data can be biased, the formulation of models can be biased, estimation procedure may be biased, and interpretation may be biased. These biases may also be connected to each other. For example, in psychology, researchers often tailor models to specific demographics (e.g., the middle class). Their theories may be also biased by the units from which they can collect data - how well can researchers infer effects observed in a student population to the general public? In medicine, clinical trials are typically randomized controlled experiments, reducing one form of bias. However, clinical trials are almost always biased by the groups that researchers can recruit to participate in the studies, which biases the predictions on the unstudied groups. This has enormous implications for trust, as we see right now with vaccines.

In the context of machine learning model creation, bias can emerge in different ways: bias during feature generation, bias within the training data, bias within existing external data sources that are used by a machine learning algorithms, e.g. dictionaries or word embeddings, and bias within the algorithm that may cause overfitting of subsets of the data---to name a few. Machine learning models require training data. Training data is a sample of data used to build a mathematical model for prediction. This training data is expected to be representative of the population of interest. If the model is designed to predict political affiliation, the training data needs to (1) have a reasonable sample of people who have different political affiliations in terms of sample size and political affiliation distribution, (2) have a reasonable sample of people who have different demographic characteristics within the different political affiliations, and (3) be a sufficient size sample. Machine learning algorithms typically optimize for classification or prediction accuracy. Therefore, if the training data is too small or some form of selection bias exists, the algorithm will "learn" that bias and propagate it. A classic example of this is the Amazon AI Recruiting tool that was designed to identify applicants for interviews. Because the training data consisted of existing employees, and the existing employees were overwhelmingly male, the AI tool discriminated against female applicants by making predictions using features that were highly correlated to gender (Dastin, 2018). This is an example of a historical bias that does not reflect current societal goals. Other examples include recidivism prediction with a high false positive rate for black people relative to white people (Angwin et al., 2016), facial recognition software mislabeling people with darker skin tones (Buolamwini & Gebru, 2018), and racial disparities in automated speech recognition (Koenecke et al., 2020).

While bias in artificial intelligence has been an area of study for decades, fairness in machine learning is a new sub-area focusing on measuring algorithmic fairness and addressing different types of inequities (Mehrabi et al., 2019). Many different definitions of fairness have emerged. The ones that are being used more extensively within the computer science community include: unawareness, demographic parity, equalized odds, and individual fairness (Chouldechova & Roth, 2018). Unawareness requires that sensitive attributes are removed from the training data. While a good first step, it does not address the issue of correlates of sensitive attributes existing in the data, as in the Amazon example above. These correlates could still be used to discriminate against protected groups. Demographic parity and equalized odds are metrics that capture notions of group level fairness, with demographic parity capturing independence between the output and the sensitive variables (protected and unprotected groups should have close to equal representation) (Zafar et al., 2017), and a classifier satisfies equalized odds if every individual in the protected and unprotected group have equal odds of being in the true positive rate and in the false positive rate (McNamara, 2019). Finally, individual fairness focuses on measuring a desired trait in a way that ensures that similar individuals are treated similarly by the algorithm. All of these measures have drawbacks, but they are the building blocks to quantifying unfair treatment of groups of people by algorithms. These different definitions of fairness can also be in conflict with other societal goals, leading to theoretical limitations on our ability to address all of them (Corbett-Davies et al., 2017). Work on correcting for algorithmic biases is emerging, but it is still in its infancy. The general approaches that are being used attempt to create intermediate, transformed versions of the data that are beneficial to the learning task, but that remove information about sensitive attributes. In order to correct model bias, regularization techniques are being used more extensively within machine learning models. Also, for some learning tasks, post-processing of learning results can be done to reduce bias and improve fairness.

Finally, we note that there is a tradeoff between bias and variance. Bias can occur when our model is too simple and "underfits" the data (Vapnik, 1998; German et al., 1992). In this scenario, bias will be high and variance will be low. On the other hand, if our model is very complicated with a large number of parameters, it may "overfit" the data. In this case, the bias will be much lower, but the variance will be high, i.e., if we repeat model construction on new data drawn similarly to the original data, and we do this many times, we will see a lot of variation in the results across trials. In other words, there is a tradeoff between bias and variance associated with a model and researchers need to evaluate the bias-variance tradeoff curves for different models to understand the impact on their findings.

8. Accelerating Research in this Area

This section focuses on identifying different resources that would be useful to help accelerate research for social scientists and computer scientists. We focus on those components that meeting participants thought would be particularly useful with regards to modeling.

The first group of ideas centers around construction of labeled data sets. In order to use social media effectively for empirical studies, more labeled data sets that are useful for social science research, e.g. demographics, stance, topics, etc., need to be created, shared, and documented. While a number of data repositories and data sharing platforms exist, there is no central repository curated for different social media data sets. Two sharing archives that have emerged as important data sharing platforms across disciplines are the Harvard Dataverse Project (https://dataverse.org/) and Google Data Commons Project (https://datacommons.org/). Either of those could be leveraged for social media data sets. The other important reason for making data available is to demonstrate how to use different models and algorithms. Because of the complexities of social media data sets, showing the strengths and limitations of different models using these data is important for accelerating research in this arena. When does an algorithm work as expected? When does it not? These questions can only be answered through examination of models across data sets.

The second group of ideas focuses on developing standards for measuring the quality of the data and the model. Examples include standards for assessing reliability and validity of data (see Measurement white paper (Ladd et al., 2020)), having standard sensitivity analyses for learning and statistical models, and requiring fairness measures on models that involve human subjects. Establishing criteria for different classification and learning tasks that will use social media data is important to standardize and advance research across disciplines.

The third group of ideas focuses on products that would help researchers better understand the population participating in conversations on social media. We know that social media are a machine/human hybrid and we do not know the distribution of each on different platforms. We know that there are different distributions of humans that use these platforms and that their activity level varies based on demographic characteristics. However, we cannot capture this information. If we can work with platforms to get these distributions, we would be able to produce higher quality research.

The fourth group centers on the differences in expectations of publishable work. Computer scientists publish new methods for different descriptive and predictive tasks. They also publish results that are prescriptive based on using complex predictive models that explore different spaces of correlations and relationships among, possibly, a large number of features. Social scientists test explicit hypotheses generated from existing theory, and are less likely to publish exploratory analyses. The crux of the problem being described is that what is considered to be valuable for each of our communities is not the same. Therefore, interdisciplinary research is not just about learning each other's approaches to research, it is also about publishing different work that may not be considered as valuable in one’s home research area. In order for this type of deep work to take place, it is important that new interdisciplinary journals emerge and that contributions across disciplines are considered important endeavors for junior faculty.

The final group of ideas focus on the development of standards for understanding measurement quality and error. Having techniques for measuring and aggregating sampling, measurement, and modeling error and understanding acceptable error propagation levels would help accelerate the use of social media data. Developing standard checks for testing sensitivity of research conclusions to modeling decisions would also help social scientists consider using these data. There are a number of sensitivity checks for machine learning models. They vary for different machine learning models. Identifying the ones that map best to ideas of reliability and validity in social science is an important for accelerating social science research involving social media data.

9. Conclusions

Social media research attracts researchers from various backgrounds who bring different perspectives, methods, and research questions to this new field. The models they use and exactly how they use them have important implications for how well these research questions are addressed. In short, model construction is an integral part of social media research and its strengths and weaknesses need to be well-understood. To this end, this white paper aimed to summarize model types and how they are used in social media research. We provided an overview of commonalities and distinctions between fields in how they approach modeling.

Focusing on the distinctions first, our cross-disciplinary conversations revealed a fundamental difference in how computer and social scientists conceptualize modeling as well as how they use it. Most commonly, social scientists conceptualize modeling as the part of the research process where a researcher applies a mathematical formulation of a theoretical model to understand the relationship between some independent and dependent variables. For a computer scientist, all of the various decisions that turn raw material into measures and analyses are commonly thought of as being part of model construction. As a result, computer scientists also place a lot of emphasis on the development and optimization of computational models for different descriptive and learning tasks.

The second important distinction relates to the purpose of modeling. While prediction is commonly the end goal for a computer scientist, it is rarely the end goal for a social scientist, who typically has a stronger focus on explanations. This affects various stages of the modeling process. There are two important steps we highlight: model construction and validation. For the model construction step, we observe a higher level of comfort with inductive modeling for computer scientists and deductive modeling for social scientists. Over-reliance on either approach can be seen as a potential point of improvement for both perspectives. There are cases where theory should dictate the model, and cases where existing theory is lagging so letting the data speak can open new doors. This is where we think convergence in thinking will benefit all disciplines.

The distinctions are obvious in the model validation as well. For instance, computer scientists are generally comfortable defining and relying on “ground truth” data to assess the

quality of their predictive models. Given labeled data and clear performance measures, computer scientists are often comfortable assessing a model according to its ability to predict that "ground truth." Cross-disciplinary discussions revealed an important point of convergence here. For many social science quantities, there is no "ground truth", but rather a concept that is being measured. And computer scientists ought to be more skeptical of the "ground truth" data they use to assess their models. These relate to important concepts such as construct validity, and data sampling biases that we covered in previous white papers. Conversations also reveal a whole new way to assess models that are generally ignored by social scientists (e.g. the ability of the model to predict future behavior) and these methods can be explored not as a substitution but perhaps a complement to current approaches used in social sciences.

Despite these distinctions, approaches to modeling are not completely distinct. There are important issues that apply across all disciplines. One of these relates to the purpose of modeling. Not carefully thinking through this question will end up in important unintended consequences, ranging from a model that is not properly evaluated to a model that is used for goals not aligned with researchers' goals, or even ethics. We list this as a common problem across disciplines based on our discussions in a mixed group of social and computer scientists, revealing an important need to focus attention on this issue across all fields. As researchers, it is our responsibility to construct models that help humanity and the societies in which we are embedded. Thinking through what each model succeeds in doing, and more importantly thinking through what it does not, are crucial for achieving that goal.

Other common issues were revealed; while not unique to social media, they are certainly critical in this context. For instance, given the fast pace with which things change on social media, models need to account for time, algorithmic drift, as well as issues of algorithmic bias and model fairness. Similarly, given how platforms shape behavior, it is important to account for the platform effects and study behavior across different platforms. Furthermore, social media present to us found data, where variables are not as carefully constructed. Uncertainties in measures derived from social media are rarely well understood. We suggest that more emphasis needs to be paid to these errors to improve modeling endeavors involving social media data.

Acknowledgements

We would like to thank the National Science Foundation and the McCourt School's Massive Data Institute (MDI) at Georgetown University for supporting this collaborative meeting. We would also like to thank Shweta Bansal and Thomas Leeper for participating in this meeting. Your insights were invaluable. This white paper is an output of that meeting and is co-authored by those who participated in the meeting. This series of white papers is funded by National Science Foundation awards #1934925 and #1934494 as a collaboration between Georgetown University and University of Michigan. A special thanks to Rebecca Vanarsdall from MDI for her support in planning the meeting and helping prepare this white paper. We also want to thank all the students and staff who helped facilitate and take notes during the meeting. Those notes were invaluable when putting this document together.

References

Abercrombie, G., & Hovy, D. (2016). Putting sarcasm detection into context: The effects of class imbalance and manual labelling on supervised machine classification of Twitter conversations. In *Proceedings of the ACL Student Research Workshop* (pp. 107-113). Berlin, Germany.

Aggrawal, C. C. (2018). *Neural networks and deep learning: A textbook.* Springer.

Agrawal, D., Budak, C., & El Abbadi, A. (2011). Information diffusion in social networks: observing and affecting what society cares about. In *Proceedings of the ACM International Conference on Information and Knowledge Management* (pp. 2609-2610). Glasgow, Scotland, UK.

Aldarwish, M. M., & Ahmad, H. F. (2017). Predicting depression levels using social media posts. In *IEEE International Symposium on Autonomous Decentralized System* (pp. 277-280). Bangkok, Thailand.

Anderson, J., Chaturvedi, A., & Cibulskis, M. (2007). Simulation tools for developing policies for complex systems: Modeling the health and safety of refugee communities. *Health Care Management Science*, *10*(4), pp. 331-339.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). Machine Bias. *ProPublica*. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Aral, S., Muchnik, L., & Sundararajan, A. (2009). Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, *106*(51), pp. 21544-21549.

Bachrach, Y., Kosinski, M., Graepel, T., Kohli, P., & Stillwell, D. (2012). Personality and patterns of Facebook usage. In *Proceedings of the Annual ACM Web Science Conference* (pp. 24-32). Evanston, IL, USA.

Bamman, D., & Smith, N. (2015). Contextualized sarcasm detection on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media, 9*(1) (pp. 574-577). Oxford, England, UK.

Banda, J. M., Tekumalla, R., Wang, G., Yu, J., Liu, T., Ding, Y., Artemova, K., Tutubalina, E., & Chowell, G. (2021). A large-scale COVID-19 Twitter chatter data set for open scientific research--an international collaboration [Data set]. *https://doi.org/10.5281/zenodo.3723939*

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *The Journal of Machine Learning Research*, *3*, pp. 993-1022.

Bollen, K. A. (2012). Instrumental variables in sociology and the social sciences. *Annual Review of Sociology*, *38*, pp. 37-72.

Boulianne, S. (2015). Social media use and participation: A meta-analysis of current research. *Information, Communication & Society*, *18*(5), pp. 524-538.

Bozarth, L., & Budak, C. (2020). Beyond the eye-catchers: A large-scale study of social movement organizations' involvement in online protests. *New Media & Society*.

Bozarth, L., Saraf, A., & Budak, C. (2020). Higher ground? How ground truth labeling impacts our understanding of fake news about the 2016 US presidential nominees. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 14, pp. 48-59). Atlanta, Georgia, USA.

Brown, P.F., Desouza, P. V., Mercer, R. L., Della Pietra, V. J., & Lai, J. C. (1992). Class-based n-gram models of natural language. *Computational Linguistics, 18*(4), pp. 467–479.

Budak, C., Agrawal, D., & El Abbadi, A. (2011). Limiting the spread of misinformation in social networks. In *Proceedings of the International Conference on World Wide Web* (pp. 665-674). Hyderabad, India.

Budak, C., Garrett, R. K., Resnick, P., & Kamin, J. (2017). Threading is sticky: How threaded conversations promote comment system user retention. In *Proceedings of the ACM Conference on Human-Computer Interaction*, *1*(CSCW), (pp. 1-20). Cancun, Mexico.

Budak, C., & Watts, D. J. (2015). Dissecting the spirit of Gezi: Influence vs. selection in the Occupy Gezi movement. *Sociological Science, 2*, pp. 370-397.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the Machine Learning Research Conference on Fairness, Accountability, and Transparency* (pp. 77-91). New York, NY, USA.

Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on twitter. In *Proceedings of the International Conference on World Wide Web* (pp. 675-684). Hyderabad, India.

Cataldi, M., Di Caro, L., & Schifanella, C. (2010). Emerging topic detection on twitter based on temporal and social terms evaluation. In *Proceedings of the International Workshop on Multimedia Data Mining* (pp. 1-10). Washington, DC, USA.

Chen, X., Wang, Y., Agichtein, E., & Wang, F. (2015). A comparative study of demographic attribute inference in Twitter. In *Proceedings of the International AAAI Conference on Web and Social Media, 9*(1), (pp. 590-593). Oxford, England, UK.

Chouldechova, A., & Roth, A. (2018). The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810.*

Chu, Z., Widjaja, I., & Wang, H. (2012). Detecting social spam campaigns on twitter. In *International Conference on Applied Cryptography and Network Security* (pp. 455-472). Springer, Berlin, Heidelberg.

Churchill, R., Singh, L., & Kirov, C. (2018). A temporal topic model for noisy mediums. In *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 42-53). Melbourne, VIC, Australia.

Churchill, R., & Singh, L. (2020). Percolation-based topic modeling for tweets. In *Proceedings of the KDD Workshop on Issues of Sentiment Discovery and Opinion Mining* (pp. 1-8). San Diego, CA, USA.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 797-806). Halifax, Nova Scotia, Canada.

Cowell, R. G., Verrall, R. J., & Yoon, Y. K. (2007). Modeling operational risk with Bayesian networks. *Journal of Risk and Insurance*, *74*(4), pp. 795-827.

Dastin, J. (2018, October 9). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters.*

De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting depression via social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, *7*(1), (pp. 128-137). Cambridge, MA, USA.

De Marchi, S., & Page, S. E. (2014). Agent-based models. *Annual Review of Political Science, 17,* pp. 1-20.

De Francisci Morales, G., Monti, C., & Starnini, M. (2021). No echo in the chambers of political interactions on Reddit. *Scientific Reports, 11*(1), pp. 1-12.

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 248-255). Miami, FL, USA.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805.*

Dilrukshi, I., De Zoysa, K., & Caldera, A. (2013). Twitter news classification using SVM. In *Proceedings of the International Conference on Computer Science & Education* (pp. 287-291). Colombo, Sri Lanka.

Dua, D. & Graff, C. (2019). *UCI Machine Learning Repository.* http://archive.ics.uci.edu/ml. Irvine, CA: University of California, School of Information and Computer Science.

Facebook. (2021). *Our Work on COVID-19*. Facebook Data for Good. https://dataforgood.fb.com/docs/covid19/.

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.

Ferrara, E., Jafari Asbagh, M., Varol, O., Qazvinian, V., Menczer, F., & Flammini, A. (2013). Clustering memes in social media. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 548-555). Niagara, Ontario, Canada.

Ford, A. (1999). *Modeling the environment: an introduction to system dynamics models of environmental systems.* Island press.

Frigg, R. & Hartmann, S. (2020). Models in Science. In E. N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy* (Spring 2020 Edition), Retrieved from: https://plato.stanford.edu/archives/spr2020/entries/models-science/

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation, 4*(1), pp. 1–58.

Ghosh, S., Singhania, P., Singh, S., Rudra, K., & Ghosh, S. (2019). Stance detection in web and social media: a comparative study. In *Proceedings of the International Conference of the Cross-Language Evaluation Forum for European Languages* (pp. 75-87). Avignone, France.

Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, *99*(12), pp. 7821-7826.

Guillaumin, M., Mensink, T., Verbeek, J., & Schmid, C. (2009). Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 309-316). Kyoto, Japan.

Guille, A., Hacid, H., Favre, C., & Zighed, D. A. (2013). Information diffusion in online social networks: A survey. *ACM Sigmod Record*, *42*(2), pp. 17-28.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Heath, B., Hill, R., & Ciarallo, F. (2009). A survey of agent-based modeling practices (January 1998 to July 2008). *Journal of Artificial Societies and Social Simulation*, *12*(4), pp. 9.

Hoi, S. C., Sahoo, D., Lu, J., & Zhao, P. (2018). Online learning: A comprehensive survey. *arXiv preprint arXiv:1802.02871*.

Homer, J. B., & Hirsch, G. B. (2006). System dynamics modeling for public health: background and opportunities. *American Journal of Public Health*, *96*(3), pp. 452-458.

Hong, L., & Davison, B. D. (2010). Empirical study of topic modeling in Twitter. In *Proceedings of the Workshop on Social Media Analytics* (pp. 80-88). Washington, DC, USA.

Hruschka, E. R., Campello, R. J., & Freitas, A. A. (2009). A survey of evolutionary algorithms for clustering. *IEEE Transactions on Systems, Man, and Cybernetics, Part C, 39*(2), pp. 133-155.

Hughes, D. J., Rowe, M., Batey, M., & Lee, A. (2012). A tale of two sites: Twitter vs. Facebook and the personality predictors of social media usage. *Computers in Human Behavior*, *28*(2), pp. 561-569.

Joshi, A., Bhattacharyya, P., & Carman, M. J. (2017). Automatic sarcasm detection: A survey. *ACM Computing Surveys (CSUR), 50*(5), pp. 1-22.

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759.*

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, *4*, pp. 237-285.

Kaiser, B., Wei, J., Lucherini, E., Lee, K., Matias, J. N., & Mayer, J. (2020). Adapting Security Warnings to Counter Online Disinformation. *Supporting materials. URL: https://osf.io/qf8e5*.

Kawintiranon, K., & Singh, L. (2021). Knowledge Enhanced Masked Language Model for Stance Detection. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Virtual.

Kempe, D., Kleinberg, J., & Tardos, É. (2003). Maximizing the spread of influence through a social network. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 137-146). Washington, DC, USA.

King, G., Lam, P., & Roberts, M. E. (2017). Computer-Assisted Keyword and Document Set Discovery from Unstructured Text. *American Journal of Political Science, 61*(4), pp. 971-988.

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J. R., Jurafsky, D. & Goel, S. (2020). Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, *117*(14), pp. 7684-7689.

Kollman, K., Miller, J. H., & Page, S. E. (1992). Adaptive parties in spatial elections. *American Political Science Review*, pp. 929-937.

Kudugunta, S., & Ferrara, E. (2018). Deep neural networks for bot detection. *Information Sciences, 467,* pp. 312-322.

Ladd, J., Ryan, R., Singh, L., Bode, L., Budak, C., Conrad, F., Cooksey, E., Davis-Kean, P., Dworak-Fisher, K., Freelon, D., Hopkins, D., Jensen, J. B., Kelley, K., Miller, R., Mneimneh, Z., Pasek, J., Raghunathan, T., Gresenz, C. R., Roy, S., Soroka, S., & Traugott, M. (2020). Measurement considerations for quantitative social science research using social media data. *PsyArXiv Preprint: https://doi.org/10.31234/osf.io/ga6nc*

Leavitt, A., & Robinson, J. J. (2017). The role of information visibility in network gatekeeping: Information aggregation on Reddit during crisis events. In *Proceedings of the ACM*

Conference on Computer Supported Cooperative Work and Social Computing (pp. 1246-1261). Portland, OR, USA.

Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J., & Glance, N. (2007). Cost-effective outbreak detection in networks. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 420-429). San Jose, CA, USA.

Li, C., Resnick, P., & Mei, Q. (2016). Multiple queries as bandit arms. In *Proceedings of the ACM International on Conference on Information and Knowledge Management* (pp. 1089-1098). Indianapolis, IN, USA.

Liu, X., Huang, Q., & Gao, S. (2019). Exploring the uncertainty of activity zone detection using digital footprints with multi-scaled DBSCAN. *International Journal of Geographical Information Science, 33*(6), pp. 1196-1223.

Liu, Y., Singh, L., & Mneimneh, Z. (2021). A comparative analysis of classic and deep learning models for inferring gender and age of Twitter users. In *Proceedings of the International Conference on Deep Learning Theory and Applications.* Virtual.

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, *28*(2), pp. 129-137.

Luceri, L., Giordano, S., & Ferrara, E. (2020). Detecting troll behavior via inverse reinforcement learning: A case study of Russian trolls in the 2016 US election. In *Proceedings of the International AAAI Conference on Web and Social Media*, *14* (pp. 417-427). Atlanta, GA, USA.

Lundberg, I., Johnson, R., & Stewart, B. M. (2021). What is your estimand? Defining the target quantity connects statistical evidence to theory. *SocArXiv Preprint: https://doi.org/10.31235/osf.io/ba67n*

McCord, M., & Chuah, M. (2011). Spam detection on Twitter using traditional classifiers. In *International Conference on Autonomic and Trusted Computing* (pp. 175-186). Springer, Berlin, Heidelberg.

McNamara, D. (2019). Equalized Odds Implies Partially Equalized Outcomes Under Realistic Assumptions. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 313-320). Honolulu, HI, USA.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635.*

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781.*

Miller, Z., Dickinson, B., Deitrick, W., Hu, W., & Wang, A. H. (2014). Twitter spammer detection using data stream clustering. *Information Sciences*, *260*, pp. 64-73.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. (2013). Playing Atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.

Moody, C. E. (2016). Mixing dirichlet topic models and word embeddings to make lda2vec. *arXiv preprint arXiv:1605.02019*.

Mooney, C. Z. (1996). Bootstrap statistical inference: Examples and evaluations for political science. *American Journal of Political Science,* pp. 570-602.

Murtagh, F., & Contreras, P. (2011). Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *2*(1), pp. 86-97.

Murtagh, F., & Contreras, P. (2017). Algorithms for hierarchical clustering: an overview, II. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *7*(6), e1219.

Ng, A. & Jordan, A. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. *Advances in Neural Information Processing Systems*, *14*(2002), pp. 841.

Nguyen, D. Q., Billingsley, R., Du, L., & Johnson, M. (2015). Improving topic models with latent feature word representations. *Transactions of the Association for Computational Linguistics*, *3*, pp. 299-313.

Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, *103*(23), pp. 8577-8582.

Orchard, L. J., Fullwood, C., Galbraith, N., & Morris, N. (2014). Individual differences as predictors of social networking. *Journal of Computer-Mediated Communication*, *19*(3), pp. 388-402.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods In Natural Language Processing* (pp. 1532-1543). Doha, Qatar.

Quan, X., Kit, C., Ge, Y., & Pan, S. J. (2015). Short and sparse text topic modeling via self-aggregation. In *Proceedings of the International Joint Conference on Artificial Intelligence* (pp. 2270-2276). Buenos Aires, Argentina.

Qiang, J., Qian, Z., Li, Y., Yuan, Y., & Wu, X. (2020). Short text topic modeling techniques, applications, and performance: A survey. *IEEE Transactions on Knowledge and Data Engineering.*

Rajadesingan, A., Resnick, P., & Budak, C. (2020). Quick, community-specific learning: How distinctive toxicity norms are maintained in political subreddits. In *Proceedings of the International AAAI Conference on Web and Social Media, 14* (pp. 557-568). Virtual.

Reece, A. G., & Danforth, C. M. (2017). Instagram photos reveal predictive markers of depression. *EPJ Data Science, 6,* pp. 1-12.

Sallab, A. E., Abdou, M., Perot, E., & Yogamani, S. (2017). Deep reinforcement learning framework for autonomous driving. *Electronic Imaging*, *2017*(19), pp. 70-76.

Schelling, T. C. (1971). Dynamic models of segregation. *Journal of Mathematical Sociology*, *1*(2), pp. 143-186.

Shepitsen, A., Gemmell, J., Mobasher, B., & Burke, R. (2008). Personalized recommendation in social tagging systems using hierarchical clustering. In *Proceedings of the ACM Conference on Recommender Systems* (pp. 259-266). Lausanne, Switzerland.

Singh, L., Wahedi, L., Wang, Y., Wei, Y., Kirov, C., Martin, S., ... & Kawintiranon, K. (2019). Blending noisy social media signals with traditional movement variables to predict forced migration. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 1975-1983).

Smith, N. A. (2020). Contextual word representations: putting words into computers. *Communications of the ACM*, *63*(6), pp. 66-74.

Sun, S., Zhang, C., & Yu, G. (2006). A Bayesian network approach to traffic flow forecasting. *IEEE Transactions on Intelligent Transportation Systems*, *7*(1), pp. 124-132.

Surian, D., Nguyen, D. Q., Kennedy, G., Johnson, M., Coiera, E., & Dunn, A. G. (2016). Characterizing Twitter discussions about HPV vaccines using topic modeling and community detection. *Journal of Medical Internet Research, 18*(8), e232.

Talukdar, D., Sudhir, K., & Ainslie, A. (2002). Investigating new product diffusion across products and countries. *Marketing Science, 21*(1), pp. 97-114.

Tan, P. N., Steinbach, M., & Kumar, V. (2005). *Introduction to data mining, First Edition*. Pearson.

Tang, J., Alelyani, S., & Liu, H. (2014). Data classification: algorithms and applications. *Data Mining and Knowledge Discovery Series, CRC Pres*s, pp. 37-64.

Tian, H., Gao, C., Xiao, X., Liu, H., He, B., Wu, H., ... & Wu, F. (2020). SKEP: Sentiment knowledge enhanced pre-training for sentiment analysis. *arXiv preprint arXiv:2005.05635.*

Ulges, A., Schulze, C., Koch, M., & Breuel, T. M. (2010). Learning automatic concept detectors from online video. *Computer vision and Image understanding, 114*(4), pp. 429-438.

Van Huffel, S., & Lemmerling, P. (Eds.). (2013). *Total least squares and errors-in-variables modeling: Analysis, algorithms and applications.* Springer Science & Business Media.

Vapnik, V. N. (1998). *Statistical learning theory. Adaptive and learning systems for signal processing, communications and control series.* John Wiley & Sons.

Vraga, E. K., & Bode, L. (2020). Defining misinformation and understanding its bounded nature: Using expertise and evidence for describing misinformation. *Political Communication*, *37*(1), pp. 136-144.

Wang, Y., Agichtein, E., & Benzi, M. (2012). TM-LDA: efficient online modeling of latent topic transitions in social media. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 123-131). Beijing, China.

Wang, J., Yang, Y., Mao, J., Huang, Z., Huang, C., & Xu, W. (2016). CNN-RNN: A unified framework for multi-label image classification. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (pp. 2285-2294). Las Vegas, NV, USA.

Weisbuch, G. (2018). *Complex systems dynamics.* CRC Press.

Xu, D., & Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science, 2(*2), pp. 165-193.

Xu, R. & Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3), pp. 645–678.

Yin, J., & Wang, J. (2014). A dirichlet multinomial mixture model-based approach for short text clustering. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 233-242). New York, NY, USA.

Zafar, M. B., Valera, I., Rogriguez, M. G., & Gummadi, K. P. (2017). Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics* (pp. 962-970). PMLR.

Zhang, A. F., Livneh, D., Budak, C., Robert Jr, L. P., & Romero, D. M. (2017a). Crowd development: The interplay between crowd evaluation and collaborative dynamics in wikipedia. In *Proceedings of the ACM on Human-Computer Interaction* (pp. 1-21). New York, NY, USA.

Zhang, A. F., Livneh, D., Budak, C., Robert, L., & Romero, D. (2017b). Shocking the crowd: The effect of censorship shocks on Chinese Wikipedia. In *Proceedings of the International AAAI Conference on Web and Social Media, 11*(1) (pp. 367-376). Montreal, Canada.