SOCIAL NETWORKS AND ARCHIVAL CONTEXT OPENREFINE PLUGIN

THE SHORTCOMINGS OF METHODOLOGIES TO ELIMINATE BIAS IN MACHINE LEARNING

A Thesis Prospectus
In STS 4500
Presented to
The Faculty of the
School of Engineering and Applied Science
University of Virginia
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

By
Jessica Xu

October 31, 2019

Technical Project Team Members
Michael Chang, Sandy Gould, Mark Jeong, John Perez, Victor Shen, Peter Tran, Grace Wu

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

Signed: _____ Date: 12/4/2019

Approved: Catherine D. Baritaud Date: Dec. 5, 2019
Catherine D. Baritaud, STS Division, Department of Engineering and Society

Approved: Ahmed Ibrahim Date: 11/23/2019
Ahmed Ibrahim, Department of Computer Science

Within the past ten years, machine learning has experienced a phenomenal growth. It is being adopted as a means to use data to make decisions and judgements by many organizations without explicit programming (Langley, 2011). Machine learning has a plethora of applications, but the STS project will focus on its social applications. When an algorithm makes decisions about topics like social welfare and policy, several moral issues come up. It is well known that machine learning has implicit bias, whether it is from the developers that formed the algorithm itself, or from the training data that was used. Several different methodologies have been developed that attempt to mitigate the bias in machine learning, but each of the methodologies has its own flaws (Calmon, Wei, Vinzamuri, Ramamurthy, and Varshney, 2018). The STS research will focus on investigating if there should be a standard for labeling data, and if creating a standard will help mitigate bias in machine learning.

The STS project is loosely coupled with the technical project, which will focus on developing a tool to benefit Social Networks and Archival Context (SNAC), an archival organization. SNAC's current workflow for adding new data to its archives requires deep and extensive knowledge about the data model of SNAC itself. The technical project will seek to create a custom plugin that will provide a more intuitive way to refine and update data according to a model. A minimally functioning product will be completed by the end of the fall semester. This will include basic functionality such as uploading simple data to the plugin and processing it. During the spring semester, improvements will be made to the plugin, including the ability to process more complex input data that contains relationships between entities. The technical team will also receive feedback from the client and make changes accordingly. A finalized product with all extra constraints implemented as well as a technical report will be completed by the end

of the spring semester. A basic timeline of the tasks that need to be completed in the fall semester for both the STS and technical projects are detailed below in Figure 1.

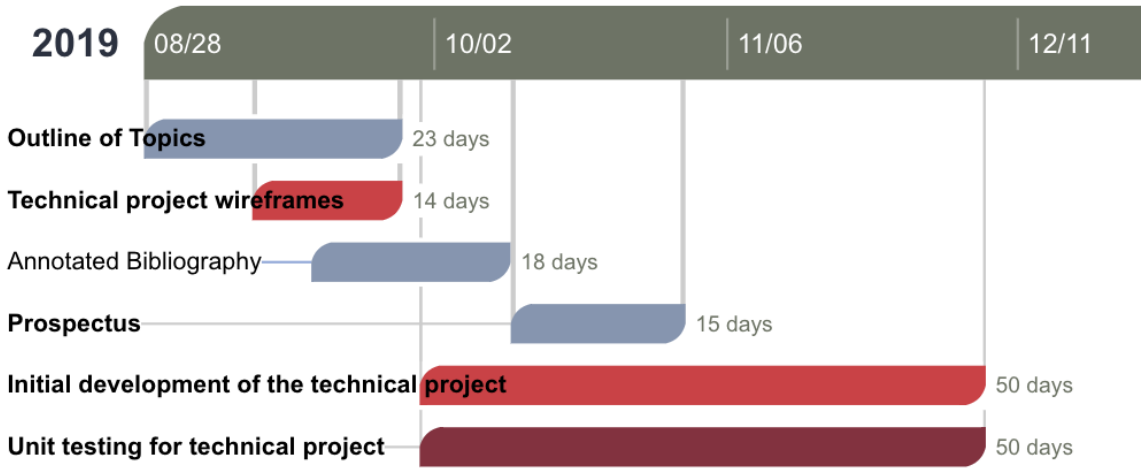

Figure 1: GANTT Chart: A basic initial roadmap and timeline of work to be completed for the capstone project for the fall semester (Xu, 2019).

Both the technical and STS projects are involved with large volumes of data. Machine learning algorithms must be trained with extensive amounts of data, and Social Networks and Archival Context (SNAC) often deals with a heavy traffic of incoming data (Social Networks and Archival Context, n.d.). The STS and technical projects both heavily focus on data labels as well. The technical project utilizes data labels to help reconcile data within SNAC, and the STS project investigates the effects of data labels on bias in machine learning. Although the technical and STS projects focus on different areas of computer science, they both emphasize the importance of accurate data.

SOCIAL NETWORKS AND ARCHIVAL CONTEXT OPENREFINE PLUGIN

Social Networks and Archival Context (SNAC) is a free, online resource that allows users to discover information about the people and organizations that are documented in primary

source documents and the connections between them (Social Networks and Archival Context [SNAC], n.d.). SNAC is used to locate archived collections as well as related resources held around the world. As an international cooperative, SNAC works to "build a corpus of reliable descriptions" of people and artifacts that link to and "provide a contextual understanding" of historical records (SNAC, n.d., para. 1). In order to create these contextual connections, SNAC sources its information from many different libraries and archival institutions. SNAC cooperates with over 4000 institutions to gather and reconcile data (SNAC, n.d.). Each of these institutions has a different structure for storing records. Relationships between different entities, labels for certain types of data, and the hierarchy of the data itself are inconsistent from each outside institution. SNAC needs to reconcile the differences between the outside data and its own data storage structure before importing the data into its database. It is extremely impractical to clean up the data manually or with simple tools (Ham, 2013). The reconciliation of this data is vital to the functionality of an archival organization such as SNAC because it is crucial for efficient and accurate querying (Park, 2008).

The technical project seeks to develop a standalone plugin for Social Networks and Archival Context (SNAC) using OpenRefine. OpenRefine is an open source software that is community-maintained designed specifically for data normalization, transformation, and cleaning (Hill, 2016). It allows users to import and normalize data with a series of pre-existing default user interfaces after connecting to a target resource. OpenRefine provides a "powerful yet user-friendly interface" for experimenting with and querying data (Hill, 2016, p. 228).

With over 700 edits occurring to its data schema in week, Social Networks and Archival Context (SNAC) is no small data archive (SNAC, n.d.). The current workflow for refining and updating data in SNAC is quite difficult and inaccessible to inexperienced users. It involves users hitting SNAC's APIs for refining data on their server from the user's local machine. The technical project aims to greatly simplify this process by creating a streamlined plugin that will have all the functionalities needed to refine and upload data in one location. The logical flow and components needed for the project are illustrated in Figure 2. The plugin will serve as a connection between the user's local data and SNAC's server. It will allow users to import external data in the form of comma-separated values (CSV) files and make use of APIs provided by SNAC to reconcile and refine that data with SNAC's unique JavaScript Object Notation (JSON) data structure. The plugin will have two main user groups: privileged and

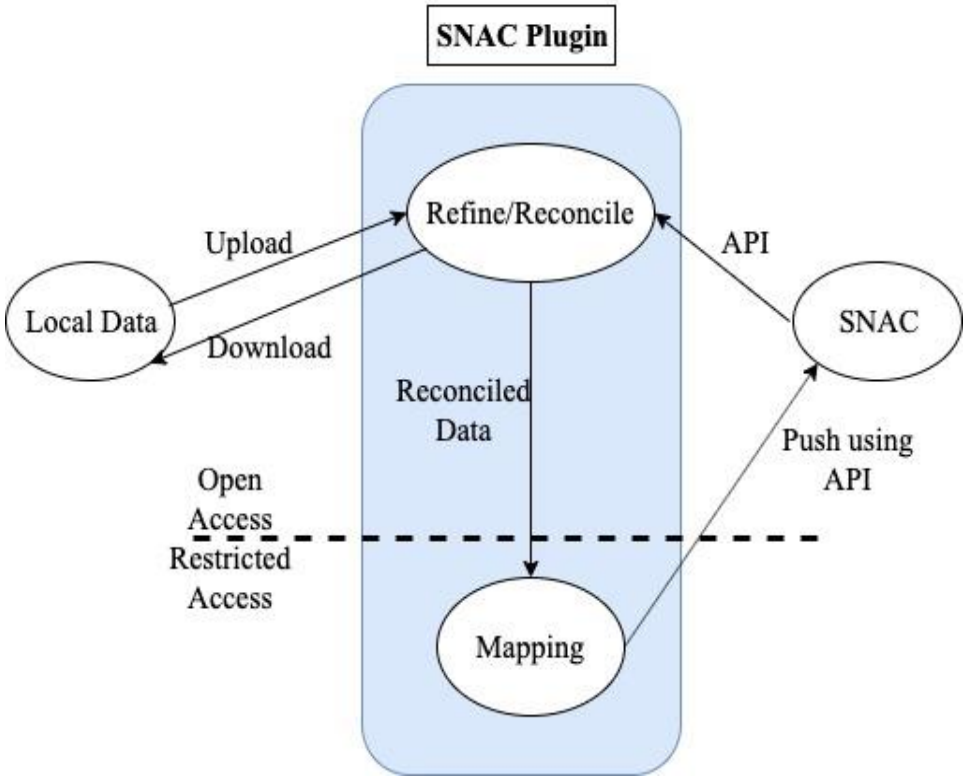

Figure 2: SNAC Plugin Model: An overview of the design of the plugin, depicting the different processes and functions that will be made available by the plugin (Xu, 2019).

unprivileged users. Both types of users will be able to use the plugin to format any data ported in using SNAC's organizational schema. Only privileged users will be able to then push the formatted data into SNAC's own database utilizing the APIs provided by SNAC. The technical project will provide an easy way to reconcile outside data with SNAC's existing data in addition with an improved user interfaced for an enhanced user experience.

The development will conduct biweekly customer meetings with the client in order to gather system requirements and get feedback about ongoing work. The minimum requirements for the plugin to be completed by the end of this semester include:

- Allowing users to import CSV data into the plugin
- Connect the data fields with different SNAC IDs
- Search for constellations in SNAC and match them to the imported data
- Allow a human editor to choose from several options to match for when the plugin is unsure
- Reconcile the imported changes based on the connection and matches
- Download the data that is now reconciled with SNAC's structure
- Users with privileges will be able to publish the data to SNAC

Desired requirements include:

- Users will be able to reconcile more complex data items like relationships and geolocations
- Users will be able to edit already existing resources and constellations

So far, no optional requirements have been specified by the client. The goal is to have a completed plugin that is fully integrated into the system by the end of the academic school year in twelve sprints.

The technical project will be developed over the course of the two-semester capstone series led by Professor Ahmed Ibrahim from the Computer Science department, and will result in a technical report. To create this plugin, OpenRefine will be used, as it is a powerful tool for working with disorganized data that can "[transform] it from one format into another; extending it with web services and external data" (OpenRefine, n.d., para. 1). A similar project exists already for WikiData, but the technical project will create a new implementation specifically for Social Networks and Archival Context (SNAC). The plugin will hopefully provide a faster and more intuitive way for SNAC users to reconciliate and update data.

THE SHORTCOMINGS OF METHODOLOGIES TO ELIMINATE BIAS IN MACHINE LEARNING

Machine learning is currently experiencing a huge growth. In 2019 adopting machine learning is "among the highest priority projects for enterprises" (Columbus, 2019, para. 1). Companies are investing more time and money in machine learning than ever, and the integration of machine learning into systems is on the rise as well. Today, decisions about jail sentences, hiring, loans, finances, advertisements, and much more are either made or influenced by machine learning algorithms (Ark, 2018). As machine learning models become more integrated into daily life, organizations that use them must be increasingly cautious of any potential flaws. Some of the many merits of machine learning and artificial intelligence (AI) are reduced labor and costs, faster turnaround, and sometimes higher quality results (Shein, 2018). There is a huge demand for AI since the number of people who require social services is growing at a rapid rate, and AI can simplify the application process significantly. AI also has the potential to streamline several services in different locations into one enhanced experience. Despite all the

merits of incorporating AI into social programs, Shein argues that AI also has several fatal flaws, one of which is inherent bias (Shein, 2018).

A Machine Learning or AI model is "a mathematic equation that uses data to produce a calculation such as a score, ranking classification, or prediction" to deliver a decision, action, or cause to support a business process (Sammy, 2019, Why Models Need to be Fair section, para. 1). It is imperative that machine learning models be fair because the decisions that their calculations end up supporting often impact many people, and if the criteria they use to make decisions is considered unethical or unacceptable, machine learning can expose organizations to risk and criticism. Examples of such unethical or unacceptable criteria include race, gender, age, religion, and sexual orientation (Sammy, 2019).

COUNTERING BIAS IN MACHINE LEARNING

Machine learning and artificial intelligence (AI) were developed in order to reduce bias in comparison to humans, however, they can unfortunately also increase bias (Silberg & Manyika, 2019). One great shortfall of machines is that they have no "common sense". If a machine received erroneous input, it will continue to apply that error because it has no capacity to (Shein, 2018). AI is meant to see detect signals or patterns in data, and in order to do so, the

machine learning system must be trained extensively. Illustrated in Figure 3 is a simple machine learning workflow. Training data is used to prime and train the machine learning model. Then, that model is applied to input data to create a prediction that can also lead to a definitive result. Rather than the algorithms of the machine learning model themselves, it is often the underlying training data that introduces bias to the whole system (Silberg & Manyika, 2019).

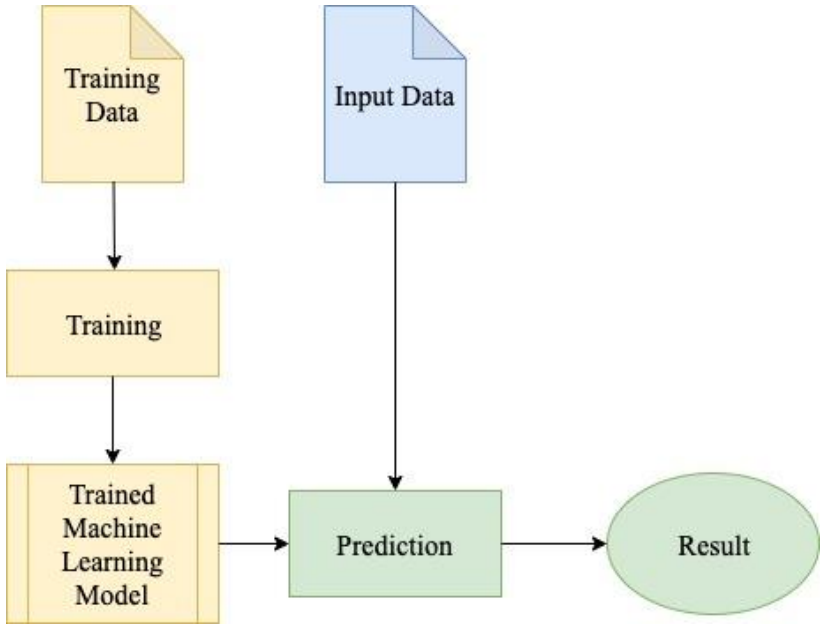

Figure 3: Simple Machine Learning Workflow: a diagram demonstrating how a machine learning model makes a decision (Xu, 2019).

Several proposals for elimination bias in machine learning already exist, including pre-processing data. In 2018, Calmon, Wei, Vinzamuri, Ramamurthy, and Varshney concluded, the goal of pre-processing the training data before feeding it to a machine learning model is to "[control] discrimination, [limit] distortion, and [preserve] utility (p. 1106). Calmon et al. (2018) state that using pre-processing, discrimination in machine learning can be reduced by a significant amount, but it comes at the cost of classification accuracy. Following the same idea, post-processing can also be used to reduce bias. After the system has made decisions, the decisions are processed and manipulated to account for bias (Calmon et al., 2018). For both of

these approaches, the challenge is to create a technique that can accurately recognize bias, and in many cases both approaches fall short of what they aim to do.

A NETWORK OF BIAS AROUND MACHINE LEARNING MODELS

The relationships between machine learning models, the engineer that developed them, society, and bias are best examined through the lens of Actor Network Theory (ANT) (Law & Callon, 1988). My interpretation of the actor network around machine learning and its bias are illustrated in Figure 4 below. Each actor in the network influences the others either directly or indirectly. For example, only engineers and data directly influence the machine learning model. However, the bias from the engineer will affect the influence the engineer has on the model, and the bias coming from society will affect the data that is going into the model.

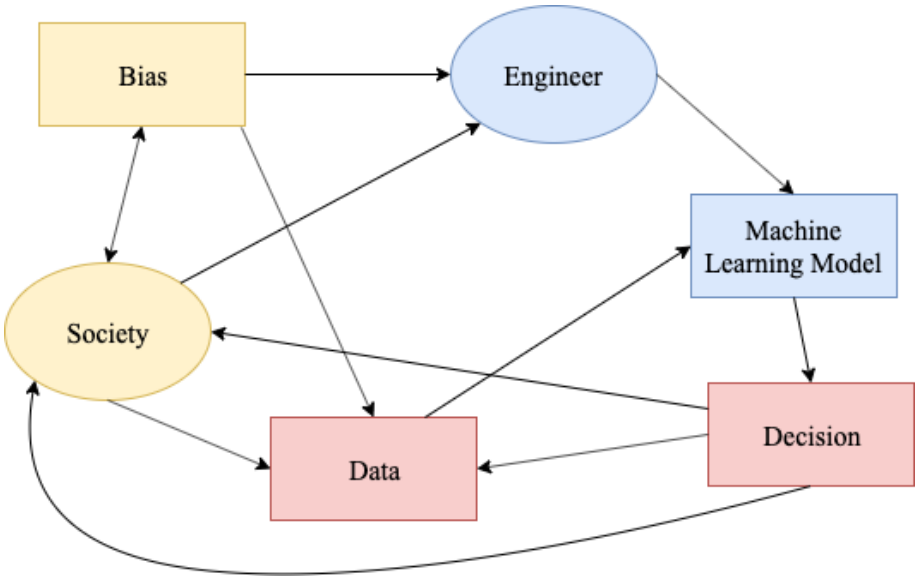

Figure 4: Machine Learning Bias Actor Network Theory: a diagram depicting the relationships between the actors (Xu, 2019).

Most of the existing methodologies to combat bias in machine learning come in the form of more algorithms that adjust either the decisions that come out of the machine learning model or the data that is put into the model. Under the guidance of professor Catherine Baritaud from the Science, Technology, and Society Department, the STS project will be a scholarly article scrutinizing the different ways that organizations try to eliminate bias in their machine learning

algorithms, and analyzing their downfalls. There is no perfect solution to eliminate bias from machine learning yet, but the methods that exist today focus too heavily on trying to use another piece of technology to correct the faults of another piece of technology. Perhaps creating a standardized labeling system to be used across all archival organizations could help with created more unbiased data and therefore mitigate some of the bias present in machine learning systems.

WORKS CITED

Ark, T. V. (2018, November). Why social studies is becoming AI studies. *Forbes*. Retrieved from http://forbes.com/

Calmon, F.P., Wei, D., Vinzamuri, B., Ramamurthy, K. N., & Varshney, K.R. (2018). Data pre-processing for discrimination prevention: Informatic-theoretic optimization and analysis. *IEEE Journal of Selected Topics in Signal Processing*, 12(5), 1106-1119. Doi:/ 10.1109/JSTSP.2018.2865887

Columbus, L. (2019, September). State of AI and machine learning in 2019. *Forbes*. Retrieved from http://forbes.com/

OpenRefine. (n.d.). *Introduction to OpenRefine*. Retrieved from http://openrefine.org/

Ham, K. (2013). Free, open-source tool for cleaning and transforming data. *Journal of the Medical Library Association*. Retrieved from https://www.ncbi.nlm.nih.gov/

Hill, K. M. (2016). In search of useful collection metadata: Using OpenRefine to create accurate, complete, and clean title-level collection information. *Serials Review*, 42(3), 222-228. Retrieved from https://www.sciencedirect.com/journal/serials-review

Langley, P. (2011). The changing science of machine learning. Machine Learning, 82(3), 275-279. doi:10.1007/s10994-011-5242-y

Law, J. & Callon, M. (1988). engineering and sociology in a military aircraft project: A network analysis of technological change. Oxford University Press, 35(3), 284-297. doi:10.2307/800623

Park, J-R. (2008). Metadata quality in repositories: A survey of the current state of the art. *Cataloging & Classification Quarterly*, 47(3), 213-228. doi:10.1080/01639370902737240

Sammy, A. (2019). Bias in the machine. *Internal Auditor*, 76(3), 42-46. Retrieved from https://na.theiia.org/Pages/IIAHome.aspx

Shein, E. (2018). The dangers of automating social programs: Is it possible to keep bias out of a social program driven by one or more algorithms? *Communications of the ACM*, 61(10), 17-19

Silberg, J., Manyika, J. (2019, June). Tackling bias in artificial intelligence (and in humans). *McKinsey Global Institute*. Retrieved from http://mckinsey.com

Social Networks and Archival Context. (n.d.). *About SNAC*. Retrieved from https://portal.snaccooperative.org/about

Xu, Jessica (2019). Figure 1: GANTT chart.

Xu, Jessica (2019). Figure 2: SNAC plugin model.

Xu, Jessica (2019). Figure 3: Simple machine learning workflow.

Xu, Jessica (2019). Figure 4: Machine learning bias actor network theory.