

# **Thesis Project Portfolio**

## **Software Development: Start-Up Development Cycle**

(Technical Report)

## **Misinformation in Government and Society**

(STS Research Paper)

An Undergraduate Thesis

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

**Samer Nahed Kadih**

Spring, 2023

Department of Computer Science

## **Table of Contents**

Sociotechnical Synthesis

Software Development: Start-Up Development Cycle

Misinformation in Government and Society

Prospectus

## **Sociotechnical Synthesis**

With great data comes great responsibility. Used correctly, data is an instrument for better understanding and assisting humankind. However, in the age of the internet and social media, time and time again we find ourselves prompting cultural discussions on individual privacy, autonomy, and integrity. The technical report details an effort to develop tools that commodify data quality and analytics, so that the barriers to entry to working with high quality data for practical applications is lessened. The Science, Technology, and Society research paper explores how user data on social media platforms can be better employed to mitigate the spread of misinformation. Both papers highlight the importance of data quality and data discovery for responsible use in practical applications.

Garbage in, garbage out. Whether or not it is purposeful, bad quality data can have unfair, prejudicial, and biased consequences. Data scientists ought to spend 80% of their time on cleaning and actively monitoring data that is used in their processing pipeline. Even then, many teams do not have the luxury of enterprise-level infrastructure for accurately assessing the quality of their data. The technical report describes the development of services that provide data engineers a better understanding and profiling of their data, as well as a means to actively monitoring that data for anomalies.

I worked on this project over the summer as an intern, and actively participated in agile methodologies that support the reliable deployment and testing of new technologies. Different services are split up into separate modules of code and communicate with one another over API calls. The “Protect” service monitors against data drift over a single live stream of data, the “Compare” service uses proprietary algorithms to infer metadata among disparate data sources and signal any inconsistencies, and the “Integrate” service gives data engineers the freedom to choose where in their pipeline they would like to moderate data quality. We had small to mid-size level businesses test our suite of services in their development environment. They loved the idea that their data can have a centralized source of truth and documentation among groups associated with different projects, and the seamless methods for documenting sources of bad data give insight into where their pipeline is failing.

In the context of data used in practical applications, social media platforms have a role in ensuring that the unfathomable amount of data they have on their users is used responsibly. In the past few years, discussions about fake news have resurfaced, as Biden's efforts to maintain a healthy rate of vaccinations in the U.S. have been stifled by the spread of misinformation that misrepresents the side effects of the vaccines. Actor-Network Theory is used to model the different actors involved in the spread of disinformation. At the center of this mapping is social media, whose platforms are an attractive tool for self-expression by the masses, and thus an attractive tool for bad actors to spread false claims. Bad actors take advantage of human psychology and biases, giving misinformation a better degree of virality over social media networks. Steps to classify fake news either by its content or by its power in a network can potentially mitigate the spread of misinformation. These steps are backed by studies outlining the specific human psychologies involved in the spread of misinformation, with a nod to proof-of-concept machine learning models and algorithms for unbiased propagation of information over a network.

The major forms of human psychology that play a role in accepting misinformation are confirmation bias and repetition bias. Traditional social media algorithms aimed at maximizing user retention tend to confine users to echo chambers that propagate sensational news. The actor-network involves local and national political entities whose interdisciplinary policies are responsible for administering educational programs to equip people with the skills for better identifying and filtering fake news. Machine learning models, however, are an effective first line of defense to classify misleading news which is otherwise too voluminous for experts to personally sift through. Furthermore, algorithms that suggest posts based on variables like user political affiliation can help to diversify the news content for that user, subverting the power of misinformation in appealing to their biases.

The quality of data and the insights gained from it are instrumental to building a platform that gives all users a fair and responsible means of using that platform. Although biases can never be fully eliminated, steps can certainly be made to mitigate them by maintaining that data quality and pursuit for

fairness. In particular, data used in influential platforms such as social media can have dire political consequences, and should take active measures to using that data responsibly.