Classifying Disinformation by Domain Name

Trevor J McGlynn

Western Governors University

**Table of Contents**

[**Proposal Overview**](#_30j0zll) **4**

[Problem Summary](#_1fob9te) 4

[IT Solution](#_3znysh7) 7

[Implementation Plan](#_2et92p0) 8

[**Review of Other Work**](#_tyjcwt) **9**

[Disinformation Sources and Strategies](#_3dy6vkm) 9

[GEC Special Report: Pillars of Russia’s Disinformation and Propaganda Ecosystem](#_1t3h5sf) 10

[Where Does Fake News Come From?](#_4d34og8) 11

[The Tactics & Tropes of the Internet Research Agency](#_2s8eyo1) 11

[ML implementations against disinformation](#_17dp8vu) 12

[The Moral Permissibility of Automated Responses during Cyberwarfare](#_3rdcrjn) 13

[Fighting misinformation on social media using crowdsourced judgments of news source quality](#_26in1rg) 14

[Fabula AI is using social spread to spot fake news](#_lnxbz9) 15

[The Role of Technology in Online Misinformation](#_35nkun2) 15

[Peering Under the Hood of Fake-News Detectors](#_1ksv4uv) 16

[Fake News Detection: A Deep Learning Approach](#_44sinio) 17

[IT Industry Responses](#_2jxsxqh) 17

[ICANN Org’s Multifaceted Response to DNS Abuse](#_z337ya) 17

[Measures Taken by Registries to Help Tackle COVID-19 Related Online Abuse](#_3j2qqm3) 19

[Who’s Behind the “Reopen” Domain Surge?](#_1y810tw) 20

[**Project Rationale**](#_4i7ojhp) **21**

[**Current Project Environment**](#_2xcytpi) **23**

[**Methodology**](#_1ci93xb) **24**

[**Project Goals, Objectives, and Deliverables**](#_3whwml4) **25**

[Goals, Objectives, and Deliverables Table](#_2bn6wsx) 25

[Goals, Objectives, and Deliverables Descriptions](#_qsh70q) 26

[Goal 1: Collect & analyze the data](#_3as4poj) 26

[Objective 1.a. Define requirements and collect data](#_1pxezwc) 26

[**Objective 1.b. Data analysis**](#_8bep5j8hf1) **27**

[Objective 1.c. Database design & modeling](#_49x2ik5) 27

[Goal 2: Implement a supervised binary classification algorithm](#_2p2csry) 27

[Objective 2.a. Implement machine learning algorithm](#_147n2zr) 27

[Goal 3: Document & make publicly available](#_3o7alnk) 28

[**Project Timeline with Milestones**](#_ihv636) **28**

[**Outcome**](#_32hioqz) **29**

[**References**](#_1hmsyys) **30**

# Proposal Overview

## Problem Summary

Disinformation is a severe problem and continues to accelerate due to strategic manipulation of social media algorithms (U.S. Department of State, 2020; 116th Congress 1st Session Senate, 2019; DiResta, 2019). Until social media companies make available more of their data, an exact understanding of the issue will remain elusive. The problem, however, is not exclusive to social media, nor is it contained to it.

A frequent feature of disinformation campaigns involves promoting posts on social media which link to dubious shell websites pushing certain narratives, narratives with ties to foreign or domestic influence operations. These shell websites can sometimes appear to be legitimate local or national news outlets intended to deceive end users(Levin, 2019). Some websites are not deceptive at all—having extremist and alarming content meant to divide and confuse(DiResta, 2019). In fact, the predominant strategy behind much of the content produced is to “confuse, polarize, and entrench,” with the consequence “that citizens tune out the political discourse or tune into their own, politically congenial filter bubble” (Krebs, 2020).

Most problematic—during the time in which this paper was authored—are influence operations that deceive on matters of public health. Common narratives of these operations include: questioning the efficacy of face masks, stoking fears about public health officials, correlating vaccinations with trans-humanist robotics (where vaccines have microchips used to control or monitor recipients), and the promotion of anti-lockdown protests(Federation of American Scientists, 2020; Krebs, 2020). As a report by STAT News across 93 sites publishing false harmful information about the outbreak found: “Many of their posts are being exponentially more widely shared than those from the health authorities trying to deliver real and reliable information” (Gregory, 2020). The dissemination of this content, and engagement with it, has undoubtedly made the COVID-19 pandemic worse. These operations are clearly working and clearly causing harm.

Some content is meant to destabilize unity and trust through stoking racial division, lying about scientific consensus (including global warming and the safety of certain vaccines), and promoting false information about elections(Kreps, 2020; Broniatowski, 2018; Martin-Rozumiłowicz & Kužel, 2019).

Political systems meant to protect, and media systems meant to inform, have allowed this issue to grow worse despite thoroughly acknowledging its existence (116th Congress 1st Session Senate, 2019). There is interest, however, into classifying disinformation as a computer security issue (Morgan & DiResta, 2018).. Doing so welcomes the multi-disciplinary world of IT into finding better solutions, solutions which come from those who have an ethical and professional stake in preserving the sanctity of information online. Recently within the industry, there exists recognition of disinformation as an issue of computer security, and that companies have a stake in preventing its spread.

Researchers have determined certain ways to curtail the spread of disinformation, including more advanced ways of detecting it. Machine learning (ML) implementations offer a solution, given clear transparency of data sets used to train learned functions and a clear, ethical definition of precisely what is considered disinformation. Thus, this paper explicitly defines it in the following context, with detailed further explanation of bias explored later:

Disinformation is defined as *false information which is intended to mislead, especially propaganda issued by the media or government*. Specifically, it is content meant to misinform on issues of public health, election integrity, scientific consensus, or any coordinated political propaganda effort where disinformation is not created “in good faith”(as defined by accepted rules of war).[[1]](#footnote-0)

This definition ends matters of subjectivity. It accepts information which has faced scientific scrutiny and it appeals to democratic principles, while addressing influence operations without clear attribution from the source or a formal declaration of war.

The viral nature of the Internet adds complexity to the current problem. First, it allows content to be produced and deployed instantly from anywhere in the world. Second, current network infrastructure cannot guarantee non-repudiation of a source, making attribution difficult. Finally, recent evidence shows that these campaigns are using AI to generate content faster than ever before (Kreps, p. 4).

Advancement in AI will only make this problem worse. Furthermore, it is plausible that the same technology which created the problem cannot solve it—more AI will not make this go away. Computer scientists that research computational propaganda work in conjunction with social media platforms to understand the tactics and tropes of these campaigns. This project aims to apply the data and research collected in a new direction which works exclusively with open source intelligence (OSINT) gathered and analyzed given only a domain name. A supervised ML implementation will be used to classify domains given this data and a feature set determined through analysis of the curated dataset.

Disinformation campaigns exist in the United States (Kreps, 2020). These are sometimes state sponsored. Russia has been linked to online influence operations for decades(Earley, 2007). Speaking on the issue in the1984 work *Dezinformatsia*, a former Czechoslovak intelligence officer stated: “The primary target was the United States. The objective was to damage the United States wherever possible, and to weaken its position in Western Europe,” and that the KGB was consulted for “all directives” (p. 172).Disinformation has long been part of Russian political activity in the United States and has created overt and covert campaigns over the last 50 years to do exactly this. Disinformation created by Russians may contain “both true and false information, leaked to an opponent to deceive him… [the targets are] the decision makers rather than the public at large” (p. 37). Principally, influential people were targeted for agent of influence operations with the goal “not to recruit solely on the left, but rather across the political spectrum” (p. 167). Investigation into interference in the 2016 and 2018elections have further connected Russia’s use of influence operations in the United States(DiResta, 2019). Even the etymological origin of the word “disinformation” is Russian—*dezinformatsiya*—and was coined by Joseph Stalin in 1923, giving it a French-sounding name to make its origin appear Western (Pacepa & Rychlak, 2013). The United States has even engaged in covert influence operations against its own citizens, such as with COINTELPRO (Hoover, 1968) and the issue of Iraq and Weapons of Mass Destruction (Jamieson, 2007).

This project makes frequent discussion of Russian’s involvement in influence operations. The background research which follows, and data analyzed during the first phase of this project, will offer proof on why that is.

## IT Solution

The purpose of this paper is to explore the implementation of a supervised binary classification algorithm to find whether a given URL could be part of a disinformation campaign. This is a partial solution to the problem because such an algorithm could be further implemented for a more complete solution, such as a publicly available reporting application. This application could be used to detect, track, and deter campaigns by making its results public, its data sources, and determinations transparent, and by supplying a ledger of actions taken by domain registrars in deterrence of propaganda online. This ledger may help better inform the public about the extent of the problem.

This paper focuses purely on the implementation of an ML algorithm which uses OSINT about domain names in its determinations. Included within the scope of this is a curated data set composed of domain names from Alexa’s most trafficked websites, similar datasets with confirmed sources of disinformation, and other sources tracking the issue on the web. It will try tokenizing a given URL using NLP and decide whether text aligns with those used by confirmed influence operations. It will then output a decision on whether the given URL could be suspicious.

While outside the scope of this solution, the algorithm could be implemented as a part of a web interface which allows users to send suspected sources of disinformation. The algorithm is extensible, and its source code and data sets (including a data sheet) will be made publicly available. In line with all great solutions in IT, collaboration is encouraged.

## Implementation Plan

Implementation will be done using a waterfall methodology which starts with requirements gathering and analysis, and concludes with communication and documentation. During the first phase, requirements are defined. Part of this requires research into the nature of influence operations on the web, including common tactics and tropes pushed. This research will help the design of the ML algorithm by uncovering what might be of interest in the feature set. The requirements will provide a framework for what data need to be collected. The next phase is design: it establishes the logical relationships among the data collected and chooses the model used by the learned function. The implementation phase features building, testing, and refining the algorithm, making alterations to the algorithm's design and its feature set based on feedback received during this phase. Following implementation is verification, which will evaluate the outputs of the learned function and optimize its hyperparameters. The last phase is communication of results and publication of all relevant documentation, including datasheets for curated datasets.

# Review of Other Work

## Disinformation Sources and Strategies

This section explores the disinformation ecosystem, and what measures are exploited to ensure its spread. The *GEC Special Report* notes that multiple ecosystems exist and can have contradictory narratives. However, its model uses a discernible hub-and-spoke structure, in which state-sponsored propaganda is propagated out to the broader web through proxy sites. The proxy site is the hub and extending from it are the spokes leading to other shell websites, all of which are channeled through social media (p. 23).

Next, the Center for Information Technology and Society in its article *Where Does Fake News Come From?* recognizes that the mechanisms used to host and send disinformation work through four major factors: the right domain name, a willing host, content plagiarizing, and velocity of social media spread.

Finally, *The Tactics and Tropes of the Internet Research Agency* explores the topics, domains, groups, and strategies used by Russia’s Internet Research Agency (IRA). It offers the most detailed analysis to date on the methods used by the IRA to push their narratives on social media.

### GEC Special Report: Pillars of Russia’s Disinformation and Propaganda Ecosystem

A report from the U.S. Department of State’s Global Engagement Center (GEC) details some of the tactics used by Russians in spreading disinformation. Russian disinformation channels develop propaganda ecosystems that are unique for each group they are targeting. This allows for “varied and overlapping approaches that reinforce each other even when individual messages within the system appear contradictory” (U.S. Department of State, 2020). Attribution and identification using textual clues become difficult unless the narratives are known, and the content of the site aligns perfectly with those. Overlapping narratives may offer a clue. Of note, however, is the fact that narratives function in groups, and knowing some of the narratives could play a part in finding key words used in domain names. Furthermore, it builds a stronger case for looking at metadata than content itself.

The report made note of the successes of Russia’s strategy. It does so by the creation of multiple information ecosystems which are sometimes contradictory to one another with minor exceptions. The US Department of State found the following:

First, it allows for the introduction of numerous variations of the same false narratives. This allows for the different pillars of the ecosystem to fine tune their disinformation narratives to suit different target audiences because there is no need for consistency, as there would be with attributed government communications. Second, it provides plausible deniability for Kremlin officials when proxy sites peddle blatant and dangerous disinformation, allowing them to deflect criticism while still introducing pernicious information. Third, it creates a media multiplier effect among the different pillars of the ecosystem that boost their reach and resonance. (p. 5)

Investigating the “media multiplier effect” which flows from these proxy sites could be key in the identification of them.

### Where Does Fake News Come From?

According to the Center for Information Technology and Society (CITS), disinformation is “a multi-step process that involves making or taking content that others have produced, passing it off as real news, and capitalizing on social media to get as much attention as possible” (Center for Information Technology and Society, 2018). According to CITS, these sites have both ideological and commercial motivations to get as many users as possible to click their links. They use legitimate sounding domain names to trick users into believing they are real sources. These domains will generate a ton of content (artificially generated, stolen, or paid) to further make themselves seem legitimate, and then links to them are posted on social media and promoted by real and bot accounts. The process is repeated *ad infinitum*.

This model, according to CITS, “is so successful because it can be easily replicated, streamlined, and requires little expertise to operate. Clicks and attention are all that matter, provided you can get the right domain name, hosting service, stolen content, and social media spread.” Those four qualities can hold a lot of information for the identification of fake news content. Thus, characteristics of these campaigns can be seen through common domains, hosting service, plagiarized content, and velocity of spread on social media.

### The Tactics & Tropes of the Internet Research Agency

At the request of the U.S. Senate Select Committee on Intelligence (SSCI), New Knowledge reviewed an expansive data set of social media posts and metadata provided to SSCI by Facebook, Twitter, and Alphabet, plus a set of relational data from additional platforms (DiResta, 2019). The study reports on common themes present, dissemination through social media, and the extensive engagement Americans had with content.

It details the extensive operations targeting black Americans, which was the most prolific effort of the IRA, and that page owners on Facebook and Instagram “exploited the trust of their Page audiences to develop human assets, at least some of whom were not aware of the role they played” (p. 8). The report goes on to say: “This tactic was substantially more pronounced on Black-targeted accounts. The degree of integration into authentic Black community media was not replicated in the otherwise Right-leaning or otherwise Left-leaning content.”

Aligning with the GEC report mentioned previously, the themes targeted by the IRA were broad and sometimes contradictory (promoting both pro- and anti-Muslim content). “The themes selected by the IRA were deployed to create and reinforce tribalism within each targeted community; in a majority of the posts created on a given Page or account, the IRA simply reinforced in-group camaraderie.” (p. 12). Two themes were identical across all targeted communities: “narratives to erode trust in mainstream media, and narratives to convey Russian’s state-sanctioned talking points on the Syrian conflict.”

## ML implementations against disinformation

This section discusses the moral implications of using ML algorithms to illustrate influence campaigns, and some case studies on its application. In *The Moral Permissibility of Automated Responses during Cyberwarfare,* Danks & Danks supply the foundational definition for the algorithm’s classification. Automated responses, in this context, are morally permissible to classify disinformation when it is made in “bad faith” and does not fall within the bounds of the “rules of warfare.” Information meant to deceive, published by a foreign entity without clear attribution, would be considered within this context.

In the study *Fighting Misinformation on Social Media using Crowdsourced Judgments of News Source Quality*, the implications of relying solely on ML implementations is discussed. Aggregate trust scores on sources of information can play a discernible role in ML algorithms targeting the issue. Oftentimes one’s intuition about the reliability of a source, when prompted and regardless of political persuasion, is usually a correct judgment on its accuracy.

*Fabula AI is Using Social Spread to Spot Fake News* discusses a successful ML implementation, whose classification uses metadata points. Fabula’s successes show the usefulness of looking at metadata instead of the content itself. Connections between accounts and the velocity of material shared serve as key giveaways. Furthermore, it supplies a benchmark (classification with 93% accuracy) for a ML classification problem involving this topic.

*The Role of Technology in Online Misinformation* focuses on the tools used by influence operations to flood the web with narrative-based content. It discusses the possibility that sources of this issue use A.I. to generate large volumes of content quickly to meet production quotas, and that a solution for this issue would be to use A.I. in defense, especially by looking at metadata.

### The Moral Permissibility of Automated Responses during Cyberwarfare

This paper ethical definitions of classifying deception using automated responses, such as ML. Generally, the argument made by Danks & Danks is that automated responses are almost always permissible in when employed defensively, and likewise are almost never ethically permissible when applied offensively. Danks & Danks discuss the ethical tolerance of propaganda and that is sometimes, but not always, morally permissible. Morally permissible propaganda, in their view, occurs when the propaganda is made in “good faith”:

Accounts in this space typically focus on whether the disinformation is in “good faith” (Mattox, 1998): that is, disinformation is ethically permissible when it falls within the “rules of warfare.” For example, placing inflatable tanks on Pacific islands during World War II to give the appearance of greater troop numbers is morally acceptable; having a weapons factory masquerade as a hospital is not. (p. 22)

Arguably, this argument is a slippery-slope with potential for bias, and lacking in clarity about what are the acceptable “rules of warfare”. For the sake of this project, the example provided by Danks & Danks in the excerpt relates to the problem of web propaganda because it shows clear attribution of source. In this case, the inflated enemy troops were clearly meant to be allied forces along with a formal acknowledgment of war by both sides

### Fighting misinformation on social media using crowdsourced judgments of news source quality

This study focuses on the feasibility and impact of using crowdsourced judgments on the accuracy of a piece of content. “Our results indicate that using crowdsourced trust ratings to gain information about media outlet reliability—information that can help inform ranking algorithms—shows promise” (Pennycook & Rand, 2019). The study noted that participant ratings were, in aggregate, quite successful at differentiating mainstream media outlets from hyper partisan fake news websites. Furthermore, ratings given by participants “were very strongly correlated with ratings provided by professional fact-checkers.” It concludes that incorporating trust ratings of everyday users into social media ranking algorithms could help stem the spread of misinformation. Furthermore, it begs the necessity of integrating aggregate human judgment to bolster the accuracy of an ML application.

### Fabula AI is using social spread to spot fake news

Fabula AI is using concrete data points to detect problematic content, instead of detecting “fakeness.” “The approach it’s taking to detecting relies not on algorithms parsing news content to try to identify malicious nonsense but instead looks at how such stuff spreads on social networks — and also therefore who is spreading it” (Lomas, 2019). It does so through implementing ML through a process called geometric deep learning (GDL). GDL is used to generalize non-Euclidian domains (in this case, graphs) in deep neural modules. Fabula co-founder and chief scientist Michael Bronstein explains its function:

The essence of geometric deep learning is it can work with network-structured data. So here we can incorporate heterogenous data such as user characteristics; the social network interactions between users; the spread of the news itself; so many features that otherwise would be impossible to deal with under machine learning techniques.

Its proprietary algorithm is capable of classifying disinformation with about 93% accuracy, which the organization believes could be used in conjunction with human supervision to further false negatives.

### The Role of Technology in Online Misinformation

According to interviews from former workers with the New York Times (MacFarquhar, 2018), IRA employees, often students, work in 12-hour shifts generating content which matches assigned narratives. Kreps theorizes in *The Role of Technology in Online Misinformation* that AI could be used in the creation of content that can be rapidly and uniquely generated. GPT-2, a text-prediction tool, could be used by savvy actors to filter distorted output and churn out more credible sounding text (Kreps, 2020). Fortunately, Kreps argues, AI can easily combat artificially generated content, since neural networks are familiar with the habits and quarks of these. One implementation referenced is Grover, which was built by the Allen Institute for AI. Grover “not only generates neural fake news but also spots its own fake news and that of other AI generators” with a 92% accuracy.

The paper also discusses looking at metadata to identify synthetic text:

Another tech-based solution involves analyzing metadata to identify synthetic text. Algorithms can be trained to identify the markers of malicious documents — such as the time written to produce the text, the number of accounts associated with a particular IP address, or the website itself — to identify malicious or inauthentic text. (p. 7)

Deciding suspicions about disinformation campaigns using textual clues is tricky. But a common trend among the works cited in this report is becoming clear: using metadata to connect disinformation campaigns is much stronger than relying purely on NLP content analysis.

### Peering Under the Hood of Fake-News Detectors

An article from MIT News reports on textual content analysis and the challenges presented to it for ML detection of disinformation. It illustrates the “black box” problem of AI—the inability to know exactly what the algorithm is doing or what methods it is using—and shows its implications in training ML algorithms to detect it, particularly involving new topics. It reports on the development of “a deep-learning model that learns to detect language patterns of fake and real news” (Matheson, 2019). But researchers also conclude that it is difficult to control for the “many different types of biases in language,” posing the question: “How do we make sure that a system trained on this dataset would not learn that real news must necessarily follow the writing style of these two specific news outlets?” Its model was trained on content mostly from *The New York Times* and *The Guardian*.

### Fake News Detection: A Deep Learning Approach

Researchers at Southern Methodists University developed neural network architecture to address the shortcomings of binary classification problems in ML detection of deception. Their model uses stance detection between a news article’s content and its headline and defines whether the stances between them “agree, disagree, discuss, or unrelated.” The approach admits an “automated solution requires understanding the natural language processing which is difficult and complex,” and that this is quite a daunting task. Like many others studied, it focuses highly on NLP to evaluate the information it is fed against the stance of the article content and its headline.

## IT Industry Responses

Stakeholders in the world of IT have a personal stake in preserving the sanctity of information online. This section discusses some of the efforts made by domain registrars and supplies a case study on how DNS architecture can expose suspicions about deception sources and methods of operation.

### ICANN Org’s Multifaceted Response to DNS Abuse

The idea of exploiting DNS registrations and domain names themselves is an enticing way to classify sources of propaganda. ICANN (Marby, 2020) is doing exactly this when it comes to fraudulent domain names registered during the start of the COVID-19 pandemic. In an article posted to Internet Corporation for Assigned Names and Numbers (ICANN) website, these domains lure in users promising cures for the coronavirus or are spreading disinformation or infecting users’ devices with malware.

ICANN developed a system which “system looks for domain names like or incorporating terms such as “coronavirus,” “covid,” “pandemic,” “ncov,” and others, and once identified, assesses them against multiple high-confidence threat intelligence sources to determine whether or not they are involved in phishing and/or malware distribution.” These are collected and shared with parties able to act against the site: registrars and registries, and in some cases national and international law enforcement organizations. This system is being developed and audited by ICANN to ensure that “the reports generated by the system meet their reporting requirements so that appropriate action can be taken in a prompt fashion.”

It is important to note that ICANN is doing this in response to the pandemic and has elevated its domain screenings to be processes with high priority, saying: “ICANN org was never granted, nor was it ever intended that it be granted, authority to act as a regulator of Internet content.” They are responding with their power and authority to prevent the spread of health-related lies. ICANN is recognizing the threat posed by this and is acting against it, despite never intending to be a regulator of Internet content.

ICANN also has the authority to revoke a registrar’s accreditation (the relationship between registrars and ICANN is decided by a registrar accreditation agreement, or RAA) if they do not follow ICANN’s policies. However, DNS abuse reports are closed by an informal process by the registrar themselves, with little oversight or accountability on how long the process took, what the findings may have been, and ultimately whether the registrar acted against the domain. But for matters of public health is this the best way? ICANN reasons as follows:

ICANN Compliance urges parties who come across domain names that appear to be used to perpetrate DNS abuse, particularly related to COVID-19, to report them to the relevant registry and registrar and to submit complaints to ICANN Compliance if they believe that the contracted parties have failed to adequately address their abuse report in a timely and reasonable manner.

ICANN is proving that IT entities can take collaborative action against threats to the Internet community. An Internet which was designed to be wild and free is now facing the widespread threat of bad actors exploiting that characteristic for ill-will. Authorities are finding they can use collaborative power for good. The question of transparency is one which must be asked: if domain registrations are transparent, why not the same for withdrawal of a registration? What can be done to hold parties accountable, and what can be done for users to know when and where action was taken against a domain, especially one that perpetrated fraudulent content?

### Measures Taken by Registries to Help Tackle COVID-19 Related Online Abuse

The Council of European National Top-Level Domains (CENTR) published in May 2020 an article on their blog about what country-code top-level domain (ccTLD) operators were doing to tackle COVID-19 related online abuse. ccTLD operators manage or administer country-specific top-level domains, such as .si or .eu. They found that COVID-19 related abuse was low in the realm of ccTLDs. “As the managers of top-level domains, ccTLDs are technical operators that do not have control over the content of websites. They rely on close cooperation with public authorities in responding to COVID-19 related abuse” (Council of European National Top-Level Domain Registries, 2020). CENTR stated:

EURid TLD, the registry for .eu (and its variants in other scripts), pro-actively monitors newly registered domain names, involving its machine learning based APEWS (Abuse Prevention and Early Warning System). In order to protect end-users from possible misuses of domain names associated with the current COVID-19 pandemic, EURid has… [been] adapted to prevent the registration of suspicious domain names, and EURid performs additional checks on the registration data of newly-registered domain names that contain keywords relating to the current pandemic.

Another instance is in Denmark, where the registry operator for .dk supplies national police with a list of newly registered .dk domain names every day.

Of note for this project is that many of the European ccTLDs are regulated and subject less to influence operations. Furthermore, domain names often supply revealing clues into the intentions surrounding a given domain.

### Who’s Behind the “Reopen” Domain Surge?

This article, published on the popular security blog *Krebs on Security*, brilliantly shows the ability to correlate coordinated misinformation campaigns using domain names and associated WHOIS query data. In this instance, a slew of similar domain names was registered in succession around the time President Trump sent a series of all-caps tweets urging citizens to liberate themselves from social distancing restrictions. The author details strategies used in their research to link various domains. This includes looking at the Google Analytics tracker in the source code for some of the websites. A coordinated campaign would use the same tracker across multiple websites.

Importantly, Krebs recognizes the utility of finding clues and making connections using information supplied by registrants. Domains “includes a date and timestamp down to the second that the domain was registered” (Krebs, 2020). Grouping these together (for domains with obfuscated registration details) and “comparing them to domains that do include ownership data, we can infer more information.”

In his research, Krebs found 50 “reopen” domains were registered with the registrant name Michael Murphy on 17 April between 3:25 p.m. ET and 4:43 ET. Krebs goes on to say: “No one responded to the email addresses and phone numbers tied to Mr. Murphy, who may or may not have been involved in this domain registration scheme. Those contact details suggest he runs a store in Florida that makes art out of reclaimed or discarded items.”[[2]](#footnote-1)

This is the exact premise of the ML algorithm implemented in this project. DNS registration information will be used to connect ownership across domains, by looking for patterns of registrations based on time, domain name, HTML header information, and other patterns detected in domain registration information. It will make its classification based on this information alone.

# Project Rationale

Disinformation is a genuine problem. It exploits social media algorithms designed to promote engagement and erodes collective trust. Disinformation during a pandemic has hastened its spread, and, leading into an election, threatens to confuse an already convoluted American voting system. Most research conducted thus far relies on NLP to make classifications. Using NLP to make determinations on the trustworthiness of content is not enough—detecting this issue requires human intuition and crowdsourced grunt work *in conjunction* with ML algorithms. Therefore, this project is being implemented to address the shortcomings of these methods and offers an alternative that uses metadata to connect sources in coordinated campaigns. The Internet was intended to ensure the wide scale distribution of knowledge. Those who built the architecture to make this possible, while acknowledging that information could be spread in bad faith, could never have foreseen the role algorithmic recommendation systems would play in this perpetuation.

Had systems of non-repudiation been included in the architecture, or systems which could manage and accredit sources of information made widely available, the problem may not be as bad as it is. Furthermore, it is still possible—with collective will—that such systems *be* implemented. But the point of this paper is not to fault these systems; in fact, fault lies with social media companies which receive tremendous benefit, in growth and engagement, when their users routinely interact with disinformation. Meanwhile governments, most notably the United States, have been useless in controlling the problem. Insight published by Texas Representative Will Hurd found that “of the 535 members of Congress, BGR has only been able to identify four whose education background includes a computer science degree” (Meek, 2016). Not only does the legislature lack the experience to understand or dictate action against this problem, but its median age does not skew towards those who have spent a broader part of their lives using the technology.

Something needs to be done to address the issues and that something must be multi-faceted, involving representatives and work from all IT groups, as well as government leadership, oversight, and recognition of the problem. Bad actors and foreign governments are consciously waging a cold war against the American population in ways that are meant to destabilize. This is an act of aggression. It is not wise to ignore it or blow it off. This paper stresses the need of those denying the existence of the problem to consider the theoretical implications of widespread psyops warfare on confidence in the economy, the destruction of social fabric, and trust in technology.

Nothing will be accomplished through the implementation of one algorithm, or any number of them. Addressing this requires collective will and some regulation.

# Current Project Environment

Disinformation has grown more complex than ever: social media algorithms, which trend toward polarizing content to increase engagement from users, are being routinely manipulated to spread fabricated information. The problem is not a matter of subjective interpretation of the narratives of the disinformation—the problem is the accuracy and power these social media algorithms have in fundamentally changing their users’ behaviors.

The AI driving these algorithms, and the billions of users from which they have harvested personal information about, have become exceptionally good at what they do. For all their complexity and engineering these algorithms are not without manipulation from bad actors pushing narratives that fit well within the business scope of serving advertisements through these platforms.

Arbitration of truth becomes necessary when the content being served has the power to fundamentally harm society. At the time of authoring this paper, misinformation produced and spread across social media in the US has no doubt fueled the pandemic. Disinformation produced about anthropogenic climate change, a scientific consensus agreed upon by at least 97% of climate scientists, has undoubtedly contributed to a failure to stem its future impacts and threatens a hotter, harsher, and less stable world for any generations following. Diseases which were once eradicated are reappearing thanks to this same effect, as the efficacy of vaccines is brought into question not just by random users on the internet but coordinated attempts to undermine the health and safety of entire populations.

The project environment exists in the unstructured world of the Web. The spread of disinformation most often occurs through media and not text, setting up added complications in the fight against it. Its spread, however, does occur through the sharing of links which point to fraudulent or deceptive websites.

The data analysis and ML algorithm produced in this project are built off the myriad research performed on misinformation on the Internet and, specifically, social media. From this research certain characteristics of campaigns appear, such as where they originate, patterns in their narratives, methods of transmission, and time series information for determining trends. Disinformation is being combated by disparate sources around the country and world. This project builds off, and contributes to, this effort.

The primary deliverables of this project will be its curated datasets and its ML classification algorithm. The scope of the project, however, leaves it orphaned upon completion. All published material resulting from this project will detail an intended roadmap for future implementation. With little effort, this project could easily be implemented into a web application where users can report domain names and, with proper accuracy, the algorithm should be able to make confident classifications (as shown in the research above) with help from human intervention.

# Methodology

This project uses a waterfall methodology. The first phase is centers around gathering requirements and analysis. After framing the problem and defining requirements, collect the data (ideally a lot of it). Following data collection, the data are tidied and summarized in such a way that a detailed analysis becomes accessible. Following data munging, the data are explored and analyzed using descriptive statistics to understand what features may compose the ML algorithm. Visualizations are used to understand the data more.

The next stage will be system design, which includes detailing entity relationships, specifying the hardware and software requirements for the system, and reducing dimensionality of the data. This is also the stage when the proper ML model is chosen for the data.

Following the design stage is implementation: during this stage, the learned function will be developed. The verification phase, which follows implementation, will be devoted to evaluating the outputs of the ML algorithm and optimizing hyperparameters.

The last stage is interpretation and communication of results, with emphasis on documentation. Datasets will be accompanied with detailed datasheets explaining the purpose of the curation, sources used, and detailed information about the data contained within.

# Project Goals, Objectives, and Deliverables

## Goals, Objectives, and Deliverables Table

|  |  |  |  |
| --- | --- | --- | --- |
|  | Goal | Supporting objectives | Deliverables enabling the project objectives |
| 1 | Collect & analyze data on disinformation | 1.a. Define requirements and collect data | 1.a.i. Define the data requirements |
| 1.a.ii. Data collection: Using various sources, perform research and find instances of structured and unstructured data sets like this problem |
| 1.b. Analyze the data collected | 1.b.i. Data cleansing: perform necessary cleanup of the data, including null values, duplicates, mistaken data, or is incomplete |
| 1.b.ii. Perform an exploratory data analysis on the data collected, creating visualizations where appropriate |
| 1.b.iii. Modeling: using statistical inference, decide properties of the population |
| 1.c. Database design & modeling | 1.c.i. Divide the data into subject-based tables, divide information items into columns, and specify the primary keys |
| 2 | Implement a supervised binary classification algorithm | 2.a. Implementation | 2.a.i. Illustrate the ER diagram and create relationships among tables |
| 2.a.ii. Write corresponding SQL schema for use with Oracle DBMS |
| 2.a.iii Refine and normalize the design of the database |
| 2.a.iv. Implementation: set up a ML pipeline comparing each algorithm on the data set; evaluate each performance |
| 2.a.v. Hyperparameter optimization: choose the set of best hyperparameters for a learning algorithm |
| 3 | Document & make publicly available | 3.a. Write documentation & upload content | 3.a.i. Write associated README files for source code; organize and cleanly present on GitHub |
| 3.a.ii. Write associated data sheet for the datasets; organize other relevant data and cleanly present on Kaggle |
| 3.b. Post-implementation planning | 3.b.i. Draft report on how this ML algorithm could be used, improved upon, and integrated into a broader web-based reporting mechanism for suspect misinformation websites |

## Goals, Objectives, and Deliverables Descriptions

### Goal 1: Collect & analyze the data

The first goal is to curate datasets finding various sources of fake news. Data sources include governmental research, datasets publicly posted, and from other sources as believed necessary during this stage. Iterative by nature, there is an expectation that data collection will swell and contract until the right data points are gathered, wrangled, and analyzed. The final deliverable from this goal will be an RMarkdown notebook plotting and detailing the analysis of findings, including detailed graphical analysis.

#### Objective 1.a. Define requirements and collect data

The first aim of this goal is towards the research and collection of the various data sources. Two deliverables will complete the objective. First, define the data requirements: What are the dimensions and measures one could hope to collect? What is the end goal? Second, collect the data from various sources and organize for analysis. These data sources may be already published, or they may be custom-curated.

#### Objective 1.b. Data analysis

Now that requirements are clear, analysis of the data gathered will occur. The final objective of the goal is the exploratory data analysis (EDA) step, including munging the data collected. Scripts written for this purpose will be written in R with some Python scripting as necessary. This step will implement both descriptive and inferential statistics about the data collected, presented in an RMarkdown notebook. Tableau may be used as a supplement for visualizations and integrated into the final notebook output.

#### Objective 1.c. Database design & modeling

After data collection and analysis comes the design and modeling aspect of the project. The first deliverable focuses on defining primary keys and the subject of the tables. After defining requirements, the ER model for the database will be created. Next, the SQL schema will be authored, and finally normalized to the 2nd normal form.

### Goal 2: Implement a supervised binary classification algorithm

The second goal and its sole objective is the implementation of the machine learning algorithm behind the binary classification of a given domain name, deciding whether it is associated with a known influence operation (or has characteristics like other influence operations). The final deliverable will be the completed algorithm.

#### Objective 2.a. Implement machine learning algorithm

This objective is fluid. The models researched during this stage can have different results, and the definition of feature vectors, at this time, is unknown. Furthermore, setting a benchmark for accuracy is also difficult, though prior work defined in this paper has categorized success as >90% accuracy in classification problems for disinformation. Worth noting is the fact that these ML problems involved textual analysis, and while this project will integrate textual analysis and NLP in a small measure, it focuses more on classification from OSINT metadata.

### Goal 3: Document & make publicly available

The final goal is aligned with the communication step of the data science process: results are published and discussed, documentation is written and made available, and a plan for implementation is discussed. Objective 3.a. governs the publishing of all content created during the project and their publication with appropriate datasheet. Objective 3.b. focuses on post-implementation planning, which discusses further implementation or improvement of the algorithm, its uses and intentions, and lessons learned during execution of the project.

# Project Timeline with Milestones

|  |  |  |  |
| --- | --- | --- | --- |
| Milestone or deliverable | Duration  (hours or days) | Projected start date | Anticipated end date |
| Definition of requirements and data collection | 2 days | 10/19/2020 | 10/20/2020 |
| Data tidying and analysis | 2 days | 10/22/2020 | 10/23/2020 |
| Database design & modeling | 6 hours | 10/26/2020 | 10/26/2020 |
| Implementation of the ML algorithm | 2 days | 10/27/2020 | 10/29/2020 |
| Writing of documentation & uploading content | 1 day | 10/30/2020 | 10/30/2020 |
| Post-implementation write-up | 1 day | 11/2/2020 | 11/2/2020 |

# Outcome

The author expects a statistically significant, but not within the top decile, rate of classification for the algorithm. However, for the project to be considered successful, the accuracy of the classification algorithm must be above 90% accuracy. The completed dataset used to train the algorithm will be at least *N* = 10000 with at least 700 domains classified as disinformation, and the rest selected from Alexa’s web traffic ratings. Overall, this project is but another attempt to answer the call to solve the problem of misinformation which is only likely to get significantly worse considering the multiplicative nature of networks. With sincere acknowledgment, and with a successful implementation which approaches 100% classification and widespread user contributions, this model would still pale in comparison to the effect that social networks, with their incredible amount of relevant data, could have over this. With a loss of trustworthy local news outlets (due in no short part with their inability to compete with free media on the social web), the broader public conscience has lost touch with the old saying: “Don’t believe everything you read on the Internet.” Unfortunately, the lack of skepticism has now resulted in countless preventable deaths from the coronavirus and crumbling social cohesion.

# References

116th Congress 1st Session Senate. (2019). *Report 116-XX.* Washington, D.C.: Select Committee on Intelligence United States Senate.

Broniatowski, D. A. (2018). Weaponized Health Communication: Twitter Bots and Russian Trolls Amplify the Vaccine Debate. *American Journal of Public Health*, 108(10), 13.

Center for Information Technology and Society. (2018, August). *Where Does Fake News Come From?* Retrieved from Center for Information Technology and Society: https://www.cits.ucsb.edu/fake-news/where

Council of European National Top-Level Domain Registries. (2020, May 04). *Measures taken by registries to help tackle COVID-19 related online abuse*. Retrieved from Council of European National Top-Level Domain Registries: https://www.centr.org/news/blog/registries-and-covid-abuse.html

Danks, D., & Danks, J. (2013, March 17). *The Moral Permissibility of Automated Responses.* Retrieved from https://www.andrew.cmu.edu/user/ddanks/papers/AutomatedResponses-Final.pdf

DiResta, R. S. (2019). *The Tactics & Tropes of the Internet Research Agency.* Retrieved from http://www.reneediresta.com/ira-report-4e8d0ff684.pdf

Earley, P. (2007). *Comrade J: The Untold Secrets of Russia's Master Spy in America After the End of the Cold War.* Penguin.

Federation of American Scientists. (2020, July 23). *Weekly COVID-19 Disinformation and False Propaganda Report.* Retrieved from Federation of American Scientists: https://fas.org/wp-content/uploads/2020/07/COVID-19-Disinformation-Report-for-July-20.pdf

Gebru, T. (2018, March 23). *Datasheets for Datasets.* Retrieved from arXiv: https://arxiv.org/abs/1803.09010

Gregory, J. (2020, February 28). *The coronavirus ‘infodemic’ is real*. Retrieved from STAT News: https://www.statnews.com/2020/02/28/websites-spreading-coronavirus-misinformation-infodemic

Hoover, J. E. (1968, March 4). *The FBI Sets Goals for COINTELPRO*. Retrieved from HERB: Resources for Teachers: https://herb.ashp.cuny.edu/items/show/814

Jamieson, K. (2007). Justifying the War in Iraq: What the Bush Administration's Uses of Evidence Reveal. *Rhetoric & Public Affairs*, 10(2), 249-273.

Krebs, B. (2020, April 20). *Who’s Behind the “Reopen” Domain Surge?* Retrieved from Krebs on Security: https://krebsonsecurity.com/2020/04/whos-behind-the-reopen-domain-surge/

Kreps, S. (2020, June). *The Role of Technology in Online Misinformation.* Retrieved from https://www.brookings.edu/wp-content/uploads/2020/06/The-role-of-technology-in-online-misinformation.pdf

Levin, D. (2019, October 21). *Mimicking Local News, a Network of Michigan Websites Pushes Politics*. Retrieved from The New York Times: https://www.nytimes.com/2019/10/21/us/michigan-metric-media-news.html

Lomas, N. (2019, February 6). *Fabula AI is using social spread to spot ‘fake news’*.’Retrieved from TechCrunch: https://techcrunch.com/2019/02/06/fabula-ai-is-usual-spread-to-spot-fake-news/

MacFarquhar, N. (2018, February 18). *Inside the Russian Troll Factory: Zombies and a Breakneck Pace*. Retrieved from The New York Times: https://www.nytimes.com/2018/02/18/world/europe/russia-troll-factory.html

Manning, J. E. (2018). *Membership of the 115th Congress: A Profile.* Washington, D.C.: Congressional Research Service.

Marby, G. (2020, April 20). *ICANN Org's Multifaceted Response to DNS Abuse*. Retrieved from ICANN: https://www.icann.org/news/blog/icann-org-s-multifaceted-response-to-dns-abuse

Martin-Rozumiłowicz, B., & Kužel, R. (2019, August). *Social Media, Disinformation and Electoral Integrity: IFES Working Paper.* Retrieved from International Foundation for Electoral Systems: https://www.ifes.org/sites/default/files/ifes\_working\_paper\_social\_media\_disinformation\_and\_electoral\_integrity\_august\_2019\_0.pdf

Matheson, R. (2019, February 6). *Peering under the hood of fake-news detectors*. Retrieved from MIT News: https://news.mit.edu/2019/opening-machine-learning-black-box-fake-news-0206

Meek, A. (2016, March 23). *We asked every member of Congress with a computer science degree about Apple’s war with the FBI*. Retrieved from Congressman Will Hurd: https://hurd.house.gov/media-center/in-the-news/we-asked-every-member-congress-computer-science-degree-about-apple-s-war

Morgan, J., & DiResta, R. (2018, July 10). *Information Operations are a Cybersecurity Problem*. Retrieved from Just Security: https://www.justsecurity.org/59152/information-operations-cybersecurity-problem-strategic-paradigm-combat-disinformation/

Pacepa, I., & Rychlak, R. (2013). *Disinformation: Former Spy Chief Reveals Secret Strategies for Undermining Freedom, Attacking Religion, and Promoting Terrorism.* WND Books.

Pennycook, G. E. (2019, November 13). *Understanding and reducing the spread of misinformation online.* Retrieved from https://doi.org/10.31234/osf.io/3n9u8

Pennycook, G., & Rand, D. (2019, February). Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences*, pp. 116 (7) 2521-2526.

Shultz, R. H., & Godson, R. (1984). Dezinformatsia: Active measures in Soviet strategy. Washington, DC: Pergamon.

Thota, A., Tilak, P., Ahluwalia, S., & Lohia, N. (2018). Fake News Detection: A Deep Learning Approach. *SMU Data Science Review*, Vol. 1 : No. 3 , Article 10.

U.S. Department of State. (2020). *GEC Special Report: Pillars of Russia’s Disinformation and Propaganda Ecosystem.* Washington, D.C.: U.S. Department of State.

1. Clarification on disinformation as acceptable warfare in nuanced; for a deeper explanation see Danks & Danks (2013). [↑](#footnote-ref-0)
2. Anecdotally, the author found similar patterns while researching suspicious domains registered in the US, where shell companies were used, or address information pointed to, businesses which were closed sharing the same address as a given company and had no other markers of engaging in business of any kind. [↑](#footnote-ref-1)