Classifying Disinformation by Domain Name

Trevor J. McGlynn

Western Governors University

**Table of Contents**

[Summary 3](#_Toc441471473)

[Review of Other Work 3](#_Toc441471474)

[Changes to the Project Environment 3](#_Toc441471475)

[Methodology 3](#_Toc441471476)

[Project Goals and Objectives 3](#_Toc441471477)

[Project Timeline 4](#_Toc441471478)

[Unanticipated Requirements 4](#_Toc441471479)

[Conclusions 4](#_Toc441471480)

[Project Deliverables 4](#_Toc441471481)

[References 6](#_Toc441471482)

[Appendix A: 7](#_Toc441471483)

[Appendix B: 8](#_Toc441471484)

[Appendix C: 9](#_Toc441471485)

[Appendix D: 10](#_Toc441471486)

# Summary

The purpose of this project was to see with what accuracy a domain name could be classified as disinformation using only publicly-available information. A list of domain names was curated from several sources: legitimate domains were scraped from Alexa web rankings, while domains classified as sources of disinformation were curated from per-compiled reports from librarians at Merrimack College, the New York Times, security and computer science researchers, and other sources, which were repeatedly validated with set intersections throughout the curating process. From these domain names and their applied trust ratings, WHOIS data was queried, analyzed, and summarized, with clustering of connected websites validated through Google Analytics codes scraped from HTML head tags. The data were fed through a random forest classifier (RFC) which resulted in a 99.4% accurate classification against its test data—far exceeding expectations.

Data collection started with data sets currently available on Kaggle.com, all of which were curated for the intention of spotting influence operations on the web. Per the requirements for the project, the most interesting attribute were domain names in these data sets. Next, a recently released data set from the New York Times which included fake local news websites was brought in, followed by web scrapes and a PDF scrape of confirmed sources of disinformation from Merrimack College.

Alexa web rankings fed the data marked “initial trust”. An assumption was made that trustworthy websites, for the most part, filter themselves over time. Due to internet connection errors here in northern Vermont during the querying, the sample size was reduced from n = 60,000 to n = 25,301; this was still an acceptable size for the purposes of the project. Domains which were labeled “fake” were validated and audited against this sample. 32 domains were re-categorized based on the results—this involved partial bias on my part, and for some domains the learned function ended up classifying some of them back to their original state.

Once the domains were collected, scripts were written to query and process WHOIS data tied to the domains. Attributes collected included: the registrar, number of name servers, and update, creation, and expiration information. From these, date information were analyzed (such as the age of a domain in days). Domain names split into keywords and were calculated for their total length, number of keywords, and were scored according to what keywords were used in the domain and weighed against the frequency of those keywords in the fake domain sample. Registrars were scored in a similar fashion.

During this stage, each domain that was labeled “fake” was also scraped to see what, if any, Google Analytics codes they may have been using. This revealed the intricate but obvious connections among domestic influence operations were shell “local news” websites are created to drive narratives (all of these sites were pro-Trump, pro-GOP and none of them are legitimate media sources).

After collecting the data, an RMarkdown notebook was created to explore the finished data set. During this stage the data were again validated, and relationships were explosed which revealed the strengths of the domain name analysis and metrics computed during the collection process. This step provided meaningful insight into the data and prepared for the success seen during the modeling step.

A RFC was chosen as a model for the machine leaning aspect of the project, due to its accuracy in classification problems with multiple features. The final features chosen were the following: the length of the domain name, the length of keywords in the domain, the number of name servers, the age of the domain (in days), the computed domain name score, the computed registrar score, and the domain’s Alexa web ranking. The classifier reported an accuracy rate of 99.4%, which exceeded the expectation of 93% set in the proposal.

# Review of Other Work

In this section, provide an expanded review of the Review of Other Work section in task 2, including three additional third-party artifacts on the topic that supported the development of the project, and explain how the artifacts supported the implementation.

https://www.nytimes.com/2020/10/20/technology/timpone-network-pay-to-play-local-news.html?action=click&module=Top%20Stories&pgtype=Homepage

https://www.nytimes.com/2020/10/18/technology/timpone-local-news-metric-media.html

https://library.southtexascollege.edu/false-misleading-clickbait-y-and-satirical-news-sources/

This module segments text according word frequency using the Viterbi algorithm. Probably

due to Peter Norvig somehow.

Three sources of frequency information is provided.

The third is from a webcrawl dataset of anchor text provided

by Vinay Goel of the Internet Archive.

# Changes to the Project Environment

The project environment has worsened since the project began, at least under the context that disinformation on the web has probably gotten worse. This project, however, does entertain a new way of identify disinformation using only data that is publicly available, and one that apparently works quite well. Most research investigated during the proposal required information given by social media companies; this project addresses domain-based disinformation that is entirely OSINT.

That said, this project does offer a promise of another solution for the problem, but not a terribly meaningful one. It best uses would be in finding coordinated domestic fake news operations who on the web, but not on social media.

# Methodology

This project used a waterfall methodology. In the first phase, requirements were established and data were gathered. Following data collection, the data were tidied and summarized to make analysis easier. Following data munging, the data were explored and analyzed to audit, explore, and understand its underlying features.

The next stage was system design, including detailing entity relationships and reducing the dimensionality of the data. During this stage, research was also done on how to best model the problem for a ML implementation.

Following design, the learned function was developed and verified against the test data and the optimization of its hyperparameters.

The last stage is interpretation and communication of results, with emphasis on documentation. Data sets are accompanied with detailed data sheets explaining the purpose of the curation, sources used, and detailed information about the data contained within. An RMarkdown notebook was published showing some of the findings within the data set.

# Project Goals and Objectives

In this section, provide a detailed explanation on how some goals and objectives were met and why some goals and objectives for the project were not accomplished. Identify the objectives that were met and explain how they were met, and then explain the reasons why some objectives were not accomplished.

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

In this section, compare the projected and actual timelines of the milestones or deliverables of the project and explain why the differences occurred. Explain the reasons for each deviation of the actual time frame from the estimated time frame.

Note: All timeline dates MUST be in the past as this document is an after-action report that should reflect a project that is completed.

# Unanticipated Requirements

In this section, describe the requirements or components that were not anticipated at project initiation but emerged during implementation. Describe the problems encountered and the unanticipated requirements, and then explain how they were resolved or why they were not solved.

# Conclusions

In this section, provide an explanation of the actual results and potential effects of the completed project. Describe the actual project accomplishments and discuss the immediately observable effects and potential future impacts of the completed project on the project environment. Explain why the project is or is not considered successful using the evaluation framework from the Outcome section in the project proposal.

# Project Deliverables

In the Project Deliverables section, explain and detail the project key deliverables. The actual project development will be documented by the key deliverables. The project includes some sort of formal report. The deliverables should provide a detailed logical explanation of what the project provided to substantiate the work and completion of such. Describe the artifacts being used to show evidence of the project’s completion and use the appendices to include the actual artifacts. Actual project artifacts may include code samples or screen shots; flowcharts, UML, or other process diagrams; charts, tables, and graphs; network diagrams (before and after); training materials; and/or the technical IT product itself.

# 

# References

List all the outside sources that the narrative refers to in-text. For in-text and reference list citations, please refer to the web link in or visit the WGU Writing Center.

Smyth, A. M., Parker, A. L., & Pease, D. L. (2002). A study of enjoyment of peas. Journal of Abnormal Eating, 8(3), 120-125. Retrieved from

http://www.articlehomepage.com/full/url/

Bernstein, M. (2002). 10 tips on writing the living Web. A List Apart: For People Who Make Websites, 149. Retrieved from http://www.alistapart.com/articles/writeliving

Bell, T., & Phillips, T. (2008, May 6). A solar flare. Science @ NASA Podcast. Podcast retrieved from http://science.nasa.gov/podcast.htm

OLPC Peru/Arahuay. (n.d.). Retrieved April 29, 2011 from the OLPC Wiki: http://wiki.laptop. org/go/OLPC\_Peru/Arahuay

Plath, S. (2000). The unabridged journals. K. V. Kukil (Ed.). New York, NY: Anchor.

# Appendix A

# Title of Appendix

Put any supporting material in these appendices. Add additional or delete superfluous appendices as needed.

# Appendix B

# Title of Appendix

Put any supporting material in these appendices. Add additional or delete superfluous appendices as needed.

# Appendix C

# Title of Appendix

Put any supporting material in these appendices. Add additional or delete superfluous appendices as needed.

# Appendix D

# Title of Appendix

Put any supporting material in these appendices. Add additional or delete superfluous appendices as needed.