Classifying Disinformation by Domain Name

Trevor J. McGlynn

Western Governors University

**Table of Contents**

[Summary](#_30j0zll) **4**

[Review of Other Work](#_1fob9te) **6**

[Changes to the Project Environment](#_3znysh7) **8**

[Methodology](#_2et92p0) **8**

[Project Goals and Objectives](#_tyjcwt) **9**

[Goal 1: Collect & analyze the data](#_9vi39chsi5qo) 9

[Goal 2: Implement a supervised binary classification algorithm](#_6f55jv18bj9v) 9

[Goal 3: Document & make publicly available](#_g8d9q3n3ptum) 10

[**Project Timeline**](#_mml5insgr3h6) **11**

[Unanticipated Requirements](#_1t3h5sf) **11**

[Conclusions](#_4d34og8) **12**

[Project Deliverables](#_2s8eyo1) **12**

[Data sets](#_krdna2m4ffpa) 13

[Exploratory data analysis](#_kq7xstdn7sr) 13

[Binary Classification Algorithm](#_oqsxdqx1inxt) 14

[References](#_3rdcrjn) **17**

[**Appendix A**](#_ediyykmdr28k) **18**

[**Analysis of domain data for suspected sources of disinformation**](#_fekzf2l7e1ak) **18**

[**Analysis**](#_k5nql0ksngor) **18**

[Univariate analysis](#_yfw96sz80mq2) 18

[Countries](#_lu4bovmnudhd) 19

[TLDs](#_w7t6a9yqmrn9) 20

[Registrars](#_8062nj49sag3) 21

[Google Analytics](#_kq9dxgtwu5is) 21

[Words used in domain name](#_a0zkkw2pxvve) 22

[Bivariate analysis](#_nqiygwi4fc9s) 23

[Density of domain registrations over time](#_s53h8vsk4w0n) 23

[Density of registration update to expiration](#_d7tc7p7557rq) 24

[Domain age by Alexa site ranking](#_tz17ccittiju) 25

[Number of name servers by domain trustworthiness](#_204dngzcjri5) 26

[Multivariate analysis](#_stkx8tx01pxf) 27

[Domain keyword length by domain age](#_yvwlf2b3y7xn) 27

[Domain length by domain keyword length, by trust](#_cez2d059nqgr) 28

[GoDaddy.com, fake registrations](#_67a2gcfu16xx) 29

[**Conclusion**](#_ggek0b71w8mq) **29**

[**Appendix B**](#_3as4poj) **31**

[Evaluation of the learned function](#_1pxezwc) **31**

# 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 pre-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 99.6% 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. 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 some bias, though interestingly for some domains the learned function ended up classifying some of them *back* to their original state, the state before they were editorialized.

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 was analyzed (such as the age of a domain in days). Domain names were 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. This was 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 where 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 exposed 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 learning 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 99.6%, which exceeded the expectation of >90% set in the proposal.

After careful analysis of the facts and deep study of the issue, the author encourages the reader of this paper to consider a distinct form of online disinformation which will be called *domain-based disinformation*. This project has succeeded in proving that this form of influence operation can be modeled accurately using ML.

# Review of Other Work

**Here Are the Hundreds of Sites in a Pay-to-Play Local News Network**

On the day the data collection process began, the New York Times (Alba & Nicas, 2020) released a list of almost 1,000 fake local news sites that were operating under the control of Metric Media; an organization which the author of this paper had been following in Michigan for nearly a year. These websites are examples of the uniquely American manufacturing of disinformation.

The majority share of content on these sites is, apparently, auto-generated statistics taken from public records (such as salaries of state employees). Some of the content is region specific—generally, Metric Media targets democratic governors in the regions they serve and distributes these attack pieces across all of their sites linked to that state. Some content, though, is shared across all of the network, with minor variations. In one case, articles in Wisconsin stated that 92% of state residents “disapproved of new lockdowns for the coronavirus”, while an article in Arizona, containing all of the same words as the Wisconsin article (with the state name altered, of course) reported that 89% of AZ residents disapproved of new lockdown measures. Neither cited the study the headline quoted, though the primitive pie chart on each article was adjusted.

Websites can be activated depending on political circumstances in the regions they serve. On one of the sites in this data set, the Kenosha Reporter, the New York Times said this:

Some of the sites are dormant, and we culled ones from our list that are now defunct. In the past, dormant sites have sprung to life when news hit the region they target, like what happened with the Kenosha Reporter site after protests broke out in Kenosha, Wis., over the police killing of an unarmed Black man there. (Alba & Nicas, 2020)

The primary benefit of these websites is their appearance as legitimate news sites (the owner of Metric Media, according to the article, has been trying to capitalize from the decline of local news for decades) when shared on Facebook; their continuation depends on the ignorance of site visitors who do nothing to verify the source they are visiting.

The danger, through the ability to mislead people at scale, is apparent. How are these sites allowed to function with impunity? It would seem that maybe, as news publications finalize their transition to fully online, legitimate news sites may need to identify themselves via some kind of certificate-based authentication to differentiate themselves as legitimate sources. Either way, the scale which Metric Media has attained should be alarming.

**False, Misleading, Clickbait-y, and Satirical “News” Sources.**

Melissa Zimdars, associate professor of communications at Merrimack College, compiled a list of untrustworthy websites in the 2016 document *False, Misleading, Clickbait-y, and Satirical “News” Sources*. This document was scraped and many of the domains listed were incorporated into the data set with the label “fake”. The evaluation criteria set in this document served as a starting point for the features uncovered during the data analysis step, and also provided guidance for the 32 domains that were categorized as disinformation during an audit of the data. Her first evaluation criteria for a source is on the domain name—this showed up as an important feature during analysis, and close to 25% of the learned function’s evaluation of a domain was based on its name alone.

**Viterbi Segmenter**

The Viterbi algorithm is an efficient means of segmenting strings of text into their most probabilistic component words. For this project, a segmenter developed by Will Fitzgerald (Fitzgerald, 2016) and trained on a corpus of anchor tags was used to provide the keywords of a domain. This segmenter was able to determine, with a fair degree of accuracy and uniformity to human interpretation, what were the component keywords in a domain name. These were used in the scoring process, based on a weighted average of their frequency in the sample of fake domain names, to provide quantification of the keywords used in a given domain name.

# 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 notably gotten worse as the election approaches. This project, however, does entertain a new way of identifying domain-based 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.

# 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 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.

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

## Goal 1: Collect & analyze the data

The first goal was to curate data sets finding various sources of fake news. Data sources collected include governmental research, data sets publicly posted on Kaggle.com, fact-checking organizations, and a few custom web scrapes. After extraction, the data were transformed and cleaned, then loaded into .csv files—this completed objective 1.a. After this process was complete, an exploratory data analysis took place, and an RMakdown notebook was published to explore and validate the data set. This step completed objective 1.b. Objective 1.c., which included database design and modeling, was completed and normalized to 2NF.

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

The second goal and its sole objective was the implementation of the machine learning algorithm. The deliverable was the completed algorithm. Research in the project proposal set a benchmark of >90% accuracy with its classifications. Two ML implementations classify disinformation, though with much different applications, achieved a 93% accuracy in classification. In the end the RFC algorithm for this project produced a 99.6% accurate classification rate, far exceeding expectations. A much broader data set, and some additional validation steps, would produce a more meaningful result, but if one thing is clear, this implementation could be extremely successful in classifying domain-based disinformation campaigns using only OSINT. Implementing the ML algorithm completed objective 2.a.

### Goal 3: Document & make publicly available

The final goal was aligned with the communication step of the data science process: results were published and discussed, documentation was 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 data sheets. All source code is available on Github with an accompanying README.md file. Github also hosts the cleaned data sets. 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—this is included in the README.md file on Github.

# Project Timeline

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Milestone or deliverable** | **Duration**  **(hours or days)** | **Projected start date** | **Anticipated end date** | **Actual end date** |
| Definition of requirements and data collection | 2 days | 10/19/2020 | 10/20/2020 | 10/21/2020 |
| Data tidying and analysis | 2 days | 10/22/2020 | 10/23/2020 | 10/28/2020 |
| Database design & modeling | 6 hours | 10/26/2020 | 10/26/2020 | 10/29/2020 |
| Implementation of the ML algorithm | 2 days | 10/27/2020 | 10/29/2020 | 10/30/2020 |
| Writing of documentation & uploading content | 1 day | 10/30/2020 | 10/30/2020 | 10/30/2020 |
| Post-implementation write-up | 1 day | 11/2/2020 | 11/2/2020 | 11/1/2020 |

The project ended ahead of schedule, though specific work units missed their anticipated end dates. The first milestone, on data collection, took 1 day longer than estimated. Then, the data analysis and data tidying milestone (and its RMarkdown notebook), took five days to complete (instead of two). This was drastically underestimated; it is said that data analysts spend up to 80% of their time cleaning data, and it certainly applies here. Implementation of the ML algorithm only took one day (instead of two), and the documentation and diagramming steps moved a little faster than expected.

# Unanticipated Requirements

One unanticipated requirement was a reliable internet connection. During the data collection phase, scripts were written which required repeatedly querying, and locally processing, tens of thousands of domains from a DSL connection in northern Vermont. Even left to run overnight, this would have been made faster, and with wider reach, with a better internet connection. Little could be done about this (especially due to COVID restrictions during the project environment), though it did make technology, especially network, limitations a consideration for any project in the future.

# Conclusions

The learned function form the ML algorithm was correct in its classifications 99.6% of the time—in other words, it was wrong 4 times per 1,000. This number is the average correct classification rate for both the test and validation sets (combined n = 10,120). False positives occurred more often than false negatives with both sets. Due to the nature of the problem, false negatives would ideally be minimized—for example, misidentifying legitimate health information as disinformation, and in both instances this was the case. The algorithm’s accuracy suggests a fairly accurate model for classifying domain-based disinformation, *especially* if its outputs are being validated by human supervision.

Overall, the project was successful because its classification algorithm performed >90% (99.6%) on a sample n > 10,000 (n = 25,299).

# Project Deliverables

### Data sets

Two data sets were produced: datasets/fake\_real\_domains\_combined.csv, which is the final data set that was fed into the learned algorithm, and datasets/fake\_domain\_word\_freq.csv, which was used to evaluate the domain name scoring. These data sets are rectangular tables which were used to train the learned algorithm, and are composed of domain names with their WHOIS attributes and computed statistics. The data included in the main set contains 27,261 observations of 23 variables; this set is larger than the one on which the ML algorithm was trained and tested due to issues caused by non-unicode characters which resulted in incomplete data collection.

### Exploratory data analysis

The complete EDA, with its findings, is available in Appendix A and also submitted as an HTML file with this report. This analysis served as the basis for the model that was created for the ML algorithm. Its discoveries led directly to the domain scoring model and determination of the feature vectors from which the learned function would be trained. Some of the conclusions produced in the report include the following:

* Though GoDaddy LLC is the most popular registrar on the Web, it is disproportionately exploited as a registrar, especially within the past 5 years.
* The length of a domain name and the number of keywords it uses seem to have a clear bearing on the trust score of a website, though part of this is likely to be influenced by a lesser availability of short name domains. In general, domains labeled with a trust rating of “fake” attempt to qualify themselves with more information about their purpose in the domain name (e.g., eastarizonanews.com, centraliowatimes.com, constitutionstatenews.com).
* The United States seems to be the most responsible for domain-affiliated disinformation campaigns, especially those involving fake local news organizations. Russian APTs are not being used to drive this content, most of which seems decidedly American. Though Russia serves as the most cited example of pushing disinformation, their tactics are better served on social media, which are arguably much more successful influential operations.
* Keyword analysis may be a driving factor in the categorization of disinformation websites—this will be improved as more data are brought in.
* For records management, established sites do consistently better at updating and maintaining their registrations. Furthermore, and for untrustworthy domains, there is more of a tendency for expiration to cluster around a mean–this is due to coordination of registrations, such as registering a bunch of fake local news sites all around the same time frame. This was not used in the current feature, but is likely to be considered in the future.
* Legitimate sites use appropriate amounts of name servers for their needs. Non-legitimate sites use (more or less) the same number across the board (provided by the host).

### Binary Classification Algorithm

The source code for the classification algorithm is found in the file classifier.py, though for best results use classifier.ipynb. Appendix B contains the confusion matrices for both the test set and the validation set, along with their computed recall, precision, and F1 scores. For full display of results, including which sites were mis-classified by the algorithm, please consult the Jupyter notebook.

The feature importance as reported from by the learned function is also listed in Appendix B. The heavy weighting on domain ranking is of slight concern and likely a result of the limitations of the data set. Also interesting was algorithm’s rejection of one of the features supplied to it, one which was understandably weak to begin with—the registrar ranking. This should hold some importance, and has the opportunity for better modeling in the future.

Also interesting with the learned algorithm was that it kicked back trust ratings that were corrected during the data cleansing process—web sites that had shown reasonable suspicion to be part of organized influence operations that the algorithm’s author editorialized and corrected the “trust” value to fake. Some of these were returned as false positives by the function (e.g., pjmedia.com). This gets into the murky nature of the problem at hand: to what extent can disinformation be defined and to what extent can a website be classified as such? These are real concerns that all ML implementations which deal with real-world problems should consider.

Furthermore, and beyond the question of modeling these data to ML applications, is the question of bias. Personal biases played a role in the categorization of 32 domains as disinformation—the author of the algorithm did this, mostly depending on a heuristic interpretation of the editorial content, specifically focusing on the narratives driven. During the authoring of this paper, the Hunter Biden email “bombshell” broke. To anyone studying influence operations, *this was clearly one of them—*which is distinctly different from saying it was “Russian disinformation”. Generally, if the only editorial content on the homepage of the site, during this research process was on Hunter Biden’s emails the domain was categorized with a trust rating of “fake”. This brings up the question of whether smut and other rumor mill publications can be considered disinformation—is it only when attribution occurs? (In the Hunter Biden emails, the original source was an AI-generated profile.)

The project was successful, and a confident interpretation on the quality of a domain as rated by the algorithm is within reach. This project has proven that that the problem can be modeled. With wider data collection, improved computational power, and deeper statistical validation, there is definitely potential for this to work at scale. It would still need a human validation step for some of its categorizations—something which the proposal showed is necessary for these types of problems. The accuracy of this model suggests that this validation step could scale easily.

# References

Alba, D., & Nicas, J. (2020, October 20). Here Are the Hundreds of Sites in a Pay-to-Play Local News Network. Retrieved from https://www.nytimes.com/2020/10/20/technology/timpone-network-pay-to-play-local-news.html

Melissa Zimdars. (2016). *False, Misleading, Clickbait-y, and Satirical “News” Sources*. Retrieved from: https://library.southtexascollege.edu/false-misleading-clickbait-y-and-satirical-news-sources/

Fitzgerald, W. (2016). *Segment*. <https://github.com/willf/segment>

Collins, B. & Zadrozny, B. (2020). How a fake persona laid the groundwork for a Hunter Biden conspiracy deluge. Retrieved November 3, 2020, from https://www.nbcnews.com/tech/security/how-fake-persona-laid-groundwork-hunter-biden-conspiracy-deluge-n1245387

# Appendix A: *Analysis of domain data for suspected sources of disinformation*

**Introduction**

The following EDA explores the data collected and lays the groundwork for the modeling to follow. This is part exploratory, part auditing of the data to ensure its accuracy. We’ll try to see what initial signals may be worth looking more deeply at and provide a starting point for how the eventual learned function may operate.

# Analysis

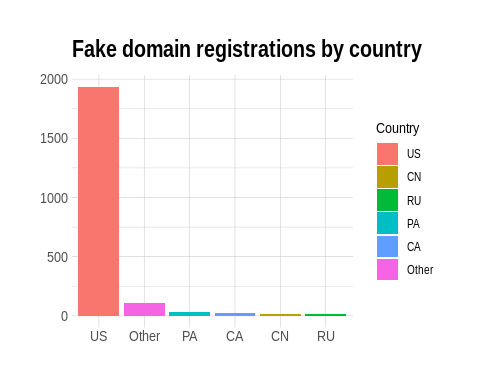
## Univariate analysis

The following plots consider single-variable investigations into the data collected. Most of the data points investigated deal exclusively with domains that have an initial trust score of “fake” as the classification of these types of domains are exactly what the learned function will need to categorize. DNSSec was also analyzed, mostly for identification of false positives, and while a write up is not included, the plot can still be seen.

## # A tibble: 2 x 3  
## `Trust rating` Count Proportion  
## <fct> <int> <dbl>  
## 1 initial trust 24870 0.912   
## 2 fake 2391 0.0877

Out of the sample of domains that were collected, 24870 were classified “initial trust”, while 2391 were classified as “fake” (or 8.77% of the sample). Limitations on collection have more to do with internet connectivity in northeastern Vermont than limitations set by the project. Ideally, more data will be added as the project continues

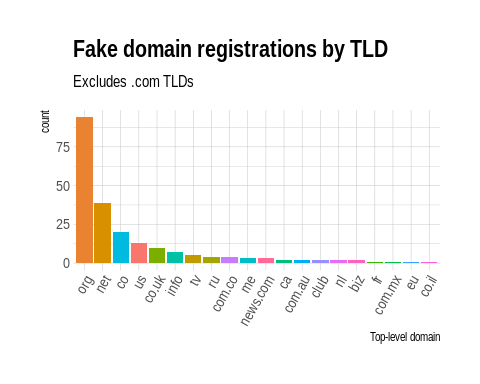
### Countries



Overwhelmingly, registration records point to the United States for being the country with the most registrations of disinformation. There are a few factors to consider with this in mind:

* A large portion of the fake domains are shell local news websites which focus on US metropolitan markets. The organization responsible for these shell publications–Metric Media–is based in the United States.
* It is very plausible (and easy) for those registering the sites to say that they are a US-based operation, and it does not seem like the United States (or its registrars) really seem to care, or do anything about influence operations.
* The United States is a frequent target for influence operations, partly because country leadership does not care or does not recognize the issue, and partly because influencing Americans is both very beneficial to stakeholder nations and very easy for them to do.
* Some countries, such as Russia, prefer targeting social media, especially influential users, to drive their narratives.

### TLDs



Top-level domains were overwhelmingly by .com domains, making up 34.42% of the sample. This is not surprising considering most domain names are registered in the US with .com TLDs. Worth considering, however, are the more deceptive styled domains which end com.co and news.com–these may be helpful in performing domain name analyses.

### Registrars

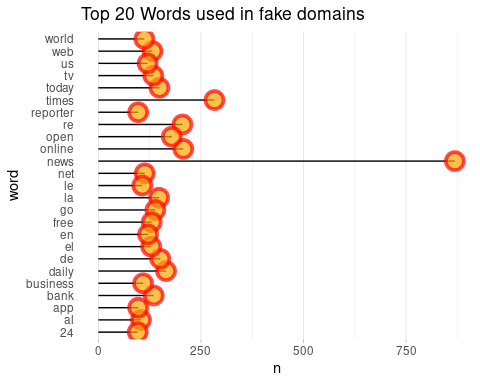
## # A tibble: 20 x 3  
## Registrar Count Proportion  
## <fct> <int> <dbl>  
## 1 godaddycom llc 1739 0.727   
## 2 namecheap inc 79 0.0330   
## 3 enom inc 38 0.0159   
## 4 network solutions llc 35 0.0146   
## 5 tucows domains inc 31 0.0130   
## 6 dynadot llc 20 0.00836  
## 7 fastdomain inc 17 0.00711  
## 8 wild west domains llc 17 0.00711  
## 9 11 ionos se 16 0.00669  
## 10 pdr ltd dba publicdomainregistrycom 16 0.00669  
## 11 namesilo llc 14 0.00586  
## 12 turncommerce inc dba namebrightcom 14 0.00586  
## 13 csc corporate domains inc 13 0.00544  
## 14 cloudflare inc 12 0.00502  
## 15 launchpad inc hostgator 11 0.00460  
## 16 google llc 10 0.00418  
## 17 registercom inc 9 0.00376  
## 18 domaincom llc 8 0.00335  
## 19 enom llc 8 0.00335  
## 20 markmonitor inc 8 0.00335

Most legitimate domains seem to use a widespread number of registrars for their needs. Domains pushing influence operations, however, overwhelmingly use GoDaddy.com. The difference is pretty telling: for domains labeled “initial trust”, 17.31% were registered with GoDaddy; for domains labeled “fake”, 72.73% were registered with GoDaddy. There may be a correlation with the popularity of a domain provider and its trustworthiness.

### Google Analytics[[1]](#footnote-2)

A cursory glance of the domain clusters reveals groupings which more of less exist as topical containers for their targeted state of industry. Many of these, it appears, come from the data set of fake local news companies. Apparently when this was setup by metric media, it seems like they had a goal of up to 50 shell sites for each of the fifty states. All of these are tied to the same organization, with a unit ID for the locations or industries targeted in the operation. This analysis identified 39 clusters of influence operations accounting for 48.98% of all fake domains captured.

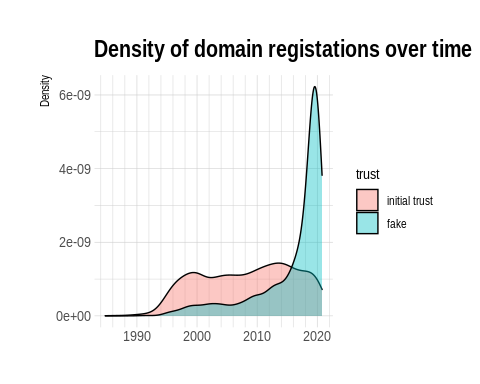
### Words used in domain name



Perhaps the most interesting (and revealing) plot is the tokenization of words which make up the domain. The science behind this algorithm is not perfect and was a general purpose adaptation for this analysis. We will be testing some hypotheses about the words used in fake domains–the belief is that properly handling these data manipulations sheds revealing information on the nature of the operation. It is not hard to consider this one of the most revealing signs of an influence operation. The data frame proves this: these sites qualify themselves as “news” c('news', 'times', 'daily', 'business', ...), use very simple English words, and appeal to geography familiar to Americans c('china', 'american', ...) in a way that might make the site appear more professional or local on social media.

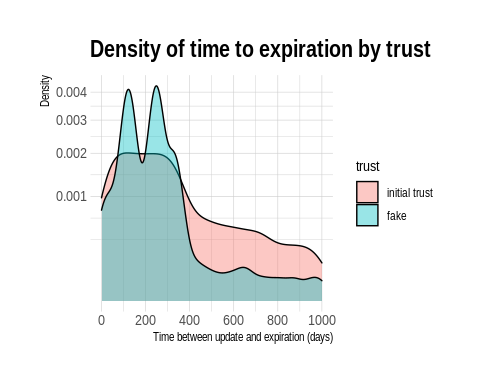
## Bivariate analysis

### Density of domain registrations over time



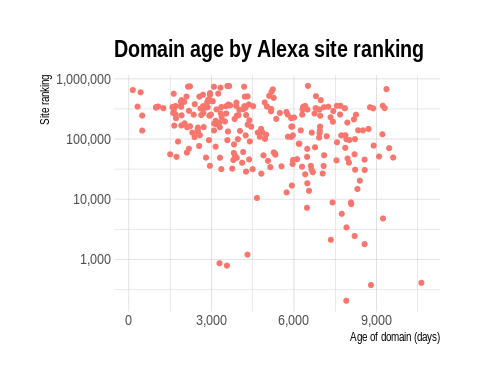
Where trustworthy domain registrations have remained more or less evenly distributed across time, registrations for domains labeled fake have very clearly spiked over the course of the past four years, with increasing acceleration starting after 2010 and trending sharply upward starting around 2014. This goes in line with much of the research conducted on the issue, so there’s even less potential to have concern about the data included within the sample. The downward trend stating around the turn of 2020 could be explained as 1.) Influence operations preparing for the 2020 elections, and/or 2.) lack of detection because as new domains are activated.

### Density of registration update to expiration



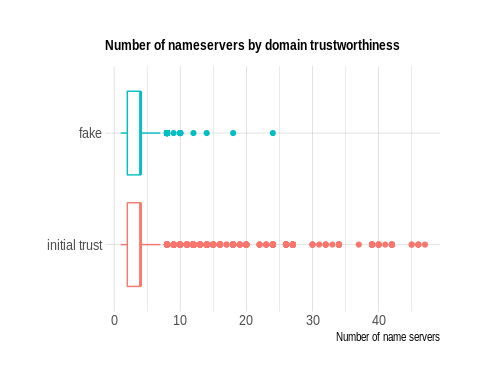
The second plot measures the density of times between an update to a site’s registration and its expiration. This time frame is influenced predominantly by the registrar; trustworthy sites will more likely have long-term management of this while fake sites registered with, say, GoDaddy.com, will typically be using a 1 year registration period. If that's so, we could expect to align the peak of registrations with the density peaks in this plot. A cursory glance at this plot seems to confirm some of the timelines.

### Domain age by Alexa site ranking



This plot seems to have a fair amount of statistical noise and, purely from an auditing perspective, confirms information known already: many sites labeled “fake” do not have domain ranking through Alexa (unless, of course, they become enormously popular), and those that do seem to have some amount of staying power. The domains in this sample are mostly more than 5 years old, but it is also known that the bulk of the domains in the sample were registered within that time.

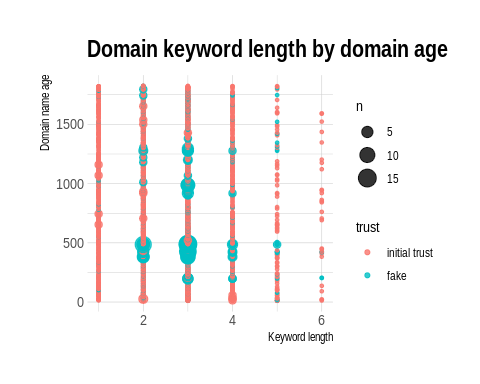
### Number of name servers by domain trustworthiness



Number of name servers could have some indication on the reliability of a site. Again, a lot of this comes down to the registrar used by the domain, but it doesn’t seem unreasonable to associate sites with >= 5 name servers to be functioning in a legitimate capacity.

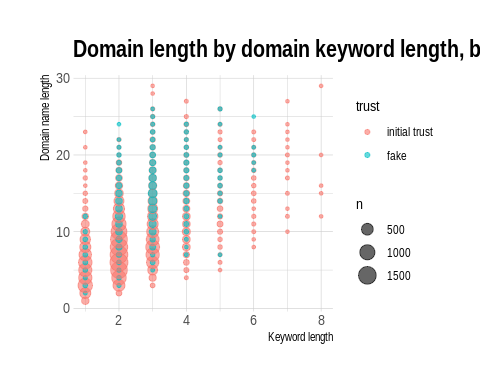
## Multivariate analysis

### Domain keyword length by domain age



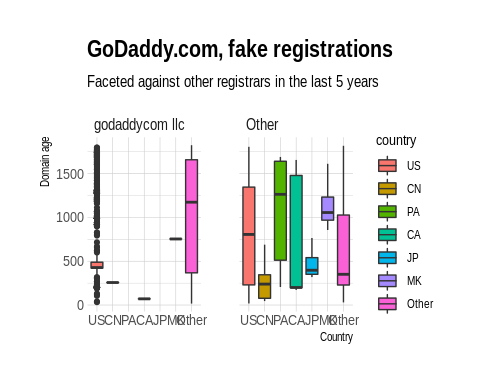
Of domains registered in the past 5 years, domain registration by veracity seems to cluster around certain periods, while keyword length seems to have some bearing on veracity. This seems obvious as a disinformation campaign domain name might attempt to validate itself by explaining more (in the domain name itself) what its intentions are. Also worth noting, as far as a correlation of date of registration, that a single influence operation (by Metric Media) accounts for a large portion of fake domains. These coordinated campaigns are signaled by the clustering of their registrations

### Domain length by domain keyword length, by trust



Looking at domain length by keyword length shows a more clear relationship between the length of a domain, its keywords, and its trustworthiness. There appears to be a visible pattern in the chart where fake domains have congregated with 2 <= Keyword length <= 5 and 10 <= Domain name length <= 25.

### GoDaddy.com, fake registrations



GoDaddy is emerging as a perfectly complacent registrar for domain fake domain registrations, disproportionate to its market share. Registrations in this chart cluster perfectly with the densities of fake registrations from above, and it might not be a reach to say that the successes in these campaigns may replicate in Canada after the US. Interestingly, when filtering for fake domains registered in the past five years, Russia fell off the map of countries represented. However, the author stresses that Russia’s strategies, and their historic successes, have been in targeted agents of influence operations, for which social media is much more effective and far more exploitable than domain names.

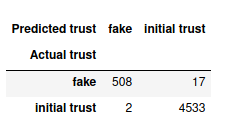
# Conclusion

There is more to be revealed as the data enter the modeling stages. However, there do seem to be a few initial signals in the data worth noting

* Though GoDaddy is the most popular registrar on the Web, it is disproportionately exploited for the use of propagating misinformation, especially within the past 5 years
* The length of a domain name and the number of keywords it uses seem to have a clear bearing on the trust score of a website
* The United States seems to be the most responsible for domain-affiliated disinformation campaigns, especially those involving fake local news organizations. Russian APTs are not being used to drive this content, most of which seems decidedly American
* Keyword analysis may be a driving factor in the categorization of disinformation websites
* For records management, established sites do consistently better at updating and maintaining their registrations. Furthermore, and for untrustworthy domains, there is more of a tendency for expiration to cluster around a mean–this is due to coordination of registrations, such as registering a bunch of fake local news sites all around the same time frame
* Legitimate sites use appropriate amounts of name servers for their needs. Non-legitimate sites use (more or less) the same number across the board (provided by the registrar).

# Appendix B: *Evaluation of the learned function*

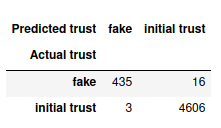
**Test**

True positive rate (Recall) = 0.99626

Positive predictive value (Precision) = 0.99956

F1 score = 0.9979

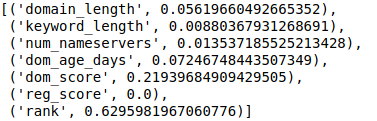
**Validate**

True positive rate (Recall) = 0.99654

Positive predictive value (Precision) = 0.99935

F1 score = 0.99794

**Feature importance**



1. N.B. – The data frame rendered in RMarkdown does not render in Word; please see the HTML rendering to explore the resulting table. [↑](#footnote-ref-2)