Proactive Discovery of Fake News Domains from Real-Time Social Media Feeds

Zhouhan Chen
New York University
zhouhan.chen@nyu.edu

Juliana Freire
New York University
juliana.freire@nyu.edu

ABSTRACT

The proliferation of web sites that disseminate fake news is a growing problem in our society. Not surprisingly, the problem of identifying whether a web page contains fake news has attracted substantial attention. However, the problem of discovering new sources of fake news has been largely unexplored. Timely discovery of such sources is critical to combat misinformation and minimize its potential harm. In this paper, we present an automatic discovery system that proactively surfaces fake news domains before they are flagged by humans. Our system operates in two-steps: first, it uses Twitter feeds to uncover user co-sharing structures to discover political websites; then it uses a topic-agnostic classifier to score and rank newly discovered domains. To demonstrate the effectiveness of our system, we conduct an experimental evaluation in which we collect tweets related to the 2020 presidential impeachment process in the United States, and show that not only our system is able to discover new sites, but that a large percentage of these sites are indeed publishing fake news. We also design an integrated user interface to support fact-checkers and leverage their knowledge. Through this interface, fact-checkers can visualize domain interaction networks, query domain fakeness score, and tag incorrectly predicted results. Our proactive discovery system will expedite fact-checking process and can be a powerful weapon in the toolbox to combat misinformation.

CCS CONCEPTS

• **Computing methodologies** → **Cluster analysis**; • **Information systems** → *Social networks*; **Web searching and information discovery**.

KEYWORDS

misinformation, fake news discovery, social network analysis

ACM Reference Format:
Zhouhan Chen and Juliana Freire. 2020. Proactive Discovery of Fake News Domains from Real-Time Social Media Feeds. In *Companion Proceedings of the Web Conference 2020 (WWW '20 Companion), April 20–24, 2020, Taipei, Taiwan.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3366424.3385772

WWW '20 Companion, April 20–24, 2020, Taipei, Taiwan

ACM ISBN 978-1-4503-7024-0/20/04.
https://doi.org/10.1145/3366424.3385772

1 INTRODUCTION

The rise of fake news and the use of misinformation campaigns is an increasing threat to our society. Much research has been devoted to improve our understanding of this problem, from how to detect whether an article contains fake information [4, 17, 18] to exploring how fake news are propagated [8, 24]. However, the problem of discovering new sources of misinformation has been largely unexplored. Timely discovery of such sources is crucial for combating fake news and minimizing their effects before they become too widespread on the Internet.

In this paper, we propose and implement a proactive fake news domain discovery system that leverages unlabeled but structured real-time social media data. The unit of detection of our discovery system is domain. We adopt the definition of fake news domains as used in [22]: a fake news domain is a web site "that entirely fabricates information, disseminates deceptive content, or grossly distorts actual news reports." Our system discovers new suspicious domains from one of the most active online platforms: Twitter. The intuition behind our approach is that domains that cover similar topics will be tweeted/retweeted by similar users. We leverage Twitter feeds to create a domain interaction graph based on user co-sharing similarities [21]. To do so, we first map each domain to a set of Twitter users that tweet about the domain. We then construct an unweighted and undirected graph where each node is a domain, and two nodes are connected if the jaccard similarity of their corresponding user sets is above a threshold. Given this graph, we extract the largest connected component. We show that with a proper similarity threshold, the largest connected component contains more than 95% of domains from the input collection and thus is sufficient for further analysis.

The domains discovered are potential sources of fake news. User input is needed to further classify the sites. Because human resource is limited, in order to help fact-checkers explore this information, we need a robust detector to score and prioritize the unlabeled domains. Detecting domain *fakeness* is an active research area [4, 17]. For our system, we adopt a topic-agnostic fake news classifier proposed by [2]. The classifier captures the style of fake news sites, rather than the topic, as predicting future news topic is very difficult.

We evaluate the ability of our system to discover fake news by collecting real-time tweets related to the 2020 process to "impeach Donald Trump". Because our tweet collection is fresh and unlabeled, we design a novel framework to evaluate our system performance. We introduce two parameters: one controls the domain similarity for the unsupervised clustering component, and the other controls the decision boundary for the supervised component. By tuning those two parameters we can achieve different levels of precision and recall. We discuss guidelines for how to choose the best model configuration with systematic parameter tuning.

There are two indispensable elements during the spread of a fake news domain: the domain itself, and the social media accounts that tweet about the domain. Using our tweet corpus and supervised classifier, we then proceed to answer the question: what are characteristics of Twitter accounts that share more fake news? We define a *fakeness score* for each Twitter account, and separate accounts into several buckets based on their scores. We find that collectively, accounts with high fakeness score are more likely to use pro-Trump phrases in their account descriptions. Our findings corroborate the results reported by Bovet and Makse [1], who analyzed individual fake account descriptions.

Finally, we describe a Web interface we designed for our system to support fact-checkers to both explore and actively label the discovered domains. The interface allows fact-checkers to both visualize domain network and submit labels for unchecked domains.

The rest of the paper is organized as follows. Section 2 discusses recent research in fake news. Section 3 details the components of our system. Section 4 shows a real-world use case of our system and presents a characterization of Twitter accounts that share fake news. We discuss the limitations of our work in Section 5. We conclude the paper in Section 6.

2 RELATED WORK

To the best of our knowledge, ours is the first approach that attempts to address the problem of timely discovery of fake news sources. Our system combines unsupervised discovery, supervised detection, and visualization into a unified system.

Supervised detection. The majority of fake news detection work relies on supervised methods to label or score news sources. Existing detection methods focus on different modalities, such as text [11, 14], image [12], or multi-modalities [7]. The granularity of detection also varies from sentence-level claim [11, 14] to page-level article [5] to a single domain [2]. Extensive summaries of features, machine learning models and datasets used by different fake news detectors are provided in [4, 17, 18]. Detection is a key component of a discovery system: once a suspicious site is found, we need to ascertain whether it is a potential propagator of fake news. The unit of detection of our discovery system is domain and we use the detector proposed in [2].

Unsupervised detection. Unsupervised detection refers to the process of identifying potential fake news sources from unlabeled or partially labeled data. Guo et al. [10] states that early discovery of fake news is very crucial and remains a challenge. Several unsupervised or semi-supervised methods have been proposed to tackle this challenge. Qian et al. [13] generates synthetic user engagement to improve fake news detection. Yang et al. [27] utilizes a probabilistic graphical model to estimate trustworthiness of news. [9] distinguishes different categories of false news using tensor decomposition on the content. [19] uses weak labeling functions to expand training set, and [25] leverages users' reports as weak supervision to enlarge the amount of training data.

Fake news visualization An intuitive visualization and human-computer interaction system enhances the practicality of underlying fake news detection or discovery algorithms. For example, [14, 16, 26] build interactive applications to visualize the spread of fake news. Miranda et al. [11] takes a claim as input, and outputs predictions (true or false) and supporting evidence. Our user interface supports both a network view to visualize domain interactions, and a tabular view for fact-checkers to sort and label discovered domains.

3 SYSTEM ARCHITECTURE

Our system consists of a front end user interface, and a back end stack of execution units. To bootstrap our discovery system, a user simply submits a list of keywords. The choice of keywords can be politically related, such as "impeachment", "government", or "election." After receiving keywords, our system triggers a Twitter data collection, web page resolving, domain clustering and prediction pipeline. Figure 1 illustrates our discovery pipeline. In this section we explain each back end component in detail. Section 4 covers the design of front end interface.

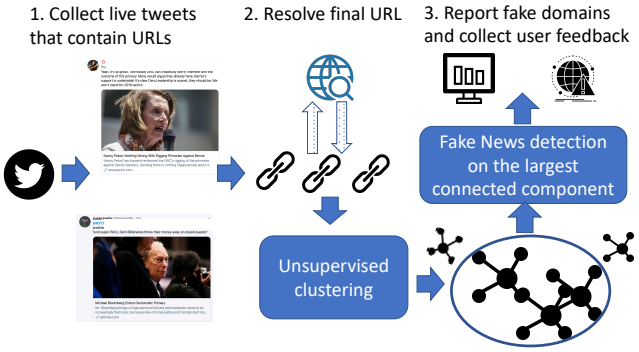

Figure 1: Fake news domain discovery pipeline.

3.1 Tweet collector

We employ a two-step data collection process to uncover domains related to a certain topic. We first use Twitter Streaming API to collect live tweets based on a list of user-specified keywords. The collection process stops when a certain number of tweets is collected or a certain number of minutes have passed, based on custom configuration.

We then extract all Twitter users who has generated at least two tweets with external URLs, i.e., URLs whose domain is not twitter.com. We then use Twitter REST API to collect the past 200 tweets of each user. We keep tweets that contain external URLs. We now have the raw dataset on which subsequent steps depend. The second collection process greatly expands the URL coverage and captures domains that are relevant but not covered by input keywords.

3.2 Web page resolver

The goal of this step is to extract embedded URLs from tweets, resolve final landing URLs, and collect features that will later be used by our supervised machine learning model. We use a headless Chrome browser to visit each URL. We trace the entire URL redirection chain, and store the HTML file of the final landing URL. Even though browser simulation is a CPU-heavy operation, it is more reliable then simple scripting, due to the wide use of shortened URLs and unpredictable redirection behaviors.

3.3 Unsupervised domain discoverer

The domain discoverer identifies unknown domains by leveraging abundant but unlabelled social network structure. Shu and Liu [17] point out that "users on social media tend to form groups containing like-minded people where they then polarize their opinions, resulting in an echo chamber effect. The echo chamber effect facilitates the process by which people consume and believe fake news." This observation drives us to cluster domains based on user co-sharing similarity. We use Jaccard similarity[1] as our similarity measure. Previous work [15, 20] demonstrate that user co-sharing networks reveal the media ecosystem that surround political conversations.

The unit of our discovery algorithm is a domain. To construct a domain interaction network, we first map each domain d_i to a set of Twitter user ids $\Gamma(d_i)$ whose tweets contain URLs to d_i. We construct an undirected graph $G < V_D, E >$, where V_D are domain nodes. For $d_1, d_2 \in V_D$, the weight of edge between d_1 and d_2 is defined by a step function:

$$E_{d_1,d_2} = 0 \text{ if } similarity(d_1, d_2) < \alpha$$
$$= 1 \text{ otherwise}$$
$$similarity(d_1, d_2) = \frac{|\Gamma(d_1) \cap \Gamma(d_2)|}{|\Gamma(d_1) \cup \Gamma(d_2)|}$$
$$\Gamma(d_i) := \text{a set of user ids who have tweets containing } d_i$$
$$\alpha : \text{similarity threshold}$$

In another word, an edge between two domains is removed if their co-sharing similarity is below a threshold, defined as α. A low α results in a densely connected graph with more irrelevant domains, which in turn increases the overall recall but lowers the precision. A high α has the opposite effect. In Section 4 we show how to systematically choose the optimal α.

After the graph construction, we run connected component algorithm to extract all clusters in the network. In our real world experiment, the largest cluster is the only interesting one that contains more than 95% of domains from the input collection.

3.4 Supervised detector

The final step is to score, rank and report domains just discovered. We adopt and improve a topic-agnostic fake news classifier (TAG) developed by [2]. TAG takes a web page as input and outputs a numerical value to indicate the *fakeness* of that page. We summarize our reasons for being topic agnostic, features used by the classifier, our improvements, training data and test accuracy.

3.4.1 Reasons for choosing TAG. As its name suggests, TAG does not rely on the topic discussed in a web page, but focuses on the writing style and page layout style. Future fake news topics are highly unpredictable and are likely to differ from topics in the training set. We argue that it takes time and money to create professional websites and write in a professional way, both are big disincentives for miscreants running on a budget. Therefore, while the news topic may change day by day, the layout and writing style of a website do not change as frequent.

[1]Jaccard similarity of sets A and B is $\frac{|A \cap B|}{|A \cup B|}$.

Table 1: Learned top features associated with fake news and real news.

Category	Feature Name	Feature Type	Explanation
real news	svg	web markup	scalable vector graphics
	noscript	web markup	defines an alternate content for users that have disabled scripts in their browser
	coleman-liau-index	readability	the higher the index, the more complex an article is
fake news	ins	web markup	underscore
	br	web markup	blank space
	i	web markup	italic text

3.4.2 Required features. Our classifier uses three categories of features: web markup, readability and morphological. Together they capture the aesthetic of a web page as well as the writing style and language usage of article writers. The full list of features can be found in the original paper [2]. Table 1 shows a subset of features positively linked to fake news and real news. Top features associated with real news include advanced HTML tags and higher readability score (for example, pages from *The New York Times* usually have a high readability score because the sentences are longer and more sophisticated). Top features associated with fake news include visual enhancers such as italic fonts and underscore.

3.4.3 Improvements. To enhance the accuracy and make the classifier compatible with our discovery pipeline, we modify the source code of [2] and make following improvements:

(1) We add Quantile Transformer[2] to transform each feature to a normal distribution. Quantile Transformer is a robust prepossessing schema that reduces the impact of outliers.
(2) We identify anomalies in training data. Specifically, we discard web pages whose total number of words is less than 200 or more than 2000. The former are pages with 404 errors and the latter are directory pages that are not relevant to a single piece of news.
(3) We remove psychological features, which are used in the original paper to capture words' semantic patterns (anger, fear, happy, etc,.). This group of features require manual processing, which does not fit into our automated prediction pipeline.

3.4.4 Dataset and accuracy. We use the same **PoliticalFakeNews** training set introduced in [2]. The training set consists of 7,136 pages from 79 fake sites, and 7,104 pages from 58 real sites. The fake sites come from an aggregated list of Politifact, Buzzfeed 2016, 2017 sets and Opensources.co. The real sites come from a subset of Alexa's top 500 news sites. We train a Support Vector Machine (SVM) with linear kernels. Our model achieves an average accuracy of 89% over a five-fold cross validation, 7% higher than the previous best model accuracy reported in [2]. Figure 2 shows the Receiver

[2]https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.QuantileTransformer.html

Operating Characteristic (ROC) curve and Area Under the Curve (AUC) value.

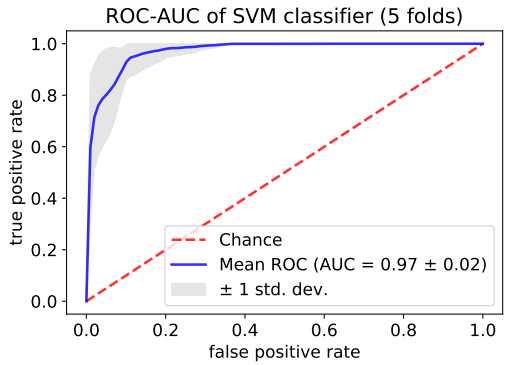

Figure 2: ROC-AUC plot of improved topic-agnostic fake news classifier.

3.4.5 Page-level score to domain level-score. Our TAG classifier takes a web page as input. In order to assign a fakeness score to a domain with multiple pages, we create a custom headless Chrome crawler to first visit the domain home page, parse its HTML content, and randomly sample five hyperlinks with the same domain. The fakeness score of a domain is the average scores from those five pages. Finally, to make a decision of *fake*, *real*, or *unknown*, we introduce a parameter β. Given a domain d, a pre-trained TAG classifier C that returns a numerical value, and β, our decision rule is:

(1) if $C(d) > \beta$, d is fake.
(2) if $C(d) < -\beta$, d is real.
(3) Otherwise, d is unknown.

Similar to the first parameter α, β also controls the relative importance between precision and recall. A high β increases the precision but lowers the recall, and a low β does the opposite. By default, and during training and testing, $\beta = 0$. In Section 4 we show how to tune β for our real world application.

4 SYSTEM DEMONSTRATION – A REAL WORLD STUDY TO DISCOVER FAKE NEWS DOMAINS ON TWITTER

In this section, we demonstrate a real world discovery experiment to prove the ability of our system to discover fake news domains. We show how we evaluate and select the best model configuration. We visualize interesting patterns behind the network of fake news domains. We end this section with a characterization of Twitter accounts who tweet more fake domains.

4.1 Real-time data collection

Our discovery pipeline starts with keywords. We are interested in keywords that are political, unique, and likely to appear in news headlines. The impeachment inquiry of Donald Trump is a tensely debated political event in the United States. The relevancy and newsworthiness of this topic make it possible to discover domains that are never present in our training data. We trigger our discovery pipeline by providing keywords *impeach*, *impeached* and *impeachment*.

In the 24 hours beginning October 29, 2019, we collect 220,909 tweets from the Twitter Streaming API. We use 24 hours to capture all conversations happened in a day. From this initial collection, we extract 39,230 distinct Twitter account, and collect past 200 tweets from each account. This expands our collection to 2,284,544 tweets.

We then extract 4,042 unique domains from our expanded collection, and calculate pair-wise user co-sharing similarity, as described in Section 3. This completes our data collection process.

4.2 Model evaluation

We now introduce evaluation metrics, parameters, trade-offs we consider, and configurations of our final discovery model.

4.2.1 Evaluation metrics. Evaluating a fresh dataset is challenging because there is no complete ground truth. Our goal is to use limited ground truth to approximate global ground truth. We address this problem by leveraging labels from another domain-based, actively maintained fact-checking dataset. We choose to adopt labels provided by MediaBiasFactCheck [3](MBFC), an independent online fact-checking outlet. MBFC publishes and updates all labeled domains on Github [4], enabling us to check factness and bias of hundreds of domains programmatically. Other fact-checking services such as PolitiFact[5] and Snopes[6] mostly focus on claims and statements made by officials, columnists, and political analysts [17], where MBFC focuses on domains. As of February 5, 2020, MBFC has 2,793 unique domains that are human-labeled.

For each domain, MBFC provides seven labels, from *VERY HIGH* to *VERY LOW*. To map seven labels to just *real* and *fake*, we define fake domains as those with labels *LOW*, *VERY LOW*, *MIXED*, and real domains as those with labels *VERY HIGH*, *HIGH*, and *MOSTLY FACTUAL*. We include *MIXED* in the fake category because MBFC assigns labels conservatively and 19% of fake news domains in our training data have *MIXED* labels.

To select the optimal discovery configuration, we consider two parameters introduced in Section 3: α and β. α is the similarity threshold used during domain network construction, and β is the decision threshold for our topic-agnostic classifier. To evaluate a configuration, we consider three metrics, which we call partial precision p, partial recall r, and partial $f1$ score, defined as:

$$p = \frac{\#\text{ domains predicted fake } \textbf{and} \text{ labeled by MBFC as fake}}{\#\text{ domains predicted fake by our model}}$$

$$r = \frac{\#\text{ domains labeled by MBFC as fake } \textbf{and} \text{ discovered by our model}}{\#\text{ domains labeled by MBFC as fake}}$$

$$f1 = 2 \times \frac{(p \times r)}{(p + r)}$$

4.2.2 Grid search. We use grid search to find the best configuration of α and β, with $\alpha \in [0.4, 0.6, 0.8]$ and $\beta \in [0, 0.5, 1.0]$. We also consider three "no-network" cases ($\beta \in [0, 0.5, 1.0]$) where the

[3]https://mediabiasfactcheck.com
[4]https://raw.githubusercontent.com/drmikecrowe/mbfcext/master/docs/revised/csources.json
[5]politifact.com
[6]snopes.com

model does not leverage the network structure to filter out domains, but predicts on all 4,042 domains in the collection. The results are listed in Table 2.

We choose $\alpha = 0.4$ as the lower bound because when $\alpha \leq 0.3$, the domain interaction network is too weakly connected. The resulting partial precision and recall are not different from those obtained from "no-network" cases, which means that the model does not extract extra information from the domain interaction network. We choose $\alpha = 0.8$ to be the upper bound because when $\alpha \geq 0.9$, the domain interaction network is broken. The largest connected component only contains 47.6% of all domains, which does not capture a complete conversational structure.

We choose $\beta \in [0, 0.5, 1.0]$ by inspecting the fakeness score distribution of domains from our training set, shown in Figure 3. The greater the β, the more conservative the decision boundary is: 0 is liberal, 0.5 is moderately conservative, and 1 is very conservative.

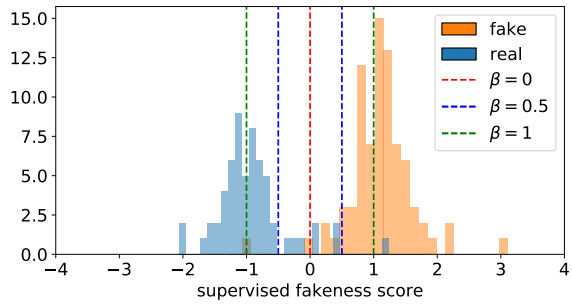

Figure 3: Fakeness score distribution for domains in the training set. We experiment with three decision boundary thresholds (β): 0, 0.5 and 1. A larger β corresponds to a higher precision and a lower recall.

4.2.3 Optimal model configuration. The criteria of an optimal model depends on the goal of end users. Table 2 shows that using network pattern increases partial precision, degrades partial recall, but improves partial $f1$ score.

For our use case, we want to achieve a balance between precision and recall, because we are often constrained by limited human resource to fact-check discovered domains. Therefore we choose $\alpha = 0.8, \beta = 0.5$, as this configuration yields the highest partial $f1$ score.

4.3 Discovered domain interaction network

After we decide the optimal system configuration, we first construct the domain interaction network using $\alpha = 0.8$. Figure 4 visualizes the network structure of the largest connected component. This component contains **2,238** domains, which accounts for 95.3% of domains from all connected components. The rest of connected components has less than 10 domains each. Therefore we are confident that the largest connected component is representative of the domain interaction network and we only focus on the largest one in further analysis.

We can identify two distinct clusters of domains related to the topic "impeachment." Domains from the top left cluster are biased toward liberal causes. Two high degree nodes in this cluster are *washingtonpost.com* and *cnn.com*. Domains from the bottom right cluster are biased towards conservative causes. Two high degree nodes in this cluster are *breitbart.com* and *thegatewaypundit.com*. A few domains, notably *wsj.com*, stays in the middle, which suggests that people from both parties tend to share information from those sources.

Table 2: Evaluation of different model configurations. *nn* means "no-network", i.e., we only use supervised score classifier over all domains in our collection. We achieve the best partial $f1$ score when we leverage network information to filter out irrelevant domains, and set $\alpha = 0.8$, $\beta = 0.5$.

Metric	partial precision			partial recall			partial f1		
α&β	0.4	0.6	0.8	0.4	0.6	0.8	0.4	0.6	0.8
0	0.12	0.13	0.17	0.72	0.70	0.53	0.21	0.22	0.26
0.5	0.16	0.17	0.24	0.48	0.47	0.37	0.24	0.25	**0.29**
1.0	0.20	0.21	0.29	0.24	0.24	0.19	0.22	0.22	0.23
nn 0		0.12			0.73			0.21	
nn 0.5		0.16			0.48			0.24	
nn 1		0.21			0.24			0.22	

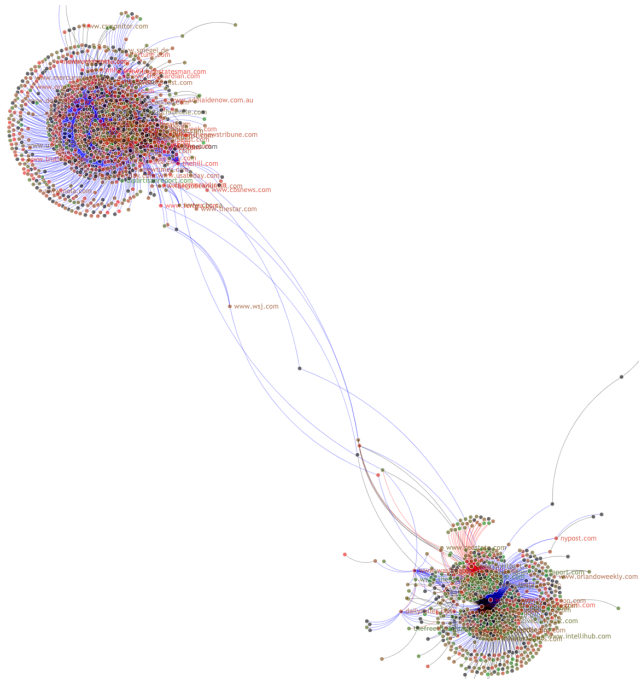

Figure 4: Fake News Discoverer Interface (network view). We visualize the domain interaction network using d3js. We can identify two clusters of domains related to the topic "impeachment." Domains from the top left cluster are biased toward liberal causes. Domains from the bottom right cluster are biased towards conservative causes.

4.4 Performance of best model: intersection accuracy and top-k accuracy

We now evaluate our supervised detector component. Using $\beta = 0.5$, we filter out domains whose score is between -0.5 and 0.5. We also remove domains that appear in our training set. This left us with **890** newly discovered domains, out of 2,238 domains.

To evaluate our best model, we breakdown discovered domains into two categories – those that have been fact-checked by MBFC, and those that have not been fact-checked.

4.4.1 Discovered domains, fact-checked. Among **890** discovered domains, 278 (31%) have been fact-checked. We now focus on this intersection. We use labels provided by MBFC as ground truth. Figure 5 shows the distribution of MBFC's factual ratings for domains predicated fake and domains predicated real. We summarize three major findings:

(1) Our model's predictions mostly agree with fact-checkers' labels. Among domains predicted real, 83% have a factual reporting of *HIGH*, *VERY HIGH* or *MOSTLY FACTUAL*. Only 3% have a *LOW* rating.
(2) Among domains predicted fake, 24% have a factual reporting of *HIGH* or *MOSTLY FACTUAL*, 31% have a factual reporting of *LOW* or *VERY LOW*. Overall, 76% of domains predicted fake have ratings below *Mostly Factual*.
(3) 45% of domains predicted fake are *MIXED*, versus 12% of domains predicted real. The high percentage of *MIXED* label reflects the constraint of domain-level classification, as a domain can host real pages and fake pages at the same time.

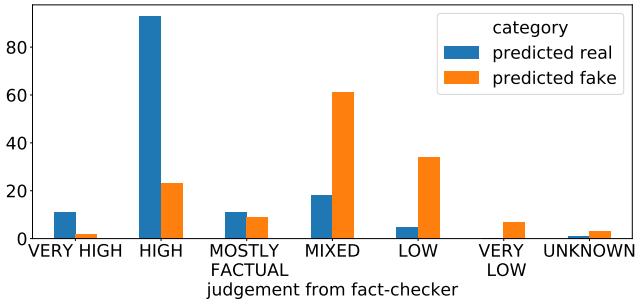

Figure 5: Factual ratings provided by MBFC. Our models' predictions mostly agree with fact-checkers' labels: only 3% of domains predicted real have *Low* ratings, while 76% of domains predicted fake have ratings below *Mostly Factual*.

4.4.2 Discovered domains, not fact-checked. For domains in this category, we measure the top-k accuracy, where $k = 20$. Specifically, we sort domains according to their fakeness scores, and manually verify the top 20 domains that have not been fact-checked. This practice is consistent with how we intend our system to be used: fact-checkers first sort domains, then label the one with the highest score first.

Table 3 shows predicted score, our manual factual rating and bias rating of top 20 domains. In summary, 30% of domains are fake, 30% of domains have a mixed factual rating with certain political bias, and 10% of domains can't be determined. Overall, 70% of domains are suspicious and require investigation. This percentage is consistent with the percentage for discovered domains that are fact-checked, where 76% of domains predicted fake have ratings below *Mostly Factual*.

To streamline the fact-checking process, we implement a user interface shown in Figure 6. When a domain is not fact-checked, a label of "unchecked" is shown. To label those domains, a user first ranks domains by fakeness score, then clicks "Report" to assign labels. This feedback loop will enable our discovery system to constantly improve itself by retraining models on a growing and more accurate dataset.

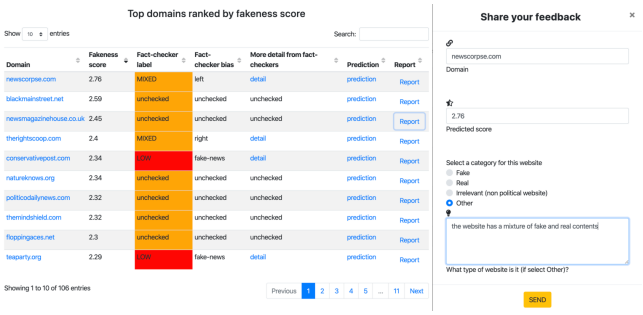

Figure 6: Fake News Discoverer User Interface (tabular view). A user can rank domains by fakeness score. If a domain exists in mediabiasfactcheck.org, a user can cross-check human-rated factual rating and bias rating. If a domain is not fact-checked, a label of "unchecked" is shown. A user can click "Report" to assign a label to the domain. A sample report window is shown on the right.

4.5 From domain to account: characterizing Twitter accounts using fakeness score

4.5.1 Why are Twitter account fakeness important? In this section, we demonstrate how to use domain-level fakeness score to infer account-level fakeness score. We define an account's *fakeness score* as the average score of domains tweeted by this account in its most recent 200 tweets. How is account fakeness score useful? First of all, if domains are the source, Twitter accounts are the carriers that spread the source. Account fakeness score quantifies the relative propensity of an account to share fake news. Second, we can identify predictive features by segregating accounts according to different score range and look for distributional difference. We believe that our account fakeness score, together with any derived features, is valuable for downstream tasks such as social bot detection, troll detection or sentiment analysis.

4.5.2 How many bots are in our "impeachment" collection? Recent research shows that bots, or accounts automated by software, are prevalent on Twitter feeds [6], especially among tweets with embedded URLs [3]. To estimate the percentage of bot accounts in our collection, we leverage Botometer [23], a machine learning model that predicts how likely an account is a bot based on more than 1,200 features. For each account, Botometer calculates a complete

Table 3: Manual factual ratings for the top 20 discovered domains that are not fact-checked by the time of discovery. For domains that are labeled "Low," a link to a misleading page is provided. For domains labeled "Mixed," a bias judgement is provided. For domains that are labeled "Irrelevant," a short description of the website is given. Domains with "Unknown" label require more investigation.

Domain	Score	Factual rating	Comment
blackmainstreet.net	2.59	Unknown	
newsmagazinehouse.co.uk	2.45	Low	right bias https://www.newsmagazinehouse.co.uk/hunters-paternity-case-spills-into-impeachment-judge-orders-biden-to-hand-over-his-burisma-financial-records/
natureknows.org	2.34	Irrelevant	about nature
politicodailynews.com	2.32	Mixed	right bias
themindshield.com	2.32	Irrelevant	personal blog
floppingaces.net	2.30	Low	right bias http://www.floppingaces.net/2020/02/02/adam-schiffs-lifetime-of-lies/
breakthematrix.com	2.20	High	
leadpatriot.com	2.19	Low	exaggeration https://leadpatriot.com/hillary-clinton-bashes-sanders-and-then-threatens-us-all/5806/
mydaughtersarmy.org	2.14	Mixed	left bias
nycpost.pro	2.09	Low	right bias https://www.nycpost.pro/judge-declares-omar-is-guilty-orders-her-to-repay-it-all-then-finds-another-skeleton-in-her-closet/
heartlanddiaryusa.com	2.04	Mixed	right bias
dmlnewsapp.com	2.03	High	
betshort.com	1.97	Low	left bias http://betshort.com/collusion/
tammybruce.com	1.91	Mixed	right bias
joemygod.com	1.90	Mixed	right bias
iotwreport.com	1.89	Mixed	right bias
churchandstate.org.uk	1.89	Irrelevant	website that covers church-state separation and free speech.
thedcpatriot.com	1.88	Unknown	
pantsonfirenews.com	1.88	Low	right bias https://pantsonfirenews.com/the-new-york-times-latest-conspiracy-theory-is-so-insane-it-will-make-your-head-hurt/
secureourvote.us	1.87	Irrelevant	website about making the voting process more transparent

automation probability (CAP). This metric uses Bayes' theorem to "take into account an estimate of the overall prevalence of bots, so as to balance false positives with false negatives [23]."

We query all 39,230 accounts via Botometer API. Figure 7 shows the distribution of complete automation probability for those accounts. The average probability of an account being a bot is 5.41%. If we use 0.5 as a bot/not-bot threshold, 1.90% accounts are bots. If we use 0.8 as a threshold, only 0.14% accounts are bots. We conclude that the majority of accounts in our collection are operated by humans, and that our account characterization reflects humans' online behaviors.

4.5.3 Discovered feature: account description. Using our "impeachment" collection, we characterize accounts with different fakeness scores and investigate what makes one account more likely to share fake news than others. We focus on active accounts that have at least three tweets with URLs among the past 200 tweets. Having more URLs reduces the variance of account fakeness score. There are 39,230 Twitter accounts in our "impeachment" collection, and 37,503 are active accounts. We first calculate fakeness score for each account. Figure 8 is the histogram of all account fakeness scores.

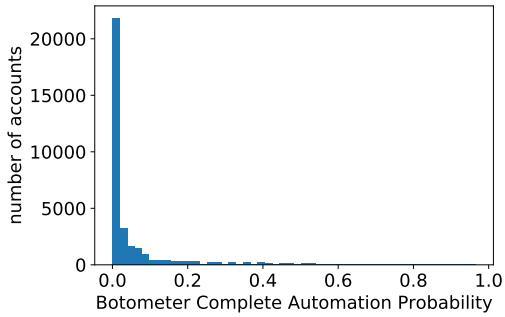

Figure 7: Botometer Complete Automation Probability (CAP) distribution for accounts in our "impeachment" collection. 1.90% accounts are above the 0.5 (50%) likelihood threshold.

Based on the shape of the distribution, we break down accounts into three categories: likely to share fake news (score > 0), might

share fake news (-1 ≤ score ≤ 0), and not likely to share fake news (score < -1).

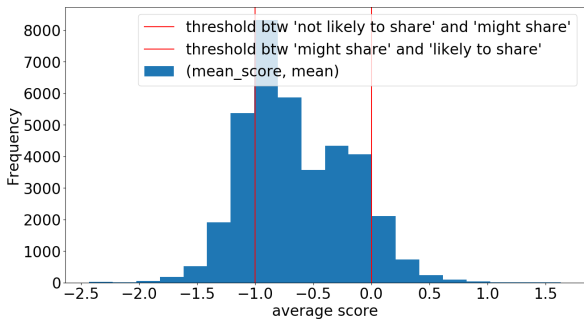

Figure 8: Histogram of fakeness score for all accounts in our "impeachment" collection. Because most scores are between -2 and 1, we segregate accounts into three categories: likely to share fake news (score > 0), might share fake news (-1 ≤ score ≤ 0), and not likely to share fake news (score < -1).

We then look at feature distribution within each category to identify discriminating features. We do not find difference in the distribution of number of tweets sent, number of friends or number of followers. One very interesting linguistic feature we find is account description. Specifically, we extract all account descriptions, remove stop words and punctuations, and generate bigrams. Table 4 shows the top 10 most commonly used bigrams in account descriptions in each category. Accounts that are likely to share fake news tend to be strong "Trump supporters," prefer to use campaign hashtags such as "#maga," "#kag," and the word "god." Accounts that are not likely to share fake news tend to label themselves as "political junkie" and "news junkie," who are probably more willing to read news from multiple sources.

Other than linguistic difference, we also notice demographic difference in each category: accounts that are more likely to share fake news are "happily married " or "husband father;" accounts that might share or are less likely to share fake news label themselves as "animal lover" and "wife mother." To be more concrete, we present screenshots of two Twitter account profiles in Figure 9. The image on the left is a user with low fakeness score. In contrast, the image on the right is a user with high fakeness score. Its profile is anonymous and its description is more provocative. We caution that those differences do not imply a causal relationship: whether certain demographics are more susceptible to fake news (i.e., they share fake news without realizing the fact that the news is fake), or are more actively involved in sharing fake news requires further investigation.

5 LIMITATIONS AND FUTURE WORK

5.1 Sample bias and selection bias

There are two major types of bias from our system – sample bias and selection bias. Sample bias come from our US-centric training dataset. Selection bias come from two parts: one is that our system focuses on Twitter exclusively, the other is that our data collection process requires input keywords which are subject to human choice.

Table 4: Top 10 most commonly used bigrams in account descriptions. Users who are likely to share fake news tend to be strong "Trump supporters," and prefer to share campaign hashtags such as "#maga." Users who are not likely to share fake news label themselves as "political junkie" and "news junkie," who are probably more willing to read news from different sources.

rank	likely to share fake news	might share fake news	not likely to share fake news
1	trump supporter	#maga #kag	animal lover
2	#maga #kag	trump supporter	political junkie
3	president trump	president trump	wife mother
4	happily married	animal lover	husband father
5	god bless	happily married	wife mom
6	god family	wife mother	new york
7	love god	follow back	mother grandmother
8	trump 2020	love god	dog lover
9	husband father	god family	news junkie
10	family country	husband father	mom wife

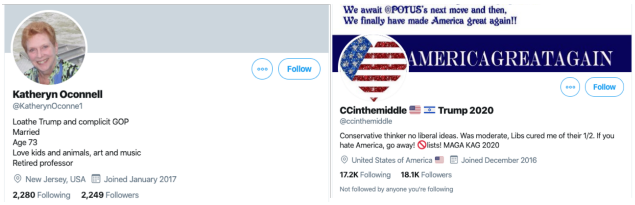

Figure 9: Twitter account profiles. The left one has a low fakeness score. The right one has a high fakeness score: its profile is anonymous and its description more provocative.

Our training data are predominantly English websites covering news in the United States. As a result, our topic agnostic classifier associates fake domains with unprofessional website designs and simple writing styles. This might not hold true in other regions of the world. One example is www.saamana.com, a regional website in India. The topic of the website is trustworthy but the design is shabby due to low budget.

We will reduce the sample bias by collecting feedback from fact-checkers who interact with our system. We will also collect fake and real domains in different countries and different languages.

We plan to reduce the selection bias by collecting data from multiple social media feeds, and using a wide variety of keywords, hashtags, user handles to capture potential news originators. For example, instead of listening to certain keywords, we can collect real-time tweets from accounts with high fakeness scores, because those accounts are more likely to share fake news.

5.2 Lack of unified dataset and evaluation framework

Research in cyber security suffer from a lack of unified dataset and evaluation framework. Unification is difficult because the target of detection (for example: fake news, social bot, or computer virus)

adapts constantly. There is a risk of using previous dataset, as adversaries can use exactly the same dataset to circumvent the detection. We are aware of this problem and therefore propose a combination of supervised and unsupervised approach.

On the evaluation side, evaluating newly discovered domains is time consuming. This is partly why we are not able to evaluate all discovered fake domains that are not fact-checked. Our next step is to introduce our Web interface to research communities, fact-checking groups, and social media companies to speed up the labeling process all together. We will also introduce application programming interface (API) to allow researchers to query our growing database of newly discovered domains with ease.

6 CONCLUSIONS

We present a discovery system that proactively surfaces fake news domains by leveraging domain network structures reconstructed from real-time social media feeds. Our system combines unsupervised clustering, supervised prediction, and human-in-the-loop interaction together. We provide a Web interface to allow users to visualize, search and label fake news domains. We show that our system is able to discover suspicious domains that have not been fact-checked before. We also show discriminating features of Twitter accounts that are likely to share fake news source. As much of today's political debates and social conversations have shifted to online social media, we are expecting to see more websites created to spread misinformation. For this reason, we hope that our work can improve early detection rate and discovery capability. We plan to open our system access to more research communities to facilitate fact-checking process and to leverage more crowd intelligence.

REFERENCES

[1] Alexandre Bovet and Hernán A. Makse. 2019. Influence of fake news in Twitter during the 2016 US presidential election. *Nature Communications* 10, 1 (2019), 7. https://doi.org/10.1038/s41467-018-07761-2

[2] Sonia Castelo, Thais Almeida, Anas Elghafari, Aécio Santos, Kien Pham, Eduardo Nakamura, and Juliana Freire. 2019. A Topic-Agnostic Approach for Identifying Fake News Pages. In *Companion Proceedings of The 2019 World Wide Web Conference* (San Francisco, USA) *(WWW '19)*. ACM, New York, NY, USA, 975–980. https://doi.org/10.1145/3308560.3316739

[3] Zhouhan Chen, Rima S. Tanash, Richard Stoll, and Devika Subramanian. 2017. Hunting Malicious Bots on Twitter: An Unsupervised Approach. In *Social Informatics*. Springer International Publishing, Cham, 501–510.

[4] Fernando Cardoso Durier da Silva, Rafael Vieira, and Ana Cristina Garcia. 2019. Can Machines Learn to Detect Fake News? A Survey Focused on Social Media. In *HICSS*.

[5] Chris Dulhanty, Jason L. Deglint, Ibrahim Ben Daya, and Alexander Wong. 2019. Taking a Stance on Fake News: Towards Automatic Disinformation Assessment via Deep Bidirectional Transformer Language Models for Stance Detection. *CoRR* abs/1911.11951 (2019). arXiv:1911.11951 http://arxiv.org/abs/1911.11951

[6] Emilio Ferrara, Onur Varol, Clayton A. Davis, Filippo Menczer, and Alessandro Flammini. 2016. The rise of social bots. *Commun. ACM* 59 (2016), 96–104.

[7] Maria Glenski, Ellyn Ayton, Josh Mendoza, and Svitlana Volkova. 2019. Multilingual Multimodal Digital Deception Detection and Disinformation Spread across Social Platforms. *CoRR* abs/1909.05838 (2019). arXiv:1909.05838 http://arxiv.org/abs/1909.05838

[8] Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. 2019. Fake news on Twitter during the 2016 U.S. presidential election. *Science* 363, 6425 (2019), 374–378. https://doi.org/10.1126/science.aau2706 arXiv:https://science.sciencemag.org/content/363/6425/374.full.pdf

[9] Gisel Bastidas Guacho, Sara Abdali, Neil Shah, and Evangelos E. Papalexakis. 2018. Semi-supervised Content-Based Detection of Misinformation via Tensor Embeddings. *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (Aug 2018). https://doi.org/10.1109/asonam.2018.8508241

[10] Bin Guo, Yasan Ding, Lina Yao, Yunji Liang, and Zhiwen Yu. 2019. The Future of Misinformation Detection: New Perspectives and Trends. *CoRR* abs/1909.03654 (2019). arXiv:1909.03654 http://arxiv.org/abs/1909.03654

[11] Sebastião Miranda, David Nogueira, Afonso Mendes, Andreas Vlachos, Andrew Secker, Rebecca Garrett, Jeff Mitchel, and Zita Marinho. 2019. Automated Fact Checking in the News Room. In *The World Wide Web Conference* (San Francisco, CA, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 3579–3583. https://doi.org/10.1145/3308558.3314135

[12] Kai Nakamura, Sharon Levy, and William Yang Wang. 2019. r/Fakeddit: A New Multimodal Benchmark Dataset for Fine-grained Fake News Detection. *ArXiv* abs/1911.03854 (2019).

[13] Feng Qian, Chengyue Gong, Karishma Sharma, and Yan Liu. 2018. Neural User Response Generator: Fake News Detection with Collective User Intelligence. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*. International Joint Conferences on Artificial Intelligence Organization, 3834–3840. https://doi.org/10.24963/ijcai.2018/533

[14] Chengcheng Shao, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. 2016. Hoaxy: A platform for tracking online misinformation. In *Proceedings of the 25th International Conference Companion on World Wide Web*. International World Wide Web Conferences Steering Committee, 745–750.

[15] Kai Shu, Ahmed Hassan Awadallah, Susan Dumais, and Huan Liu. 2019. Detecting Fake News with Weak Social Supervision. arXiv:arXiv:1910.11430

[16] Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. dEFEND: Explainable Fake News Detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Anchorage, AK, USA) *(KDD '19)*. ACM, New York, NY, USA, 395–405. https://doi.org/10.1145/3292500.3330935

[17] Kai Shu and Huan Liu. 2019. Detecting Fake News on Social Media. *Synthesis Lectures on Data Mining and Knowledge Discovery* (2019).

[18] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake News Detection on Social Media: A Data Mining Perspective. *SIGKDD Explor. Newsl.* 19, 1 (Sept. 2017), 22–36. https://doi.org/10.1145/3137597.3137600

[19] Kai Shu, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Mining Disinformation and Fake News: Concepts, Methods, and Recent Advancements. *CoRR* abs/2001.00623 (2020). arXiv:2001.00623 http://arxiv.org/abs/2001.00623

[20] Kate Starbird, Ahmer Arif, and Tom Wilson. 2019. Disinformation As Collaborative Work: Surfacing the Participatory Nature of Strategic Information Operations. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 127 (Nov. 2019), 26 pages. https://doi.org/10.1145/3359229

[21] Leo Graiden Stewart, Ahmer Arif, and Kate Starbird. 2018. Examining Trolls and Polarization with a Retweet Network.

[22] Maciej Szpakowski. 2020. *FakeNewsCorpus*. https://github.com/several27/FakeNewsCorpus

[23] Onur Varol, Emilio Ferrara, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2017. Online Human-Bot Interactions: Detection, Estimation, and Characterization. https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15587

[24] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151. https://doi.org/10.1126/science.aap9559 arXiv:https://science.sciencemag.org/content/359/6380/1146.full.pdf

[25] Yaqing Wang, Weifeng Yang, Fenglong Ma, Jin Xu, Bin Zhong, Qiang Deng, and Jing Gao. 2019. Weak Supervision for Fake News Detection via Reinforcement Learning. *arXiv e-prints*, Article arXiv:1912.12520 (Dec 2019), arXiv:1912.12520 pages. arXiv:1912.12520 [cs.SI]

[26] Fan Yang, Shiva K. Pentyala, Sina Mohseni, Mengnan Du, Hao Yuan, Rhema Linder, Eric D. Ragan, Shuiwang Ji, and Xia (Ben) Hu. 2019. XFake: Explainable Fake News Detector with Visualizations. In *The World Wide Web Conference* (San Francisco, CA, USA) *(WWW '19)*. ACM, New York, NY, USA, 3600–3604. https://doi.org/10.1145/3308558.3314119

[27] Shuo Yang, Kai Shu, Suhang Wang, Renjie Gu, Fan Wu, and Huan Liu. 2019. Unsupervised Fake News Detection on Social Media: A Generative Approach. In *AAAI*.