Proceedings of the Fourteenth International AAAI Conference on Web and Social Media (ICWSM 2020)

Toward a Better Performance Evaluation Framework for Fake News Classification

Lia Bozarth, Ceren Budak
University of Michigan, School of Information
105 S State St
Ann Arbor, MI 48109
{lbozarth, cbudak}@umich.edu

Abstract

The rising prevalence of fake news and its alarming downstream impact have motivated both the industry and academia to build a substantial number of fake news classification models, each with its unique architecture. Yet, the research community currently lacks a comprehensive model evaluation framework that can provide multifaceted comparisons between these models beyond the simple evaluation metrics such as accuracy or f1 scores. In our work, we examine a representative subset of classifiers using a very simple set of performance evaluation and error analysis steps. We demonstrate that model performance varies considerably based on i) dataset, ii) evaluation archetype, and iii) performance metrics. Additionally, classifiers also demonstrate a potential bias against small and conservative-leaning credible news sites. Finally, models' performance varies based on external events and article topics. In sum, our results highlight the *need* to move toward systematic benchmarking.

Introduction

In the United States, many political pundits and media scholars alike have cautioned against the rising influence of fake news (Silverman 2017; Balmas 2014), stressing that the spread of false information weakens the legitimacy and public trust in the established political and media institutions. Outside of the U.S., fake news has been tied to Brexit in Europe (Kucharski 2016), and the rising hate, violence, and nationalism in Indonesia (Kwok 2017). It has also been linked to the endangerment of election integrity of European and Latin American nations (Fletcher et al. 2018; Alimonti and Veridiana 2018). Indeed, fake news, backed by armies of social bots, disseminates significantly faster and deeper than mainstream news (Shao et al. 2017). Additionally, subsequent research also suggests that it is difficult for the general public to distinguish fake news from credible content. Equally alarming is that repeated exposure causes readers to perceive false content as more accurate (Balmas 2014). Past work also shows the importance of detecting and combating misinformation in its early phases of spread (Budak, Agrawal, and El Abbadi 2011). Thus, timely, scalable, and high-performing fake news detection automatons become a vital component in combating fake news.

Thus far, researchers have leveraged linguistic attributes, user network characteristics, temporal propagation patterns of news articles, and various machine learning paradigms to build effective models that separate fake news from traditional news content (Ruchansky, Seo, and Liu 2017; Liu and Wu 2018; Shu et al. 2019a; Horne et al. 2018; Castelo et al. 2019; Yang et al. 2019). These are all valuable contributions. Some of these novel approaches led to high performing classifiers with exceptionally high accuracy and F1 scores. Yet, our review also reveals a key gap: many papers lack comprehensive model performance evaluation and error analysis steps. Here, we first review 23 distinct classifiers from related work and consolidate each according to 4 major components: data source, data types, feature engineering techniques, and machine learning paradigms. We then reached out to the authors of 17 papers and acquired a total of 5 classifiers. Next, we evaluate these models using a few very simple procedures. Our paper makes the following contributions:

- We show that model performance may vary drastically based on the choice of dataset. As such, results from individual papers, especially those that use a single dataset for evaluation, should be taken with a grain of salt.
- Additionally, classifiers generally have significantly higher performance when trained and validated using the common 5-fold 80/20 data split evaluation archetype compared to validation using a set of domains never before encountered in training. This suggests that classifiers might be learning domain-specific features as opposed to actual distinctions between fake and traditional news.
- We show that all classifiers studied here demonstrate significantly higher false-positive rates for right-leaning mainstream news sites. This bias raises an important concern for trust in fake news detection systems. Similarly, articles from small credible news sites are also classified as fake news more often than those from large sites.
- Next, the performance of classifiers can be worse following external shocks such as scandals. This indicates that temporal variations in classifier performance need to be

taken into consideration when taking actions based on these predictions.

- Finally, models generally have a higher false-positive rate when classifying news articles involving scandals, and a higher false-negative rate for articles focused on the 2016 election (e.g., polling results).

In sum, our simple evaluation approach reveals potential biases and significant weaknesses in existing classifiers. It also provides a cautionary tale for real-world applications that use these models. Our work sufficiently demonstrates that using simple metrics such as accuracy, AUC, or F1 to evaluate and compare models is insufficient. As a community, we need to collectively investigate and construct a more comprehensive performance evaluation framework for fake news classification.

Fake News Classifiers—Review & Selection

We first review a total of 23 existing fake news classifiers in Section . We observe that text, relational and temporality data are the most commonly used data to construct feature-sets: within the 23 classifiers reviewed, 17 (or 74%) exclusively use 1 or more of the 3 data types. The remaining 6 use at least 1 additional data type (e.g., images). Next, we reach out to authors of all 17 papers and obtained 5 code repositories. We describe this subset of models in detail in Section .

Meta-review

As shown in Figure 1, the process of building a fake news classifier consists of 4 major decisions: i) choosing data-sources, ii) selecting a subset of data from all available data types, iii) deciding on the techniques that transform raw data into features (i.e., feature engineering). These features are bundled closely with iv) the specific machine learning algorithm(s) one adopts.

Data Sources: News sites and social media platforms are the two primary data sources. Our review shows that 12 (or 52.1%), and 15 (or 65.2%) out of the 23 classifiers use data from news sites and social media platforms respectively, with four (or 17.4%) using both sources.

Data Types: We show *text* data are by far the most common with 21 (or 91.3%) classifiers using at least some text-based data. It's followed by *relational* data (e.g., follower, friend) at 47.8% and *temporality* data at 26.0% usage. A small number of classifiers also use additional data types including multimedia images and videos (Gupta et al. 2013; Boididou et al. 2018). Other data types include author age, gender, and credibility (Long et al. 2017; Shu, Wang, and Liu 2018); website DNS records (Ma et al. 2009); web markup and advertising (Castelo et al. 2019); and geo location (Deligiannis et al. 2018).

Feature Engineering: A large arsenal of techniques are available to transform raw data of varied types into usable features.

Text Data: First, features can be extracted from text using existing theories and domain knowledge such as psycholinguistic theories and frameworks (Horne et al. 2018; Castelo et al. 2019; Zhou and Zafarani 2019). Focusing on news articles, these features include i) quality, complexity and style (e.g. word count, lexicon diversity, readability), and ii) psychological attributes (e.g., sentiment, subjectivity, biases). See (Zhou and Zafarani 2018) for a detailed literature review. Intuitively, fake news articles are likely to include more "clickbaity" elements—capitalization of all words in the title, use of many exclamation marks, or adoption of sharp and sentimental words (e.g., "poisoning") in the text. Within a social context, user profile descriptions or user posts can be used to derive implicit features such as a user's personality, gender, and age (Shu et al. 2019b). These features are also used to detect fake news. As such, this feature extraction approach often leads to highly explainable and transparent classifiers.

Additionally, text can be transformed into i) *ngrams*, commonly combined with tfidf weighting (Ahmed, Traore, and Saad 2017; Qazvinian et al. 2011); ii) *vectors*, i.e., converting content to numeric vectors using variations of GloVE, Skipgram, CBOW, and then word2vector, sent2vec, or doc2vec (Riedel et al. 2017; Gravanis et al. 2019); and iii) *tensors* (Guacho et al. 2018; Papanastasiou, Katsimpras, and Paliouras 2019), which are 3-dimensional vector representations of words or documents. Extracted features can also undergo additional transformation steps including feature reduction.

Finally, researchers also use text to build networks (e.g., hashtag-hashtag or text similarity-based networks) and derive graph-based features (Rubin 2018).

Relational Data: Relational data are generally used to construct networks (e.g., follower-followee network, retweet network, user-article bipartite network). Researchers then use these networks to derive usable features including communities, clustering coefficient, and network motifs. (Coletto et al. 2017; Volkova et al. 2017). Networks can also be represented as matrices which can then be reduced into a low-dimensional representation and adopted as features (Ruchansky, Seo, and Liu 2017). Finally, these networks can be used by semi-supervised label propagation classifiers (Tacchini et al. 2017).

Temporality Data: Many existing studies use temporality data to build classifiers (Jin et al. 2014; Liu and Wu 2018; Ruchansky, Seo, and Liu 2017; Do et al. 2019). For instance, Do et al. (2019) and Ruchansky et al. (2017) partition user interactions (posts) about news based on the timestamps. Posts within the same partition are treated as a single text document. Liu and Wu (2018) model the propagation path of each news story as a multivariate time series which are then used as features. Additionally, similar to relational data, propagation networks can also be used to extract useful graph-based features such as motifs.

Machine Learning Paradigms: Reviewed classifiers can be categorized as *supervised*, *semi-supervised* or *unsupervised*. Work by Katsaros et al. (2019) provides an overview of existing classifiers categorized with respect to paradigms. Further, supervised models can be subcategorized as neural network (Volkova et al. 2017; Riedel et al. 2017; Ruchansky, Seo, and Liu 2017; Wang et al. 2018; Ma, Gao, and

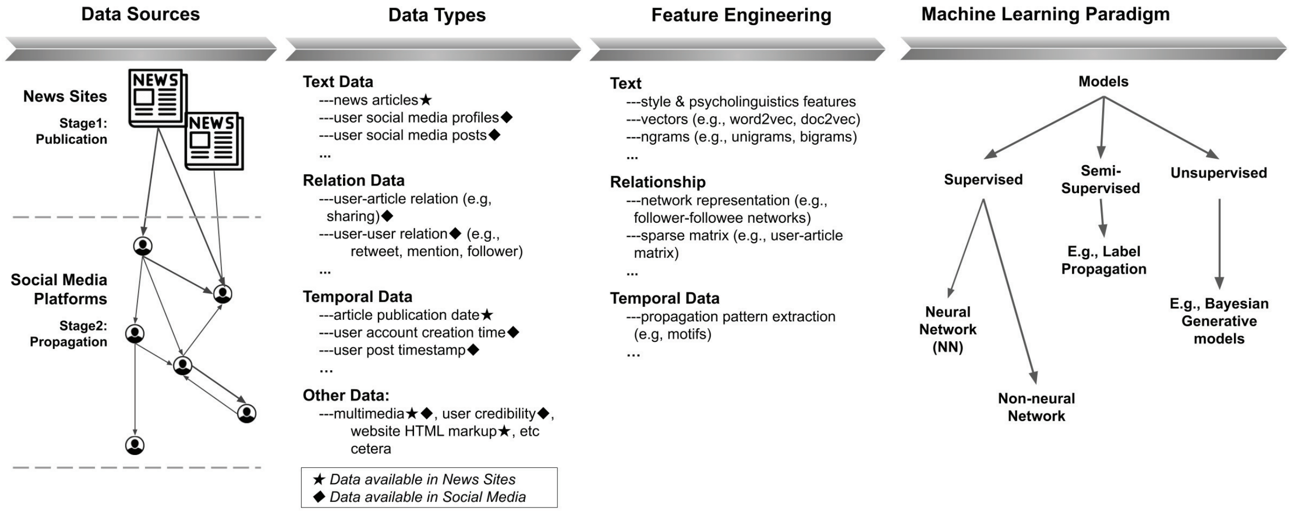

Data Sources
Data Types
Feature Engineering
Machine Learning Paradigm
News Sites
Stage1: Publication
NEWS
NEWS
Social Media Platforms
Stage2: Propagation
Text Data
---news articles★
---user social media profiles◆
---user social media posts◆
...
Relation Data
---user-article relation (e.g, sharing)◆
---user-user relation◆ (e.g., retweet, mention, follower)
...
Temporal Data
---article publication date★
---user account creation time◆
---user post timestamp◆
...
Other Data:
---multimedia★◆, user credibility◆, website HTML markup★, etc cetera
★ Data available in News Sites
◆ Data available in Social Media
Text
---style & psycholinguistics features
---vectors (e.g., word2vec, doc2vec)
---ngrams (e.g., unigrams, bigrams)
...
Relationship
---network representation (e.g., follower-followee networks)
---sparse matrix (e.g., user-article matrix)
...
Temporal Data
---propagation pattern extraction (e.g, motifs)
...
Models
Supervised
Semi-Supervised
Unsupervised
E.g., Label Propagation
E.g., Bayesian Generative models
Neural Network (NN)
Non-neural Network

Figure 1: An Overview of the Fake News Detection Process. This process consists of 4 major choices in: i) datasources, ii) data types, iii) feature engineering techniques, and iv) machine learning paradigm.

Wong 2018; Zhang, Dong, and Yu 2019) or non-neural network based approaches (Ahmed, Traore, and Saad 2017; Horne et al. 2018; Castelo et al. 2019; Gravanis et al. 2019). *Semi-supervised* models use label propagation techniques (Jin et al. 2014; Tacchini et al. 2017; Rubin 2018; Guacho et al. 2018). Last, the *unsupervised* paradigm is very rare–we identified only one related prior work (Yang et al. 2019).

Representative Fake News Classifiers

From the 17 models, we see that 3 include a code repository link in their original publication. We then emailed authors of the remaining 14 papers, and 5 of them responded. Finally, we include 2 of the 5 models in our subsequent analysis [1] In sum, we collect a total of 5 distinct classifiers: BTC, CSI, HOAX, NELA, and RDEL. The code repositories (including our code) are available at https://github.com/lbozarth/fakenewsMPE.

BTC: This classifier (Gravanis et al. 2019) uses only news article text data. The authors extract 70 stylistic and psycholinguistic features (e.g., number of unique words, sentence readability, and sentiment) and geo-location features. Additionally, they transform each word into a vector using GLOVE and then sum the vectors to generate a vector representation of each article. The authors use Adaboost (Pedregosa et al. 2011) and concatenation of listed features to model fake news.

CSI: For each news article, this paper (Ruchansky, Seo, and Liu 2017) first partitions user engagements (e.g., tweets) with an article based on timestamps of the posts. All engagements within a partition are treated as a single document. They then use LSTM to capture the temporal patterns of the documents. Additionally, the authors also build a user-user network with the edge weight being the number of shared articles between pairs of users. This network's corresponding adjacency matrix is then used to generate lower dimensional features that capture the similarity of users' article sharing behavior. Finally, both sets of features are integrated together using another neural network layer.

HOAX: The authors (Tacchini et al. 2017) construct a user-article bipartite graph based on whether a user liked or shared an article or a post. They then use semi-supervised harmonic label propagation to classify unlabeled articles. This approach is based on the hypothesis that users who frequently like or share fake or low-quality content can be used to identify the quality of unlabeled content.

NELA: This classifier (Horne et al. 2018) uses the following 3 distinct dimensions of text-based features to predict fake news: i) style features (e.g. exclamation marks, verb tense, pronoun usage), ii) psycholinguistic features such as sentiment scores using LIWC, SentiStrength (Thelwall 2017) and iii) content complexity features including readability (Mc Laughlin, 1969), dictionary size, and average word length. We refer readers to the original paper for the complete list of 100+ features. The authors use Linear Support Vector Machine (SVM) and Random Forest as their classification algorithms.

[1] Two of the classifiers (Bourgonje, Schneider, and Rehm 2017; Guacho et al. 2018) are omitted due to our lack of experience with the programming language (e.g., MatLab). For the 3rd, our performance analysis reveals that it likely has an over-fitting issue due to its HTML-based features. To elaborate, we use the pre-generated HTML-based features and model provided by the paper. We train the model on 90% of the domains and validate on the remaining 10% (see the *bydomains* archetype in Section). The accuracy score provided in the paper using 5-fold training and validation is 0.86. In comparison, here, accuracy scores are 0.84 and 0.75 for training and validation respectively. Further, the AUC scores are 0.92 and 0.69, suggesting overfitting. We also used the leave-one-out training and validation approach. Observations are similar. Given the overfitting and that collecting HTML features for 0.7M webpages in our dataset (see Section) is also a costly task, we choose to omit this model from our analysis.

CLF	Source	Text	Relational	Temporal	Other	Number of Features	Machine Learning Paradigm
BTC	News Sites	stylistic; psycholinguistics; co-ntent complexity; word2vec	X	X	geotext features (from text)	70 (e.g., stylistic); 300 (word2vec); N (geotext)	supervised; non-NN; AdaBoost
CSI	Social Media	doc2vec	user-user	YES	X	122	supervised; NN; LSTM
HOAX	Social Media	X	user-article	X	X	X	semi-supervised; propagation
NELA	News Sites	stylistic; psycholinguistics; co-ntent complexity	X	X	X	122	supervised; non-NN; RandomForest
RDEL	News Sites	ngram (tfidf); cosine similarity between title and text	X	X	X	4001	supervised; NN; Multi-layer Perceptron

Table 1: Overview of Classifiers based on i) data source, ii) data types, iii) feature engineering, and iv) machine learning paradigm

RDEL: This model (Riedel et al. 2017) first tokenizes text from news articles and extracts the most frequent ngrams (unigram, bigram). Then, for each news article, it constructs the corresponding term frequency-inverse document frequency (TF-IDF) vectors for article title and body separately, and computes the cosine similarity between the 2 vectors. Finally, the authors concatenated the features together and use Multilayer Perceptron (Pedregosa et al. 2011) to classify fake and real news articles.

The data source, data types, feature engineering process, and machine learning paradigm for each of the 5 classifiers are summarized on Table 1. As shown, this set of classifiers encompasses both neural network and non-neural network based supervised learning paradigms, as well as the semi-supervised paradigm. Additionally, it includes all 3 most common data types: text, relational, and temporality. Focusing on feature engineering, the classifiers collectively cover the most common text feature engineering approaches: theory-driven, ngrams, word2vec, and doc2vec. Similarly, they also cover some common relational data feature engineering techniques: user-article network, user-user network. Notably, these classifiers do not include popular approaches such as leveraging user profile descriptions, or follower-followee networks. Overall, however, we argue that this set of classifiers is sufficiently representative.

Data

We use two datasets in this study. The summary statistics are available on Table 2. Both datasets use *Media Bias Fact Check* (Van Zandt 2018) as ground truth to label whether a domain is a fake news site. *Media Bias Fact Check* contains one of the most comprehensive list of fake and mainstream news sites and is also used by many related work (Starbird 2017; Main 2018). Prior research shows that the choice of ground truth can have a significant effect on downstream analysis (Bozarth, Saraf, and Budak 2020). Therefore, we note that our study is only the first step towards building a comprehensive fake news model evaluation framework. Future work should consider multiple ground truth labels.

	Election-2016	NELA-GT
Time Period	12/2015-01/2017	02/2018-10/2018
# Domains	1390	130
# Fake News Domains	335 (24%)	38 (29%)
# Articles	231.6K	304.1K
# Fake News Articles	31.3K (16%)	37.4 (14%)
# Tweets	1.16M	
# Fake News tweets	141.9K (12.2%)	
# Users	215.2K	
# Fake News Users	37.9K (18%)	

Table 2: Basic Statistics for Datasets

Election-2016: This dataset is primarily focused on the 2016 U.S. presidential candidates and consists of both social media data and news articles. Social media data collection is described in detail in Bode et al. (2020). The data collection was performed using Sysomos MAP. For any given day between December, 2015, and January 1, 2017, this dataset includes i.) 5,000 tweets randomly sampled from all tweets that included the keyword “Trump”, and ii) 5,000 tweets similarly sampled from all that mentioned “Clinton”. The webpages dataset (Budak 2019) includes the content of the webpages shared in the Twitter dataset described above. For each tweet with an external URL, the dataset includes a record with: i) the shortened URL, ii) the original URL, iii) domain name, iv) title of the document, v) body of the document, (vi) the date of the tweet, and vii) Twitter account id of the user sharing the URL. Here, we use *Media Bias Fact Check* to identify the list of fake and mainstream news sites present in *Election-2016*, and filter out non news-related sites. As shown on Table 2, *Election-2016* contains 231.6K unique articles (16% of which are fake news), and 1.16M tweets (12.2% of which contain links to fake news articles) shared by 215.2K unique users (18% users shared at least 1 fake news article).

NELA-GT: This dataset (Nørregaard, Horne, and Adalı 2019) contains articles scraped from 194 news sites between 02/2018 and 11/2018. The list of domains is collected by aggregating existing lists of fake and mainstream news sites provided by other researchers and organizations. News content is scraped via the RSS feed of these sites. Each domain has source-level veracity labels from 1 or more independent assessments (e.g., *Media Bias Fact Check, News Guard*). We note that 130 out of 194 domains have labels from *Media Bias Fact Check*.

As shown on Table 2, *NELA-GT* contains 304.1K unique articles (14% of which are fake news).

There are several key distinctions between *Election-2016* and *NELA-GT*. First and foremost, news articles in *Election-2016* are collected through tweets mentioning Trump or Clinton. As such, these articles are almost exclusively about one or both candidates. In comparison, creators of *NELA-GT* directly access and scrape the websites which likely resulted in more diverse news topics. Next, *Election-2016* contains 1.4K distinct news sites, 10 times that of *NELA-GT*, yet the latter has 72.5K more news articles. Indeed, the median number of articles per domain is 13 for *Election-2016* and 1.12K for *NELA*. Finally, over 30% of all fake news domains active in 2016 have since become defunct (Bozarth, Saraf, and Budak 2020), thus they are in *Election-2016* but not *NELA-GT*.

Additional Auxiliary Data: We also obtain the following data for each news site: i) *average monthly traffic* using similarweb.com, a popular web analytics platform (SimilarWeb 2019; Singal and Kohli 2016); ii) *age* using whois.com, a domain name registrar database (Mueller and Chango 2008); and finally, iii) *ideology* using *Media Bias Fact Check*. Additionally, for each article, we also include its publication date (For *Election-2016*, an article's date is approximated using the timestamp of the earliest tweet that included it).

Data Preprocessing: For each dataset, we first filter out all news articles i) without a title, or ii) with fewer than 10 words in the article body. Next, we aggregate 4.1K words and phrases representing news sites names [2] (e.g., "Daily Dot", "CNNPolitics"); for each news article, we remove matching words and phrases from the article.

Analysis

In this section, we first compare and contrast classifier performance using different evaluation archetypes. Then, we conduct an in-depth error analysis focusing on domains and articles of varied attributes. Finally, we examine model performance and bias tradeoffs.

Performance Overview

We first underline the 3 most common training and performance evaluation archetypes `basic`, `forecast`, and `bydomains`. We then present our evaluation metrics and results.

(1) Basic N-folds (`basic`): Using this common approach, data are split into *N*-folds for training and cross-validation. Here, for each dataset, we use 5-fold (i.e., 80/20) training and validation data split.

(2) Forecasting into the Future (`forecast`): Here, given a time t, classifiers are trained on fake and mainstream news articles that were written before t, and tested against those that were written after t. For each dataset, we randomly sample 10 dates within its data collection period and split data into training and validation accordingly.

(3) Predicting Never Before Encountered Domains (`bydomains`): In this archetype, classifiers are trained on articles from 90% randomly sampled domains and tested against articles from the remaining 10% that are not present in training. We repeat this sampling process 10 times for each dataset.

Evaluation Metrics: We use *AUC*, *F1* (fake news articles as the positive label), *precision*, and *recall* to evaluate model overall performance. We omit *accuracy* due to label imbalance in the datasets (Huang and Ling 2005).

Results: Results are summarized in Figure 2. As shown, we denote classifiers using different colors. Additionally, each grid represents a distinct dataset and performance metric combination. Within a given grid, outer-rings represent higher performance. Note that *NELA-GT* doesn't provide social media data, so analyses for `CSI` and `HOAX` are not available for this dataset. Overall, we see that classifier performance varies considerably based on the i) dataset, ii) evaluation archetype, and iii) metric.

(i) Dataset Effects: The average *AUC* scores for `BTC` under the `basic` evaluation archetype are 0.78 and 0.96 when the datasets are *Election-2016* and *NELA-GT* respectively, a considerable difference. We also observe a similar but less significant effect for `RDEL` and `NELA`. One possible explanation is that *Election-2016* is specifically focused on Donald Trump and Hillary Clinton. Thus, news content from fake and mainstream publishers in this dataset is presumably much more similar compared to *NELA-GT*, which scrapes the entire RSS feed of news sites daily. The higher article similarity likely contributes to a performance drop in content-based models.

(ii) Archetype Effects: `RDEL` and `BTC` both perform considerably better under `basic` evaluation compared to predicting new domains (i.e., `bydomains`). For instance, when dataset is *NELA-GT*, the *AUC* score for `RDEL` is 0.97 using `basic` archetype but 0.72 using `bydomains` (similar patterns for `BTC`). One possible explanation is that, despite removing news site name-related tokens (e.g., "daily beast", "nytimes") from data, certain remaining word tokens may still be indicative of a domain and its practices (e.g., New York Times has the custom of using the word "Mr." when addressing the President of the United States (Corbett, P. 2017)). Thus, if word2vec or ngram-based classifiers such as `BTC` and `RDEL` rely heavily on site-specific features for training, they may perform poorly when validating on newly encountered domains. Similarly, we also see that `BTC` and `RDEL` perform worse in `forecast` than `basic`. Findings here are complementary to prior research (Horne, Nørregaard, and Adali 2019) which demonstrates that the performance of text-based models decreases over time due to changes in news content. Interestingly, `NELA`, the only other exclusively text-based classifier, has more comparable *AUC* scores across the archetypes—the *AUC* scores are 0.80,

[2] Here, we use the list of news site names provided by *Media Bias Fact Check*.

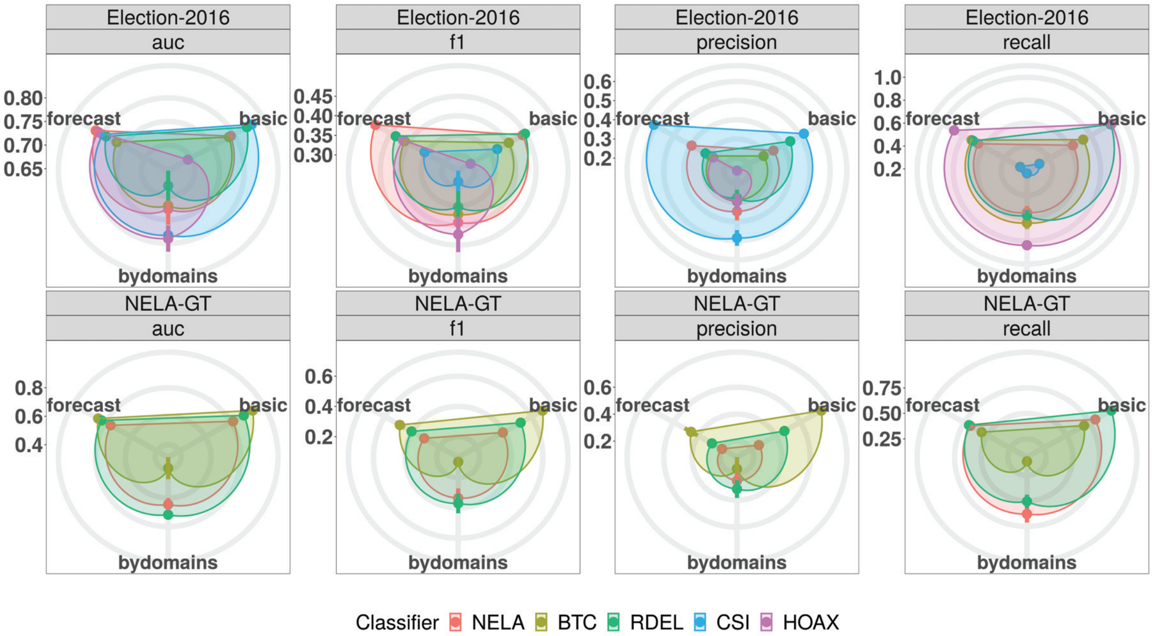


Figure 2: Performance Overview For All Classifiers. As shown, we separate the classifiers by colors. Additionally, each grid represents a distinct dataset and performance metric combination. Within a given grid, datapoints that lie in the outer-rings represent higher performance. For instance, the upper left corner grid contains each classifier's AUC scores for the *Election-2016* dataset. We see that `RDEL`, colored in green, has a significantly higher AUC (data point lies on the outer-ring) when validated against the `basic` archetype compared to `forecast` and `bydomains`.

0.73, and 0.82 for `basic`, `bydomains`, and `forecast` respectively when dataset is *Election-2016*. We note that `NELA` only uses the stylistic and psycholinguistic features of articles. As such, it may be more robust to site-specific linguistic eccentricities. Compared to text feature-only classifiers, both `CSI` and `HOAX`—the classifiers that only partially or not-at-all rely on text content— have more comparable *AUC* scores across all archetypes.

Out of the 3 archetypes, `basic` is the most common. That is, model performance is typically examined using 5-fold cross-validation with 80/20 training and validation data split. Yet, if researchers only assess classifiers' performance based on *basic*, they may not discover the potential weakness their models have against never before encountered domains. In fact, our results demonstrate that the word2vec and ngram text feature-based models maybe especially need to be evaluated against `bydomains`.

Evaluation Metric Effects: The ranking of classifiers change based on whether we use *AUC* or *F1*. For instance, when the dataset is *Election-2016* and the archetype is `bydomains`, we show that `CSI` has *rank* = 2 using *AUC* and *rank* = 4 using *F1*. Additionally, all classifiers excluding `CSI` generally have significantly higher *recall* than *precision*. In fact, *CSI* has the highest *precision* across all classifiers and the lowest recall. As such, the definition of "best-performing" is dependent on whether one prefers precision over recall. Generally, high precision is preferred in high stake circumstances. For instance, Google and Facebook both have banned hundreds of fake news sites from using their Advertising services (Paresh 2016). In this context, a credible mainstream news site would suffer considerable economic drawbacks if it's falsely labeled as fake. As such, `CSI` should be the preferred model despite having a low recall score. Alternatively, classifiers with a high recall can serve as a useful filter. As an example, researchers and organizations interested in identifying new fake news domains can first apply a high recall model to obtain a list of presumptive fake news domains, and then manually review each site to label true fakes.

In sum, we demonstrate that the performance of classifiers varies considerably from case to case as a result of the difference in i) dataset, ii) evaluation archetype, and iii) metric.

Domain and Context-specific Error Analysis

In this section, we conduct in-depth error analysis using false-negative rate **(FNR)** and false-positive rate **(FPR)**. Here, for each $c \in \{$`BTC, CSI, HOAX, NELA, RDEL`$\}$, we first i) assess whether models perform better or worse on classifying domains of particular subcategories. Next, we then explore ii) errors that are of significant interest in the context of the 2016 election.

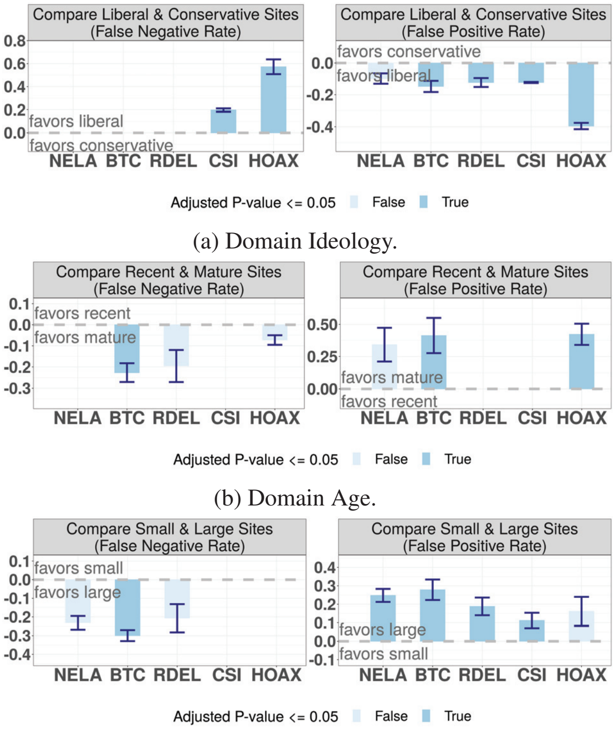
Domain-level Error Analysis: Here, we examine error rates based on domain i) ideological-leaning, ii) age, and iii) popularity. To elaborate, conservative elites have long criticized both the academia and tech firms for having a liberal bias (Gross and Simmons 2006). Thus, we aim to determine whether classifiers indeed contain biases such as having a higher *fnr* for liberal-leaning fake news sites, and/or a higher *fpr* for conservative-leaning mainstream news sites. Similarly, a model's performance may also vary based on a domain's age and popularity (e.g., incorrectly classifying recently created websites with small viewerships as fake news at a higher rate compared to mature domains with heavy traffic). We note that analysis here is focused on the *Election-2016* [3] dataset and `bydomains` archetype. We focus on this archetype for potential impact. While tech giants and online platforms have blacklisted hundreds of fake news sites, reports show that owners of these domains are ramping up for the 2020 election by creating new sites (Wingfield, Isaac, and Benner 2016; Soares 2019).

Ideological Biases: For each article i in a given validation set, we first assign i to a bin using i's corresponding domain's ideology $\{unknown, conservative, center, liberal\}$. Then, for each classifier c, we calculate c's *fpr* and *fnr* for each bin separately. Here, we denote *fpr* for the liberal(left)-leaning and conservative(right)-leanings bins as $fpr(l)$ and $fpr(r)$ respectively. Next, to evaluate "liberal bias", we examine i) whether liberal-leaning fake news sites on average are significantly more often classified by c as credible news (i.e., $fnr(l) > fnr(r)$), and ii) whether articles by conservative-leaning mainstream news sites on average are significantly more often classified by c as fake news (i.e., $fpr(l) < fpr(r)$).

To elaborate, we apply Student's T-test (Gibbons and Chakraborti 1991) on the distributions $fpr(V,l,c)$ and $fpr(V,r,c)$. Here, $fpr(V,l,c)$ is classifier c's false-positive rates for the liberal-leaning bin l across all the validation sets V of the `bydomains` archetype. For results that are statistically significant ($p - value <= 0.05$), we compute the mean differences between the distributions and then plot the values. We repeat this process for false-negative rates. As shown in Figure 3a, all classifiers have a higher false-positive rate for articles from conservative-leaning credible news sites. Furthermore, all results remain significant, except for `NELA`, even after adjusting the p-values using the Holm-Bonferroni method (Hochberg 1988) to account for multiple hypothesis testing. One possible explanation for this bias is that there are significantly fewer liberal-leaning fake news sites available for training. Another explanation is provided by Benkler, et al. (2018). They argue in a recent book that some traditional right-leaning news outlets participated in the dissemination of fake news by echoing and giving platform to false claims initially produced and campaigned by fake news sites. This might bring into question the ground truth definition as opposed to the classifier results. As a whole, results in this section demonstrate that researchers should actively consider potential ideological biases when building and evaluating fake news classifiers.

Domain Age: Similar to the previous section, we first partition each domain i based on its age into 3 bins: i) unknown (i.e., DNS record is not available), ii) recent ($<= 3years$), and iii) mature ($> 3year$). We then compare *fnr* and *fpr* between the bins and assess whether the mean differences are statistically significant. As shown in Figure 3b, both `BTC` and `HOAX` have a significantly higher *fpr* for recent domains. In other words, the 2 classifiers more often label articles by recently created credible news domains as fake news. A potential explanation is that newly created mainstream news domains have published fewer articles and thus models have less data to train on. Alternatively, newer mainstream sites might have language and consumers that are better aligned with fake news outlets. Finally, for robustness check, we also repeat the evaluation and set the partitions into recent ($<= 5years$) and mature ($> 5years$). We observe similar patterns.

Domain Popularity (web-traffic): For each dataset, we di-



(a) Domain Ideology.

(b) Domain Age.

(c) Domain Popularity

Figure 3: Mean Differences in Error Rates.We first assign each prediction from the validation dataset into groups using its corresponding domain i) ideology, ii) age, or iii) popularity. We then compare the classifiers' false-positive and false-negative error rates for each group of predictions. Here, the x-axis denotes the classifiers, and the y-axis denotes each classifier's mean differences in error rates between a selected group and the baseline group (e.g., subtract a classifier's average false-positive rate for liberal-leaning sites by the average for conservative-leaning sites). The baseline groups are $\{conservative, mature\ site, large\ site\}$ for domain ideology, age, and popularity respectively. Note*, results that are not statistically significant are removed. Further, results that are insignificant after adjusting p-values using the Holm–Bonferroni method (Hochberg 1988) are colored in light-blue; finally, results that remain significant after adjustment are in darkblue.

[3] We also repeat the analysis using the *NELA-GT* dataset. Results are largely insignificant given that *NELA-GT* only contains 130 domains. In comparison, *Election-2016* has 1.4K.

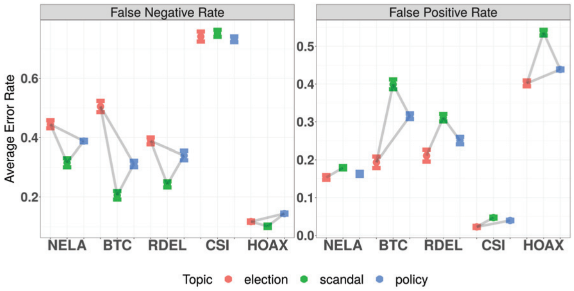


Figure 4: Article Topic-level Error Rates. The x-axis denotes the classifiers, and the y-axis denotes a topic's average *fpr* (or, *fnr*). Topic are differentiated by color. Further, if a pairwise comparison between 2 topics is statistically significant even after adjustment for multiple hypothesis testing, the pair is linked by a gray line. For instance, the mean differences in false-positive rate between the pairs (election, scandal) and (scandal, policy) are significant for *RDEL*.

vide domains based on web-traffic into: i) unknown (no web-traffic data on similarweb.com), ii) small (web-traffic percentile calculated to be between 0%-33% percentile), iii) median (33%-66% percentile), and iv) large. We again compare *fnr* and *fpr* across the bins. As shown in Figure 3c, all classifiers, except for `HOAX` have a significantly higher false-positive rate for small websites. This bias potentially causes small but legitimate news sites to be more often incorrectly labeled as fake news domains. Finally, for robustness check, we also repeat the evaluation and set the partitions into small (0%-50% percentile) and large (50%-100% percentile). We observe similar patterns.

Context-specific Error Analysis *(2016 election)*: As previously stated in Section , the *Election-2016* dataset is collected specifically to study the 2016 U.S. presidential candidates. Here we identity 2 additional error analyses that are of significance to the election. First, prior research (Bozarth, Saraf, and Budak 2020) has demonstrated that the prevalence of fake news temporarily decreases after scheduled high-profile events (e.g. presidential debates). Though, results are inconclusive for scandals (e.g., Trump Hollywood tape). We expect the fake news articles produced shortly after these shocks to differ from those published in other time periods. Accordingly, we examine model performance for such articles. Furthermore, political communications studies show that news coverage of different topics have varied impact on a voter's knowledge, decision-making, and trust in the government (Praino, Stockemer, and Moscardelli 2013; Van der Meer, Hakhverdian, and Aaldering 2016). As such, we also aim to determine whether model performance differs across article topics. In sum, we conduct error analysis based on i) the types of external events, and ii) article topics. We note that analysis here is focused on the `forecast` archetype given shocks are temporal events and news coverage of different topics varies over time.

External Events: We first obtain a list of scandals and planned key events of Trump, Clinton, or both that occurred in the general election from *ABC News* and *The Guardian*. The list, ordered chronically, includes: Republican nomination (07/18), Democrat nomination (07/28), Clinton "deplorable" and "pneumonia" scandals (09/09), first debate (09/26), Clinton email involving Wikileaks and Trump Hollywood tape scandals (10/07), second debate (10/09), Clinton email scandals involving the FBI (10/28, 11/06), and finally, the election day (11/08). Here, nominations, debates, and election day are assigned to *scheduled* and others to *scandal*. We also randomly select 10 dates and assign them as *baseline* for comparison purposes. Next, given classifier c, for each day $t \in \{07/18/2016, 07/28/2016...\}$, we train c using articles before t, and validate c on the articles that were published and shared within x days after t where $x \in \{3,5,7\}$. We again compute *fpr* and *fnr* for articles published right after *scheduled, scandal* events and compare these error rates to that of *baseline*. Surprisingly, we see that error rates are generally comparable across different types of external shocks for all classifiers except for `HOAX`, which has a considerably higher false-positive rate for predicting articles published shortly after scandals (the mean difference between *fpr* for *scandal* and *baseline* is 10.0%). In our paper, `HOAX` is the only model that exclusively adopts a user-article network-based classification approach. It's possible that users may have temporarily altered their news-sharing behavior right after scandals. In comparison, results from text-based models suggest that the linguistics features of articles published after shocks are not significantly different from those in *baseline*.

Article Topics: We obtain the topic for each article in the *Election-2016* dataset from related work (Bozarth, Saraf, and Budak 2020). We refer readers to the original paper to review the detailed topic-modeling process which assigns 49% of total articles into 15 unique topics (the remaining 51% is labeled as *other*). Here, we further cluster documents into broader topics $\{election, scandal, policy, other\}$ [4]. We observe that 20.3%, 8.9% and 19.2% of all articles belong in *election* (e.g., news involving polling results), *scandal* (e.g., news involving Clinton's pneumonia incident), and *policy* (e.g., the economy) respectively.

Next, we calculate and compare the average *fpr* and *fnr* rates for the topics. Results are summarized in Figure 4. The x-axis denotes the classifiers, and the y-axis denotes a topic's average *fpr* (or, *fnr*). Topics are differentiated by color. Further, if a given pair of topics, e.g. $(scandal, election)$, has a significant difference in mean error rates even after p-value adjustment, the pair are then linked together by a gray line. As shown, all text-based models and `HOAX` have a significantly higher *fpr* for *scandal* compared to both *policy* and *election*. That is, articles written by credible sites about scandals are significantly more often mislabeled as fake. We also see that, for 4 out of the 5 models, fake news articles focused on the election are more often labeled as credible compared to those about scandals or policy.

What does this mean? Past studies in voter decision-making demonstrate that scandals have a strong and long

[4] We merge {'email', 'clinton-health', 'sexual', 'clinton-wst', 'benghazi'} as *scandal*, and {'russia', 'economy', 'abortion', 'climate', 'mid-east', 'd&i', 'security', 'religion'} as *policy*.

lingering impact on voter choice (Praino, Stockemer, and Moscardelli 2013). As such, perhaps misclassifying articles involving scandals would have a more detrimental downstream impact. On the other hand, erroneous election-related coverage such as fake polling results and endorsements also affect voter behavior (Van der Meer, Hakhverdian, and Aaldering 2016). It is important to note one caveat. The ground truth labels are provided at the outlet level. It is entirely possible that fake news sites generally publish accurate election-related articles. Regardless, our analysis shows that one might need to adjust prediction w.r.t. the proportions of topics covered to determine the quality of news outlets. Further, we highlight that researchers invested in building effective models should consider the impact/weight of different types of misclassifications (e.g., misclassifying a credible celebrity gossip piece as fake news may be harmless, but incorrectly labeling vaccine-related misinformation as credible is harmful).

Performance and Bias Trade-off

In this section, we examine error bias and performance trade-off with respect to i) domain ideology, ii) domain age, iii) domain popularity, iv) external events and v) article topic [5]. In other words, we ask if a classifier has the highest $F1$ (or, AUC) score, but significantly favors liberal-leaning news sites, does there exist another classifier which has a slightly worse $F1$(or, AUC) score but lower or no significant bias with respect to domain ideology?

Here, we first rank all classifiers based on their average $F1$ scores for the validation datasets, and denote the one with the highest $F1$ as $c\star$. We then add models with worse $F1$ scores but lower bias (with respect to domain ideology, domain age, etc.) than $c\star$ into the set C_{alt} (these are the alternative options). Models that have worse $F1$ scores in addition to higher bias when compared to $c\star$ are excluded from analysis. Next, we plot the $F1$ scores and error bias measurements of $c\star$ and C_{alt} in Figure 5. Here, the z-axis denotes $F1$ scores, the x-axis and y-axis denotes fnr and fpr based bias respectively. That is, the x-axis and y-axis values are equivalent to the mean differences calculated in Section and summarized in Figure 3. A higher absolute value of x and/or y implies a higher bias. We note that results for event shocks are omitted given that none of the alternative models provides a notable reduction in bias.

As shown in Figure 5a which focuses on ideological bias, *HOAX* has the highest $F1 = 0.42$, yet it has $fpr = -0.40$ and $fnr = 0.57$. In other words, *hoax* erroneously classifies articles published by mainstream conservative-leaning news sites as fake news 40% more often compared to articles by mainstream liberal-leaning publishers. Further, it also misclassifies articles by liberal-leaning fake news sites as credible 57% more often compared to articles by conservative-leaning fake sites. We can reduce this bias by choosing the alternative `NELA`, which has $F1 = 0.39$, $fpr = -0.10$, and $fnr = 0.0$. In other words, by trading a small reduction of $F1$, we can significantly reduce ideological bias. Focusing on Figure 5b, we see that trade a small drop in $F1$ from 0.42 to 0.39 can lead to a modest reduction in domain age-based bias. For domain popularity (Figure 5c) and article topic (Figure 5d), however, any reduction in bias requires a substantial drop in performance.

Discussion

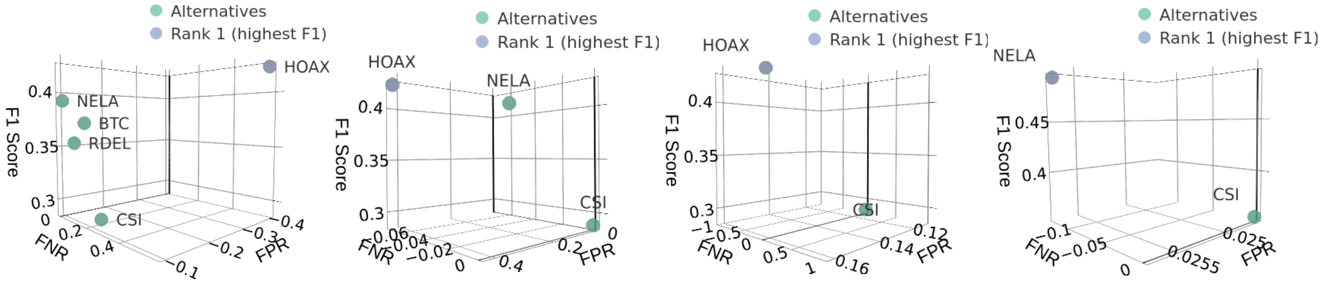
We reviewed 23 existing fake news classification models and provided a comprehensive overview of the current state of this research field. Furthermore, by reaching out to the authors of 17 papers, we collected a representative set of 5 classifiers that we used for additional performance evaluation and error analysis. The results reveal important concerns about generalizability. Performance of fake news classifiers varies significantly from one dataset to another, from one evaluation archetype to another, and from one evaluation metric to another. We also observed important bias: articles from small and/or conservative-leaning mainstream sites, for example, were more often labeled incorrectly as fake news. Furthermore, we also showed that model error rate varies across different topics. Finally, we illustrated that, in some cases, we can trade the model that has the best overall performance but high-bias with another that has a slightly worse performance but a substantially lower bias.

There are several limitations to our work. First, we used the list of fake and mainstream news sites provided by *Media Bias Fact Check*. Yet, many other sources such as *News Guard* provide domain-level veracity labels. Related work (Bozarth, Saraf, and Budak 2020) has highlighted that the choice in ground truth labels affects downstream observations. As such, future work should evaluate models using different ground truth of fake and mainstream news sites to ensure robustness. Similarly, certain sources (Zimdars, M 2018) also partition fake news domains into more fine-grained subcategories (e.g. junk science, state-sponsored misinformation sites, clickbait). Understanding whether existing models perform better on some subcategories compared to others can also provide valuable insights into potential model bias and weaknesses.

Despite these limitations, we believe this work takes an important step towards reproducibility and replicability in fake news classification. We invite other scholars to build on this effort and help collectively build towards a well-formulated and comprehensive evaluation framework for fake news detection. As a very first step, we argue that our community needs to make datasets and code repositories available to others. In our case, we were able to acquire only 5 out of the 17 published classifiers. The code repositories for the 5 classifiers and our own code is available at https://github.com/lbozarth/fakenewsMPE. Easier access would allow our community to compare and contrast different datasets and algorithms. Such access would also enable our community to develop a robust evaluation framework more quickly.

A simple guideline on model evaluation: Here, we provide a simple checklist for researchers interested in building robust and effective models. (1) Models should be evaluated using multiple datasets and ground truth labels provided by various parties and may be of varying granularity.

[5] We focus on the topics *scandal* and *election* given that mean differences in error rates between the pair are the highest.



(a) Domain Ideology. Trade performance to reduce bias that favors *liberal-leaning* sites compared to *conservative* sites.

(b) Domain Age. Trade performance to reduce bias that favors *mature* sites compared to *recently created* sites.

(c) Domain Popularity. Trade performance to reduce bias that favors *large* sites compared to *small* sites.

(d) Article Topic. Trade performance to reduce bias that favors articles about *election* compared to those about *scandal.*

Figure 5: Error Bias and Performance Trade-off. Here, the z-axis denotes F1 scores, the x-axis and y-axis denote false-positive and false-negative error-based bias respectively. Further, the model $c\star$ with highest $F1$ score is colored in blue, and alternative models with worse $F1$ scores but lower bias than $c\star$ are in green.

(2) Individuals interested in creating a model with the intent of long-term use should adopt variations of the *forecast* archetype to train and validate the model. Similarly, individuals should use the *bydomains* archetype if the aim is to label never-before-encountered websites with unknown factualness. (3) Focusing on metrics, researchers can also consider using the precision-recall curve (Boyd, Eng, and Page 2013) to optimize precision over recall (or vise versa) if a model's particular usage warrants it (e.g., favoring precision over recall if the impact of a false-positive is high). (4) Ingrained prejudices are shown to be present in machine learning models used in real-world applications (O'neil 2016). When such biases come to light, they can significantly raise the public's distrust in automatons. As such, using domain-expertise to identify potential high-cost biases and evaluate the model for these biases are crucial. Finally, (5) if the highest-performing model is indeed biased, researchers should consider a trade-off between maximizing overall performance and reducing potential biases.

We note that this guide is far from being a comprehensive framework. Indeed, it is commonplace machine learning best-practice guidance. Yet, our analysis revealed that these simple guides have yet to be fully followed in this important area of research so far.

Acknowledgement

This research was partly supported by Michigan Institute for Data Science (MIDAS) at the University of Michigan and the National Science Foundation (Grant IIS-1815875).

References

Ahmed, H.; Traore, I.; and Saad, S. 2017. Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques. In Traore, I.; Woungang, I.; and Awad, A., eds., *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments*, Lecture Notes in Computer Science, 127–138. Springer International Publishing.

Alimonti, K. R., and Veridiana. 2018. "fake news" offers latin american consolidated powers an opportunity to censor opponents.

Balmas, M. 2014. When fake news becomes real: Combined exposure to multiple news sources and political attitudes of inefficacy, alienation, and cynicism. *Communication Research* 41(3):430–454.

Benkler, Y.; Faris, R.; and Roberts, H. 2018. *Network propaganda: Manipulation, disinformation, and radicalization in American politics*. Oxford University Press.

Bode, L.; Budak, C.; Ladd, J. M.; Newport, F.; Pasek, J.; Singh, L. O.; Soroka, S. N.; and Traugott, M. W. 2020. *Words That Matter: How the News and Social Media Shaped the 2016 Presidential Campaign*. Washington, D.C.: Brookings Institution Press.

Boididou, C.; Papadopoulos, S.; Zampoglou, M.; Apostolidis, L.; Papadopoulou, O.; and Kompatsiaris, Y. 2018. Detection and visualization of misleading content on Twitter. *International Journal of Multimedia Information Retrieval* 7(1):71–86.

Bourgonje, P.; Schneider, J. M.; and Rehm, G. 2017. From Clickbait to Fake News Detection: An Approach based on Detecting the Stance of Headlines to Articles. In *NLPmJ@EMNLP*.

Boyd, K.; Eng, K. H.; and Page, C. D. 2013. Area under the precision-recall curve: point estimates and confidence intervals. In *Joint European conference on machine learning and knowledge discovery in databases*, 451–466. Springer.

Bozarth, L.; Saraf, A.; and Budak, C. 2020. *Higher Ground? How Groundtruth Labeling Impacts Our Understanding of Fake News About the 2016 U.S. Presidential Nominees*. ICWSM 2020.

Budak, C.; Agrawal, D.; and El Abbadi, A. 2011. Limiting the spread of misinformation in social networks. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, 665–674. New York, NY, USA: ACM.

Budak, C. 2019. What happened? the spread of fake news publisher content during the 2016 u.s. presidential election. WWW '19.

Castelo, S.; Almeida, T.; Elghafari, A.; Santos, A.; Pham, K.; Nakamura, E.; and Freire, J. 2019. A Topic-Agnostic Approach for Identifying Fake News Pages. *Companion Proceedings of The 2019 World Wide Web Conference on - WWW '19* 975–980. arXiv: 1905.00957.

Coletto, M.; Garimella, K.; Gionis, A.; and Lucchese, C. 2017. A Motif-based Approach for Identifying Controversy. *arXiv:1703.05053 [cs]*. arXiv: 1703.05053.

Corbett, P. 2017. Why the times calls trump 'mr.' (no, we're not being rude).

Deligiannis, N.; Huu, T. D.; Nguyen, D. M.; and Luo, X. 2018. Deep Learning for Geolocating Social Media Users and Detecting Fake News.

Do, T. H.; Luo, X.; Nguyen, D. M.; and Deligiannis, N. 2019. Rumour Detection via News Propagation Dynamics and User Representation Learning. *arXiv:1905.03042 [cs, stat]*. arXiv: 1905.03042.

Fletcher, R.; Cornia, A.; Graves, L.; and Nielsen, R. K. 2018. Measuring the reach of "fake news" and online disinformation in europe. *Reuters Institute Factsheet*.

Gibbons, J. D., and Chakraborti, S. 1991. Comparisons of the mann-whitney, student'st, and alternate t tests for means of normal distributions. *The Journal of Experimental Education* 59(3):258–267.

Gravanis, G.; Vakali, A.; Diamantaras, K.; and Karadais, P. 2019. Behind the cues: A benchmarking study for fake news detection. *Expert Systems with Applications* 128:201–213.

Gross, N., and Simmons, S. 2006. Americans' views of political bias in the academy and academic freedom. In *annual meeting of the American Association of University Professors*.

Guacho, G. B.; Abdali, S.; Shah, N.; and Papalexakis, E. E. 2018. Semi-supervised Content-based Detection of Misinformation via Tensor Embeddings. *arXiv:1804.09088 [cs, stat]*. arXiv: 1804.09088.

Gupta, A.; Lamba, H.; Kumaraguru, P.; and Joshi, A. 2013. Faking Sandy: Characterizing and Identifying Fake Images on Twitter During Hurricane Sandy. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13 Companion, 729–736. New York, NY, USA: ACM.

Hochberg, Y. 1988. A sharper bonferroni procedure for multiple tests of significance. *Biometrika* 75(4):800–802.

Horne, B. D.; Dron, W.; Khedr, S.; and Adali, S. 2018. Assessing the News Landscape: A Multi-Module Toolkit for Evaluating the Credibility of News. In *Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW '18*, 235–238. Lyon, France: ACM Press.

Horne, B. D.; Nørregaard, J.; and Adali, S. 2019. Robust fake news detection over time and attack. *ACM Transactions on Intelligent Systems and Technology (TIST)* 11(1):1–23.

Huang, J., and Ling, C. X. 2005. Using auc and accuracy in evaluating learning algorithms. *IEEE Transactions on knowledge and Data Engineering* 17(3):299–310.

Jin, Z.; Cao, J.; Jiang, Y.-G.; and Zhang, Y. 2014. News Credibility Evaluation on Microblog with a Hierarchical Propagation Model. In *2014 IEEE International Conference on Data Mining*, 230–239. ISSN: 2374-8486.

Katsaros, D.; Stavropoulos, G.; and Papakostas, D. 2019. Which machine learning paradigm for fake news detection? In *2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, 383–387. ISSN: null.

Kucharski, A. 2016. Post-truth: Study epidemiology of fake news. *Nature* 540(7634):525.

Kwok, Y. 2017. Where memes could kill: Indonesia's worsening problem of fake news. *Time, January* 6.

Liu, Y., and Wu, Y.-F. B. 2018. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Long, Y.; Lu, Q.; Xiang, R.; Li, M.; and Huang, C.-R. 2017. Fake News Detection Through Multi-Perspective Speaker Profiles. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 252–256. Taipei, Taiwan: Asian Federation of Natural Language Processing.

Ma, J.; Saul, L. K.; Savage, S.; and Voelker, G. M. 2009. Beyond blacklists: learning to detect malicious web sites from suspicious urls. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1245–1254. ACM.

Ma, J.; Gao, W.; and Wong, K.-F. 2018. Rumor Detection on Twitter with Tree-structured Recursive Neural Networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1980–1989. Melbourne, Australia: Association for Computational Linguistics.

Main, T. J. 2018. *The Rise of the Alt-Right*. Brookings Institution Press.

Mueller, M., and Chango, M. 2008. Disrupting global governance: the internet whois service, icann, and privacy. *Journal of Information Technology & Politics* 5(3):303–325.

Nørregaard, J.; Horne, B. D.; and Adalı, S. 2019. Nela-gt-2018: A large multi-labelled news dataset for the study of misinformation in news articles. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, 630–638.

O'neil, C. 2016. *Weapons of math destruction*. Broadway Books.

Papanastasiou, F.; Katsimpras, G.; and Paliouras, G. 2019. Tensor Factorization with Label Information for Fake News Detection. *arXiv:1908.03957 [cs]*. arXiv: 1908.03957.

Paresh, D. 2016. Without these ads, there wouldn't be money in fake news.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research* 12(Oct):2825–2830.

Praino, R.; Stockemer, D.; and Moscardelli, V. G. 2013. The lingering effect of scandals in congressional elections: Incumbents, challengers, and voters. *Social Science Quarterly* 94(4):1045–1061.

Qazvinian, V.; Rosengren, E.; Radev, D. R.; and Mei, Q. 2011. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the conference on empirical methods in natural language processing*, 1589–1599. Association for Computational Linguistics.

Riedel, B.; Augenstein, I.; Spithourakis, G. P.; and Riedel, S. 2017. A simple but tough-to-beat baseline for the Fake News Challenge stance detection task. *arXiv:1707.03264 [cs]*. arXiv: 1707.03264.

Rubin, V. L. 2018. Semi-supervised Content-based Fake News Detection using Tensor Embeddings and Label Propagation.

Ruchansky, N.; Seo, S.; and Liu, Y. 2017. CSI: A Hybrid Deep Model for Fake News Detection. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management - CIKM '17* 797–806. arXiv: 1703.06959.

Shao, C.; Ciampaglia, G. L.; Varol, O.; Flammini, A.; and Menczer, F. 2017. The spread of fake news by social bots. *arXiv preprint arXiv:1707.07592* 96–104.

Shu, K.; Cui, L.; Wang, S.; Lee, D.; and Liu, H. 2019a. dEFEND: Explainable Fake News Detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining - KDD '19*, 395–405. Anchorage, AK, USA: ACM Press.

Shu, K.; Zhou, X.; Wang, S.; Zafarani, R.; and Liu, H. 2019b. The Role of User Profile for Fake News Detection. *arXiv:1904.13355 [cs]*. arXiv: 1904.13355.

Shu, K.; Wang, S.; and Liu, H. 2018. Understanding User Profiles on Social Media for Fake News Detection. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 430–435. ISSN: null.

Silverman, C. 2017. The fake news watchdog.

SimilarWeb. 2019. Website Demographic Data.

Singal, H., and Kohli, S. 2016. Trust necessitated through metrics: estimating the trustworthiness of websites. *Procedia Computer Science*.

Soares, I. 2019. The fake news machine: Inside a town gearing up for 2020.

Starbird, K. 2017. Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on twitter. In *ICWSM*, 230–239.

Tacchini, E.; Ballarin, G.; Della Vedova, M. L.; Moret, S.; and de Alfaro, L. 2017. Some Like it Hoax: Automated Fake News Detection in Social Networks. *arXiv:1704.07506 [cs]*. arXiv: 1704.07506.

Thelwall, M. 2017. The heart and soul of the web? sentiment strength detection in the social web with sentistrength. In *Cyberemotions*. Springer. 119–134.

Van der Meer, T. W.; Hakhverdian, A.; and Aaldering, L. 2016. Off the fence, onto the bandwagon? a large-scale survey experiment on effect of real-life poll outcomes on subsequent vote intentions. *International Journal of Public Opinion Research* 28(1):46–72.

Van Zandt. 2018. Media bias/fact check (mbfc news) about.

Volkova, S.; Shaffer, K.; Jang, J. Y.; and Hodas, N. 2017. Separating Facts from Fiction: Linguistic Models to Classify Suspicious and Trusted News Posts on Twitter. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 647–653. Vancouver, Canada: Association for Computational Linguistics.

Wang, Y.; Ma, F.; Jin, Z.; Yuan, Y.; Xun, G.; Jha, K.; Su, L.; and Gao, J. 2018. EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, 849–857. New York, NY, USA: ACM. event-place: London, United Kingdom.

Wingfield, N.; Isaac, M.; and Benner, K. 2016. Google and facebook take aim at fake news sites. *The New York Times*.

Yang, S.; Shu, K.; Wang, S.; Gu, R.; Wu, F.; and Liu, H. 2019. Unsupervised fake news detection on social media: A generative approach. In *Proceedings of 33rd AAAI Conference on Artificial Intelligence*.

Zhang, J.; Dong, B.; and Yu, P. S. 2019. FAKEDETECTOR: Effective Fake News Detection with Deep Diffusive Neural Network. *arXiv:1805.08751 [cs, stat]*. arXiv: 1805.08751.

Zhou, X., and Zafarani, R. 2018. Fake News: A Survey of Research, Detection Methods, and Opportunities. *arXiv:1812.00315 [cs]*. arXiv: 1812.00315.

Zhou, X., and Zafarani, R. 2019. Network-based fake news detection: A pattern-driven approach. *ACM SIGKDD Explorations Newsletter* 21(2):48–60.

Zimdars, M. 2018. False, misleading, clickbait-y, and/or satirical "news" sources. Accessed: 2018-01-30.