Robust Fake News Detection Over Time and Attack

BENJAMIN D. HORNE, Rensselaer Polytechnic Institute, USA
JEPPE NØRREGAARD, Technical University of Denmark, Denmark
SIBEL ADALI, Rensselaer Polytechnic Institute, USA

In this study, we examine the impact of time on state-of-the-art news veracity classifiers. We show that, as time progresses, classification performance for both unreliable and hyper-partisan news classification slowly degrade. While this degradation does happen, it happens slower than expected, illustrating that hand-crafted, content-based features, such as style of writing, are fairly robust to changes in the news cycle. We show that this small degradation can be mitigated using online learning. Last, we examine the impact of adversarial content manipulation by malicious news producers. Specifically, we test three types of attack based on changes in the input space and data availability. We show that static models are susceptible to content manipulation attacks, but online models can recover from such attacks.

CCS Concepts: • **Information systems** → **World Wide Web**; • **Computer systems organization** → Maintainability and maintenance; Reliability;

Additional Key Words and Phrases: Fake news, biased news, misleading news, fake news detection, misinformation, disinformation, concept drift, robust machine learning, adversarial machine learning

ACM Reference format:
Benjamin D. Horne, Jeppe Nørregaard, and Sibel Adali. 2019. Robust Fake News Detection Over Time and Attack. *ACM Trans. Intell. Syst. Technol.* 11, 1, Article 7 (December 2019), 23 pages.
https://doi.org/10.1145/3363818

1 INTRODUCTION

Preventing the spread of false and misleading news has become a top priority for researchers and practitioners. This rise in relevance is due to the alarmingly high societal cost of misinformation [37, 39] and the scale at which it spreads on social media platforms [2, 56]. As Lewandowsky et al. stated in their seminal paper, Misinformation and Its Correction, "democracy relies on a well-informed populace [37]." If even a fraction of the population is misinformed, then sociopolitical decisions can be made that are contrary to what is best for the public. A clear example of this is the rise of preventable diseases due to false claims that vaccinations lead to autism [37] or the increasing opposition to policies that address climate change despite little to no benefit to members of society [53]. This cost has become even more prevalent with the popularity of social networks, where news can spread without the information being vetted by a trained journalist [2, 39]. Information on social media is often consumed passively, hence mental shortcuts are often used

Authors' addresses: B. D. Horne, Rensselaer Polytechnic Institute, 110 8th Street Troy NY, 12180, USA; email: horneb@rpi.edu; J. Norregaard, Technical University of Denmark, Anker Engelunds Vej 1 Building 101A, 2800 Kgs. Lyngby, Denmark; email: jepno@dtu.dk; S. Adali, Rensselaer Polytechnic Institute 110 8th Street Troy NY, 12180, USA; email: adalis@rpi.edu.

2157-6904/2019/12-ART7 $15.00
https://doi.org/10.1145/3363818

to assess information, making this unchecked information spread even more dangerous. Further, users can be trapped in like-minded "echo chambers," where both hyper-partisan news and conspiracy-based news can be normalized [30, 39]. If these isolated social groups are not enough for malicious news producers, then bots can be introduced to spread news to the general public and make information appear more widely believed than it really is [18].

Due to these concerns, many methods have been proposed to automatically detect or approximate the veracity of news. These methods are based on various signals including article content, source meta-data, external fact-checkers, and crowd sharing behavior. In experimental settings, content-based methods have been shown to be highly accurate in detecting various types of news, including hyper-partisan news [4, 10, 21, 28] and unreliable news [4, 25, 44, 45, 49]. Specifically, Horne et al. [28] achieved near 0.90 ROC AUC on a large weakly labeled set of news articles that come from unreliable and biased news sources using only content-based features. These features are not dependent on the topic of an article, making them fairly robust. Similarly, Baly et al. [4] obtained high performance in detecting articles that come from unreliable and biased news sources using features from Wikipedia, Twitter, and article content. Popat et al. [44] achieved 0.80 ROC AUC using both content and source level features on textual claims. These studies and more lead us to conclude that news sources that produce biased, misleading, or false claims generate content that is easily distinguished from well-established mainstream news sources. Given the strength of these methods, the main question we ask in this article is the following:

"Do the content-based differences between mainstream, unreliable, and hyper-partisan sources change over time?"

The likely answer to this is "Yes," given the highly dynamic nature of the news cycle and the changing attention of news consumers. News topics change dramatically over time due to fast pace of external events. In addition, the need for engagement (often called the attention economy [7]) impacts which stories are prioritized and how they are covered by media sources. These externalities likely lead to a steady drift of the style in which news is reported. This notion could be considered a"natural" drift with the content in news cycle.

In addition to this natural drift, there may be a more "artifical" drift due to the adversarial environment in which sources spreading misinformation and biased information operate. Such sources have two competing objectives: to increase the effectiveness of their message while avoiding detection of any algorithmic solutions in place to detect their existence. In essence, it is possible for such sources to produce content that is indistinguishable from mainstream news in terms of writing style and language. However, this content change is likely a costly task in multiple ways. Such a task requires these sources to hire news editors or develop algorithms designed to explicitly evade classification methods. A second cost to such evasion methods is the loss in the effectiveness of the underlying message. Past research shows us that highly engaging, viral social-media content and conspiracy theories have similar properties: They are written with simpler language, are more negative in tone [25, 57], and introduce novel ideas [54]. As a result, we expect certain features of misinformation to change very little.

Given that any misinformation detection method needs to operate in such a complex environment, where potentially both natural and artificial changes in the news can take place, it is not clear how such methods should be designed and maintained over time. It is likely that both natural and artificial changes in the news can shift the input data distribution, causing previously trained models to perform sub-optimally. This notion is generally referred to as *concept drift* in machine-learning literature. In this article, we begin to fill this important gap in the literature. Specifically, we ask the following research questions:

Q1 Does concept drift impact the performance of state-of-the-art classifiers?
Q2 If concept drift does impact these classifiers, then can we reduce that impact?

Q3 How generalizable are distant labeled models on sources that have not been seen in training?
Q4 How do attacks from malicious news producers impact the learning algorithm over time?

We first examine the effectiveness of state-of-the-art, distant labeled (or weakly labeled) classifiers over a time period of 10 months with multiple types of sources. We identify to which degree the effectiveness of classifiers diminish across different feature groups. Then, we examine how well this drift can be remedied with different types of retraining methods. To the best of of our knowledge, this work is the first explore the impact of concept drift on "fake news" classifiers. We find that the performance of content-based models does decrease over time, but this decrease is considerably slower than hypothesized. While individual feature groups, such as the style of writing or the affect of writing, can be inconsistent over time, the combination of many content features is relatively stable. These models become even more stable over time when using an online learning method, taking advantage of the weak labeling used in the state-of-the-art models.

We then consider how well distant labeled models generalize by testing each model on unseen news sources over time, simulating a real-life scenario where newly founded sources may emerge. We find that, on average, classifiers trained on one set of sources and tested on another have a slight decrease in performance compared to those tested on a traditional article test set. However, this decrease is minimal and the average performance is consistent over time.

Last, we consider an adversarial setting where the sources providing unreliable content are explicitly manipulating their content to evade classification. We consider methods that are easily implementable by such sources and discuss their effectiveness in evasion. We also discuss the potential cost of each attack in terms of the potential loss of engagement in the content. To the best of of our knowledge, this work is the first to simulate adversarial attacks on classifiers by malicious news producers. Surprisingly, we find that content-based methods are fairly robust to attacks on the input space if an online or incremental model is used. Yet, we also find that if access to malicious news data is blocked, the performance of each classifier suffers greatly, demonstrating the need for continually collected data from a broad variety of alternative news sources.

We then conclude the article with suggestions for the design of automated misinformation classification methods in the light of our findings. Overall, we find that content-based methods are robust to both natural and artificial changes in the news cycle when online or incremental learning is used. Saliently, these online learning methods require source-level labels to work in a timely manner. While many researchers have been focused on finding article-level granularity in labeled data, we show that articles from extreme sources (extremely unreliable or extremely biased) can be robustly detected using a much less costly labeling method.

2 RELATED WORK

Just as many of the previous works have done [4, 28], we focus on two dimensions of veracity: the reliability of news and the bias in news. While both concepts can be highly related, they represent two distinct ways the news can misinform consumers. A reliable source tends to report factually correct and properly vetted information, while a biased source tends to report unbalanced or unnecessarily partisan information. For example, an unreliable article may contain fabricated information or partially false information, typically in a malicious manner. In contrast, a biased (hyper-partisan or extremely biased) article may contain true information that is presented partially, in a misleading way, or decontextualized by subjective opinions [16, 17, 19]. The two concepts are often mixed in news articles, as one may motivate the other (e.g., political partisan motivating a false claim) [39, 45].

Automated methods have been designed to detect both concepts in news. Horne et al. focused on detecting reliability and bias in news articles using features based on writing style and language use [28]. Baly et al. provided a similar analysis of automatic classification of reliability and bias, but focused on source-level rather than article-level classification [4]. Several other works have focused on more granular ground truth, such as fact-checked news articles or claims. Horne and Adalı explored the content feature space of fact-checked fake and real articles, as well as satire [25]. Likewise, Ahmed et al. built fake news classifiers using n-gram features on the article content [1], and Potthast et al. employed a meta-learning approach on fact-checked fake news articles produced by hyper-partisan news sources [45]. Focusing on claims rather than articles, Popat et al. used fact-check claims as ground truth for distant supervised credibility detection using both content features and source features [44]. Singhania et al. [49] obtained 96% accuracy detecting fake news articles using a deep learning model based on words, sentences, and the news headline. The experimental setting used fact-checked fake and real articles from politifact.com rather than using weak labeling. Similarly, Wang explored claim-based classification using deep learning on labeled claims from politifact.com [55]. These labels included varying levels of veracity: *pants-fire, false, barely true, half-true, mostly true,* and *true*. In addition to these machine learning methods, there have been several works that utilize knowledge graphs to automatically fact-check claims [11, 14, 23]. While fact-checking-based methods, knowledge graphs, or learning from fact-checked articles offer effective solutions, they do not scale well to methods needing continuous retraining, do not capture novel information well, and do not capture the article bias properly. In all of these studies, no matter the labeling granularity, the most successful method for news classification has been supervised machine learning models using features extracted from the article content, with accuracy ranging from 70% to well over 90% in each lab setting.

Despite the recent success in automated news veracity classification, there has been no work on how robust these methods are over time. If these automated methods were placed in a real-life setting, then they would need to be resilient to various changes in the news cycle as well as adversarial efforts to evade classification.

This concern of algorithm performance over time is not unique to the context of news veracity. There are many general works in machine learning that focus on the learned target changing in unforeseen ways as time progresses. This notion is commonly referred to as concept drift [48]. Mathematically, concept drift can be defined as follows [20]:

Let $p_t(x, y)$ be the joint distribution of target variable y and input space x, for a given time t. *Concept drift* between times t_0 and t_1 can be formally defined as a change in this joint distribution:

$$\exists x \ : \ p_{t_0}(x, y) \neq p_{t_1}(x, y). \tag{1}$$

By expressing the joint distribution as $p_t(x, y) = p_t(y \mid x)p_t(x)$, concept drift can be distinguished into two types:

$$p_{t_0}(y \mid x) \neq p_{t_1}(y \mid x) \qquad \text{Real concept drift, the target changes over time.} \tag{2}$$

$$p_{t_0}(x) \neq p_{t_1}(x) \qquad \text{Virtual drift, the input space changes over time.} \tag{3}$$

It is likely that both types of drift happen at the same time [20], and that these changes are unpredictable. In content-based news classification, virtual drift[1] is more likely to happen than real drift, as changes in topic, entities, or writing style will change the input space. However, real concept drift in news would mean malicious news producers begin producing proper news articles or vice versa.

[1]We use the terminology from Reference [20], but this has also been referred to as covariate shift, sampling shift, and temporary drift.

There are some commonly studied methods to reduce concept drift. In general, concept drift can be combated through retraining the machine learning algorithm as time progresses. However, depending on the speed and intensity of the drift, how often retraining should occur and how much of previous data should be remembered or forgotten can change [20]. Another commonly proposed technique for handling concept drift is online ensemble learning, where multiple decision makers vote for the classification. Minku et al. proposed online bagging for ensemble machine learning algorithms to handle concept drift [41], illustrating the usefulness of ensemble diversity in handling sudden target changes. Kolter and Maloof introduced an intelligent ensemble method for handling concept drift called Dynamic Weighted Majority (DWM) [34]. This method monitors the performance of each base classifier in an ensemble and adjusts their weights depending on their performance. When a base learner's weight decreases to below a given threshold the learner is removed. Likewise new base learners are created when the ensemble overall makes incorrect classifications. Thus the DWM has a dynamic number of learners in the ensemble. Several other researchers propose algorithm specific ways to handle concept drift, including using Support Vector Machines (SVM) [33] and application-specific case-based systems [15].

These methodological works illustrated viable ways to handle concept drift on synthetic data sets but do not address the higher uncertainty of drift in many real-life problems. In particular, there is no work that investigates concept-drift in the misinformation detection problem and examines the effectiveness of methods in the presence of adversarial manipulation of content. This is the topic of this article.

3 DATA

To understand news classifiers' performance over time, we use the NELA-GT-2018[2] dataset, which is a political news article data covering 10 months in 2018 [42]. The dataset contains 194 sources in both mainstream and alternative media from multiple countries. In this study, we only use sources that cover U.S. political news, removing the potentially confounding variable of writing styles across countries.

Just as previous studies have done [4, 25, 28, 44], we utilize distant labeling (or weak labeling) for our machine learning classifiers by labeling articles based on the sources that published them. Hence, we start our data extraction by identifying sources that fall into three categories: unreliable (UR), biased (B), and mainstream (M). We identify these sources using two sets of labels provided in Reference [42], namely, NewsGuard[3] and Open Sources.[4] NewsGuard uses a group of trained and experienced journalists to assess credibility and transparency of news sources based on a strictly developed rating process. NewsGuard is transparent about who is rating news sources, the process of rating news sources, and funding. Open Sources ratings are similarly done by a group of experts and the criteria for source labels is clearly available. Both have been used in previous studies [4, 5, 28, 30, 31]. For more information on how NewsGuard and Open Sources label media sources, we ask that you refer to information provided in Reference [42].

In this study, the mainstream category contains sources that have a credibility score above 90 according to NewsGuard, meaning they do not repeatedly publish false content, gather and present information responsibly, and handle the difference between news and opinion responsibly. The unreliable category contains sources were marked as repeatedly publishing false content by NewsGuard or marked as unreliable/conspiracy/fake by Open Sources. Last, the biased category contains sources that were marked as not handling the difference between news and opinion

[2] https://dataverse.harvard.edu/dataverse/nela.
[3] www.newsguardtech.com.
[4] opensources.co.

Table 1. Sources Used in Data Set Construction

(M) Mainstream sources	(UR) Unreliable sources	(B) Biased sources
Reuters	True Pundit	News Busters
NPR	Natural News	Bipartisan Report
USA Today	Infowars	Crooks and Liars
CNN	Veterans Today	Shareblue
The New York Times	Activist Post	Conservative Tribune
CBS News	Mint Press News	The Conservative Tree House
PBS	Waking Times	Delaware Liberal
The Hill	Intellihub	Daily Kos
CNBC	NODISINFO	FrontPage Magazine
Washington Examiner	21st Century News Wire	Freedom Outpost
Mercury News	The Political Insider	The Right Scoop
The Guardian	Newswars	CNS News
Politico	Prison Planet	Palmer Report
The Denver Post	The Gateway Pundit	Western Journal
BBC	The Daily Stormer	Bearing Arms
Chicago Sun-Times	LewRockwell	RedState

Note: while BBC is a British news source, we only extract articles from their U.S. news feed, not the U.K. news feed.

responsibly by NewsGuard or marked as biased by Open Sources. Once sources with these criteria are extracted, we randomly select 16 sources in each of the three categories, for a total of 48 sources. Sources that fall into each category can be found in Table 1.

Note, while we create disjoint categories of unreliable and biased news sources, this does not necessarily mean the bias of an article and the reliability of an article are always mutually exclusive. An article can be both unreliable and biased. However, our goal in creating ground truth sources is to capture the "extreme ends" of each of these two dimensions. For example, according to NewsGuard, *Daily Kos* is a strongly left-leaning blog that may report information in a misleading way due to their hyper-partisan viewpoints, but they are unlikely to report eccentric conspiracy theories like *Infowars*. One "repeatedly reports completely false information" (*Infowars*),[5] while the other "does not present information and opinion responsibly" (*Daily Kos*).[6] This subtle, but important difference is one we want our classifiers to learn. Furthermore, these distinctions have been used in previous automated news veracity studies [4, 28], allowing us to provide some comparison to the literature.

Once each of our source sets is created, we extract every article by each source for a total of 40 weeks between February 1 and November 6 of 2018 from the NELA-GT-2018 dataset. This dataset captures nearly every article published by each source during this time period, as it was collected using live RSS feed scraping. In total our extracted data set contains over 158.5K articles from 48 media sources.

We hypothesize that in addition to natural concept drift in the news cycle, large events may change the separation between reliable news and unreliable news. This hypothesis is supported by the literature that showed an increase in false news spread during the 2016 U.S. Presidential Election [2, 36], illustrating a change in news reporting behavior during an event. Thus, we ensure

[5] https://bit.ly/2sZl0vy.
[6] https://bit.ly/2RuFJBs.

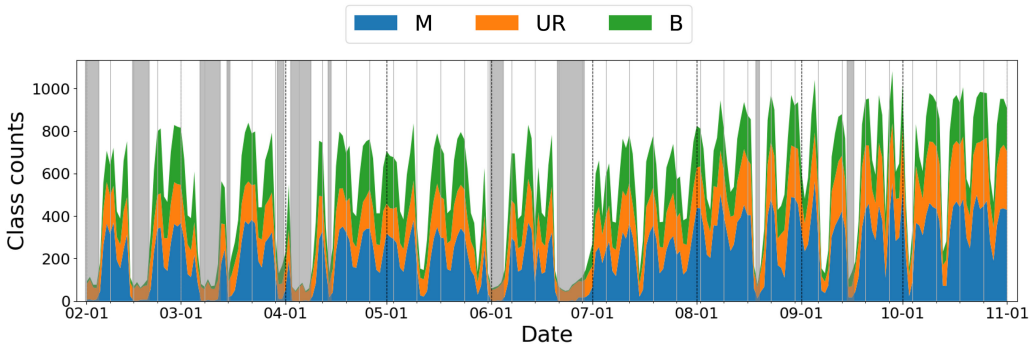


Fig. 1. Above is a stacked histogram of the article counts for the three classes each day (M class on the bottom, UR class in the middle, and B class on the top). There are two conclusions to draw from this graph: (1) The amounts of data are fairly consistent over time, with a slight increase in data toward the end of the timeline. (2) A few regions have very low amount of data, likely due to problems with article scraping. These "low-data" regions have been marked by grey shading in the back ground. These regions will be shown throughout figures in the article to indicate when performance metrics may be influenced by the low data in those regions.

this extracted data set covers time with major events, such as the 2018 U.S. Midterm Elections and the Kavanaugh Hearings.[7]

The number of articles published and scraped each day can vary quite a bit as a result of the news cycle itself and due to occasional problems encountered in the collection of the NELA-GT-2018 dataset, such as changes in a news sources' RSS feed. Figure 1 shows the number of articles extracted from the three classes, M, UR, and B, for each day in the time period. As expected, there is a noticeable fluctuation on a weekly basis due to the news cycle, and there are some regions where only a very small amount of data was collected, likely due to challenges with the news scraper. To show these regions, we have fitted a truncated normal distribution over the number of publications each day for each class and marked out the regions where the number of publications for any class is less than the second percentile. These regions have very little data and can therefore create artificial fluctuations in the performance metrics. In particular, the region 06/21–06/28 may inflate or deflate our performance metrics, as there are eight low-data days in a row, and the training and testing are done on a weekly basis. Keep in mind, our metric of "low data" here does not mean zero data, just significantly less data than expected. As these low data periods are small in number and total length (less than 10% of total time period), we have high confidence in the general trends we report in this article.

4 FEATURES

Using this data set, we extract natural language features from each article. Specifically, we compute the following feature groups from each article's content and title independently:

(1) **Style**—This feature group captures the style and structure of the article. It includes POS (part of speech) tags and simple linguistic features such as number of quotes, punctuation, and all capitalized words. In total this group contains 55 features.
(2) **Complexity**—This feature group captures how complex the writing in the article is. It includes lexical diversity (type-token ratio), reading difficulty, length of words, and length of sentences. In total this group contains 6 features.

[7] https://en.wikipedia.org/wiki/Brett_Kavanaugh_Supreme_Court_nomination.

(3) **Bias**—This feature group captures the overall bias and subjectivity in the writing. This feature group is strongly based on Recasens et al. work [46] on detecting bias language. It includes number of hedges, factives, assertives, implicatives, and opinion words. It also include number of biased words according to the biased word lexicon in Reference [46] and how subjective the text is according to the subjectivity classifiers used in Reference [28]. In total this group contains 13 features.
(4) **Affect**—This feature group captures sentiment and emotion used in the text. It includes LIWC emotion features such as anger, anxiety, affect, and swear words [52]. It also includes positive and negative sentiment measures using VADER sentiment [32]. In total this group contains 12 features.
(5) **Moral**—This feature group is based on Moral Foundation Theory [22] and lexicons used in [38]. While this feature group has been used in previous studies, it has not been shown to perform well in the news setting or capture much meaningful signal. We include this group for completeness in our feature group analysis. In total this group contains 11 features.
(6) **Event**—This feature group captures two concepts: time and location. Specifically, we expect trained journalists to state both the time and location of the event being reported on. Since we assume many bloggers and malicious news producers have little to no formal training or are reporting on unverifiable events on purpose, they may not report the time and location of the event. In total this group contains 2 features: the number of locations in the article and the number of dates or times in the article.
(7) **Wiki**—This feature is was introduced in Reference [4]. Specifically, Baly et al. suggest that if a news source does not have a Wikipedia page, it may not be as credible as one that does. They showed some ability to separate unreliable news from proper news using this feature. In total this group contains 1 binary feature.

In total, we compute 99 features on the content of the article, 97 features on the title of the article, and 1 feature on the source of the article.

These same feature groups, with the exception of the event feature group, have been used in several previous "fake news" studies and have been tested on various granularity of labels, including source-level (distant labels) and document-level (fact-checked article labels) [4, 25, 28, 44]. At both levels, these features have been shown to provide strong signal in news veracity tasks. Further, many of these features have been well-studied in other contexts [32, 38, 43, 46, 52]. The reason for their effectiveness is likely due to many different factors as discussed below.

The Wiki feature captures how well established a source is. For example, many unreliable sources are not well-established enough to have a Wikipedia page. Even though sites can create a page themselves, the length of their Wikipedia history can also be used as a stand-in for this feature. The Event features capture the nature of false claims, which are often vague by default and do not report any verifiable time or location to support the claim. The Affect features have been used to detect viral content on social-media, specifically it has been shown that content that contains high emotion (typically negative emotion) also receives high engagement [27, 47]. Literature on misinformation also shows that many viral conspiracy theories tend to be more negative overall [6, 8, 57]. The Complexity features are based on research in cognition that shows that individuals tend to find information more credible if it is easier to read. The Style features make up the largest group of features in our study. In general, they capture the writing style of a news article, including the use of capitalization to draw attention, the use of exclamation points to express emotionally charged content, and use of quotes. Journalistic venues tend to quote information from outside sources while unreliable sources rarely attribute information to a source. Also the use of

specific pronouns and use of tenses tend to differ significantly among the categories of sources we study, as well as claims structure captured by noun phrases and verb phrases [25]. Most of these differences are due to the nature of the articles: mainstream news articles are targeted toward giving information while unreliable/biased articles tend to personalize the stories, attach emotional overtones, and calls for action to them.

Our past work shows that the differences between reliable and unreliable sources are even more significant when we compare the titles of the articles [25]. Journalistic venues tend to use titles as a way to get consumers to read the articles. They are shorter and do not make complete statements. In contrast, many unreliable article titles are long and make claims about individuals and events. They likely operate on the assumption that many individuals will not read the articles and try to simply convey information through the titles. It is likely that opinions will be formed through passive, repeated exposure to these claims over time.

Similar features have been used to detect clickbait articles as well [12]. Contrary to common perception [13], we find that unreliable or biased articles are not necessarily clickbait articles as some are written with the explicit purpose of disseminating false or misleading information, not getting revenue through clicks. In short, many of these hand-crafted features capture intentional differences between mainstream, unreliable, and biased articles. If these differences diminish, then it may hinder the dissemination of certain claims that rely on cognitive shortcuts to form opinions instead of factual claims and well-formed arguments.

5 DOES CONCEPT DRIFT IMPACT THE PERFORMANCE OF STATE-OF-THE-ART CLASSIFIERS?

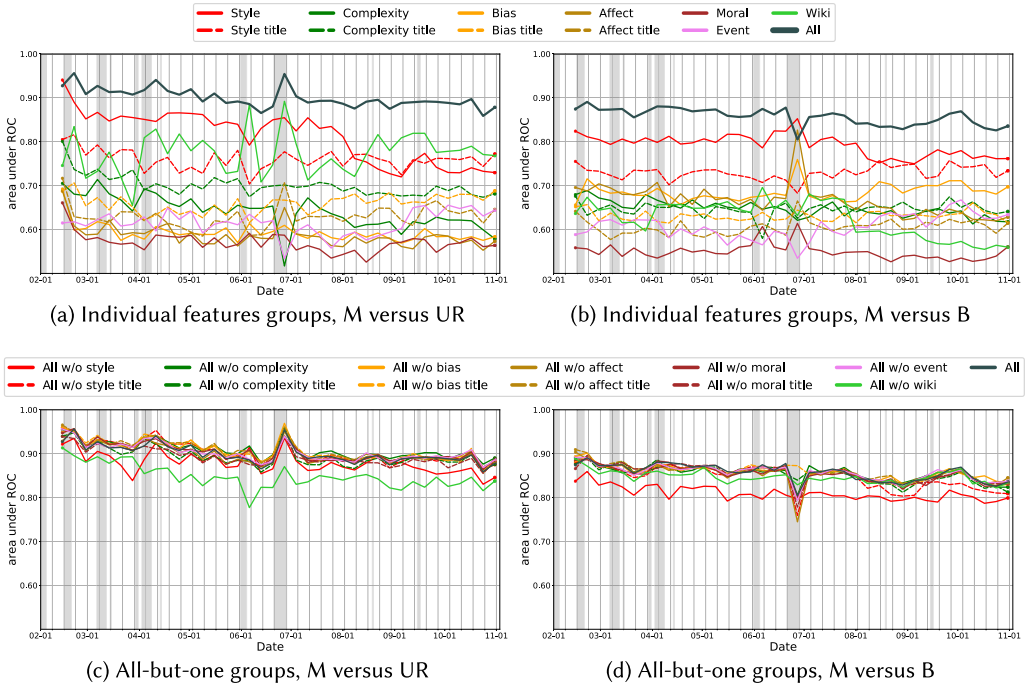
5.1 Testing on Different Feature Groups

Using these previously studied features, we first want to understand how well each feature group works in the classification of different classes of articles and how this performance changes over time. To do this, we trained a model on two weeks of data (first two weeks of February) using each feature group and computed the predictive performance of each model in each week moving forward (from the third week of February to the first week of November). Each week that we test, we compute the average ROC AUC over 20 samples of 20% of the test set. When testing on the same time frame in which the classifier was trained, we test on a random 20% of the data and train on a random 80%.

Each model is built using a Random Forest (RF) classifier with hyper-parameters tuned using 20-fold cross-validation. We choose to use a Decision Tree-based model for several reasons: (1) Decision Trees, specifically Random Forest, have been used successfully in previous news veracity studies [28]. (2) We have found that these hand-crafted features provide the best separation when they are not scaled. Tree-based methods do not require scaling. In smaller studies, distant-based algorithms that require scaling, such as Support Vector Machines (SVM), have provided some signal on similar hand-crafted features [25], but in general, we do not find this to be the case. (3) Enemble methods, like Random Forest, have been shown to handle concept drift better than single classification models [40]. We want our starting baseline to be the best stationary models for handling concept drift.

As previously mentioned, we use distant labeling to train our classifiers. Thus, for example, if an article comes from a source that is deemed unreliable, we label that article as unreliable. Since our goal is to understand the changes in this signal over time due to both natural and artificial concept changes, we also test our classifiers on these distant labels.

Style of writing feature group has the best overall performance. In Figure 2, we show the ROC AUC score for each model over time. Out of each individual feature group, the style features computed



(a) Individual features groups, M versus UR

(b) Individual features groups, M versus B

(c) All-but-one groups, M versus UR

(d) All-but-one groups, M versus B

Fig. 2. In panels (a) and (b), we show a comparison of the ROC AUC performance of each feature groups using a RF classifier. Each feature group uses a different color and features computed on the title use dotted lines. Again, the shaded areas represent time periods with unusually low data. In panels (c) and (d), we show performance of all feature combined while leaving one group out, commonly known as an ablation study. The core conclusions to draw from panels (c) and (d) is not which feature combinations are performing best but which feature combinations perform worse, as most feature combinations perform the same.

on the content of the articles works best overall, performing between 0.94 ROC AUC and 0.74 ROC AUC for classifying M versus UR and between 0.83 and 0.78 for classifying M versus B. This performance is significantly better than other individual feature groups. In Figures 2(c) and 2(d), we see that while each feature group adds some signal to the models, the most important feature groups are the style feature group and the wiki feature group. While this is true for both the unreliable classification task and the bias classification task, the decrease in performance over time is much larger when classifying M versus UR, with a drop of 20 ROC AUC over the 10 months. This drop is much smaller when classifying M versus B, only dropping 5 ROC AUC. This difference in performance over time is likely due to the type of articles being written by UR sources and B sources. For example, the sources in our UR category "break news" often, while the sources in our B category write opinion or discussion pieces on news often. In general, the higher performance from the style feature group for both tasks may not only be caused by the features themselves but also the higher dimensionality of the feature group.

While the style feature group performs best, other feature groups provide a more consistent signal. For both classifiers, we find the complexity of the title and bias in the title provide a consistent signal over time, each only decreasing small amounts over time. When classifying M versus B, we find the affect feature group performs consistently over time. When classifying M versus UR, we find the Wikipedia feature to work well, but very inconsistently over time. Overall, we find that combining all feature groups provides the best performance for both tasks, boosting both the

initial ROC AUC for both classifiers, as well as provide a more consistent signal over time. We will use this combined feature group throughout the rest of the article.

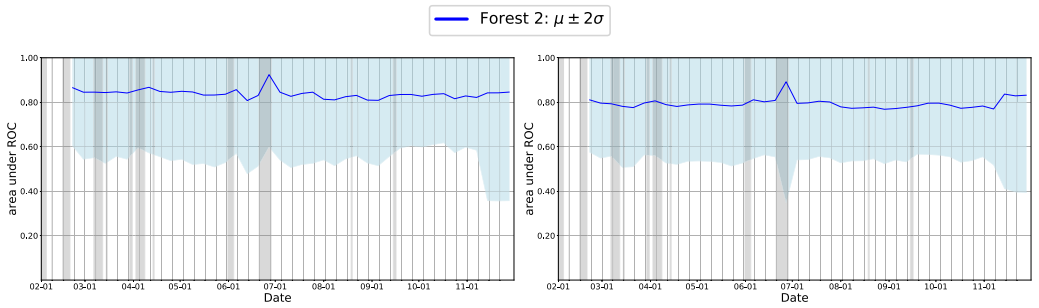
Natural concept drift happens slowly. Just as previous studies [28], at best our classifiers perform close to 0.90 ROC AUC for classifying both unreliable news versus mainstream and for classifying biased news versus mainstream news. Note, this best performance happens when testing in the same time frame in which the classifier was trained. As we test the classifier on each week after the training period, we see a slight degradation in performance. Specifically, after 38 weeks have passed, we see a drop from 0.94 ROC AUC to 0.87 ROC AUC when classifying M versus UR and a drop from 0.87 ROC AUC to 0.83 ROC AUC when classifying M versus B (using all features combined). Surprisingly, when using all features this drop in performance is a fairly smooth trend, with very few major dips or spikes in performance. However, other feature models, particularly those with fewer dimensions, have some major dips and spikes in performance over time. For the most part, these fluctuations coincide with the portions of our data that are smaller (specifically between 06/21 and 06/28). However, there are other large fluctuations in performance that may be caused by the news cycle itself. Overall, concept drift is impacting these models, but the change in performance happens slowly over time, illustrating the robustness of these simple content-based features. This drift happens slower for biased news sources than it does for unreliable news sources.

5.2 Testing on Unseen Sources

To better understand what each model is learning, we perform a secondary test. Specifically, we test each model on news sources the training algorithm has never seen. One concern with high dimensional models is that they may overfitting to the training data. In our case, the models could be overfitting to the specific set of sources we train on rather than the higher level concept that we want to learn: whether an article comes from an unreliable, biased, or mainstream news source. Having more sources represented in training set should mitigate this concern, but in a real-life setting, the algorithm will have a large variety of "out-of-sample" media sources as input, including newly established sources. In fact, new fake news sources often emerge during events of large public interest [51]. If automatic news classification is to be used confidently in a real life system, then it is important to understand how robust our models are when predicting on news sources it has never seen.

To test how well our models work on such never seen sources, we randomly select 10 sources to leave out of training. These 10 sources are then used for testing in each time slice moving forward. Again, we only train on data from the first 2 weeks of February, but only the remaining 22 sources (32 total sources for each classifier M versus UR and M versus B) are used. We perform this leave-10-sources-out scheme for 50 trials and show the mean performance with 2 standard deviations in Figure 3.

On average, the models work well on never seen sources. For both classifiers, we find that the initial mean performance drops from the initial performance when testing on sources the classifier has seen, but this drop is not drastic. When classifying M versus UR, our ROC AUC when testing on seen sources is 0.94 and when testing on not-seen sources is 0.86. When classifying M versus B, our ROC AUC when testing on seen sources is 0.87 and when testing on not-seen sources is 0.81. Interestingly, we see that the drift from the initial performance for both classifiers is actually less than the drift of performance when we have seen all news sources. So, while we have a slightly worse performance overall, that performance is more stable over time. However, we do notice a large range in performance depending on the random sources left out. Looking at the standard deviation in Figure 3, for both classifiers, the lowest performance around 0.40 ROC AUC, while the highest performance is at 1.0 ROC AUC.



(a) M versus UR (b) M versus B

Fig. 3. Cross evaluation of each model on unseen sources. For each trial, we leave out 10 randomly selected test sources and use the remaining sources used for training. Again, we only train on the first 2 weeks of data, and test on each week moving forward. We ran 50 trials and show the mean performance with 2 standard deviations. Note that the high variation is also caused by smaller numbers of articles among some randomly selected sources.

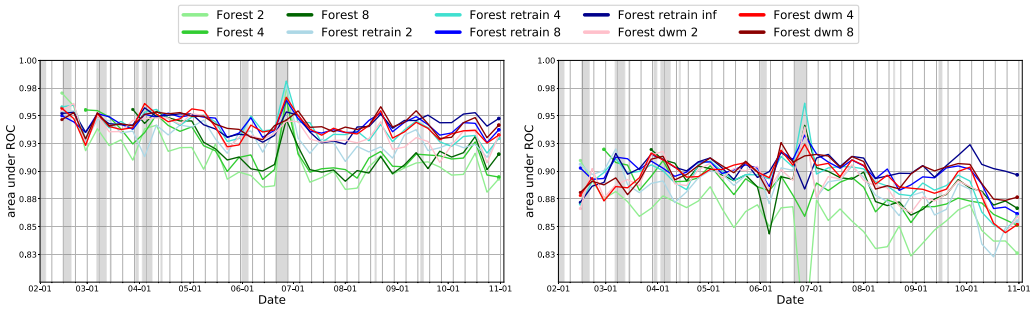
These results illustrate several things. First, on average, each model is learning a more general concept about these sources, rather than the sources themselves, which is precisely what the task requires. Furthermore, these results show that our hand-crafted, content-based features are capturing a higher level behavior across the sources. However, there may be some initial over-fitting to the sources trained on, as we do see a slight drop in overall performance. Second, somewhat unsurprisingly, there is some variation in performance based on the sources left out. A few factors may cause this high variance. First, since each source publishes a different number of articles each week, which can range very widely, each random trial may have a very different number of articles to test on. Some of the highest variance weeks align with our low data dates. Second, the random source selection of each trial may be imbalanced, selecting many more of one class over the other. This imbalance in a single trial may cause a very low performance, as the initial training model may be imbalanced. For example, it may be the case that 1 trial selects 1 M sources and 9 UR source for testing, leaving 15 M sources and 7 UR sources for training. Third, this high variance demonstrates some diversity among the articles written by each source, even within a single class, but not so much diversity as to make the average performance on never-seen sources drastically worse.

6 CAN WE REDUCE THE IMPACT OF CONCEPT DRIFT?

Overall, changes in the news over time showed only a small impact on the classifiers' accuracy. In addition, we are confident the concept we are learning is not simply related to the sources we trained on, but the higher level concept of unreliable and biased news sources. Despite this positive result, as time moves further from the point of training, there is a degrade in performance, albeit a small degrade. We expect as time moves even further from the training point the performance to continue this downward trend. Since we use distant labeling (i.e., source reliability or bias) rather than article level labeling (i.e., fact-checked articles), finding new and timely training data is not very costly. Hence, we can easily utilize online learning methods to reduce this concept drift. We test two general methods:

(1) Online machine learning with varying memory,
(2) Dynamic Weighted Majority with varying memory.

First, we test online learning using Random Forest. We initially trained on 2 weeks of data, but then retrained every week using a memory of past samples for retraining. We vary the memory



(a) M versus UR (b) M versus B

Fig. 4. Comparison of models performance over time. The static forest models are shown in shades of green, the simple retrained forest models are shown in shades of blue, and the dynamically weighted forests (DWM) are shown in shades of red. Notice the gap between the green lines and the other lines. This illustrates an improvement in performance over time by all of the online/retraining models.

from 2, 4, and 8 weeks to a model that uses all known past samples (infinite memory). For example, in the 4 week memory model, we only train the model on a 4-week window. As that window slides, the model forgets examples from over 4 weeks ago and adds the new weeks' examples in training. The interval of retraining and the length of the memory can be used to make the method adapt faster or slower, while also changing the robustness and computational load of the method.

Second, we test a simple variation of Dynamic Weighted Majority (DWM), introduced by Reference [34]. DWM monitors the performance of each base classifier in an ensemble and adjusts their weights depending on their performance. When a base learner's weight decreases to below a given threshold the learner is removed and a new learner is introduced. The goal of DWM is to exploit the traits of ensemble models to gradually keep up with a change in the target concept. While Kolter and Maloof introduced this algorithm using Naive Bayes as the base learner [34], we choose to use Decision Trees as the base classifier. This choice will create an easy comparison to our other RF models. In this variation of DWM, every week the performance of each tree is evaluated and its weight reduced if it has a low performance. If a tree's performance is below a threshold, then the tree is removed. At each iteration, we fill the ensemble with new trees until we reach the initial number of trees. The new trees are trained using a past memory of data, which is boosted to favor samples that were incorrectly classified in the past, according to the distant labeled data. Thus, trees are only trained once and not updated along the way. Again, we vary the memory for training new trees to be 2, 4, and 8 weeks.

In addition to testing these online learning methods, we also test varying initial training times for our original RF algorithm with no retraining. Specifically, we tested the model with an initial data set of 2, 4, and 8 weeks. For all models, we will resume the training and testing using all sources (as we did in Section 5.1) rather than source folds (as we did in Section 5.2).

Retraining improves performance and combats concept drift. In Figure 4, we show each models' performance over the 40 weeks using all feature groups combined, as chosen by our ablation study in Section 5.1. In Figure 4, it is clear that the retrained random forests show improvement over the once-trained random forests, showing at most a 0.08 increase in ROC AUC 38 weeks after the initial training for M versus UR and at most a 0.07 increase in ROC AUC for M versus B (e.g., the green lines versus the blue and red lines). While all of the retraining models improve upon the non-retraining models, there are some slight differences between each. Interestingly the model with 8 weeks of retraining has nearly the same performance as the model using infinite memory data. The DWM forests perform close to the simple retraining forest, but they are slightly slower

at learning the new concept than the simple retraining forest. DWM does require considerably less computational power, as it does not need retraining at every step and only retrain a subset of models at every step. However, these models have multiple hyperparameters to initially tune, unlike the simple online learning methods. Both the sliding window models and the DWM models create a more consistent performance over time than the non-retraining models.

In general, these results suggest that simply retraining the RF model every so often is enough to keep up with changes in the news. While continuously learning and remembering all previous examples does well, it is not necessary to achieve high performance.

7 HOW DO ATTACKS FROM MALICIOUS NEWS PRODUCERS IMPACT THE LEARNING ALGORITHM OVER TIME?

So far, we have explored how natural concept drift caused by changes in the news cycle impacts our news veracity algorithms. Our findings show that with retraining, it is easy to compensate for these more gradual changes. However, it is possible that changes may occur much more suddenly in real-life systems. This could be caused by unreliable sources changing their tactics significantly. There could be many reasons behind this, such as a desire to fool automated methods, to reach new readers, or to employ new information spreading tactics. For example, it has been argued that some sources may mix real and fake information to create confusion and appear like a legitimate news source [26, 31]. In this section, we simulate such purposeful changes and show their potential impact on our classification methods. We then discuss in the next section the implications of our findings.

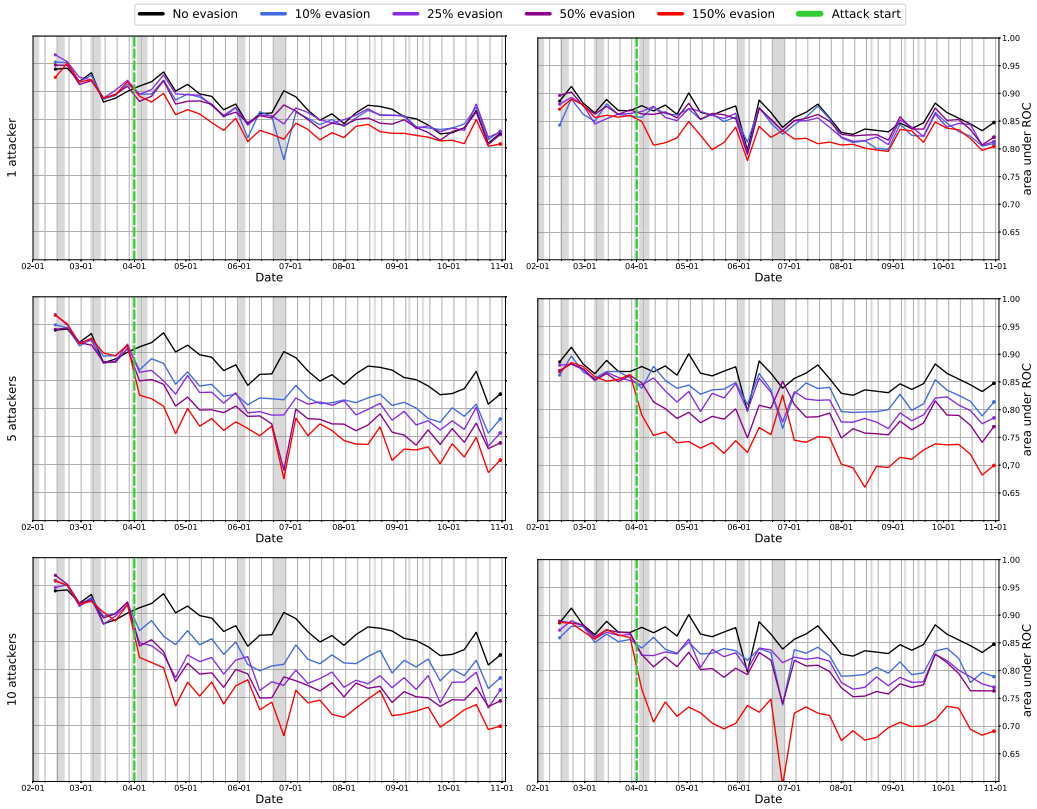
We have identified three feasible methods that could be used to counter automatic evaluation systems of news sources:

(1) Evasion attacks,
(2) Poison attacks,
(3) Blocking attacks.

7.1 Evasion

Evasion attacks are actions performed by a source to hide questionable content, making it more similar to real news. In these attacks, the source does not change the content it is providing but copies additional content from mainstream sources to appear more similar to these sources. In fact, a form of this tactic has already been discovered in previous research [26, 31, 50], illustrating conspiracy-based news sources copying proper news to gain credibility (e.g., Infowars copying news articles from The Associated Press). While, previous work has only discovered this tactic in full article form (copying a mainstream article verbatim and placing it next to a conspiracy article on a web page), it is easy to imagine a situation where the content would be mixed together in one article. From the perspective of concept drift, as malicious news change their content toward mainstream news, they shift $p(x)$ toward potentially unknown regions of the feature space, creating a forced virtual concept drift. They also cause real concept drift by changing the decision boundary $p(y|x)$ itself, as some mixed features can no longer discriminate the classes. Ultimately, this type of attack has only limited utility, as it directs the attention of readers away from the main messages or the stories the source is trying to push.

Another possible scenario is that the malicious news producers copy proper news articles, but hide the proper information from the consumer, not the data scraper. This attack could be done simply by placing real news text at the bottom of a fake news article and making the font color of the real text blend into the background. This version of the evasion attack could feasibly be done automatically and would not distract readers from the unreliable message being pushed.



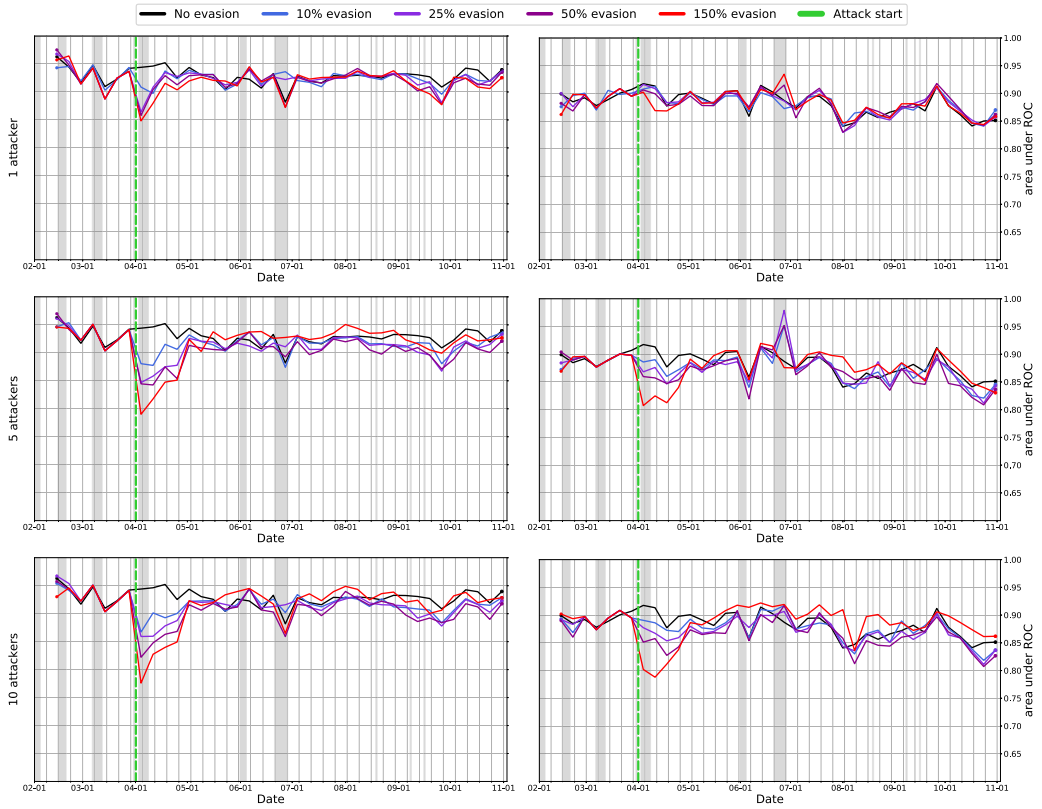
(a) M versus UR (b) M versus B

Fig. 5. **Evasion** attacks' impact on RF model with **no retraining**. The initial model is trained on 2 weeks of data starting in February and the attack starts on April 1 (indicated by the horizontal dashed, green line). Each row represents a different number of sources coordinating the attack.

To simulate this type of attack, we choose y unreliable (or biased) sources and mix content from randomly selected mainstream articles in the same week time frame. We randomly mix in n sentences from the randomly selected mainstream article, where n is equal to $x\%$ of the number of sentences in the unreliable (or biased) article. Then features are recomputed for the newly mixed articles and replace the previously non-mixed articles.

In Figures 5 and 6, we display the ROC AUC scores over time for an evasion attack starting on April 1st. We show this attack across several parameters. Specifically, we show the attack for varying numbers of attackers and varying amounts of content mixing. For example, in 10% evasion, if an unreliable article is 10 sentences long, we add 1 sentence to the article from a randomly selected mainstream article. This simulates partial copying from real news. In 150% evasion, if an unreliable article is 10 sentences long, we add 15 sentences to the article from a randomly selected mainstream article. This simulates hidden copying of real news. Last, the number of attackers simulates varying levels of collaboration among the malicious news sources. Figure 5 shows these attacks when we do not retrain the RF model and Figure 6 shows these attacks when we retrain the RF model using 8 weeks memory.

Online training models are robust to evasion attacks, static models are not. Not surprisingly the classifiers' performance is degraded as the amount of copying and number of sources copying



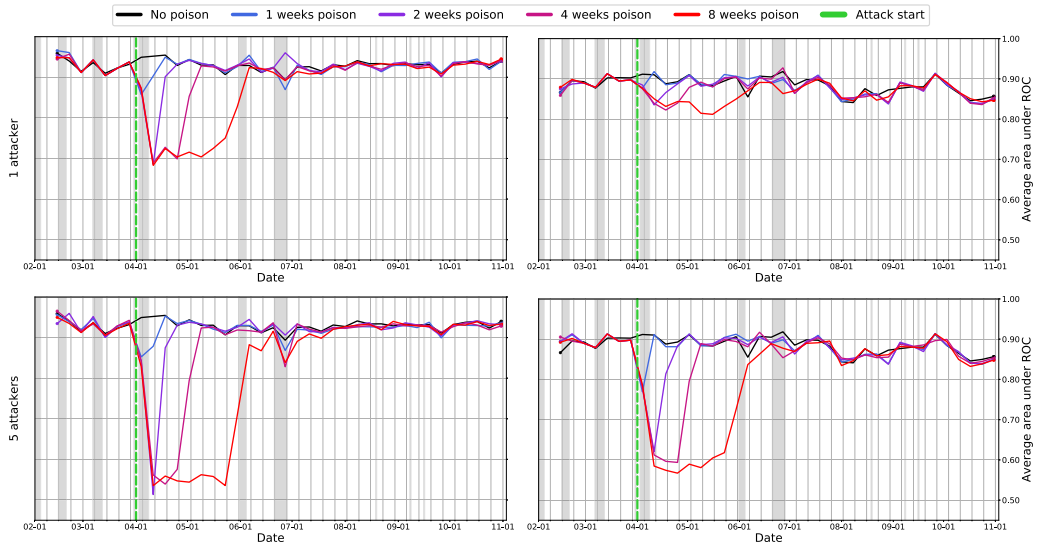
(a) M versus UR (b) M versus B

Fig. 6. **Evasion** attacks' impact on online RF model with **retraining** every week based on 8 weeks of data. The attack starts on April 1 (indicated by the horizontal dashed, green line). Performance is measured by ROC AUC score each week. Each row represents a different number of sources coordinating the attack.

increases. The instances of the unreliable (or biased) class are pushed toward the original classification boundary, making it difficult to tell them apart from mainstream articles. However, when the attack happens to the retraining algorithm, the result is very different. When the attack is initiated the performance makes a clear drop, but within weeks the classifier recovers to original performance. The classifier manages to learn a tighter classification boundary, which continues to correctly classify malicious articles, despite even large amounts of content copying. This indicates that the presence of specific features in the unreliable (or biased) news articles may be more important than the presence of specific features in mainstream news articles. For example, the presence of highly emotional writing may be a more important signal than the lack of it.

7.2 Poison

Poison attacks are attempts at deteriorating the classifiers' performance by injecting samples that the malicious source knows will be used to train the classifier [9, 35]. Traditionally, poison attacks are attempts at hiding malicious behavior (i.e., hiding in a security setting on a server). This attack is done by injecting large quantities of samples that appear in the outer region of what is considered good behavior. If enough samples are injected, then these new data points can shift the mass of good behavior until malicious behavior can no longer be separated from good behavior. In our



(a) M versus UR (b) M versus B

Fig. 7. Effect of **poison** attacks on Online RF with 8 weeks of memory. The top row shows a single source poisoning their articles for 1, 2, 4, and 8 weeks, and the bottom row shows five sources in a coordinated attack for the same time period. The horizontal dashed, green line represents the start of the attack.

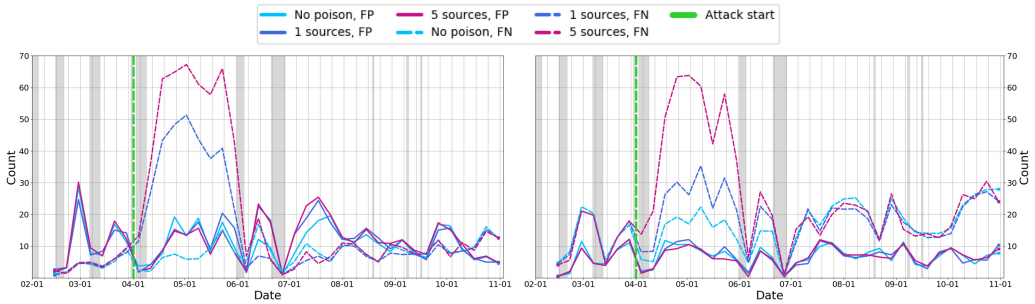
case, malicious news sources may be interested in causing the miss-classification of proper news articles to deteriorate the reputation of mainstream news sources (and in a sense weaponizing the classifier).

If for some period of time a malicious news sources behaves like proper news, whether through simply behaving better or copying proper news verbatim, then they influence the classification boundary $p(y|x)$ of the online learning algorithm. In other words, the attack forces real concept drift where the decision boundary should momentarily be different, because some mainstream news articles are distributed from a malicious source. The attack is also related to the more classical machine learning problem of mislabelled instances in the training set, as the poisoned samples from a practical point of view are mislabelled mainstream news articles. Of course, this attack does come at a cost to the attacker. First, it would require a unreliable (or biased) news source to understand how the online training algorithm works. Second, the malicious news produced could not produce their normal content for their readers for some period of time.

To simulate this type of attack, we randomly replace the feature vectors of y unreliable (or biased) sources with mainstream source feature vectors from the same week time frame. After random mainstream vectors are chosen to be used as replacements, they are removed from the mainstream data set. This removal is to simulate "bad actors behaving well" rather than copying mainstream news. The attack will happen for a set time t.

In Figure 7, we display the ROC AUC scores over time, with the poison attack starting on April 1st. We display this attack with several parameters: varying the length of the poison attack and the number of attackers. We again use the online RF algorithm with 8 weeks of memory.

Poison attacks harm the attacker more than the victim. Again unsurprisingly, a major decrease in performance is observed when the attacks begin and is maintained throughout the attack. After the attack time frame is over, the algorithm almost immediately recovers to its original performance. However, it is important to note what type of errors the algorithm is making. Remember, the



(a) M versus UR (b) M versus B

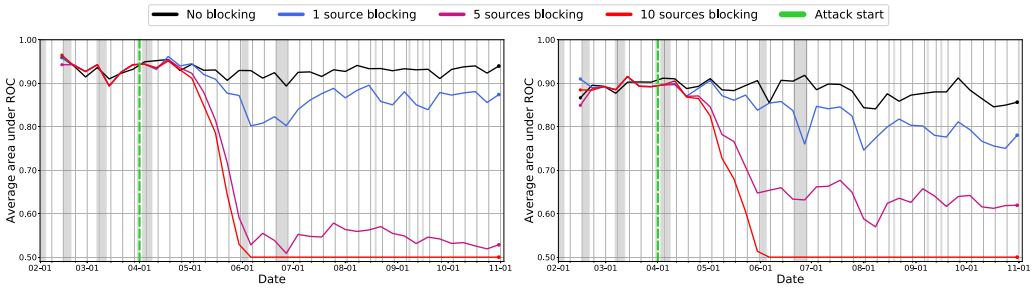
Fig. 8. False positives and false negatives during poison attacks of 8 weeks. The dashed lines are False Negatives (FN) while the solid lines are False Positives (FP). The horizontal dashed, green line represents the start of the attack. Overall attack impact on the model can be found in Figure 7.

intention of the poison attack is to decrease the performance of the system on other sources than itself—in this case to make the system classify mainstream sources as unreliable (or biased) and thus cast doubt on the system or cause uncertainty about the mis-classified mainstream sources. To explore this, we display the specific types of errors made by the classifier in Figure 8. We only show this for the 8 week poison attack. Notice in Figure 8 the steep increase in false negatives (poisoned articles are classified as mainstream) when the attack starts rather than false positives (mainstream articles classified as unreliable or biased). While there is some small amount of false positives, particularly at the end of the attack time for M versus UR, it is likely not enough to make the attack worth while. However, the poison attack may be more effective if it happens for significantly longer than the memory of the algorithm. But the longer a malicious source poisons, the longer they do not spread false or misleading news, defeating its own goals.

7.3 Blocking

Blocking attacks are when the malicious news producers make their data impossible for the classifier to scrape, hence leaving only positive samples for the supervised algorithm. The attack leaves the decision boundary untouched but causes a virtual concept drift as the sampling distribution in the input space $p(x)$ is not the same for the training and test sets. To simulate this type of attack, we simply remove some sources from the data set at some specified start time t. Blocking sources will make it harder for the classifier to predict the target of those specific sources, and can also affect its overall performance if the size of the training data is too heavily diminished. In Figure 9, we show the impact of this attack with 1, 5, and 10 attackers. Again, we use the online RF algorithm with 8 weeks of memory.

Blocking attacks are effective. After the blockade, the model no longer receives training-data from the blocked sources, but it is still tested using these sources (just as a real system would still be getting unreliable (or biased) articles from users or social network feeds). While the model has no degrade in performance for the first 8 weeks, once it forgets old examples of the unreliable (or biased) class it's performance dramatically decreases. Some simple solutions to this decrease in performance is to increase the memory size of the algorithm or to substitute blocked sources with other weakly labeled sources. Obviously, with more sources attacking, the performance hit is worse. More surprising is the large hit the algorithm takes when only 1 source is blocking. In this case, the source that is blocking (for both M versus UR and M versus B) is the source that publishes the most data (True Pundit and News Busters, respectively), which may cause the larger



(a) M versus UR (b) M versus B

Fig. 9. Effect of **blocking** attacks on Online RF with 8 weeks of memory. Attack starts on April 1 (indicated by the horizontal dashed, green line).

than expected performance decrease. There also may be some slight decay due to the testing on never seen sources, but as demonstrated in Section 5.2, this out-of-sample decay should be small.

8 DISCUSSION AND CONCLUSION

In this article, we presented a method using content analysis to classify news articles as coming from reliable, unreliable, or biased sources. Many such methods have been proposed in the literature [4, 5, 10, 24, 28, 44, 45, 49, 56]. Our work is the first study that tests all commonly used feature sets on a large amount of data over time. We show that these methods can work well over time, but they require periodic retraining. Even if sources mix content from mainstream sources, these methods can also work surprisingly well. The worst type of attack on these classifiers appears to be those that hinder access to unreliable (or biased) article data. In fact, these classifiers are very sensitive to the amount of data available. This finding suggests that these methods should detect such changes in data collection and stop retraining during times of low data or expand their memory to fall back to old training examples.

These findings support the use of source-level distant labeling, which can provide large quantities of training data quickly with little cost compared to methods based on fact-checked articles. Source-level distant labeling has been used widely in the news veracity literature. This article supports those works [4, 5, 28, 44]. These findings also indicate that our models rely on presence of features in unreliable articles (e.g., presence of personal pronouns and exclamation points) instead of lack of features present in mainstream news sources. As a result, it becomes difficult to learn an accurate classification model based on mainstream news sources alone. Hence, it is crucial that research efforts concentrate on collecting comprehensive data sets of unreliable articles that are broad-based, not for a specific event or topic. Moreover, we need to continue developing strategies to locate unreliable sources and collect data from them. This can be a challenge given unreliable sources may appear and disappear quickly. In addition, future research should explore single-class learning methods, such as positive unlabeled learning, to be robust against these potential blocking attacks.

We have shown that many simple and easy-to-compute features are effective in this classification task. The code for computing these features is available to the research community [29], as well as the data used in this study [42]. Despite the effectiveness of the features, there are many ways to improve them, and future study should continue to explore both their manipulation and generalization on never seen sources. We found that on average the models can predict consistently well on articles from news sources it has never seen, but the variance in performance of any one fold of sources can range widely.

By using features that are not topic specific, we are able to capture differences between news types that do not change drastically over time. Despite this, we do see a more significant drop in the performance of these features without retraining near the end of our timeline. This drop could be due to the major event in our data set (the U.S. midterm elections) or simply the regular content drift over time. Future work will explore the changes after large events. Further, the frequency of retraining may need to be adjusted given what is happening in the news cycle. We need to further analyze the nature of the changes in these classes and how they relate to actual events happening in the real world.

Inline with previous work [25], we show that titles provide strong signal. As many readers won't read past the title, the titles of unreliable articles are engineered to make statements. Hence, a potential improvement to these features could be to give more weight to appearance of a feature earlier in the article, as statements made earlier are more likely to receive attention [25]. Such features may be even more robust against manipulation. Another way these models can be strengthened is by adding more orthogonal features to the set, such as the expert labeling. The Wikipedia feature showed strong signal, but it is easily manipulated over time. A stronger, less noisy, version of this feature could simply be an expert/journalist curated list of unreliable and reliable sources, of which many have already been made. In a machine learning study this type of feature could be considered a "cheat" feature, as these expert lists are used as ground truth, but in a real life setting these types of list should augment the classifiers. In fact, there may be many other more complex source-level features that improve both overall the classification accuracy and the stability of static models over time.

From a methods perspective, it is important to explore other methods for news classification and how well those methods generalize over time and attack. This exploration includes the type of features used and the algorithms used. For example, features based on the behavior of the news organization or the publishing patterns of a media source may drift less over time than content-based features. Also unexplored are features such as word embedding features or content representations created by deep learning models. Similarly, other learning algorithms, such as deep learning methods, may be more robust to attack due to the obscurity of the algorithm choices, but they may sacrifice explainablity/transparency to end-users or learn topic-specific features that hurt generalizability. These questions are important, but left for future work.

Our article has used specific examples of misinformation and bias from the most common and highly shared news sites. It is also likely that misinformation can be presented in ways that are completely different than anything we have seen in these sources. We currently do not have any image- or video-based features. Text can be embedded in images to make it harder to apply methods like ours. The introduction of deep fakes provides many novel challenges as manipulated images can be especially effective. However, text content is still needed to draw attention to these images and video. We expect our methods will still provide useful signal to such methods. What if a source produces content that reads very much like one produced by a mainstream source? Our methods cannot distinguish between a well-written lie and truth. We note that there is a cost to malicious news sources by doing this, both in terms of employing staff and also reduced engagement as such articles are not as attention grabbing anymore.

We recognize that many different methods are needed to help the misinformation problem, including manual and institutional efforts. Our methods are especially useful for sources that use the "fire-hose model": produce lots of content quickly with sensationalist writing style, then copy and disseminate information to many other sources [26, 31, 50] to make information more widely available. Recent work has also shown that the truth of claims is not always the main concern of malicious news producers, but instead the goal is to create confusion and increase polarization [3]. Such efforts tend to target information with opposite points of view to specific groups [3]. As such

targeting methods focus on high engagement, our models can work well in these cases. There are many sources that lie between the extremes we study in this article. Understanding how the extreme sources are related to "in-between" sources and to what level they disseminate true and false information is part of ongoing work.

Even though our methods work well, the true cost of exposure to misinformation is hard to assess. Even a single incorrect story can have dire consequences for individuals or populations. However, as opinions are formed over time, it is important to find methods to inform readers about the sources they get information from quickly and effectively. Content-based approaches like ours are the first step in helping such efforts.

REFERENCES

[1] Hadeer Ahmed, Issa Traore, and Sherif Saad. 2017. Detection of online fake news using N-gram analysis and machine learning techniques. In *Proceedings of the International Conference on Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments*. Springer, 127–138.

[2] Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *J. Econ. Perspect.* 31, 2 (2017), 211–36.

[3] Ahmer Arif, Leo Graiden Stewart, and Kate Starbird. 2018. Acting the part: Examining information operations within# BlackLivesMatter discourse. *Proceedings ACM Hum.-Comput. Interact.* 2 (2018), 20.

[4] Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018. Predicting factuality of reporting and bias of news media sources. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'18)*.

[5] Ramy Baly, Georgi Karadzhov, Abdelrhman Saleh, James Glass, and Preslav Nakov. 2019. Multi-task ordinal regression for jointly predicting the trustworthiness and the leading political ideology of news media. *arXiv preprint arXiv:1904.00542* (2019).

[6] Alessandro Bessi, Mauro Coletto, George Alexandru Davidescu, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. 2015. Science vs. Conspiracy: Collective narratives in the age of misinformation. *PLoS ONE* 10, 2 (Feb. 2015), e0118093–17.

[7] Alessandro Bessi, Antonio Scala, Luca Rossi, Qian Zhang, and Walter Quattrociocchi. 2014. The economy of attention in the age of (mis)information. *J. Trust Manage.* 1, 1 (Dec. 2014), 105–13.

[8] Alessandro Bessi, Fabiana Zollo, Michela Del Vicario, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. 2015. Trend of narratives in the age of misinformation. *PLoS ONE* 10, 8 (Aug. 2015), e0134641–16.

[9] Battista Biggio, Blaine Nelson, and Pavel Laskov. 2012. Poisoning attacks against support vector machines. *Proceedings of the 29th International Conference on Machine Learning (ICML'12)*, vol. 2, 1807–1814.

[10] Ceren Budak, Sharad Goel, and Justin M. Rao. 2016. Fair and balanced? quantifying media bias through crowdsourced content analysis. *Public Opin. Quart.* 80, S1 (2016), 250–271.

[11] Sylvie Cazalens, Philippe Lamarre, Julien Leblay, Ioana Manolescu, and Xavier Tannier. 2018. A content management perspective on fact-checking. In *Proceedings of the Web Conference*.

[12] Abhijnan Chakraborty, Bhargavi Paranjape, Sourya Kakarla, and Niloy Ganguly. 2016. Stop clickbait: Detecting and preventing clickbaits in online news media. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM'16)*. IEEE, 9–16.

[13] Yimin Chen, Niall J. Conroy, and Victoria L. Rubin. 2015. Misleading online content: Recognizing clickbait as false news. In *Proceedings of the ACM on Workshop on Multimodal Deception Detection*. ACM, 15–19.

[14] Giovanni Luca Ciampaglia, Prashant Shiralkar, Luis M Rocha, Johan Bollen, Filippo Menczer, and Alessandro Flammini. 2015. Computational fact checking from knowledge networks. *PloS One* 10, 6 (2015), e0128193.

[15] Sarah Jane Delany, Pádraig Cunningham, Alexey Tsymbal, and Lorcan Coyle. 2005. A case-based technique for tracking concept drift in spam filtering. In *Applications and Innovations in Intelligent Systems XII*. Springer, 3–16.

[16] Stefano DellaVigna and Ethan Kaplan. 2007. The fox news effect: Media bias and voting. *Quart. J. Econ.* 122, 3 (2007), 1187–1234.

[17] James N. Druckman and Michael Parkin. 2005. The impact of media bias: How editorial slant affects voters. *J. Politics* 67, 4 (2005), 1030–1049.

[18] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2016. The rise of social bots. *Commun. ACM* 59, 7 (2016), 96–104.

[19] Frederick Fico, John D. Richardson, and Steven M. Edwards. 2004. Influence of story structure on perceived story bias and news organization credibility. *Mass Communication & Society* 7, 3 (2004), 301–318.

[20] João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. 2014. A survey on concept drift adaptation. *ACM Comput. Surv.* 46, 4 (2014), 44.

[21] Matthew Gentzkow and Jesse M. Shapiro. 2010. What drives media slant? Evidence from U.S. daily newspapers. *Econometrica* 78, 1 (2010), 35–71.

[22] Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P. Wojcik, and Peter H. Ditto. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in Experimental Social Psychology*. Vol. 47. Elsevier, 55–130.

[23] Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017. Toward automated fact-checking: Detecting check-worthy factual claims by ClaimBuster. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1803–1812.

[24] Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anil Kumar Nayak et al. 2017. ClaimBuster: The first-ever end-to-end fact-checking system. *Proc. VLDB Endow.* 10, 12 (2017), 1945–1948.

[25] Benjamin D. Horne and Sibel Adali. 2017. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In *Proceedings of the International Workshop on News and Public Opinion (NECO'17)*.

[26] Benjamin D. Horne and Sibel Adali. 2018. An exploration of verbatim content republishing by news producers. *arXiv preprint arXiv:1805.05939* (2018).

[27] Benjamin D. Horne, Sibel Adali, and Sujoy Sikdar. 2017. Identifying the social signals that drive online discussions: A case study of Reddit communities. In *Proceedings of the 26th International Conference on Computer Communication and Networks (ICCCN'17)*. IEEE, 1–9.

[28] Benjamin D. Horne, William Dron, Sara Khedr, and Sibel Adalı. 2018. Assessing the news landscape: A multi-module toolkit for evaluating the credibility of news. In *WWW Companion*.

[29] Benjamin D. Horne, Sara Khedr, and Sibel Adalı. 2018. Sampling the news producers: A large news and feature data set for the study of the complex media landscape. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM'18)*.

[30] Benjamin D. Horne, Dorit Nevo, John O'Donovan, Jin-Hee Cho, and Sibel Adalı. 2019. Rating reliability and bias in news articles: Does AI assistance help everyone?. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 13. 247–256.

[31] Benjamin D. Horne, Jeppe Nørregaard, and Sibel Adalı. 2019. Different spirals of sameness: A study of content sharing in mainstream and alternative media. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 13. 257–266.

[32] Clayton J. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media*.

[33] Ralf Klinkenberg and Thorsten Joachims. 2000. Detecting concept drift with support vector machines. In *Proceedings of the International Conference on Machine Learning (ICML'00)*. 487–494.

[34] J. Zico Kolter and Marcus A. Maloof. 2007. Dynamic weighted majority: An ensemble method for drifting concepts. *J. Mach. Learn. Res.* 8 (Dec. 2007), 2755–2790.

[35] Pavel Laskov and Marius Kloft. 2009. A framework for quantitative security analysis of machine learning. *Proceedings of the ACM Conference on Computer and Communications Security*. 1–4. DOI: https://doi.org/10.1145/1654988.1654990 00052.

[36] David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild et al. 2018. The science of fake news. *Science* 359, 6380 (2018), 1094–1096.

[37] Stephan Lewandowsky, Ullrich K. H. Ecker, Colleen M. Seifert, Norbert Schwarz, and John Cook. 2012. Misinformation and its correction: Continued influence and successful debiasing. *Psychol. Sci. Public Interest* 13, 3 (2012).

[38] Ying Lin, Joe Hoover, Gwenyth Portillo-Wightman, Christina Park, Morteza Dehghani, and Heng Ji. 2018. Acquiring background knowledge to improve moral value prediction. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM'18)*. IEEE, 552–559.

[39] Nicco Mele, David Lazer, Matthew Baum, Nir Grinberg, Lisa Friedland, Kenneth Joseph, Will Hobbs, and Carolina Mattsson. 2017. Combating fake news: An agenda for research and action. Retrieved on October 17, 2018 from https://www.hks.harvard.edu/publications/combating-fake-news-agenda-research-and-action.

[40] Leandro L. Minku, Allan P. White, and Xin Yao. 2009. The impact of diversity on online ensemble learning in the presence of concept drift. *IEEE Trans. Knowl. Data Eng.* 22, 5 (2009), 730–742.

[41] Leandro L. Minku, Allan P. White, and Xin Yao. 2010. The impact of diversity on online ensemble learning in the presence of concept drift. *IEEE Transactions on Knowledge and Data Engineering* 22, 5 (2010), 730–742.

[42] Jeppe Nørregaard, Benjamin D. Horne, and Sibel Adalı. 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*, Vol. 13. 630–638.

[43] Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 271.

[44] Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2016. Credibility assessment of textual claims on the web. In *Proceedings of ACM International Conference on Information and Knowledge Management (CIKM'16)*. ACM, 2173–2178.

[45] Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2017. A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint arXiv:1702.05638* (2017).

[46] Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. 1650–1659.

[47] Julio Reis, Fabricio Benevenuto, P. Vaz de Melo, Raquel Prates, Haewoon Kwak, and Jisun An. 2015. Breaking the news: First impressions matter on online news. In *Proceedings of the 9th International AAAI Conference on Web-Blogs and Social Media*.

[48] Jeffrey C. Schlimmer and Richard H. Granger. 1986. Beyond incremental processing: Tracking concept drift. In *AAAI*. 502–507.

[49] Sneha Singhania, Nigel Fernandez, and Shrisha Rao. 2017. 3HAN: A deep neural network for fake news detection. In *International Conference on Neural Information Processing*. Springer, 572–581.

[50] Kate Starbird, Ahmer Arif, Tom Wilson, Katherine Van Koevering, Katya Yefimova, and Daniel Scarnecchia. 2018. Ecosystem or echo-system? exploring content sharing across alternative media domains. In *Twelfth International AAAI Conference on Web and Social Media*.

[51] Laura Sydell. 2016. We Tracked Down A Fake-News Creator In The Suburbs. Here's What We Learned. Retrieved from http://www.npr.org/sections/alltechconsidered/2016/11/23/503146770/npr-finds-the-head-of-a-covert -fake-news-operation-in-the-suburbs.

[52] Yla R. Tausczik and James W. Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *J. Lang. Soc. Psychol.* 29, 1 (2010), 24–54.

[53] Sander Van der Linden, Anthony Leiserowitz, Seth Rosenthal, and Edward Maibach. 2017. Inoculating the public against misinformation about climate change. *Global Challenges* 1, 2 (2017), 1600008.

[54] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151.

[55] William Yang Wang. 2017. "Liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648* (2017).

[56] Savvas Zannettou, Michael Sirivianos, Jeremy Blackburn, and Nicolas Kourtellis. 2018. The web of false information: Rumors, fake news, hoaxes, clickbait, and various other shenanigans. *arXiv preprint arXiv:1804.03461* (2018).

[57] Fabiana Zollo, Petra Kralj Novak, Michela Del Vicario, Alessandro Bessi, Igor Mozetič, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. 2015. Emotional dynamics in the age of misinformation. *PLoS ONE* 10, 9 (Sept. 2015), e0138740–22.

Received February 2019; revised July 2019; accepted September 2019