# FADE: Detecting Fake News Articles on the Web

Bahruz Jabiyev
Northeastern University
Boston, MA, USA

Sinan Pehlivanoglu
Brown University
Providence, RI, USA

Kaan Onarlioglu
Akamai Technologies
Cambridge, MA, USA

Engin Kirda
Northeastern University
Boston, MA, USA

## ABSTRACT

Internet-based media and social networks enable quick access to information; however, that has also made it easy to conduct disinformation campaigns. *Fake news* poses a serious threat to the functioning and safety of our society, as demonstrated by nation-state-sponsored campaigns to sway the 2016 US presidential election, and more recently COVID-19 pandemic hoaxes that promote false cures, putting lives at risk.

FADE is a novel approach and service that helps Internet users detect fake news. FADE discovers multiple news sources covering the same story, analyzes their reputation, and checks the trustworthiness of cited sources. Our approach does not depend on any specific social media or news source, does not rely on costly textual content analysis, and does not require lengthy offline processing. Our experiments demonstrate above 85% detection accuracy with a practical implementation. FADE offers a path to empowering the Internet community with effective tools to identify fake news.

## CCS CONCEPTS

• **Security and privacy** → **Usability in security and privacy**.

## KEYWORDS

fake news; disinformation campaigns; internet safety

**ACM Reference Format:**
Bahruz Jabiyev, Sinan Pehlivanoglu, Kaan Onarlioglu, and Engin Kirda. 2021. FADE: Detecting Fake News Articles on the Web. In *The 16th International Conference on Availability, Reliability and Security (ARES 2021), August 17–20, 2021, Vienna, Austria.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3465481.3465751


*ARES 2021, August 17–20, 2021, Vienna, Austria*

ACM ISBN 978-1-4503-9051-4/21/08…$15.00
https://doi.org/10.1145/3465481.3465751

## 1 INTRODUCTION

Fake news is a social phenomenon which involves fabrication of false stories, malicious modification of real news for propaganda purposes, dissemination of rumors and conspiracy theories, and deception of users with unsupported claims – all with the aim of influencing user opinions and actions. Leaving aside the tremendous benefits of social media, unfortunately, these platforms have also provided fake news with the perfect breeding ground to quickly spread.

Fake news has tangible and severe consequences in real life. For example, the 2016 U.S. presidential election demonstrated that an online disinformation campaign has potential to materially impact the world. A study shows that during the elections, every American clicked on at least one fake news article related to the presidential candidates [4]. In 2020, the barrage of fake news on the COVID-19 pandemic threatens lives by spreading fear and promoting false cures [38].

Ideally, every Internet user should fact-check all of the information that they consume online and possess a healthy level of skepticism. In practice, however, constant fact-checking creates an overwhelming cognitive and manual load with the deluge of information available online, and does not seem to be a reasonable expectation. Many users might not even know where to begin in order to confirm the veracity of something that they have read, even if they are inclined to. This problem is exacerbated by the fact that fake news can spread rapidly, and such false information may even be picked up by otherwise reputable news sources [22].

Facebook, as a recognized breeding ground for fake news, has reportedly begun to address the issue on its platform [13]. By utilizing a combination of user feedback and machine learning, Facebook aims to identify potential false stories, and then sends these to third-party fact-checker organizations such as Snopes, PolitiFact, and FactCheck.org. If the fact-checker rates a story as being false, that story is then pushed to a lower position in the Facebook news feed, slowing its spread. Unfortunately, it may take more than three days for fact-checkers to rate the accuracy of a story, which provides a significant window of opportunity for fake news to spread and affect a large number of readers. Furthermore, organizations such as Facebook are businesses, and may base their decisions on parameters that are not necessarily in the best interest of Internet users. For instance, Facebook initially defended Infowars, but later banned it due to public pressure [32]. Thus, users cannot solely depend on social platforms and news providers to perform fake news detection on their behalf.

Fake news detection sits at the intersection of multiple disciplines including social media research, machine learning, and web security. For instance, Horne and Adali modeled fake news using text features such as word length and term frequency [18], whereas CSI and dEFEND proposed detection approaches using article content [33, 34]. In general, and as exemplified above, features used to classify news articles in previous work are confined to the article text itself.

In this paper, we propose a novel fake news detection system called FADE. Our approach is based on the intuition that if a news story is real, then trusted news sources are likely to cover that story. In contrast, if a story is fake, either no trusted news source will cover it, or the story will also be found in sources that are known to spread fake news. One of our central contributions over prior work is showing that news source reputation *alone* is a strong signal in identifying fake news. Moreover, we show that fake news detection can be performed independently of any single social media platform or news provider, reducing the chances that a single authority can exert undue influence over fake news classification.

We leverage a search engine to identify multiple news sources covering the same story as the tested news article. Next, we perform a similarity analysis on search results to filter out irrelevant search hits. Besides the reputation analysis of news sources covering the same story, we also analyze the reputation of cited news sources within the tested article. The results of these analyses form a reputation graph that is the basis for our fake news classification technique. In essence, we automate the actions Internet users need to perform to validate a news article, and provide machine learning support for their decision-making process.

We train our classifier on a data set of 4,750 fake and 4,750 real news articles, and validate the performance of the model on an independent data set of 500 fake and 500 real news articles. Our approach achieves approximately 87% detection accuracy in this experiment.

We build a prototype demonstrating that FADE can be deployed as a practical online service, where Internet users issue queries for suspect news articles and view detection reports. We also develop a browser extension to streamline this process for the end user. We further evaluate our implementation on-the-go in a second experiment, using news articles tested over a period of ten weeks. In this experiment, we similarly achieve a detection accuracy above 85%.

While FADE does have certain limitations fundamental to the approach, which we discuss at length in this paper, it nevertheless differentiates itself from existing literature on fake news detection by proposing a practical, highly-usable system with real-life impact.

In summary, our contributions are as follows:

- We propose a novel approach to detect fake news articles based on the reputation of multiple sources that cover the same story. Our approach is independent of any single social media platform or news provider.
- We train and test our detection model on larger data sets than previous work. We demonstrate that the news coverage and source reputation features used in our approach can detect fake news with more than 85% accuracy, without relying on costly (and sometimes intractable) story content analyses.
- In contrast to previous work, we show that FADE can be deployed in practice as a service and a browser extension.

**Availability.** The source code and data sets are available at *https://github.com/bahruzjabiyev/FADE*.

## 2 BACKGROUND

We define *news* as new information about important events that appears in published media. There is an implicit assumption that media organizations perform some vetting as to public interest in the information, although this does not necessarily translate to veracity. *Real news* describes events as they happened in reality. Such news may still include opinion or bias, but nevertheless are based on verifiable facts.

In contrast, *fake news* contains fabrications that deviate from reality in an evidence-based manner. This class of news is often disseminated with malicious intent, for instance as propaganda against a person or an organization, to spread damaging rumors, or to make money from user clicks. However, our definition focuses on facts measured through endorsement by reputable sources, and does not consider the motivation. A *fake news website* is a website that purposefully creates, publishes, and disseminates fake news.

Fake news take various forms in the wild. In its simplest form, it may contain only a string of words which makes up a false claim – e.g., "Sasha Obama Murdered in a Drive-By Shooting." In some cases, a photo or a video accompanies the claim. To continue our previous example, a photo from an unrelated crime scene could accompany the false story. In order to appear legitimate, fake news claims are often accompanied by links to fake news articles. In fact, when we analyzed widely-spread false claims identified by Snopes and PolitiFact over a three month period, we found that approximately 70% of fake news claims were supported by a fake news article. The Sasha Obama murder hoax, for example, appeared on social media as a news article [24].

Fake news articles can vary widely in content, but they still share some fundamental characteristics such as their failure to present factual evidence or meet the ethics and standards of professional journalism. We provide three examples below.

The fake story "Pope Francis Found Guilty Of Child Trafficking, Rape, Murder" found on a website called "The Pedo-GATE" falsely claims that the International Common Law Court of Justice in Brussels holds Pope Francis liable for criminal acts [9]. This story uses unsupported claims as a means to create propaganda, and to influence public opinion.

Another story titled "Las Vegas: Video Footage Confirms Multiple Shooters, Co-ordinated Attack" by a website called "Your News Wire" asserts that there was more than one shooter in the Las Vegas shooting of October 2017 [11]. This story promotes a conspiracy theory by fabricating supposed secret plans and actors behind an illegal act.

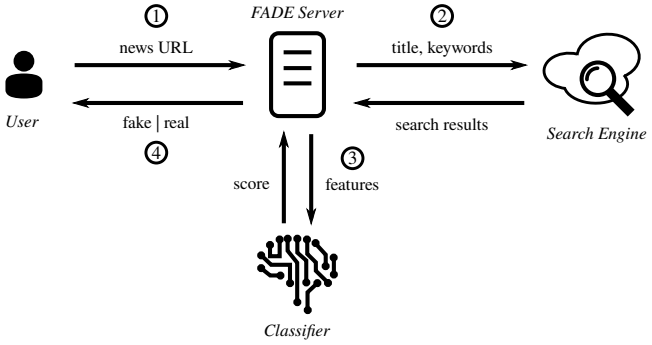


**Figure 1: High-level architecture of FADE.**

Finally, the fake story "Donald Trump Ends School Shootings By Banning Schools" from "8Satire" claims that President Trump has decided to ban all schools, saying "if there are no schools, there will be no school shootings" [1]. Note that 8Satire is a satire news and humor website according to its disclaimer. To be clear, satire is a recognized and valuable genre of media. Unfortunately, it can also be confused by users for real news, especially when it appears without context on social media platforms [4].

To reach a wider audience, fake news frequently abuses online social media through post sharing. For example, one article used shots of the actor Sylvester Stallone from a film in which he portrayed a character battling cancer, and claimed that the star had died of prostate cancer. In about a day, this hoax was shared over 2.5 million times on Facebook [15].

Sadly, the impact of fake news can reach far beyond celebrity gossip. On top of the aforementioned cases of manipulated voter opinions and the spread of false cures, fake news were involved in inciting mob violence towards the Muslim minority in Sri Lanka [16], motivating a shooting in a pizzeria [19], and spreading rumors on WhatsApp that led to the deaths of more than two dozen people in India [12].

# 3 APPROACH

Given its rise and real-life consequences, detecting fake news is an important research objective. Our goal is to accurately identify fake news articles, and enable Internet users to make informed decisions on the authenticity of news claims.

We call our approach FADE, and build it on a key insight that news source reputation can be used as a reliable signal for judging news authenticity. That is, if a news article is real, then trusted news sources will cover it. However, if the article is fake, then its story will not be covered by any trusted sources, or perhaps will only inadvertently be picked up by a small number of such sources. On the other hand, fake news will be propagated by known fake news outlets.

Figure 1 depicts the architecture of FADE. (1) A user queries the FADE service by sending it the URL of a news article. (2) FADE downloads the article, extracts its title and keywords, and uses these extracted values to perform a search in order to identify additional news sources that cover the same story. (3) The system then supplements the discovered sources with external references in the queried article, and uses the reputations of these sources as features passed to a machine learning model for classification. (4) Finally, FADE retrieves the classification score, and returns the detection result back to the user together with a summary report that lists all sources that cover the story, and their reputation.

## 3.1 Searching for Media Coverage

The purpose of a news article title is to give a concise description of the story, and hence, it is a natural search engine query term to identify additional sources covering the same story. To make the search more targeted, we use a combination of the title and keywords extracted from the article body. We discuss the details of this process in Section 4.

Fake news sources can sometimes report real stories, and real news sources can inadvertently republish fake stories. As such, in our detection methodology we do not take into account the queried source's reputation itself. We exclude the URL of the queried article from the search results, and instead only focus on the additional sources we discover.

## 3.2 Identifying Additional Sources

Once we retrieve candidate results for the queried news story from the search engine, we must narrow down the results so that recent and relevant articles comprise the additional source set, and false search hits are removed.

Based on the insight that additional sources would cover the same story at about the same date, we first filter results based on proximity to the queried article's publication date.

Next, we measure the text similarity between the remaining search results and queried article, using the term frequency – inverse document frequency (TF-IDF) algorithm. That is, we compare the frequency of words in the articles, and how defining these words are within the context of the article by considering their frequency in a larger corpus of articles. This gives more weight to common terms of two documents that are less common in the overall document corpus. Each article is represented as a normalized vector of word occurrences, and we compute the cosine similarity between vectors to determine whether the articles cover the same story.

We set text and date similarity thresholds based on empirical testing. We discuss our experiments for obtaining suitable thresholds in Section 5.

## 3.3 Forming Similarity Subsets

We group sources into subsets based on their text similarity to the queried article. The subsets are defined based on similarity ranges. We examine each similarity subset separately for the reputation of sources rather than examining the entire set of similar search results. Doing this allows us to give more weight to sources that are in higher similarity subsets when we decide on the truthfulness of the queried article. Similarity subsets are explained in more detail in Section 3.4. Listing 1 shows the algorithm for processing search results and retrieving similarity subsets.

```
def get_similar_articles(article):
    keywords = extract_keywords(article.content)
    all_results = query_search_engine(article.title)
    filtered_results = [x for x in all_results \
        if abs(x.date - article.date) < DATE_THRESHOLD]
    query_vector = vectorize(article)
    similar_articles = []
    for s in filtered_results:
        source_vector = vectorize(s)
        text_similarity = cosine_similarity(query_vector,
                                            source_vector)
        if text_similarity > TEXT_THRESHOLD:
            similar_articles.append((s, text_similarity))
    return filtered_results.group_by(similar_articles)
```

**Listing 1: Article similarity algorithm in Python.**

### 3.4 Features & Classification

FADE's machine learning classifier uses 19 features to make the article trustworthiness decision. These are the numbers of (1) highly-trusted, (2) legitimate, and (3) fake news sources in each of the $k$ similarity subsets of search results for a total of $3k$ features, and 1 additional feature corresponding to the number of fake news sources the queried article references. As we later explain in Section 5, we choose to use $k = 6$ similarity subsets based on our analysis, and therefore end up with $3k + 1 = 19$ features in total.

As we mentioned above, we group the additional news sources discovered for a queried article into three tiers, ordered by their reputation. *Highly-trusted* sources are almost universally regarded as well-established, high-profile news outlets. These sources are often cited by other news media because they have solid reputations, are known for performing original investigative reporting, and are generally trusted by readers (e.g., The Washington Post, The New York Times). Intuitively, the more this class of source appears in the search results, the more confidence FADE has in the article's veracity. *Legitimate* sources are not known for actively disseminating fake news, but they also do not enjoy the same level of reputation and recognition that highly-trusted news outlets described above do. Appearance of these sources in search results signals that the queried article may be trustworthy. Finally, *fake* sources are well-known for creating and spreading fake news. Naturally, encountering these in search results indicates that the queried article may also be fake.

When we examine the search results for additional sources covering the same story, we check for the existence of these three classes of sources among those and use their count as a classification feature. Note that, in some cases, when we search for coverage of a fake story, highly-trusted news sources meet our filtering thresholds and appear to cover the same fake news. This usually happens when trusted sources cover stories with similar topics, but not the exact same story. We observe that similarity levels of these search results to the fake queried article tend to be low. In contrast, when we search for coverage of real stories, highly-trusted news sources that cover the same story tend to have a high text similarity to the queried article.

We observe a similar trend with search hits from other tiers of news sources. This justifies our grouping of search results based on their similarity levels. We treat the numbers of news sources that appear in each similarity subset, for each tier, as separate features for classification.

In addition to the reputation of news sources covering the story, we also examine external references in the queried article for pointers to fake sources, and use this count as a classification feature. This is based on our observation that fake news articles often cite other fake news sources, whereas this rarely occurs in real news articles, except when a real article refutes false claims contained in fake news stories.

## 4 IMPLEMENTATION

### 4.1 FADE Components

We implement all components comprising FADE, except the external search engine, using Python. We leverage the library `Newspaper` [1] for extracting the title, keywords, and publication date of queried news articles. This library internally uses the TF-IDF algorithm to extract the keywords as we previously described, and uses regular expressions and heuristics to extract the article title and date.

We measure date similarity by simply calculating the difference between two publication dates. To measure text similarity between articles, we use the library `gensim` [2]. We set the date similarity threshold to two days, and the text similarity threshold to 40%. We discuss how we empirically determine these threshold values in Section 5. We filter out search results that do not meet the date proximity threshold, or that have less than 40% text similarity to the queried article. Above 40%, we form 6 subsets corresponding to 10% intervals, and group search results into these subsets depending on which interval their text similarity score falls into.

After the features are extracted, we pass them on to an SVM classifier provided by the library `scikit-learn` [3] to perform the classification. We compare the results obtained with different classifiers in Section 5.

### 4.2 Search Engine & Source Lists

Our prototype implementation leverages DuckDuckGo as its external search engine component to discover additional news sources reporting the queried story. While our approach is search engine-agnostic, our choice is motivated by the fact that DuckDuckGo makes automated access to its service easy, and does not enforce aggressive rate limiting or bot blocking defenses that could interfere with our experiments.

We compile the lists of highly-trusted, legitimate, and fake news outlets we use when computing FADE's classification features from three separate sources. We obtain the list of highly-trusted sources from surveys conducted by the Pew Research Center and Reynolds Journalism Institute [21, 26]. In this list, we have 30 different news sources: ABC News, Associated Press, BBC, Bloomberg, CBS News, CNN, Dallas

[1] https://github.com/codelucas/newspaper
[2] https://pypi.org/project/gensim/
[3] https://scikit-learn.org/

News, Fox News, Google News, Los Angeles Times, MSNBC, NBC News, NPR, PBS, Politico, Reuters, The Atlantic, The Denver Post, The Economist, The Guardian, The Kansas City Star, The New York Times, The New Yorker, The Seattle Times, The Wall Street Journal, The Washington Post, TheBlaze, Time, USA Today, Yahoo News.

For the legitimate sources list, we take the Alexa Top 500 sites in the "News" category, excluding highly-trusted sources from the list [3]. For the fake sources, we use a list published by PolitiFact that includes around 330 fake news websites [27].

### 4.3 Deployment & Use

We implement FADE as a web service that exposes an HTTP interface. Users issue queries to the service by sending a request with the URL of a suspect news article, and receive back a response with the detection result, together with a summary report of additional sources discovered and their reputations. We also implement a Chrome extension that streamlines this interaction for users, allowing them to initiate an analysis of a page with a right mouse-click.

During our experiments in Section 5, we determine the end-to-end runtime for issuing a query and receiving back the results to be median $= 12.4s$, $Q_1 = 10.0s$, $Q_3 = 18.7s$). However, we envision a deployment model where the FADE service can be accessed publicly by all Internet users, and results from previous analyses can be stored for quickly responding to duplicate queries (e.g., a model similar to online binary and URL scanning services such as VirusTotal). Our prototype has this capability, and reports for repeat queries are available to users instantly. Users can still request that the analysis be performed from scratch if they wish. This can be desirable in some cases, such as when additional news sources have picked up the story or the search engine has indexed more results since the initial query. For the same reasons, we mark articles that were queried shortly after their publication, and automatically analyze them in the background at a later date to update our cached results.

## 5 EVALUATION

### 5.1 Data Sets

To compile our training data set for fake news, we start with a Kaggle data set that contains approximately 12,000 fake news articles collected during October and November 2016, the U.S. presidential election period [31]. Unfortunately, the original data set does not contain URL metadata for the news articles it contains, which a realistic deployment of our solution requires. Therefore, for each article in the data set, we query a search engine for the title of the article and pick the URL of the search result which has at least 90% text similarity to the queried article. We successfully retrieve the URLs of 4,750 fake news articles and use these as our training data set. We refer to our fake news training data set as FAD-10.

For the training data set with real news articles, we build our own data set by collecting another 4,750 news articles published in the same data range as FAD-10, approximately 500 each from 10 highly-trusted news sources: ABC News, BBC, CBS News, CNN, NBC News, NPR, Reuters, The Guardian, The New York Times, The Washington Post, USA Today. We refer to our real news training data set as RED-10.

Similarly, we compile two test data sets to evaluate FADE. For fake news, we collect 500 articles from fact-checker organizations Snopes, PolitiFact, and FactCheck.org. These articles are labeled "false" by Snopes, "false" and "pants on fire" by PolitiFact, or "false" by FactCheck.org from January 2017 to May 2018. We call this data set FAD-5. For real news articles, we take 50 articles each from 10 highly-trusted news sources, for a total of 500. We call this data set RED-5.

### 5.2 Setting Similarity Thresholds

In order to measure text similarity distributions and choose effective thresholds for our filtering, we perform a preliminary experiment with three data sets, each consisting of 9,000 articles. The first data set comes from the same Kaggle fake news data set as before, the second consists of articles taken solely from Reuters as an exemplar highly-trusted news source, and finally, the third contains real news articles taken from the entire highly-trusted source list. For each article in each of these data sets, we conduct a search on DuckDuckGo to measure the average text similarity between the article and the top five results returned from the associated search query.

Figure 2 shows the distribution of the average similarity between each article from each data set and the corresponding top five results returned from the search engine. These results show that articles from Reuters display significantly higher similarity with their corresponding search results, compared to experiments with the other data sets. A closer look at the results reveals this may be due to the fact that every story has a different spreading pattern, and is picked up by a different number of news sources. Reuters, being more selective with their stories, mostly concentrates on breaking news that usually finds redistribution on multiple news outlets.

While our fake news data set shows a similar distribution through each average similarity range, our mixed real news data set resembles a Gaussian curve, and our Reuters news data set shows an exponential increase. Investigating these results further, we observe a contrast in distributions starting from the 30%-40% average similarity range in each of these graphs. In other words, additional news sources covering the same story, which are highly likely to appear in the top five search results, tend to have at least a 30%-40% similarity to the queried article. Thus, we chose 40% as our text similarity threshold, and end up with 6 similarity subsets corresponding to each 10% bucket from 40% to 100%.

We empirically pick two days for the date similarity threshold. In the majority of cases we observe, a story is reported by a news source on the specific day the event takes place; however, sometimes, a different news source may cover the same story a day or two later. Beyond two days, the news cycle quickly moves on to fresh information.

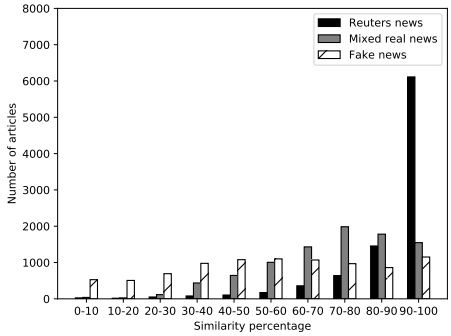


**Figure 2: Distribution of average similarity between news articles and their top five search results.**

**Table 1: Detection performance on testing data using SVM. Different classifiers provided for comparison.**

| Classifier | Accuracy | Precision | Recall | F1 score |
|---|---|---|---|---|
| SVM | 87.26% | 83.88% | 92.15% | 87.82% |
| MLP | 86.96% | 83.67% | 91.75% | 87.52% |
| Random Forest | 85.36% | 81.17% | 91.95% | 86.23% |
| Decision Tree | 81.04% | 75.84% | 90.95% | 82.71% |

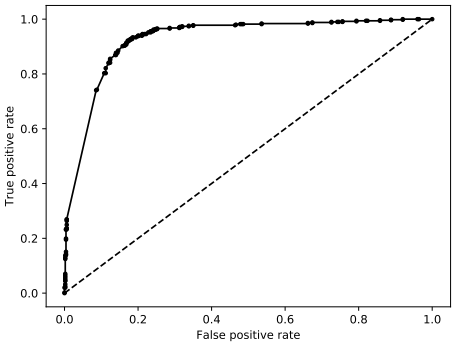


**Figure 3: ROC curve for the SVM model.**

## 5.3 Detection Performance

We train our classifier with FAD-10 and RED-10, and test FADE on FAD-5 and RED-5 to determine the accuracy, false positive, and false negative rates. We summarize the results in Table 1. 87.26% of the fake and real news articles are labeled correctly. 83.88% of real news articles are labeled correctly as "real," and 92.15% of fake news articles are labeled correctly as "fake." The overall F1 score is 87.82%. Figure 3 shows the ROC curve for our classification model. The corresponding *Area Under the ROC Curve* value is 0.926.

We also compare FADE's detection performance to a popular browser extension, B.S. Detector [36], that checks URLs in web content and marks them as unreliable if the domains match a manually-curated blacklist. On FAD-5, B.S. Detector only recognizes 139 stories as fake, compared to FADE's 460.

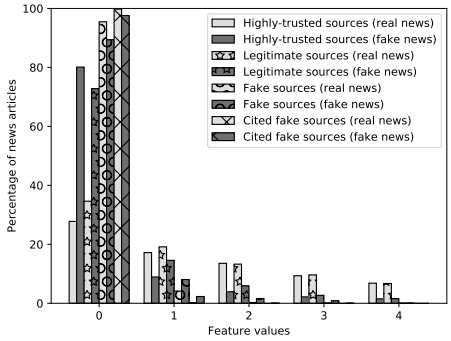


**Figure 4: Distribution of features over news articles.**

## 5.4 Feature Contributions

Next, we investigate the contribution of each feature in the classification of news articles by analyzing FAD-10 and RED-10. Figure 4 summarizes our findings by presenting a distribution of feature values over percentage of news articles. For the sake of brevity, we group together the separate features corresponding to each similarity subset, and present results for our four main feature classes collectively.

*5.4.1 Number of Highly-Trusted Sources.* Our analysis shows that a real news article is almost four times more likely to have at least one highly-trusted source covering the same story compared to a fake news article. In fact, when we search online for 80% of fake news articles, no highly-trusted source covers the same story. In contrast, only 27% of real news articles fail to have a highly-trusted source covering the same story. Clearly, this is a powerful feature.

*5.4.2 Number of Legitimate Sources.* The number of legitimate news sources covering the same story also plays an important role in the classification. 72% percent of fake news articles fail to have a legitimate source covering the same story, while this percentage falls to 34% for real news articles. Fake news articles are more than twice as likely to not have a legitimate source covering the same story compared to real news articles. Also, almost 47% of real news articles have more than one legitimate source covering the same story. In comparison, only 14% of fake news articles have more than one legitimate source covering the story.

*5.4.3 Number of Fake Sources.* 11% percent of fake news articles have at least one fake news source covering the same story, whereas only 4% of real news articles do. The probability of having a fake news source covering the same story for a fake news article is almost three times as high as for a real news article. While helpful, this feature is not as powerful as the two features discussed previously.

*5.4.4 Number of Cited Fake Sources.* 2.3% of fake news articles cite at least one fake news source; in contrast, this value is 10 times smaller (0.23%) for real news articles. This feature, seen more commonly in fake news articles, improves the automated classification; however, it makes the smallest contribution overall.

**Table 2: Detection performance in deployment.**

| Performance Measures | Values |
|---|---|
| Accuracy | 85.53% |
| Precision | 79.83% |
| Recall | 91.23% |
| F1 score | 85.15% |

### 5.5 Real-Time Detection

We conduct a final experiment to observe how FADE performs in a real deployment, where we test recently published news articles using our prototype and Chrome extension. Specifically, every day over ten weeks, we collect labeled fake news articles from Snopes, PolitiFact, and FactCheck.org from the previous day, and real news articles from highly-trusted sources. We then visit and run those articles through our FADE browser extension to perform a fact-check.

Table 2 shows the results of this experiment. Overall, we achieve an accuracy greater than 85%, a result in line with our experiments with the testing data set. However, we also observe a 4% decrease in precision. We believe this is a consequence of working with recently published stories that have not had an opportunity to be indexed by search engines, or spread as widely as those older articles in the testing data set. Otherwise, we do not observe any unexpected usability issues during this experiment, providing us with empirical evidence that FADE is suitable for everyday use.

## 6 DISCUSSION

As our evaluation indicates, FADE is an effective and practical system for online detection of fake news articles. However, our approach does not work equally well in every scenario. In this section, we explore the false positives and negatives we observe in the previous experiments, and describe FADE's limitations. We also discuss how attackers attempting to bypass FADE's detection can potentially deceive the system.

### 6.1 False Positives

FADE incorrectly labels 80 out of 500 (16%) real news articles in RED-5 as fake. We manually investigate the reasons for these false positives and explain our observations below.

Half of FADE's false positives result from *non-news* articles. These 40 articles do not describe facts, but are opinion pieces or listicles. For example, a CNN article lists the top surfing spots in Africa [41]. As these are not likely to be covered by multiple sources, they are prone to being labeled incorrectly.

A second category includes 30 articles that do not cover stories popular enough to be widely reported by other news outlets. One example is a CNN article covering a speech Adebola Williams gave at the Obama Foundation Summit in 2017 [5], which failed to receive widespread interest. As FADE's detection algorithm relies on multiple news outlets covering a story, this category of false positives stem from a fundamental limitation of our approach.

The remaining 10 false positives are due to DuckDuckGo failing to return hits for stories that are in fact reported elsewhere. For example, querying for a Guardian article about a bombing in Somalia [8] does not return any highly-trusted news sources even though The New York Times and CBS News also cover the story. This shows that FADE's performance depends on the efficacy of the chosen search engine.

### 6.2 False Negatives

FADE incorrectly marks 40 out of 500 (8%) fake news articles in FAD-5 as real. As above, we manually analyze these to gain insights into FADE's limitations.

Instead of fabricating a story from scratch, a purveyor of fake news can make subtle changes to a real news story. Furthermore, these fake stories are published around the same date as their real counterparts. For example, we find that an article in FAD-5 modified a real Fox News story. While the original story reports that California Governor Jerry Brown considered signing a bill for reducing penalties for infecting others with HIV [14], the modified version instead claims that Jerry Brown considered signing a bill which allows HIV-positive people to donate blood [2]. Because such fake news stories have a high text similarity to their real counterparts and are published at around the same date, they pass our similarity checks. Subsequently, in 15 cases, FADE mistakenly identifies a highly-trusted secondary source covering the same story and incorrectly decides that this story is real.

Interestingly, 10 false negatives are due to fake stories that are inadvertently picked up by real news sources. For example, a false story published by Newsweek claims that First Lady of Poland Kornhauser-Duda refused to shake hands with Donald Trump, including a video clip from the event in the article [29]. However, an extended version of the video shows that Kornhauser-Duda did in fact shake hands with Trump. The same story is covered by another major source, Yahoo News, which causes FADE to classify this article as real [30].

The remaining 15 false negatives stem from limitations of our date and textual similarity checks. These are not fundamental disadvantages of our approach, but implementation issues that can be addressed with further development effort.

In a recurring instance of this problem, FADE fails to correctly extract the publication date of inspected articles. We opt to skip the date check in such cases, which may lead to incorrect results. For example, a fake story claims that Pluto has been officially reclassified as a planet [39]. Two news articles from BBC [28] and USA Today [6] cover a similar story two years and seven months prior, respectively, but only present the scientific debate around Pluto's reclassification as a dwarf planet. Without proper date checks, the latter two stories from highly-trusted sources are not excluded from consideration, causing FADE to label the fake story as real.

### 6.3 Limitations of the Approach

Below, we present a general analysis of FADE's limitations based on our investigation of the false positives and negatives.

FADE bases its decisions on coverage of queried news articles by multiple sources and reputations of these sources. Certain types of articles and corner cases may diverge from a fundamental assumption we make in our approach, that multiple news outlets pick up and disseminate a given news story. In particular, purely subjective articles such as opinion pieces and non-news are likely to remain exclusive to a single source. FADE is not a good fit to label those articles, but we stress that this is by necessity and design. To reiterate our definition of fake news in Section 2, our work focuses on fabrications that deviate from reality in an evidence-based manner, often disseminated with malicious intent such as spreading damaging rumors. Opinion pieces and non-news are, by definition, speculative in nature, and they are not primary channels for fake news under our definition.

Likewise, our approach is less effective on *local* news and unpopular stories, as these may not enjoy as widespread of coverage as news items of national or international interest. We expect that errors due to this issue can be mitigated by incorporating smaller news outlets into our source lists, or perhaps using them as a separate classification feature.

*Breaking news* may initially pose a similar problem, as it may take time until a sufficient number of sources publish their versions of an emerging story to make accurate classification decisions. However, stories of interest are not likely to remain exclusive, and will rapidly spread as news outlets have strong incentives to compete with each other for attracting readers. Based on this observation, we enhance our FADE implementation to improve usability and avoid creating undue mistrust in such cases. FADE performs a check after the date extraction step to determine whether the queried article has been published very recently (e.g., minutes prior to analysis). If so, FADE warns the user that the results may not yet be accurate. We stress that, even though this remains a limitation of our approach, it is not a limitation *we* introduce. The same fundamental limitation applies when Internet users attempt to confirm the veracity of stories on their own, without FADE.

### 6.4 Limitations of the Prototype

Correctly parsing a queried article for date and term extraction is essential for FADE to perform its similarity checks, and in turn, classification. As we observe in our experiments, shortcomings of our prototype in this respect can lead to false detections. Such implementation limitations may be addressed by improving the underlying libraries we used in our prototype, or otherwise their negative impact reduced via usability enhancements. For example, our prototype displays a notification if it cannot extract the date from an article, warning users that the results may not be reliable.

The external search engine is another key component of our system. Clearly, the efficacy of the search engine used and its particular behavior in ranking results all affect FADE's overall performance. An empirical evaluation of alternative search engines used in combination with FADE is an interesting future experiment we elide in this paper. Note that the choice of search engine may require considerations beyond detection performance. For example, substituting a different search engine for DuckDuckGo could subject FADE to aggressive rate limits, and may introduce costs for paid services.

### 6.5 Attacks against FADE

A purveyor of fake news may attempt to trick FADE by taking a real article, making subtle and misinformative changes to it, and publishing it within our date similarity threshold. While having to create fake stories in this manner severely limits the options an attacker has for launching misinformation campaigns, this is nonetheless the most practical attack against FADE. However, FADE could still detect the attack if other known fake news websites republish the same article.

Other potential attacks include poisoning search engine results so that FADE cannot reliably identify additional sources, or maliciously manipulating the Alexa Top 500 to include fake news websites in FADE's legitimate websites list. Even though these are possible attacks, they are unlikely given that they require extensive effort and resources on the attacker's part.

Lastly, we clarify points that may falsely appear to be FADE limitations. It is not possible to trick FADE by creating a fake news article that references real articles from trusted sources. FADE only uses *references to fake articles* as a detection feature, but not to real articles. In addition, attackers cannot blacklist FADE to avoid detection, as FADE does not directly fetch resources itself, but relies on an external search engine to discover additional sources.

## 7 RELATED WORK

Emerging in the last decade alongside the rise of social media, online fake news has been the focus of a plethora of research studies. While most of the research performed to date focuses on the social aspects of fake news, a relatively small number of studies have investigated technical solutions to the problem.

Nørregaard et al. present a data set of 713K news articles collected from 194 sources over 9 months in 2018. Articles are labeled across multiple dimensions to assist studies of misinformation in news [25].

Karduni et al. investigate how uncertainty and confirmation bias affect users' ability to identify misinformation. They design a visual analytics system called Verifi to show various dimensions (e.g., linguistic features, network interactions) that distinguish fake news from real news accounts on Twitter, and conduct experiments to gain insights into decision-making around misinformation [20].

Zannettou et al. analyze how misleading information originating from alt-right communities ends up in mainstream social networks. Specifically, they focus on alt-right communities hosted on 4chan and Reddit, and how they spread news to Twitter. Their results show that these communities have a surprising level of influence on the Twitter ecosystem [42]. Similarly, Hine et al. analyze the "/pol/" (Politically Incorrect) section of 4chan. Their study reveals that hate speech

is predominant in /pol/, and this section plays an important role in the spread of hate speech on social platforms [17].

Conroy et al. discuss two approaches for fake news detection: A linguistic approach in which contents are analyzed for deceptive message patterns, looking at the frequency of conjunctions, pronouns, and negative emotion words, and a network approach in which factual statements and content metadata can be verified by querying various resources [10].

Shu et al. suggest approaches for fake news detection specifically for social media platforms. They discuss features in two main classes: news content features, and social context features. For news content features, they suggest lexical and syntactic features extracted from the body and title of a news article, and visual features such as the count, clarity, and coherence of media embedded in the content. For social context features, they discuss user demographics, registration age, number of friends, and the number of tweets to identify whether the user is a bot [35].

Wang introduces LIAR, a large data set that contains 12.8K statements fact-checked and labeled manually by PolitiFact, and investigates automated fake news detection on this data set with Convolutional Neural Networks. Wang achieves 27% accuracy with an approach integrating features from text and metadata [40].

Tacchini et al. present a model that detects hoaxes on social media. They use machine learning techniques and a feature set mainly based on the interaction between users and posts. They include features such as the number of users who like the post, the identity of the users, and the number of likes by a single user. They achieve 99% accuracy on their own data set of Facebook posts [37].

Ruchansky et al. propose CSI, a neural network model that leverages the article text, user comments, and users who share them. CSI achieves 89% accuracy on a Twitter data set containing approximately 1000 articles [33].

Shu et al. explore why a machine learning model rates a news article as fake by identifying the decisive sentences of news contents and user comments. They build a neural network on this framework, called dEFEND, and achieve 90% accuracy on a PolitiFact data set containing approximately 400 articles, and 80% accuracy on a Gossip Cop data set containing about 6000 articles [34].

Lin et al. also use the FakeNewsNet data set to evaluate numerous machine learning models. Their approach is based on textual analysis of articles, which includes count features, bag of words features, and sentiment analysis features. They achieve the highest F-1 score of 83% on the PolitiFact portion of the data set and 82% on the Gossip Cop portion [23].

Horne and Adali [18] use features extracted from the text body and title, and SVM classification to label articles as fake or real. They achieve 77% accuracy on a Buzzfeed election data set and 71% accuracy on their own data set.

Burfoot and Baldwin [7] use machine learning to detect satire. They use standard topic-based text and sentiment classification methods, and add features such as use of profanity and slang. They achieve an 80% F1-score on their test data set of 133 satire articles and 1495 real articles.

### Comparing FADE to Related Work

FADE's fake news article detection approach is fundamentally different. In contrast to the papers we discuss above that use textual features and article content in their classification, FADE uses a novel detection approach based on the coverage of a news story by multiple sources and their corresponding reputations. In fact, a core contribution of this paper is demonstrating that news source reputation alone is a sufficiently strong signal for fake news detection. Our approach also easily translates into a practical system for everyday use. Our source code for FADE's server-side logic and browser extension, and the data sets we use in this paper are all open-source and publicly available at *https://github.com/bahruzjabiyev/FADE*.

Unfortunately, we were not able to experimentally compare FADE's detection performance with the prior work. While the source code for CSI, dEFEND, and Tacchini et al.'s approach is publicly available, these works require analyses of user comments and interactions on social platforms. FADE does not rely on these features and is not tied to any specific platform. Therefore, running these tools on our data sets lacking social context is not possible. We also attempted to run FADE on data sets used in related work, but without success. Where data sets with full articles were released, we found that significant portions of the data were outdated with invalid URLs. Furthermore, prior works often use unspecified slices of the data appropriate for their protocol, making it impossible for us to use the same as a benchmark.

Due to the above obstacles, we cannot present a comparative evaluation, but only provide a discussion based on the reported detection numbers. Overall, FADE's 85% accuracy is a significant improvement over many other works. While there are two other approaches that perform strictly better (i.e., Tacchini et al. reports 99%, and CSI reports 89% accuracy), we note that both of these have limitations for practical use and are strongly tied to a single social platform, whereas FADE provides an implementation for everyday use and works with all news articles.

## 8 CONCLUSION

In this work, we propose an effective, practical, and usable approach to enable Internet users to make informed decisions on the veracity of news claims. FADE meets all of our design goals and achieves above 85% fake news detection accuracy in a practical setting.

FADE's novel approach to detecting fake news identifies multiple sources covering the same news story and classifies story authenticity based on those source reputations. FADE does not require costly offline processing of articles for content analysis, and offers a viable path to empowering the Internet community to fact-check news articles in real time.

## ACKNOWLEDGMENTS

We thank Jeremiah Onaolapo and the anonymous reviewers for their helpful comments. This work was supported by the National Science Foundation under grant CNS-1703454. This work was also partially supported by Secure Business Austria.

# REFERENCES


[1] 8Satire. Donald Trump Ends School Shootings By Banning Schools, 2018. https://www.8shit.net/donald-trump-ends-school-shootings-banning-schools/.

[2] Sean Adl-Tabatabai. California Governor Jerry Brown To Allow HIV Positive People To Donate Blood, 2017. https://yournewswire.com/california-jerry-brown-hiv-blood/.

[3] Alexa. Top Sites by Category: News, 2018. https://www.alexa.com/topsites/category/Top/News.

[4] Hunt Allcott and Matthew Gentzkow. Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*, 31(2):211–236, 2017.

[5] Munachim Amah. The powerful speech that got Adebola Williams a standing ovation from Barack Obama, 2017. https://www.cnn.com/2017/11/03/africa/adebola-williams-speech-barack-obama/index.html.

[6] Mary Bowerman. NASA scientists want to make Pluto a planet again, 2017. https://www.usatoday.com/story/tech/nation-now/2017/02/21/pluto-have-last-laugh-nasa-scientists-wants-make-pluto-planet-again/98187922/.

[7] Clint Burfoot and Timothy Baldwin. Automatic Satire Detection: Are You Having a Laugh? In *ACL International Joint Conference on Natural Language Processing Short Papers*, 2009.

[8] Jason Burke. Mogadishu truck bomb: 500 casualties in Somalia's worst terrorist attack, 2017. https://www.theguardian.com/world/2017/oct/15/truck-bomb-mogadishu-kills-people-somalia.

[9] Judy Byington. Pope Francis Found Guilty Of Child Trafficking, Rape, Murder, 2018. http://thepedogate.com/religion/pope-francis-found-guilty-of-child-trafficking-rape-murder/.

[10] Niall J. Conroy, Victoria L. Rubin, and Yimin Chen. Automatic Deception Detection: Methods for Finding Fake News. In *ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community*, 2015.

[11] Baxter Dmitry. Las Vegas: Video Footage Confirms Multiple Shooters, Co-ordinated Attack, 2017. https://yournewswire.com/las-vegas-video-multiple-shootes/.

[12] Elizabeth Dwoskin and Annie Gowen. On WhatsApp, fake news is fast — and can be fatal, 2018. https://www.washingtonpost.com/business/economy/on-whatsapp-fake-news-is-fast--and-can-be-fatal/2018/07/23/a2dd7112-8ebf-11e8-bcd5-9d911c784c38_story.html.

[13] Facebook Newsroom. Hard Questions: How Is Facebook's Fact-Checking Program Working?, 2018. https://newsroom.fb.com/news/2018/06/hard-questions-fact-checking.

[14] Fox News. California governor to consider bill for reducing penalties for infecting others with HIV, 2017. https://www.foxnews.com/us/california-governor-to-consider-bill-for-reducing-penalties-for-infecting-others-with-hiv.

[15] Daniel Funke. A viral fake about Sylvester Stallone highlights a major flaw in Facebook's fact-checking tool, 2018. https://www.poynter.org/fact-checking/2018/a-viral-fake-about-sylvester-stallone-highlights-a-major-flaw-in-facebook's-fact-checking-tool/.

[16] Vindu Goel, Hari Kumar, and Sheera Frenkel. In Sri Lanka, Facebook Contends With Shutdown After Mob Violence, 2018. https://www.nytimes.com/2018/03/08/technology/sri-lanka-facebook-shutdown.html.

[17] Gabriel Emile Hine, Jeremiah Onaolapo, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Riginos Samaras, Gianluca Stringhini, and Jeremy Blackburn. Kek, Cucks, and God Emperor Trump: A Measurement Study of 4chan's Politically Incorrect Forum and Its Effects on the Web. In *The International AAAI Conference on Web and Social Media*, 2017.

[18] Benjamin D. Horne and Sibel Adali. 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 *International Conference of Networks and Communications*, 2017.

[19] Cecilia Kang and Adam Goldman. In Washington Pizzeria Attack, Fake News Brought Real Guns, 2016. https://www.nytimes.com/2016/12/05/business/media/comet-ping-pong-pizza-shooting-fake-news-consequences.html.

[20] Alireza Karduni, Ryan Wesslen, Sashank Santhanam, Isaac Cho, Svitlana Volkova, Dustin Arendt, Samira Shaikh, and Wenwen Dou. Can You Verifi This? Studying Uncertainty and Decision-Making about Misinformation in Visual Analytics. In *The International AAAI Conference on Web and Social Media*, 2018.

[21] Michael W. Kearney. Trusting News Project Report 2017, 2017. https://www.rjionline.org/reporthtml.html.

[22] Kalev Leetaru. 'Fake News' And How The Washington Post Rewrote Its Story On Russian Hacking Of The Power Grid, 2017. https://www.forbes.com/sites/kalevleetaru/2017/01/01/fake-news-and-how-the-washington-post-rewrote-its-story-on-russian-hacking-of-the-power-grid/.

[23] Jun Lin, Glenna Tremblay-Taylor, Guanyi Mou, Di You, and Kyumin Lee. Detecting Fake News Articles. In *IEEE International Conference on Big Data*, 2019.

[24] News Bible Report. JUST IN: Barack Obama's daughter murdered in drive-by shootout, 2017. https://archive.is/gi7fw.

[25] Jeppe Nørregaard, Benjamin D. Horne, and Sibel Adali. NELA-GT-2018: A Large Multi-Labelled News Dataset for the Study of Misinformation in News Articles. In *The International AAAI Conference on Web and Social Media*, 2019.

[26] Pew Research Center. Trust Levels of News Sources by Ideological Group, 2014. http://www.journalism.org/2014/10/21/political-polarization-media-habits/pj_14-10-21_mediapolarization-01/.

[27] PolitiFact. PolitiFact's guide to fake news websites and what they peddle, 2017.

[28] Paul Rincon. Why is Pluto no longer a planet?, 2015. http://www.bbc.com/news/science-environment-33462184.

[29] Chris Riotta. WATCH DONALD TRUMP HANDSHAKE REJECTED BY POLISH FIRST LADY IN HILARIOUSLY AWKWARD EXCHANGE, 2017. http://www.newsweek.com/donald-trump-handshake-poland-president-wife-melania-trump-smack-video-watch-632808.

[30] Chris Riotta. Watch Donald Trump Handshake Rejected By Polish First Lady In Hilariously Awkward Exchange, 2017. https://www.yahoo.com/news/watch-donald-trump-handshake-rejected-153201001.html.

[31] Megan Risdal. Getting Real about Fake News, 2016. https://www.kaggle.com/mrisdal/fake-news.

[32] Kevin Roose. Facebook Banned Infowars. Now What?, 2018. https://www.nytimes.com/2018/08/10/technology/facebook-banned-infowars-now-what.html.

[33] Natali Ruchansky, Sungyong Seo, and Yan Liu. CSI: A Hybrid Deep Model for Fake News Detection. In *ACM International Conference on Information and Knowledge Management*, 2017.

[34] Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. dEFEND: Explainable Fake News Detection. In *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019.

[35] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake News Detection on Social Media: A Data Mining Perspective. In *ACM SIGKDD Explorations Newsletter*, 2017.

[36] Daniel Sieradski, 2017. https://gitlab.com/bs-detector/bs-detector.

[37] Eugenio Tacchini, Gabriele Ballarin, Marco L. Della Vedova, Stefano Moret, and Luca de Alfaro. Some Like it Hoax : Automated Fake News Detection in Social Networks. Technical Report UCSC-SOE-17-05, University of California, Santa Cruz, Santa Cruz, CA, 2017.

[38] United Nations, 2020. https://news.un.org/en/story/2020/04/1061592.

[39] untold-universe.org. Pluto Has Been Officially Reclassified, 2017. http://www.untold-universe.org/2017/09/pluto-has-been-officially-reclassified.html.

[40] William Yang Wang. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In *Annual Meeting of the Association for Computational Linguistics*, 2017.

[41] Olivia Yasukawa and Torera Idowu. Surfs up! 5 of Africa's best surfing spots, 2017. https://www.cnn.com/2017/10/03/africa/surfing-destinations-africa/index.html.

[42] Savvas Zannettou, Tristan Caulfield, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Michael Sirivianos, Gianuca Stringhini, and Jeremy Blackburn. The Web Centipede: Understanding How Web Communities Influence Each Other Through the Lens of Mainstream and Alternative News Sources. In *ACM Internet Measurement Conference*, 2017.