Hidden Biases in Unreliable News Detection Datasets

Xiang Zhou[1], **Heba Elfardy**[2], **Christos Christodoulopoulos**[2]
Thomas Butler[2], **Mohit Bansal**[1]
[1]UNC Chapel Hill [2]Amazon
{xzh, mbansal}@cs.unc.edu
{helfardy, tombutl}@amazon.com, chrchrs@amazon.co.uk

Abstract

Automatic unreliable news detection is a research problem with great potential impact. Recently, several papers have shown promising results on large-scale news datasets with models that only use the article itself without resorting to any fact-checking mechanism or retrieving any supporting evidence. In this work, we take a closer look at these datasets. While they all provide valuable resources for future research, we observe a number of problems that may lead to results that do not generalize in more realistic settings. Specifically, we show that selection bias during data collection leads to undesired artifacts in the datasets. In addition, while most systems train and predict at the level of individual articles, overlapping article sources in the training and evaluation data can provide a strong confounding factor that models can exploit. In the presence of this confounding factor, the models can achieve good performance by directly memorizing the site-label mapping instead of modeling the real task of unreliable news detection. We observed a significant drop (>10%) in accuracy for all models tested in a clean split with no train/test source overlap. Using the observations and experimental results, we provide practical suggestions on how to create more reliable datasets for the unreliable news detection task. We suggest future dataset creation include a simple model as a difficulty/bias probe and future model development use a clean non-overlapping site and date split.[1]

1 Introduction

The proliferation of unreliable news is widely acknowledged (Del Vicario et al., 2016; Lazer et al., 2018; Vosoughi et al., 2018), and its identification is a socially important problem. In this work we use the label unreliable news as a broad term for all unverifiable and misleading news content, regardless of whether the content is malicious (targeted misinformation) or not. Accordingly, while specific definitions vary in different datasets used in this work, we refrain from using the term "fake" since identifying the intent of the author(s) is beyond the scope of this work. To mitigate the problem of surfacing unreliable news content, various websites (e.g., PolitiFact[2], Media Bias/Fact Check (MBFC)[3], GossipCop[4], etc.) determine the reliability of news by manually fact-checking the important claims in given news articles. Beyond requiring investigative expertise, manual fact-checking is time-consuming and is thus limited to only a small set of selected news articles.

Recent research has explored automating this process using machine learning methods to automatically determine news veracity (Pérez-Rosas et al., 2018; Baly et al., 2018; Nie et al., 2019; Wright and Augenstein, 2020). These efforts were made possible due to the availability of large-scale unreliable news detection datasets (Horne et al., 2018b; Shu et al., 2017; Wang, 2017). In our work, we examine if these datasets accurately reflect the real difficulty of this task or if there are any hidden biases in the datasets. Specifically, we study different methods of dataset construction (e.g., how the data was collected, how the data was split, etc.) and show that the assessed difficulty of the task is sensitive to how carefully different factors are considered when building and using these datasets.

Our investigation begins with data collection procedures: we look at the source of news stories (news outlets, social media, fact-checking websites, etc.) as well as the annotation process (number of

[1]Our code is publicly available at https://owenzx.github.io/unreliable_news

[2]https://www.politifact.com/
[3]https://mediabiasfactcheck.com/
[4]https://www.gossipcop.com/

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 2482–2492
April 19 - 23, 2021.

Data Collection	Dataset Construction	Experiment Design
1. Collect from less biased or unbiased resources (e.g. original news outlets). (Sec. 3.2)	1. Examine the most salient words to check for biases in the datasets. (Sec. 3.2)	1. Apply debasing techniques when developing models on biased datasets. (Sec. 3.2)
2. Collect from diverse resources (in terms of sources, topics, time, etc.). (Sec. 3.2, 4)	2. Run simple BoW baselines to check how severe the bias is. (Sec. 4)	2. Check the performance on sources/dates not in your training set. (Sec. 4.2, Sec. 4.3)
3. Collect precise article-level labels if possible. (Sec. 3.1)	3. Provide train/dev/test splits with non-overlapping source/time. (Sec. 4.2, 4.3)	3. Check the performance on sources with limited examples. (Sec. 4.4)
		4. Test your model on multiple complementary datasets (e.g. with different domains, styles, etc.). (Sec. 3.2, 4.4)

Table 1: Suggestions for data collection, dataset construction and experiment design for unreliable news research.

labels, granularity of labels, article- or site-level annotation). We discuss the pros and cons of each approach and point out some hidden pitfalls. Using FakeNewsNet (Shu et al., 2017) as an example, we demonstrate how selection biases in data collection can lead to undesired biases in the created datasets.

Moving beyond data collection, we examine two commonly applied ways of splitting the dataset for training and testing that help the model achieve high performance without correctly modeling the task. Specifically, we show that using a disjoint set of sites/news outlets for training and test data significantly decreases the models' performance ($>$10%) and that the drop in performance is related to how similar (or dissimilar) the sites in both sets are (reflected by various site-level distributional distance metrics including L2, cos, EMD, etc.). Additionally, we also examine the effect of time overlap between both train and test sets. We observe that different news outlets are likely to have similar content in a small time window (i.e., the same story gets covered by multiple outlets within a day or a few days period). While we do not find any evidence that the studied models exploit this factor, we nevertheless suggest that future datasets are split both by time and site/news outlet.

In summary, our main contributions are: (1) showing how data collection procedures can lead to systematic biases in unreliable news datasets, (2) demonstrating how confounding factors——such as site/news outlet and time——in these datasets can degrade their quality and lead to underestimating the difficulty of the task, and finally (3) suggesting possible mechanisms to avoid these biases and confounding factors when building new datasets. To facilitate future research, we also provide a list of practical suggestions for data collection, dataset construction, and experiment design in Table 1.

2 Related Work

Unreliable News Detection. Unreliable news detection and other news veracity related tasks have been receiving an increasing focus as news sources have become more accessible in recent years. A lot of effort has been put into collecting high-quality datasets. Wang (2017); Shu et al. (2017) collected manually labeled statements or news articles from fact-checking websites. The NELA datasets (Horne et al., 2018b; Nørregaard et al., 2019; Gruppi et al., 2020) scrape news articles directly from news outlets and use the manually annotated labels from Media Bias/Fact Check (MBFC) as site-level annotations. Social media is also a popular resource for collecting news stories (Nakamura et al., 2020; Santia and Williams, 2018; Mitra and Gilbert, 2015). Researchers have also collected datasets for various related topics, such as rumor detection (Kwon et al., 2017; Ma et al., 2016), and propaganda detection (Da San Martino et al., 2020; Barrón-Cedeno et al., 2019). Besides classifying the veracity of news articles, researchers have also explored related problems, such as predicting the reliability of news sites (Baly et al., 2018), identifying fact-check worthy sentences (Wright and Augenstein, 2020), among other tasks. Several recent papers also focus on measuring the trustworthiness of single statements (Wang, 2017; Pomerleau and Rao, 2017; Alhindi et al., 2018). In this work, we focus on article-level classification because of its relevance to applications, like news feeds, that operate at the article level.

Dataset	Size	Article Source	Label Type
NELA[5]	136K/713K/1.12M	News outlets	Site-level
FakeNewsNet[6]	603K	Fact-checking websites	Article-level
r/Fakeddit[7]	1.06M	Social Media (Reddit)	Site-level

Table 2: Statistics and properties of three recent large-scale unreliable news datasets. The three statistics of NELA dataset sizes correspond to its three versions released in 2017, 2018 and 2019, respectively.

Pitfalls in Data Collection. Datasets collected through crowd-sourcing or scraping the Internet have the advantage of much better scalability compared to expert-annotated datasets. However, these automatic processes are prone to hidden pitfalls. Gururangan et al. (2018); Poliak et al. (2018) show that crowd-sourcing "Natural Language Inference" datasets leads to various dataset biases. Similar observations have been made for "Fact Verification" datasets (Schuster et al., 2019). Splitting data–—for training, testing, and validation——is another important procedure in creating datasets that can lead to several problems. For example, Geva et al. (2019) show that models may just learn the patterns of certain annotators in a random split. Lewis et al. (2020b) demonstrated a significant overlap in current open-domain QA datasets. When present, these unexpected biases or overlaps in datasets can significantly undermine the utility of a dataset and lead to deceptively promising results that are in part due to artifacts of flaws in the dataset rather than successfully modeling the intended task.

Automated Fact Checking for Statements. Automated fact checking is an important task closely related to unreliable news detection, yet is constructed in a more controlled manner. This task focuses on strictly judging the factuality of one single statement instead of an entire article. Vlachos and Riedel (2014) first constructed a dataset with 106 claims from fact-checking websites with paired labels. FEVER (Thorne et al., 2018) is currently the largest scale fact-verification dataset, where 185,445 claims were generated by modifying sentences from Wikipedia. Both the altered claims and the ground truth supporting evidence are included in the dataset. Existing effective approaches for fact-verification include self-attention based networks (Nie et al., 2019), large-scale pretrained transformers (Soleimani et al., 2020), neural retrieval methods (Lewis et al., 2020a), and reasoning on semantic-level graphs (Zhong et al., 2020).

3 Unreliable News Datasets

Collecting high-quality datasets plays an important role in automatic unreliable news detection research. Here we review dataset collection strategies used in constructing recent datasets and point out some hidden pitfalls in these procedures.

3.1 Data Collection Strategies

Unreliable news detection is usually formalized as a classification task. Accordingly, constructing a dataset requires collecting pairs of news articles and labels.

News Articles: Each individual news outlet has its own website where news articles are published. The easiest way to collect a large number of these articles is to simply scrape these websites. Manual annotation or some other mechanism must then be incorporated in order to collect the corresponding labels for each article. Another common way to collect articles is through fact-checking websites. While this approach provides both articles and article-level labels, it normally only provides a limited set of articles. Additionally, scraping these fact-checking websites can lead to additional selection bias in the dataset as highlighted in Section 3.2.

One other recent trend is collecting posts and corresponding labels from social media (Nakamura et al., 2020; Santia and Williams, 2018; Mitra and Gilbert, 2015). While large-scale datasets can be collected through such an approach, they are often noisier than those collected through traditional news sources, due to a more casual use of language, and a heavier dependency on the context.

News Labels: The largest challenge in collecting these datasets lies in collecting labels. Manually checking the factuality (or reliability) and bias of a single article is time-consuming and requires non-trivial expertise. Modeling such a task through a crowd-sourcing framework is diffi-

[5] dataverse.harvard.edu/dataverse/nela
[6] github.com/KaiDMML/FakeNewsNet
[7] github.com/entitize/Fakeddit

FakeNewsNet		r/Fakeddit	
Positive Features	Negative Features	Positive Features	Negative Features
season	trump	psbattle	clicks
at	brad	says	colorized
2018	pitt	sues	2018
the	jenner	accused	2019
awards	jennifier	sells	mrw

Table 3: Top five most salient features in the FakeNewsNet dataset and the r/Fakeddit dataset. The features are the highest weighted Bag-of-Word features learned by a Logistic Regression model.

News Outlets	Daily Mail
Site Label	Unreliable
Dates	2018/09/06
Title	Roy Moore sues Sasha Baron Cohen
Article	Failed Senate candidate Roy Moore is suing comedian Sacha Baron Cohen for $95 million for tricking him into appearing on his Showtime program 'Who is America?' Moore, whose bid for the Alabama failed in the wake of claims he molested a 14-year-old, filed the lawsuit in Washington DC on Wednesday...

Table 4: An example showing a reliable news article from the "*Daily Mail*' site which has a "Low" factual reporting rate on MBFC. Despite coming from a source with low reliability score, the shown article is reliable and very similar to the content on sites with high reliability scores (such as *"BBC"* and *"The Week UK"*) on the same date.

cult. As such, current research datasets almost exclusively rely on existing resources. As discussed earlier, these resources either provide article-level or site-level labels. **Article-level** labels are only available through a few fact-checking websites such as PolitiFact, GossipCop, etc., but the scale is limited since generating these labels is time-consuming and costly. **Site-/Outlet-level** labels, on the other hand, available through websites such as MBFC, provide manual labels for each site/outlet. These websites often assign reliable/unreliable or biased/unbiased labels to each news outlet. Many datasets for unreliable news detection assign these site-level labels to all articles in a given site. While these weak or distant labels are not always accurate (one example is shown in Table 4) , they provide an easy way to create large-scale datasets. In Table 2, we highlight three recent large-scale unreliable news datasets along with their data collection procedure.

3.2 Dataset Selection Biases

Datasets annotated without expert verification (e.g., through crowdsourcing, automatic web scraping, etc.) can have some undesired properties that undermine their quality (Gururangan et al., 2018; Poliak et al., 2018; Schuster et al., 2019). In the following analysis, we choose the FakeNewsNet dataset (Shu et al., 2017) as a representative example.

We first examine the most salient features in the dataset. To achieve this, we train a Logistic Regression (LR) model on the titles of FakeNewsNet using Bag-of-Words features and show the word features with the highest weights for each class in Table 3.[8] The features in the table show clear patterns: the top-features for the reliable (positive) class are either stop words (e.g., 'at', 'the', etc.) or words presumably carrying neutral semantics (e.g. 'season', '2018', 'awards', etc.) while the top features for the unreliable news (negative) class are mostly celebrity names. Using this basic model, we achieve an accuracy of ∼78% , while using a BERT-based model that uses both the article and title as input only achieves an incremental improvement yielding an accuracy of 81% (see Sec. 4.1 for detailed model descriptions). By examining the articles in the dataset, we attribute this to the selection bias exhibited by fact-checking websites. Most unreliable (negative) articles contain click-bait titles mentioning celebrities, while reliable sources usually have less sensational titles with fewer mentions of celebrities and more diverse keywords.

Another potential problem is the articles' retrieval framework. FakeNewsNet uses Google search to retrieve the original news article (Shu et al., 2017). Internet search engines have proprietary news ranking and verification processes, which means that even when using the original title and source of a given article, the search results

[8] We also calculated the PMI between the label and word features as suggested by Gururangan et al. (2018) and found the two lists to be very similar.

Label Resource	GossipCop
Title	NYC terror attack: Celebrities react on social media
Article	Celebrities are sending their love and support to New York on social media following a terror attack that left eight people dead Tuesday when a truck plowed down pedestrians on a bicycle path near the World Trade Center in Lower Manhattan...
Label	Unreliable
News URL	tinyurl.com/yxhvdne6

Table 5: One example from the FakeNewsNet dataset where it is difficult for the article content to support the label. This article contains celebrities' reactions after a terrorist attack. While the article itself does not look like a standard news piece, the reactions in the article are all paired with tweets, so the unreliable label seems to be inconsistent.

might prioritize specific sites over others leading to inaccurate data collection. While Shu et al. (2017) propose several heuristics to handle these problems, it is unlikely that this noisy process is completely fixed. As a result, we find a few mis-matched title-content pairs where the retrieved article cannot support the label, hence making the example confusing. We show one example with a questionable label in Table 5, where we suspect the inconsistency is due to the noisy retrieval step.

Finally, the informal nature of user-generated content on social media may be the source of additional biases. In our preliminary experiments, we found that in r/Fakeddit dataset, a simple Bag-of-Words(BoW)-based logistic regression model can reach equal—or even better—performance than the reported BERT-based models (86.91% vs. 86.44% in the text-only two-way classification setting), hinting at the strong correlation between the label and lexical inputs. This is also reflected in the equally confusing most salient features in this dataset shown in Table 3.

Since different collection procedures and data resources will lead to different problems, there is no uniform solution to producing a completely bias-free dataset. However, one good test is to check the performance of a simple model such as a BoW-based linear model. By analyzing the features learned by the simple model as well as measuring the gap between the performance of a state-of-the-art system and the simple model, one can get a hint of the dataset quality. Unreasonable features, together with small performance gaps, may reveal unwanted biases in the dataset. In practice, we also suggest that when developing models using biased datasets to use debiasing techniques (e.g. Schuster et al. (2019)).

4 Dataset Split Effect

In this section, we study the effect of time and site/outlet overlap between the training and the evaluation set on the model's performance and show how these confounding factors can impact it.

4.1 Baseline Models & Experimental Setup

In the following experiments, we use two models: a logistic regression baseline and a state-of-the-art large-scale pretrained Transformer-based model (RoBERTa; Liu et al. (2019)).

Logistic Regression (LR): We use scikit-learn's (Pedregosa et al., 2011) implementation of Logistic Regression along with TFIDF-based Bag-of-Words features. We add L2 regularization to the model with a regularization weight of 1.0 and train the model using L-BFGS. In our experiments, the LR model uses only the title (and not the article body) as the input.

RoBERTa: Our implementation is based on the Transformers library (Wolf et al., 2019) and AllenNLP (Gardner et al., 2017). We use RoBERTa in two different ways, one takes only the title as the input, the other takes both the title and the article content as the input and formalizes the task as pairwise sentence classification. Specifically, we concatenate the title and the article content with a `[SEP]` token in the middle and use different token type embeddings to differentiate between the title and the content. Articles are truncated to fit the 512-token length limit. In the title-only setting, the batch size is set to 32, the learning rate is set to 5e-5, and the model is trained for 3 epochs. In the article+title setting, the batch size is set to 8, the learning rate is set to 2e-5, and the model is trained for 10 epochs. These hyperparameters are set empirically, and our preliminary experiments show that the results are not sensitive to different settings of these hyperparameters.

Datasets: Here, our analysis focuses on the 2018 version of the NELA dataset (Horne et al., 2018b). Unlike FakeNewsNet, NELA gathers news directly from news outlets, so the influence of selection bias is insignificant. Thus we focus our analysis on

Model	Input	Random Split	Source Split (Article)	Source Split (Site)
Majority	/	50	50	69.29 (0.56)
LR	Title	77.45	67.18 (4.13)	79.28 (5.27)
RoBERTa	Title	85.22	70.40 (4.28)	87.83 (10.44)
RoBERTa	Title+Article	96.94	80.36 (11.91)	85.14 (8.00)

Table 6: Accuracy on validation sets with different split strategies. For "Source Split', we report the mean and standard deviation (in parentheses) of five different runs. The last column shows the aggregated site-level accuracy.

other potentially confounding factors in the dataset. We use the latest aggregated site-level labels provided in NELA-GT-2019 (Gruppi et al., 2020) and report both the article- and site-level accuracy. For article-level accuracy, we assign the site-level label to all articles from that news outlet and calculate per-article accuracy. For the Source (Site) Split setting (with no overlap between training and evaluation sites), we also report the site-level accuracy: we aggregate the predictions over individual articles for a given outlet and use the majority prediction as the site-level prediction. We use a balanced label distribution for all dataset splits.

The results in the third column of Table 6 show the models' performance on the random split, which is the default split method used in most papers, e.g. (Nakamura et al., 2020; Horne et al., 2018a). As the results show, even the simplest logistic regression model achieves an accuracy of over 77% whereas the RoBERTa model using both title and the news article as the input reaches almost 97% accuracy.

4.2 Effect of Split by Source

For this experiment, instead of using the standard random split of all the news articles in the dataset, we first randomly split all the sites in the dataset into three disjoint sets (train/dev/test) before adding all articles from each site to their assigned set (train, dev or test). We believe this setup is closer to real-world tasks. For instance, in order to block all unreliable news sources, one simple—yet useful—approach is to maintain a list of questionable sources. All the news from those sources will be automatically blocked. In this setting, the only remaining task is classifying sources with no or very few annotated examples. As the results in Table 6 show, there is a significant drop in performance for all the models when compared to the random split. The logistic regression model's performance drops from 77.5 to 67.2%, and even the more powerful RoBERTa model with both title and article as input drops from 96.9 to 80.4%, demonstrating the task's significantly increased difficulty. While aggregating article-level results to site-levels can significantly improve the accuracy, we also see a plateauing trend of the performance where adding the article as additional input brings no further improvement to the RoBERTa model. Since we subsample the original dataset and balance the number of news articles for each label, the majority baseline (at the article level) is always 50%. But the site-level majority baseline is well above random (69.29%). While a new 50% majority baseline can be achieved by re-subsampling the dataset, the current number also indicates a severe imbalance of dataset size between reliable/unreliable sites which can—potentially—be exploited by the models.

Model	Input	Gold Label	Rand. Label
Majority	/	50	50
LR	Title	77.45	66.29
RoBERTa	Title	85.22	74.37
RoBERTa	Title+Article	96.94	95.04

Table 7: Article-level accuracy for the random label experiments compared to gold site labels.

Random Label Experiments: For this experiment, we use the original random split strategy. However, we permute all the site-level labels randomly. Hence each label no longer represents the reliability of the site, and is just an arbitrary feature of the site itself. Therefore, the only way for the models to achieve good performance on this task is to memorize the arbitrary site-label mapping. The results in Table 7 show that the models achieve very high accuracy with the more powerful RoBERTa model with both title and article showing only ~2% accuracy loss when compared to the true labels. These results demonstrate the models' ability to memorize random site-labels, and the similarity between these results and the results on the random splits suggest that the models are bypassing the real task of reliable/unreliable news classification and are just memorizing the site identities.

Distance	Top 10 Sites	Bottom 10 Sites
l2	11.59	5.79
cosine	210.72	82.91
MMD	7.78	4.20
CORAL	29.07	14.95

Table 8: Average similarity score between sites in the evaluation and training sets.

Performance Variance and Site Similarity Analysis: Another interesting observation from the results in Table 6 is that while the performance on every random split is fairly stable, the performance is much more unstable with respect to splitting by source. For example, the RoBERTa (Title+Article) model results have a standard deviation larger than 10 points, with the highest accuracy reaching over 90% and the lowest one below 60%.

One potential factor behind the varying performance is the heterogeneity of different news sources (sites). News sites that are similar to those in the training set could be much easier to classify than sites with completely different styles or content. In this case, even when splitting by site, correlations between the content of similar sites in the training and evaluation sets may drive the generalization performance. To assess this hypothesis, we measure the dependence on the distances between sites in the training and evaluation sets and the model performance at the site level in the evaluation set. Given a set s in the evaluation set, we measure its similarity to all the sites in the training set $t \in S_{train}$. Below we show that higher accuracy on the site s is associated with a higher similarity between s and sites in the training set with the same label $t \in S_{same}$, providing evidence in favor of our hypothesis.

In order to measure the similarity between different sites, we take the representation learned by the RoBERTa model as the representation of the article with a focus on its reliability. Since the RoBERTa model feeds the whole sentence into the multi-layer transformer architecture and feeds the representation of `[CLS]` token to the downstream classifier (Devlin et al., 2019; Liu et al., 2019), we use the same `[CLS]` representation as the representation for the whole title+article input.

For similarity-metrics between sites, we follow Guo et al. (2020) and calculate the l2-distance, cosine distance, MMD (maximum mean discrepancy) distance (Gretton et al., 2012; Li et al., 2015) and the CORAL (correlation alignment) distance (Sun and Saenko, 2016; Sun et al., 2016). Following Guo et al. (2020), the l2 and cosine distances are calculated by first averaging all the example representations to get the site representation and calculating the distance between site representations; the MMD distance is calculated using an unbiased finite sample estimate from Li et al. (2015); and the CORAL distance is calculated by $D_{CORAL} = \frac{1}{4d^2}\|C_s - C_t\|_F^2$, where d is the feature dimension, C_s and C_t are the co-variance of two sets and $\|\cdot\|_F^2$ is the squared matrix Frobenius norm. To simplify our analysis, we filter out all the sites containing less than 100 examples (assuming the articles from these sites are too few to significantly influence the model). For every site in the evaluation set s, we calculate its distance with respect to every different site t in the training set, and then compare its minimum distance w.r.t the subset of sites with the same gold label S_{same} and the subset of sites with the opposite label S_{oppo},

$$sim\_score_s = \frac{\min\limits_{t \in S_{oppo}} \{dist(s,t)\}}{\min\limits_{t \in S_{same}} \{dist(s,t)\}}$$

We compute this ratio using all four distances above for the top and bottom 10 sites in the evaluation datasets (ranked based on their accuracy with RoBERTa) and report the mean over all the sites and over all five different random splits in Table 8. The top 10 sites always have a much larger similarity score than the bottom 10 sites, indicating that they have a much larger similarity with sites in the training sets with the same label. This trend holds across all of the distance metrics. The sensitivity of performance on the site similarity raises additional concerns about how the results in Table 6 may generalize in real-life. As newly emerged unreliable sites are likely to behave differently from old sites, the model's performance may be on the lower end of the variance.

As a natural extension, we also explored building a model that directly optimizes these site-level distance metrics in order to have better site-level generalization performance. However, in our preliminary results, our model does not show significant improvement from the baseline models. This can also hint at the fact that it is very difficult for these models to extract features that are useful to the task of reliable/unreliable news classification itself and instead learn site-specific features.

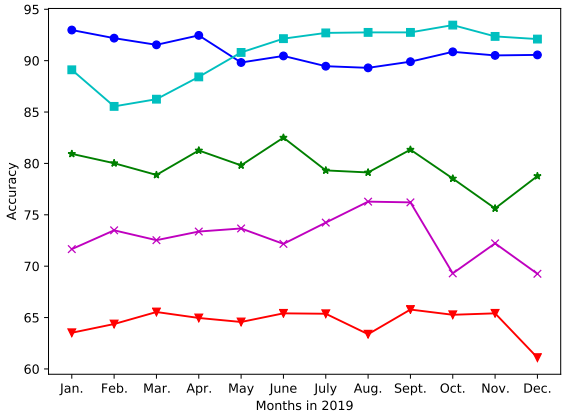


Figure 1: Accuracy of RoBERTa models trained on NELA-GT-2018 and tested on articles from the 12 months covered in the NELA-GT-2019 dataset. The five different lines in the figure represent models trained using five different random site splits.

4.3 Effect of Split By Time

Another potentially important factor to consider while creating train/test/dev splits for a news-based dataset is time. As news-worthy events happen everyday, multiple news articles from different outlets can report the same event. For example, in the NELA 2018 dataset (Nørregaard et al., 2019), within a period of two days (from 2018/10/01 to 2018/10/02), there are more than 100 news articles from over 60 sources about the US-Canada-Mexico trade accord. Therefore, by remembering the content of the event from one article, the model can easily predict the label for any related news article.

To test the effect of time, we examine the model's performance on news articles from a temporally disjoint dataset. Specifically, since all our models are trained on the NELA-GT-2018 (Nørregaard et al., 2019), we use the NELA-GT-2019 (Gruppi et al., 2020) as the evaluation dataset. We split the news articles in 2019 into twelve months and plot the performance trend in Figure 1. We can see that, unlike the significant performance drop in the source split experiments, we do not observe a clear correlation between the performance and the length of the time gap. Therefore, at least for the current models and datasets, splitting by time does not significantly influence the current results. This finding may result from that the fact that the model is not memorizing the exact events in the training set (this is not limited to the unreliable news domain), or it could be attributed to the noise in the training set (similar events can be reported both in reliable and unreliable sources). However, we do have to point out that our current observation only holds for the current models, and it is possible for more powerful models to memorize all events. In addition, the widest time gap tested here is still within a couple of years, which is still a relatively short time in terms of news events. A longer time gap (or a major event such as COVID-19) may lead to different behavior by the models. So in practice, we nonetheless suggest splitting datasets by time to avoid these issues.

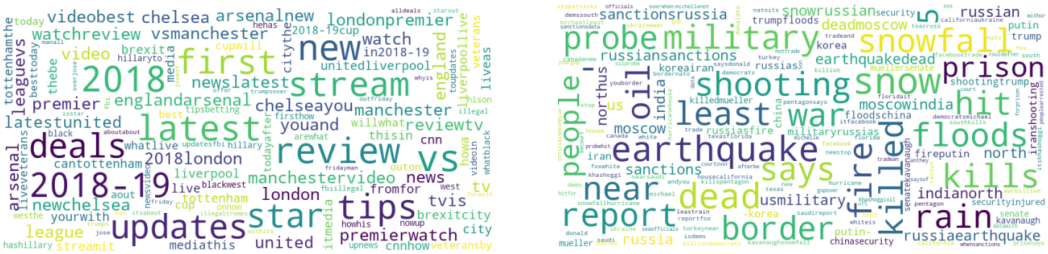
4.4 Error Analysis

Here, we conduct an error analysis to see how the model performs with respect to the variation of some other factors of practical interest, such as topic and site size.

The Influence of Topic in Article-Level Prediction: In order to gain better insight on the performance drop in the source split experiments, we perform a deeper investigation of the numbers in Table 6. We first check whether the models show different performance on different topics. To get a high-level understanding of what the topics are, we look at the titles of articles in the evaluation set and calculate words with the highest PMI with the accuracy of prediction of the RoBERTa model. We then use these PMI values as weights and plot the word cloud figures in Figure 2. In the word cloud of correct predictions, we observe many words related to sports events, while words in the incorrect predictions cloud mostly appear in political news. This is not surprising since there is much more of an incentive to interfere with political news than sports news — making the need for more robust models even more pressing for real-world applications.

The Influence of Size in Site-Level Prediction: Finally, we examine the effect of prediction aggregation from article-level to site-level. Unlike in current datasets where most sites can have hundreds or even thousands of articles, a newly-emerged news outlet waiting for classification may only have a very limited number of articles. Accordingly, while in Table 6, we see a general improvement of the aggregation, it is also important to check the aggregation effect when the number of articles in a given site is small.

In Figure 3 we plot the performance of 5 different runs of the RoBERTa (Title+Article) model



(a) Word cloud of article titles with correct predictions.

(b) Word cloud of article titles with incorrect predictions.

Figure 2: Word cloud of article titles. The words with highest PMI to the prediction correctness of the RoBERTa model are selected.

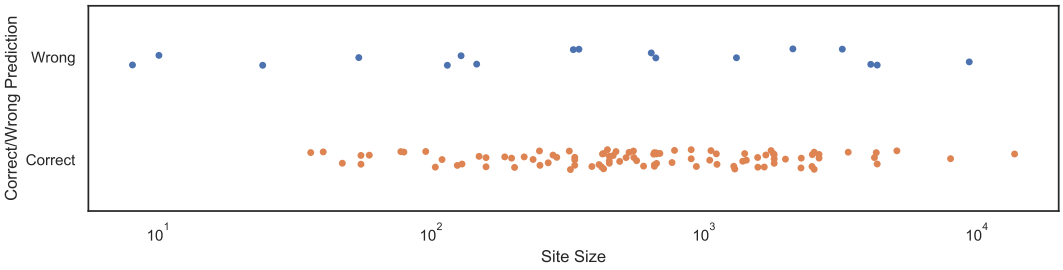


Figure 3: Site-level prediction accuracy of the RoBERTa (Title+Article) model vs. numbers of article in the site (in all five random runs). Blue circles denote wrong predictions and red circles denote correct predictions.

against the number of articles on a given site. We can see that the performance is worse when the size of the site is less than 100, demonstrating the difficulty of predicting the reliability of a site given limited resources. It is also surprising to see a significant number of errors even when the site size is over 1000. This indicates the limitation of simply aggregating the site-level prediction at test-time. Capturing the article-site hierarchy in a better way is a potential future research direction.

5 Conclusion

In this paper, we took a closer look at current large-scale unreliable news detection datasets. We studied their collection procedures and dataset split strategies, and pointed out important flaws in the current approaches. Specifically, we demonstrated that selection bias in dataset collection that often leads to undesired and significant artifacts in these datasets; highlighting confounding factors (e.g., article source, time) in news datasets that can lead to underestimating the difficulty of the task. Finally we provide suggestions on how to better create and process such datasets in the future. We hope our work leads to more high-quality news datasets and that it inspires further work in this direction.

Acknowledgments

We thank the reviewers for their helpful comments. XZ interned at Amazon. This work was also supported by ONR Grant N00014-18-1-2871, DARPA YFA17-D17AP00022, and DARPA KAIROS Grant FA8750-19-2-1004. The views contained in this article are those of the authors and not of the funding agency.

References

Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 85–90, Brussels, Belgium. Association for Computational Linguistics.

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 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3528–3539.

Alberto Barrón-Cedeno, Israa Jaradat, Giovanni Da San Martino, and Preslav Nakov. 2019. Proppy: Organizing the news based on their propagandistic content. *Information Processing & Management*, 56(5):1849–1864.

Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeno, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. 2020. A survey on computational propaganda detection. *arXiv e-prints*, pages arXiv–2007.

Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. 2016. The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, 113(3):554–559.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. 2017. Allennlp: A deep semantic natural language processing platform. *arXiv:1803.07640*.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1161–1166.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773.

Maurício Gruppi, Benjamin D Horne, and Sibel Adalı. 2020. Nela-gt-2019: A large multi-labelled news dataset for the study of misinformation in news articles. *arXiv preprint arXiv:2003.08444*.

Han Guo, Ramakanth Pasunuru, and Mohit Bansal. 2020. Multi-source domain adaptation for text classification via distancenet-bandits. In *AAAI*, pages 7830–7838.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. In *NAACL-HLT (2)*.

Benjamin D Horne, William Dron, Sara Khedr, and Sibel Adali. 2018a. Assessing the news landscape: A multi-module toolkit for evaluating the credibility of news. In *Companion Proceedings of the The Web Conference 2018*, pages 235–238.

Benjamin D Horne, William Dron, Sara Khedr, and Sibel Adali. 2018b. 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*.

Sejeong Kwon, Meeyoung Cha, and Kyomin Jung. 2017. Rumor detection over varying time windows. *PloS one*, 12(1):e0168344.

David MJ 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):1094–1096.

Patrick Lewis, Ethan Perez, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020a. Retrieval-augmented generation for knowledge-intensive nlp tasks. *arXiv preprint arXiv:2005.11401*.

Patrick Lewis, Pontus Stenetorp, and Sebastian Riedel. 2020b. Question and answer test-train overlap in open-domain question answering datasets. *arXiv preprint arXiv:2008.02637*.

Yujia Li, Kevin Swersky, and Rich Zemel. 2015. Generative moment matching networks. In *International Conference on Machine Learning*, pages 1718–1727.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks. In *25th International Joint Conference on Artificial Intelligence, IJCAI 2016*, pages 3818–3824. International Joint Conferences on Artificial Intelligence.

Tanushree Mitra and Eric Gilbert. 2015. Credbank: A large-scale social media corpus with associated credibility annotations. In *ICWSM*, pages 258–267.

Kai Nakamura, Sharon Levy, and William Yang Wang. 2020. Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6149–6157.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6859–6866.

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*, volume 13, pages 630–638.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191.

Dean Pomerleau and Delip Rao. 2017. The fake news challenge: Exploring how artificial intelligence technologies could be leveraged to combat fake news. *Fake News Challenge*.

Giovanni C Santia and Jake Ryland Williams. 2018. Buzzface: A news veracity dataset with facebook user commentary and egos. In *Twelfth International AAAI Conference on Web and Social Media*.

Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. 2019. Towards debiasing fact verification models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3410–3416.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36.

Amir Soleimani, Christof Monz, and Marcel Worring. 2020. Bert for evidence retrieval and claim verification. In *European Conference on Information Retrieval*, pages 359–366. Springer.

Baochen Sun, Jiashi Feng, and Kate Saenko. 2016. Return of frustratingly easy domain adaptation. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 2058–2065.

Baochen Sun and Kate Saenko. 2016. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pages 443–450. Springer.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22.

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

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Dustin Wright and Isabelle Augenstein. 2020. Fact check-worthiness detection as positive unlabelled learning. *arXiv preprint arXiv:2003.02736*.

Wanjun Zhong, Jingjing Xu, Duyu Tang, Zenan Xu, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. 2020. Reasoning over semantic-level graph for fact checking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6170–6180, Online. Association for Computational Linguistics.