Leveraging Exposure Networks for Detecting Fake News Sources

Maor Reuben
Ben-Gurion University of the Negev
Software and Information Systems Engineering
Beer Sheva, Israel
maorreu@post.bgu.ac.il

Lisa Friedland
Independent researcher
Boston, USA
LisaDFriedland@gmail.com

Rami Puzis
Ben-Gurion University of the Negev
Software and Information Systems Engineering
Beer Sheva, Israel
puzis@bgu.ac.il

Nir Grinberg
Ben-Gurion University of the Negev
Software and Information Systems Engineering
Beer Sheva, Israel
nirgrn@bgu.ac.il

ABSTRACT

The scale and dynamic nature of the Web makes real-time detection of misinformation an extremely difficult task. Prior research mostly focused on offline (retrospective) detection of stories or claims using linguistic features of the content, flagging by users, and crowdsourced labels. Here, we develop a novel machine-learning methodology for detecting fake news *sources* using active learning, and examine the contribution of network, audience, and text features to the model accuracy. Importantly, we evaluate performance in both offline and online settings, mimicking the strategic choices fact-checkers have to make in practice as news sources emerge over time. We find that exposure networks provide information on considerably more sources than sharing networks (+49.6%), and that the inclusion of exposure features greatly improves classification PR-AUC in both offline (+33%) and online (+69.2%) settings. Textual features perform best in offline settings, but their performance deteriorates by 12.0-18.7% in online settings. Finally, the results show that a few iterations of active learning are sufficient for our model to attain predictive performance to comparable exhaustive labeling while incurring only 24.7% of the labeling costs. These results stress the importance of exposure networks as a source of valuable information for the investigation of information dissemination in social networks and question the robustness of textual features.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning approaches**;
• **Networks** → **Network reliability**.

KEYWORDS

fake news detection, social media, network science, fact-checking, misinformation

ACM Reference Format:
Maor Reuben, Lisa Friedland, Rami Puzis, and Nir Grinberg. 2024. Leveraging Exposure Networks for Detecting Fake News Sources. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24), August 25–29, 2024, Barcelona, Spain.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3637528.3671539

KDD '24, August 25–29, 2024, Barcelona, Spain

ACM ISBN 979-8-4007-0490-1/24/08.
https://doi.org/10.1145/3637528.3671539

1 INTRODUCTION

Social media has been identified as a key avenue for the proliferation and dissemination of fake news. Prior empirical findings showed that much of the exposure to misinformation originates from social media [1, 25, 50] and that misinformation travels farther, faster, deeper, and more broadly than the truth on social media [56]. Detecting fake news has thus emerged as a crucial step towards limiting its spread, with various approaches being employed including manual labeling, crowd-sourcing, and machine learning methods [48]. Section §2 summarizes the most relevant approaches.

We focus on the problem of identifying low-credibility *sources*, referred to hereafter as fake news sources. Specifically, we focus on identifying fake news sources using multiple types of signals (network representations, audience features, and text), and evaluate their robustness over time and dependence on the labeling budget. Tackling the problem at the level of a news source, i.e., the level of a web domain, is an imperfect first-order approximation for the credibility of information with the key benefit of reducing the scale of the problem. Labeling sources instead of individual news articles, claims, or social media posts significantly reduces the number of labels needed, which is critical for assessing credibility at Web scale. Focusing on sources also has the theoretical merit of shifting the focus from the veracity of individual pieces of information to the editorial norms and journalistic standards of the processes that generated them [28]. While prior research used manually-curated lists of fake news sources extensively [2, 22, 24, 37, 43, 47], relatively little research focused on developing computational methods for detecting fake news sources.

The importance of computational methods for detecting fake news sources lies in their potential to improve the coverage and efficiency of fact-checking efforts, which has important implications for both research and practice. Fact-checking is the foundation for much of the empirical research on fake news online [2, 22, 24, 37, 43, 47], and it is at the heart of the leading efforts by social media and other information systems to limit the spread of falsehoods online [13, 16, 34]. Comprehensive and up-to-date lists of fake news sources are essential for getting an accurate read of the volume of misinformation circulating online and informing the public about

it. There are also benefits for displaying credibility ratings to users, as it was shown to reduce consumption of heavy fake news consumers, lowered people's intention to distribute misinformation, and spilled over to other sources without labels [4, 12]. Early detection methods are particularly important for making fact-checking efforts as relevant and effective as possible before falsehoods have circulated widely. Therefore, methods that enhance the coverage and responsivity of fact-checkers can greatly contribute to limiting the spread of misinformation.

Identifying fake news sources is difficult for several reasons. First, sources of fake news are rapidly changing. An analysis by Poynter found that nearly half of the sources considered as fake news in the 2016 U.S. presidential election were no longer active in 2018 [21]. Updating these lists is a costly and labor-intensive task that requires careful attention to detail by expert fact-checkers, even with the aid of crowdworkers. Wrong evaluations are costlier nowadays as fact-checkers are increasingly exposed to litigation [7]. Second, it is not clear how fact-checkers should allocate their resources or prioritize their efforts. Checking the most popular sources may lead to a low discovery rate since most people consume relatively little fake news [22, 24]. It may also miss sources that circulate among smaller communities with more avid consumers of fake news. Alternatively, evaluating new narratives from known fake news sources may miss falsehoods spread by other emerging sources. Finally, determining the credibility of a source is a complicated task. It is not clear how accurately machine learning models can perform this task and what signals are most useful and robust over time.

In this work, we develop a novel methodology for identifying fake news sources using multiple signals about a source. We formulate the problem as a binary classification task of sources and consider the relative contribution of different feature families. Core to our approach is the idea that the credibility of a source is partially determined by the credibility of other sources consumed by its audience. This idea was alluded to by prior work [3, 22] but was not empirically tested for detecting fake news sources. Relying solely on users who actively share news online has a major flaw since it overlooks a significant portion of the population, the "silent majority" who rarely posts about news but still engages with it. Potential exposure networks offer a solution by incorporating a wider range of users and information sources. Thus, we consider network features of a source derived from sharing and potential exposure networks of users and sources (§3.2) as well as aggregate characteristics of the audience engaging with a source. In addition, as described in §3.3, we use linguistic features of a source as well as measures of popularity and co-occurrence with other known fake news sources. We evaluate our methodology using a pre-existing, de-identified, and unique dataset collected by Grinberg et al. [22]. This dataset includes tweets that were shared by and were available to a panel of over 16,000 registered U.S. voters on Twitter during the 2016 U.S. presidential election (see §3.1 for details). Importantly, this dataset enables evaluation when the complete set of labels is known (offline settings), and when sources and labels emerge over time (online settings; see §3.5 for precise definitions). Finally, we test the ability of several Active Learning (AL) methods to produce accurate predictions with a limited labeling budget. AL is a semi-supervised learning approach that uses a model trained on available labels to select new data points for labeling (see §4.3 for details). AL is particularly helpful in settings, like fake news source detection, where unlabeled data is abundant and labeling is difficult, time-consuming, or costly [46].

Therefore, our contributions are the following:

- A novel network-based methodology for identifying fake news sources with active learning (§3.2-3.4).
- Empirical evidence of the predictive power of sharing- and exposure-networks in offline and online settings (§4.1).
- An analysis of the distinctive characteristics of fake news sources (§4.2).
- An evaluation of different active learning strategies for efficiently detecting fake news sources with a limited fact-checking budget (§4.3).

2 RELATED WORK

Quantitative work about fake news can largely be divided into two groups: studies utilizing benchmark datasets with veracity labels at the story level (news article or a social media post), and studies that use pre-existing lists of fake news sources (see Shu et al. [49] for a comprehensive review). Both approaches rely on fact-checked information as ground truth. Typically, large benchmark datasets consist of a few thousand stories and labels [58].

Crowdsourcing has been proposed as a way to increase the scale of fact-checking efforts and prioritize stories for fact-checking. Pennycook and Rand [37] showed that crowdworkers can produce quality credibility labels that align well with the judgment of professional fact-checkers. For example, in 2017, Facebook added a feature that allowed users to flag content they deem as False News[1]. User flags may, in turn, affect the ranking of content and guide decisions on which stories to send to fact-checkers. Similarly, Twitter's Birdwatch allowed people to identify misleading content and write notes that provide additional context [16]. Nevertheless, even with crowdworkers operating as fact-checkers and professional fact-checkers making only the final determinations, issues of scale still persist due to the cost and labor-intensive work required to fact-check the amount of flagged social media content [15, 19].

Machine learning methods offer a promising approach for scaling up the detection of fake news content. Several benchmark datasets have been developed to facilitate comparison across models [5, 53, 58]. Research in natural language processing (NLP) has considered a variety of deep neural network architectures for the task of using linguistic features of textual content for detecting fake news [17, 26, 39, 54]. Others have considered different modalities including fabricated and manipulated images [33, 45], videos [55, 59], metadata such as time stamps, formatting information or the occurrence of certain HTML tags [11], and knowledge graphs [18]. Many works examined how rumors and misinformation propagate through social networks [30, 31, 57], while others used network features for detecting false content [14, 32, 38, 51]. Some studies have even employed active learning to select the most informative posts to label, reducing labeling costs and improving performance [6, 42].

An alternative approach that is prominent in social science literature relies on source-level definitions and existing lists of fake news sources [2, 22, 24, 37]. Compared to story-level definitions,

[1] https://www.facebook.com/facebookmedia/blog/working-to-stop-misinformation-and-false-news

the use of source-level definitions offers a significant reduction in the amount of labels needed at the cost of accuracy: All content from a given source receives the same label, regardless of the veracity of individual stories. Lazer et al. [28] justify this approach as focusing on sources that "lack the news media's editorial norms and processes for ensuring the accuracy and credibility of information". Indeed, much of this line of work relies on some combination of lists produced by trusted journalists [50], fact-checking organizations [20, 27], academics [60], and commercial organizations like NewsGuard[2] that evaluate the credibility of news sources. Yet, few methods were proposed for augmenting these lists with additional sources, which is the focus of the current investigation.

Perhaps closest in nature to the current work is the study by Chen and Freire [14]. Their model constructs a co-sharing graph of sources, where unweighted edges connect pairs of sources with a minimum number of users who posted tweets linking to both sources. Then, clustering is applied to the network, and sources in the largest connected component of the graph are considered as candidate fake news sources. The final classification of sources is based on the average score obtained from applying a Linear SVM over a sample of pages from the source. To avoid over-fitting, the authors build their page classifier to be topic-agnostic (based on HTML tags, readability scores, Part-of-Speech frequencies) since "the news topic may change day by day, the layout and writing style of a website do not change as frequent" [14].

This study extends Chen and Freire's research by (1) incorporating potential exposure networks, (2) utilizing additional network representations, (3) exploring audience characteristics, and (4) evaluating the model's sensitivity to topic-aware textual features. We also extend beyond prior art by experimenting with source-detection methods in both offline and online settings.

3 METHODS

In this section, we present our methodology for identifying fake news sources. We begin by describing the dataset used to construct network and audience representations. Then, we detail the features extracted for news sources in both offline and online settings. Finally, we describe the active learning strategies used to evaluate the performance of our methodology over time and depending on labeling budget.

3.1 Dataset

In this study, we conduct a secondary analysis of the de-identified dataset made available by Grinberg et al. [22]. The primary dataset we use consists of over 10 million tweets that linked to political news during the 2016 U.S. presidential election (Aug-Nov, inclusive) and were shared by or available to a panel of 16,442 registered U.S. voters on Twitter. The panel also contains coarse sociodemographic information about voters including age, gender, state, and registration with a political party. Grinberg et al. also showed that this panel is reflective of the population of American voters on Twitter. Importantly, the dataset includes not only tweets shared by U.S. voters but also tweets posted by the accounts they follow on Twitter. This enables the examination of potential exposure, content that is available to voters from their social network, referred to hereafter as *exposure* for brevity.

In terms of credibility labels for news sources, we utilize the full set of 1,505 labels produced by Grinberg et al. [22], which have been used extensively in prior research. Grinberg et al. compiled this list by drawing on existing lists from multiple trusted sources (fact-checkers, other academics, etc.) and by labeling additional sources using fact-checked information from snopes.com, a long-standing and prominent fact-checking site. The set of labeled sources consists of 1,212 green or yellow news sources that are non-fake, i.e., having no indication of repeatedly publishing false claims. In addition, there are 293 news sources with black, red, or orange labels that have demonstrated repeated disregard for the truth, and are regarded as fake news sources. For example, the nytimes.com is considered non-fake (labeled as green), and infowars.com is considered fake (labeled as red). The panel shared links to a total of 1,006 labeled news sources (176 fake and 830 non-fake) and was potentially exposed to a total of 1,505 labeled news sources (293 fake and 1212 non-fake). All 1,006 sources shared by the panel appear in the larger set of 1,505 sources shared by their peers. Our primary results are based on the full set of 1,505 sources. Additional robustness checks on the subset of 1,006 news sources shared by the panel are reported in Appendix A. Appendix B demonstrates the robustness of our results when exposure is approximated from news cascade data.

3.2 Network Representations

Data about user interaction with news sources allows us to consider the following two network representations:

(1) **User-to-source** interaction network: A bipartite multi-graph $G^{bi} = \{U, S, E^{bi}\}$, where $u \in U$ represents a user, $s \in S$ represents a news source, and $e_{u,s} \in E^{bi}$ is an edge iff user u has interacted with the source s.
(2) **Source-to-source** network $G^{co} = \{S, E^{co}\}$: A unipartite weighted projection graph of G^{Bi} where $s, s' \in S$ represent news sources and $e_{s,s'} \in E^{co}$ iff there exists at least one user who has interacted with both sources s and s'.

For each representation, we consider two types of interactions: *sharing* and *exposure*. User u shared a source s if u posted a tweet that contained a link to source s. Similarly, user u was potentially exposed to source s if user v shared the source and u follows v. We define the networks $G^{bi-sharing}$ and $G^{bi-exposure}$ based on the respective definitions. From the $G^{bi-sharing}$ and $G^{bi-exposure}$ networks, we derive the $G^{co-sharing}$ and $G^{co-exposure}$ projection networks, which capture the respective co-interaction relationships between sources. Statistics about the number of news sources, users, and interaction edges is in Table 1. Next, we describe the features derived from these graphs.

3.3 Feature Extraction

We use four feature families to characterize a source: basic measures of popularity and connectivity with fake news sources, network features, audience composition, and textual features.

The **basic (baseline) features** include a measure of audience overlap with known fake news sources and several measures of a source's popularity. Audience overlap is computed as the fraction

[2] https://www.newsguardtech.com/

	Sources	Users	Nodes	Edges
$G^{bi-sharing}$	1,006	3,945	4,951	77,129
$G^{bi-exposure}$	1,505	15,212	16,717	7,638,054
$G^{co-sharing}$	1,006	-	1,006	172,826
$G^{co-exposure}$	1,505	-	1,505	964,075

Table 1: Statistics about the sharing and exposure networks detailing the number of sources, panel members (users), nodes, and edges in each representation.

of fake news sources among neighbors of a source according to Equation 1, where E_s^{co} is the set of edges connected to s from G^{co}, s' is a neighboring source to s, and $FakeSources$ is the set of sources currently labeled as fake.

$$\%FakeNeighbors(s, G^{co}) = \frac{\left|\{e_{s,s'} \in E_s^{co} : s' \in FakeSources\}\right|}{\left|\{e_{s,s'} \in E_s^{co}\}\right|} \quad (1)$$

This feature indicates the extent to which fake news sources are co-shared or co-consumed by the same set of people.

Another set of baseline features pertains to the source's popularity. Fake news sources typically lack mainstream popularity, but it's unclear if they differ significantly from credible non-mainstream sources. We assess source popularity through three metrics: the number of unique users interacting with the source ($|E_s|$), the posts linking to the source ($|P_s|$), and their ratio ($|E_s|/|P_s|$).

The second feature family consists of **network features**, which aim to capture the structural characteristics of sources and indicative relations in the user-to-source network. We use Node2Vec [23] to obtain 20-dimensional embeddings for news sources based on G^{bi}. Node2Vec explores the network neighborhood of each node, striking a balance between local and farther away ties, which provides a representation of the topology of the network. We also derive a User-Frequency Inverse Source Frequency (UF-ISF) weighting scheme inspired by TF-IDF, similar to the approach used in prior work [35]. The rationale for this is that users who interact with fake news sources may have a particular propensity to interact with certain sources. To compute UF-ISF, we adapt the common TF-IDF weighing as follows. First, we calculate the user frequency (UF) for each s, u pair by dividing the number of posts that connect user u with source s ($P_{u,s}$) by the total number of posts that linked to the source (P_s). Then, we calculate the inverse source frequency (ISF) for each user as the inverse proportion of unique sources that a user u engaged with on a logarithmic scale. Finally, we multiply UF and ISF to obtain the final score as shown in Equation 2. The *UF-ISF* final representation for a source is a vector. To reduce dimensionality, we used the F-test feature selection method to select the 100 most informative users.

$$\textit{UF-ISF}(u, s, G^{bi}) = \frac{|P_{s,u}|}{|P_s|} \times \log \frac{|S|}{|\{s \in S : P_{s,u} \neq \emptyset\}|} \quad (2)$$

Another feature family includes **audience features**. Our goal here is to capture the demographic and political characteristics of users who engage with fake news sources. Prior work showed that the age and the political affiliation of users who engage with content from news sources are strong signals of news source reliability [8, 22, 24, 25]. The unique panel created by Grinberg et al. [22] enables us to use high-quality sociodemographic to characterize the audience of a news source. In particular, we had access to users' age, gender, state of residence, and registration with a political party. Categorical variables were encoded using a one-hot representation. For each source, we calculate audience features as described in Equation 3, where D represents a function that returns the sociodemographic characteristics of user u, and f represents one of the following aggregation functions: mean, standard deviation, median, and the 25th and 75th percentiles.

$$Audience(s, G^{bi}, f, D) = f(\{e_{u,s} \in E_s : D(u)\}) \quad (3)$$

The final set of audience features for a source consists of all values calculated using all available combinations of sociodemographic characteristics and aggregation functions (pairs of (D, f))

Finally, our last feature family includes **text features** of posts linking to a source. We used the standard TF-IDF weighting scheme to represent posts and averaged them across posts linking to the same source. We removed stopwords and used the 5,000 most common terms as our vocabulary. Again, we reduced the dimensionality by using the F-test feature selection method and selected the 100 most informative terms. In addition to TF-IDF, we calculated text embedding for each post using the 'all-MiniLM-L6-v2' model, which is a more recent transformer-based architecture for text representation [40, 41]. The final embedding vector for a source is a 384-dimensional vector that is the average across the embeddings of posts linking to the source. Embedding-based methods may offer superior results in fake news detection [17], but their interpretability is more limited relative to TF-IDF.

The full list of features is in Table 2.

3.4 Active Learning Strategies

A prominent approach in machine learning in settings where labeled examples are scarce, unlabeled examples are abundant, and labeling costs are high is to use Active Learning (AL). AL lets the model choose the data points for labeling, i.e., the model trained on labeled examples is used to select additional examples to obtain labels for [46]. AL methods differ in the way they choose examples. For an extensive survey of active learning methods see Tharwat and Schenck [52]. We experimented with several prominent AL strategies to identify the most accurate and effective method for selecting news sources for labeling. We briefly describe these strategies next.

Uncertainty Sampling is a strategy that focuses on selecting samples that the model is most uncertain about. In the binary case, this involves choosing samples that are closest to a predicted probability of 0.5. With each iteration, the model reduces its uncertainty and improves its performance.

Certainty Sampling is a strategy that focuses on selecting samples that the model is most certain about. This involves choosing samples where the model exhibits high confidence or where the predicted probability is closest to the value of 1.

Diversity Sampling is a strategy that aims to cover as much of the input space as possible. It considers the dissimilarity or diversity between samples when selecting new data points to label. It seeks

Feature		Explanation
% Fake Neighbors		Percent of fake neighbors in the source-to-source network.
Popularity	# users	Number of unique users linking to the source.
	# posts	The total number of source's posts.
	users/posts	The ratio of unique users to the number of posts.
Network	Node2Vec	Node embedding of sources in G^{bi}, user-to-source network with {dim=20, walk_length=5, p=2.0, window=5, q=0.5, num_walks=50}
	UF-ISF	Weighting scheme where each source is a document and each user engaging with it is a term.
Audience	Age	Aggregate stat. of users' age.
	Gender	Aggregate stat. of users' gender.
	State	Aggregate stat. of users' state.
	Party	Aggregate stat. of users' registration party.
Text	TF-IDF	Averaged TF-IDF weighting of posts linking to a source.
	Embedding	Averaged sentence-transformer embeddings over posts, computed using all-MiniLM-L6-v2.

Table 2: Description of features used in predicting fake news sources. The aggregations in the Audience features are mean, standard deviation, median, and the 25th and 75th percentiles over the source users.

to capture different aspects or clusters of the data, learn diverse patterns, and improve its generalization capabilities.

We compared the above general-purpose active learning strategies to the following three baselines. **Random Sampling** is our most simple and unbiased sampling approach, where new sources for labeling are selected at random. **High Degree Sampling** chooses the most popular sources for labeling. This strategy is likely to be employed by fact-checkers who seek to cover the falsehoods that reach the largest number of people. For example, NewsGuard claim that their list of sources covers 95% of online engagement with news across multiple major media markets [36]. **Unlimited Budget** is a strategy where all new sources that emerge in a time period are labeled. While this strategy is unlikely to be feasible in practice, it serves as an upper bound for the theoretical performance when labeling resources are unlimited.

By comparing the performance of our active learning strategies against these baselines, we assess their effectiveness in terms of labeling costs and classification accuracy.

3.5 Experimental Setup

The main goal of our proposed methodology is to assist fact-checking efforts when labeling historical data as well as newly emerging sources. To the end, we evaluate our methodology in both offline and online settings, and when labeling budget is limited.

In offline settings, we gauge our method's effectiveness when the complete set of sources is known. We use five-fold cross-validation to assess performance. In online settings, the assessment is performed over time in increments of two weeks. In other words, the model has access to labels from previous periods but is challenged to predict the labels of sources that have surfaced in the most recent time segment. The decision to divide the analysis into two-week intervals over four months was motivated by a desire to balance the emergence of new sources, approximately 40 in each period, and the need for tight error bounds in the results.

In our experiments, we trained different classification models including Logistic Regression, Neural Networks, XGBoost, and Random Forests. When appropriate, we conducted a grid search over parameters space to identify the best-performing model. Our binary labels (fake or non-fake) were drawn from the dataset by Grinberg et al. [22]. We used Precision-Recall Area Under the Curve (PR-AUC) as our metric for evaluation, which is particularly suited for circumstances of class imbalance [10]. In our case, class imbalance is indeed an issue with non-fake sources dominating the label distribution with a ratio of about 4:1 compared to fake sources.

To facilitate a more stringent comparison of sharing-based [14] versus exposure-based models, we conducted additional experiments that focused only on the subset of 1,006 sources shared by the panel. While the additional signal about 499 (49.6%) sources in the exposure network is likely to improve model performance, an evaluation on the same subset of sources is important for disentangling any performance gains from being solely the result of having more labels in the exposure network.

4 RESULTS

In this section, we report the predictive performance of our models in both offline and online settings, the important features associated with fake news sources, and the results from our Active Learning experiments.

4.1 Offline and online settings

Table 3 reports the predictive performance in PR-AUC of the best-performing classification model (Random Forest with 100 trees and a depth of four) using different feature families. The table includes results for offline (columns 2-4) and online (columns 5-7) settings. In both settings, we report separately the PR-AUC of using sharing and exposure features for prediction, and the relative improvement of the exposure-based model compared to the sharing-based model (as percentages, in columns 4 and 7). The last column in the table compares the performance of the best-performing model in online settings against the best-performing model in offline settings while using the same set of features.

Feature families across settings: The results in Table 3 show that the performance of exposure-based models are superior to the sharing-based models across nearly all feature families [3]. The exposure-based model outperforms the sharing-based model by 7.6% to 59.7% in offline settings and 33.8% to 96.4% in online settings, except for the Fake Neighbors feature, which exhibits a 16.6% degradation in performance. This degradation is attributed to noise introduced by

[3] Class-wise metrics of the best performing models are detailed in https://tinyurl.com/class-wise-scores

	Offline settings			Online settings			Online vs. Offline
Features	Sharing [14]	Exposure	%	Sharing [14]	Exposure	%	%
Baseline	.529	.569	+7.6	.371	.497	+33.8	-12.7
% Fake Neighbors	.467	.390	-16.6	.363	.509	+40.3	+8.9
Popularity	.310	.495	+59.7	.237	.364	+53.9	-26.4
Network	.567	.692	+21.9	.422	.726	+71.9	+5.0
UF-ISF	.526	.670	+27.4	.369	.725	+96.4	+8.1
Node2vec	.532	.628	+18.2	.421	.597	+41.7	-5.0
Audience	.519	.642	+23.6	.409	.710	+73.6	+10.6
Text (TF-IDF)	.509	.766	+50.4	.366	.623	+70.3	-18.7
Text (Embedding)	.576	.778	+35.2	.391	.685	+75.1	-12.0
Baseline+Audience	.542	.646	+19.3	.420	.714	+70.2	+10.5
Baseline+Network	.575	.702	+22.2	.435	.726	+66.9	+3.4
Baseline+Audience+Network	.567	.708	+24.9	.436	.736	+68.9	+4.0
Baseline+Text	.513	.774	+51.0	.413	.628	+52.2	-18.9
All features	.596	.792	+33.0	.438	.740	+69.2	-6.5

Table 3: The PR-AUC of the best-performing classification model using different feature families on the 1,505 labeled sources. Offline settings results were evaluated using five-fold cross-validation over sources (columns 2-4). Online settings results were evaluated on emerging sources in increments of two weeks with the model gradually having access to more labels (columns 5-7). In both settings, sharing- and exposure-networks are reported separately, with the percentage gain of exposure models over sharing models. The last column shows the percentage improvement of the best online model over the best offline model.

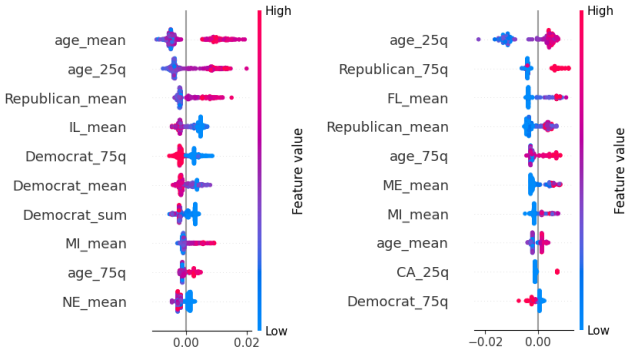
the number of fake neighbors in many non-fake sources, possibly due to co-consumption with a large number of non-fake sources. When using all available features, the exposure-based model outperforms the sharing-based model by 33.0% in offline settings and 69.2% in online settings.

Additionally, all individual feature families improve over the baseline, and combining features with the baseline yields further improvements. Notably, the performance of the baseline, text-based, and the model based on all features is lower in online settings compared to offline settings, while network and audience features improve the performance. The degradation in performance of online text-based models, both lexical (TF-IDF) and transformer-based, stands out as these models have the highest PR-AUC among all feature families in offline settings. These results suggest that text features can distinguish fake news sources, but are significantly less robust in predicting emerging sources. Network and audience features, in contrast, show considerably more robust results across offline and online settings. For example, the offline model based on network features derived from the exposure network achieves a PR-AUC of 0.692, which is comparable to the PR-AUC of 0.726 found in online settings (+5.0%).

Detecting fake news sources ahead of fact-checkers: To further demonstrate that our methodology can identify fake news sources ahead of fact-checkers, we manually identified nine fake and nine non-fake news sources that did not appear in Snopes before the end of the study period. Using this held-out set, our classifier achieved a PR-AUC of 0.924. This suggests that our method can effectively discover fake news sources before they are known to fact-checkers. It should be noted that the high PR-AUC value is partially driven by the fact that this small sample is balanced.

What is the value of exposure-network information? Repeating our experiments using only the subset of 1,006 sources in the sharing network results in a significant overestimation of performance. The full details of these experiments are in Appendix A. We find that using the subset of 1,006 sharing-network sources produces higher PR-AUC values than using the full set of sources. For example, the offline sharing-based model with all available features attains a PR-AUC of 0.762 when evaluated on the 1,006 sources and a PR-AUC of 0.596 on the full set, an overestimation of 27.8% of the true performance. We find overestimation across feature families, in both offline and online settings, and when using exposure-based features (although of a smaller magnitude). These results suggest analyzing only sharing-network information introduces a selection bias. Aside from overestimation, Table 4 also replicates the finding that exposure-based models outperform sharing-based ones in the subset of 1,006 sources. This suggests that incorporating exposure-based features not only improves predictive performance but also extends the ability to detect a wider range of sources.

Other approximations of exposure information: The superior performance of exposure-based models was further affirmed through experiments on an additional dataset of fake news cascades. Despite having only partial exposure information, exposure features significantly improved predictive accuracy. For detailed data description and results, please refer to Appendix B.



(a) Offline **(b) Online**

Figure 1: SHAP values of the best-performing exposure model in the offline (left) and online (right) settings.

4.2 Characteristics of fake news sources

The relative importance of individual features is indicative of the characteristics that distinguish fake news sources. To that end, we compute SHAP values [29] for our best-performing models, which show both the magnitude and direction of the contribution of individual features to model accuracy.

Figure 1 presents the SHAP values of both offline (left) and online (right) models computed over the exposure network with all available features. Features on the y-axis are sorted according to their importance from top to bottom. Each point represents a source and its position on the x-axis corresponds to the SHAP value of a particular feature. The color indicates the magnitude of the raw feature value. For example, the average age of people sharing a source (age_mean in panel a) shows that many fake news sources have a high average age and that this is the most important feature. Uninterpretable features such as Node2vec embeddings were omitted from the figure and text features are reported on separately below. A similar figure using the sharing network model can be found in Appendix D.

Across both online and offline settings, key features help distinguish fake news sources. Age quantiles and mean values appear prominently in Figure 1, indicating older audiences are more likely to share or consume fake news. Party registration features are also significant, showing fake news sources are less associated with Democrats and more with Republicans. These findings align with prior research indicating older, conservative individuals are more likely to engage with fake news [22, 24]. Finally, fake news sources have larger audiences in some U.S. states, but the small sample size in some states should be noted.

Next, we examined the top five text terms having the highest SHAP values. In offline settings, the top terms were hillary, soros, WikiLeaks, bombshell, and corrupt. In online settings, the top terms were gowdy, hillary, fbi, soros, and caught. The above terms represent public figures and topics that were involved and linked to fake news stories during the 2016 election. Clearly, these textual features are tied to the specific time span of the data.

4.3 Active learning experiments

Our active learning (AL) experiments followed the same setup as the online settings, using all available features, with the additional constraint of obtaining only a fixed number of labels at each time interval. We experimented with different number of labels obtained in each iteration, namely 10, 20, 40, 50, and 100. The results were largely the same, and for brevity, we only report on the findings with 40 new labels at each iteration. It is important to note that all strategies started with the same set of 100 sources and were evaluated on the same 30% of held-out sources.

The results of our AL experiments are shown in Figure 2. The figure shows PR-AUC as a function of the number of labeled sources. Each line represents the PR-AUC of a different active learning strategy. The left panel shows the results obtained using the sharing network and the right panel shows the results of the exposure network.

The results in Figure 2 show that the Uncertainty Sampling strategy outperforms all other strategies both when using sharing- and exposure-network information ($p<0.05$). After a few iterations, Uncertainty Sampling achieves comparable performance to the Unlimited Budget strategy despite having fewer labels. For example, Uncertainty Sampling using the exposure-based model reached PR-AUC of 0.790 after only four iterations (a total of 260 labels), which is a mere 24.7% of the labels used in the Unlimited Budget benchmark with roughly the same PR-AUC (0.805). Similarly, Uncertainty Sampling using the sharing-based model achieved a PR-AUC of 0.569 with just 180 labels, equivalent to 25.5% of the Unlimited Budget benchmark with PR-AUC of 0.572. As an additional robustness check, we tested the different strategies in offline settings, and found similar results (see Appendix C for more details). Taken together, these results demonstrate that our methodology can achieve near-optimal performance with significantly fewer labels, emphasizing its benefits for use in real-world applications.

In addition, Figure 2 shows that Random Sampling is a strong baseline that is superior or comparable to three other sampling strategies. When using the exposure network (right panel), the PR-AUC of the High Degree, Certainty Sampling, and Diversity Sampling strategies are significantly lower than Random Sampling. When using the sharing network (left panel), only Certainty Sampling is significantly lower, while the two other strategies are comparable to Random Sampling. This suggests that sampling strategies that focus on just one "type" of sources – whether it is the most popular, different from the labeled ones, or most likely to be fake – are not optimal for identifying fake news sources, and a different sampling strategy is needed to perform better than random.

5 DISCUSSION

In this study, we introduced a network-based methodology for identifying fake news sources. Our experiments included offline, online, and active learning settings, which demonstrate the effectiveness and robustness of our approach.

A key finding across all experimental settings is the persistent improvement of exposure-based models over sharing-based models. We observed performance gains of 33.0% in offline settings, 69.2% in online settings, and substantial improvements when using active learning. Moreover, our findings show that using sharing

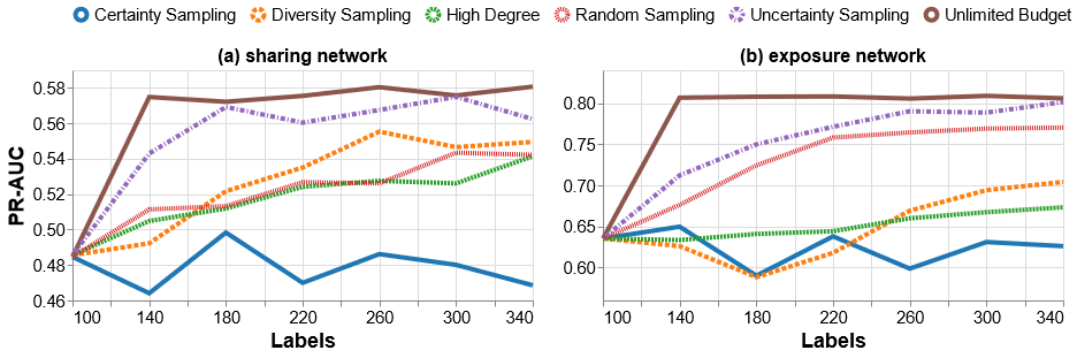


Figure 2: PR-AUC as a function of the number of labeled sources in online settings based on sharing (left) and exposure (right) networks. Each line represents the PR-AUC of the different active learning strategies.

information alone leads to a significant overestimation of model accuracy, and that exposure features produce better predictive results and more calibrated estimates when evaluated on the same set of sources. These observations suggest that exposure-based features not only enhance predictive performance but also enable more accurate predictions on a broader range of sources, which is critical for increasing the coverage of fact-checkers. Taken together, the results provide compelling evidence in support of the use of exposure information for detecting fake news sources.

The findings also highlight the risks of using text-based methods for detecting fake news sources. While text-based models, both lexical (TF-IDF) and transformer-based, produced some of the best performance in offline settings, their PR-AUC dropped by 12.0-18.7% in online settings. Examining the SHAP values of the text features provided further insight into why these drops may have occurred: the model identified relevant terms that are *ephemeral*. In contrast, network and audience features produced results that are robust over time. Another potential benefit of using network and audience features is that they are more challenging to fabricate compared to textual features.

Finally, our active learning experiments demonstrated the capacity of Uncertainty Sampling to guide cost-efficient fact-checking. Across different settings, we found that a few iterations of Uncertainty Sampling were enough to achieve nearly optimal performance while using only a small fraction of the labels. All other sampling strategies produced results that were either no better than random or worse. Particularly interesting is the result that High Degree Sampling is no better than random because it suggests that focusing on popular sources may hurt the discriminability of other sources. In other words, labeling the most popular sources may maximize reach among people, but it may be less informative about the full distribution of fake news sources.

Our methodology has a few limitations. Twitter (now called X) has increased the pricing for large-scale collection of a random sample of tweets, which enabled Grinberg et al. [22] to approximate exposure. To continue investigating exposure networks researchers will have to limit themselves to historical data, pay the increased fees, or partner with social media platforms to use exposure signals. Although not directly tested, we believe that there is a need to effectively filter out bots, organizations, and other non-individual accounts to extract high-quality signals about exposure. It is unclear how well our methodology will perform in the presence of bots and other malicious accounts or when audience features are based on inferred characteristics rather than administrative records. Finally, our findings may not generalize to all social media platforms or other political systems. To partially mitigate generalizability limitations the main claims were validated on another data set of fake news cascades obtained from Reuben et al. [44] (see appendix B).

Several avenues for future research exist to extend this work. The method's resilience to adversarial attacks by bots, coordinated manipulation, or flooding of content is a worthy topic for future research, especially in a political climate of increasing affective polarization, which is particularly strong in the U.S. but also present in other multi-party countries [9]. Future research could also explore combining source-level and story-level methods, potentially refining credibility estimates and detecting changes in source credibility. Moreover, once the list of labeled sources becomes large there is a trade-off between labeling new sources and re-evaluating existing ones. Future empirical work could also accurately quantify the coverage of fact-checking efforts and determine the extent to which their coverage is biased toward popular sources. This could better inform fact-checkers about their coverage and their blind-spots.

In conclusion, this study introduced a network-based methodology for detecting fake news sources that is both robust and effective. The findings highlight the importance of exposure-based features for model accuracy, the challenges of text-based features in online settings, and the efficiency of Uncertainty Sampling in labeling sources over time with a limited budget. These insights can inform future research and practical applications in the ongoing battle against misinformation.

6 ETHICAL CONSIDERATIONS

The study protocol was approved by the departmental ethics committee at Ben-Gurion University (protocol #SISE-2024-41). Several broader implications of this research should be considered. First, the methods developed in this research should only be used as part of a decision-support system that involves human judgment, ideally by trained journalists and qualified fact-checkers who follow clear and transparent guidelines. We advise against usage without human-in-the-loop as it raises important issues of accountability, liability, and

potentially systematic bias. Second, fact-checkers should refrain from overly relying on such algorithmic recommendations as this may lead to algorithmic blind spots. We recommend combining the approach laid out in this research with other approaches.

REFERENCES

[1] Hunt Allcott and Matthew Gentzkow. 2017. *Social Media and Fake News in the 2016 Election.* Working Paper 23089. National Bureau of Economic Research. https://doi.org/10.3386/w23089

[2] Hunt Allcott, Matthew Gentzkow, and Chuan Yu. 2019. Trends in the diffusion of misinformation on social media. *Research & Politics* 6, 2 (2019), 205316801984855 4.

[3] Ahmer Arif, Kelley Shanahan, Fang-Ju Chou, Yoanna Dosouto, Kate Starbird, and Emma S. Spiro. 2016. How Information Snowballs: Exploring the Role of Exposure in Online Rumor Propagation. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (San Francisco, California, USA) *(CSCW '16)*. Association for Computing Machinery, New York, NY, USA, 466–477. https://doi.org/10.1145/2818048.2819964

[4] Kevin Aslett, Andrew M. Guess, Richard Bonneau, Jonathan Nagler, and Joshua A. Tucker. 2022. News credibility labels have limited average effects on news diet quality and fail to reduce misperceptions. *Science Advances* 8, 18 (2022), eabl3844. https://doi.org/10.1126/sciadv.abl3844 arXiv:https://www.science.org/doi/pdf/10.1126/sciadv.abl3844

[5] Fatemeh Torabi Asr and Maite Taboada. 2019. Big Data and quality data for fake news and misinformation detection. *Big Data & Society* 6, 1 (2019), 2053951719843310. https://doi.org/10.1177/2053951719843310 arXiv:https://doi.org/10.1177/2053951719843310

[6] Giorgio Barnabò, Federico Siciliano, Carlos Castillo, Stefano Leonardi, Preslav Nakov, Giovanni Da San Martino, and Fabrizio Silvestri. 2023. Deep active learning for misinformation detection using geometric deep learning. *Online Social Networks and Media* 33 (2023), 100244. https://doi.org/10.1016/j.osnem.2023.100244

[7] Sara Bealor. [n. d.]. IFCN launches new Legal Support Fund for fact-checkers facing harassment. https://www.poynter.org/from-the-institute/2022/ifcn-launches-new-legal-support-fund-for-fact-checkers-facing-harassment/.

[8] Saumya Bhadani, Shun Yamaya, Alessandro Flammini, Filippo Menczer, Giovanni Luca Ciampaglia, and Brendan Nyhan. 2022. Political audience diversity and news reliability in algorithmic ranking. *Nature Human Behaviour* 6, 4 (April 2022), 495–505. https://doi.org/10.1038/s41562-021-01276-5

[9] Levi Boxell, Matthew Gentzkow, and Jesse M Shapiro. 2022. Cross-country trends in affective polarization. *Review of Economics and Statistics* (2022), 1–60.

[10] Paula Branco, Luís Torgo, and Rita P Ribeiro. 2016. A survey of predictive modeling on imbalanced domains. *ACM Comput. Surv.* 49, 2, Article 31 (aug 2016), 50 pages. https://doi.org/10.1145/2907070

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

[12] Tatiana Celadin, Valerio Capraro, Gordon Pennycook, and David G Rand. 2023. Displaying News Source Trustworthiness Ratings Reduces Sharing Intentions for False News Posts. *Journal of Online Trust and Safety* 1, 5 (April 2023). https://doi.org/10.54501/jots.v1i5.100

[13] Google Safety Center. [n. d.]. Google's approach to fighting misinformation online. https://safety.google/intl/en_us/stories/fighting-misinformation-online/.

[14] Zhouhan Chen and Juliana Freire. 2020. Proactive Discovery of Fake News Domains from Real-Time Social Media Feeds. In *Companion Proceedings of the Web Conference 2020* (Taipei, Taiwan) *(WWW '20)*. Association for Computing Machinery, New York, NY, USA, 584–592. https://doi.org/10.1145/3366424.3385772

[15] Yuwei Chuai, Haoye Tian, Nicolas Pröllochs, and Gabriele Lenzini. 2023. The Roll-Out of Community Notes Did Not Reduce Engagement With Misinformation on Twitter. *arXiv preprint arXiv:2307.07960* (2023).

[16] Keith Coleman. [n. d.]. Introducing Birdwatch, a community-based approach to misinformation. https://blog.twitter.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation.

[17] Ehab Essa, Karima Omar, and Ali Alqahtani. 2023. Fake news detection based on a hybrid BERT and LightGBM models. *Complex & Intelligent Systems* (2023), 1–12.

[18] Lifang Fu, Huanxin Peng, and Shuai Liu. 2023. KG-MFEND: an efficient knowledge graph-based model for multi-domain fake news detection. *The Journal of Supercomputing* 79, 16 (2023), 18417–18444.

[19] Full Fact. 2019. Report on the Facebook Third Party Fact Checking programme. http://web.archive.org/web/20190731042936/https://fullfact.org/media/uploads/tpfc-q1q2-2019.pdf.

[20] Joshua Gillin. 2018. PolitiFact's guide to fake news websites and what they peddle. *Politifact* (2018).

[21] Barrett Golding. 2018. UnNews: An index of unreliable news websites. *Poynter Institute* (Apr 2018). https://web.archive.org/web/20190501173436/http://www.poynter.org/ifcn/unreliable-news-index/.

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

[23] Aditya Grover and Jure Leskovec. 2016. Node2vec: Scalable Feature Learning for Networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 855–864. https://doi.org/10.1145/2939672.2939754

[24] Andrew Guess, Jonathan Nagler, and Joshua Tucker. 2019. Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science advances* 5, 1 (2019), eaau4586.

[25] Andrew Guess, Brendan Nyhan, and Jason Reifler. 2018. Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 US presidential campaign. *European Research Council* 9 (2018), 1–49.

[26] Ye Jiang, Johann Petrak, Xingyi Song, Kalina Bontcheva, and Diana Maynard. 2019. Team Bertha von Suttner at SemEval-2019 Task 4: Hyperpartisan News Detection using ELMo Sentence Representation Convolutional Network. In *Proceedings of the 13th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, Minneapolis, Minnesota, USA, 840–844. https://doi.org/10.18653/v1/S19-2146

[27] Kim LaCapria. 2017. Snopes' field guide to fake news sites and hoax purveyors. *Snopes.com* 8 (2017).

[28] David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. 2018. The science of fake news. *Science* 359, 6380 (2018), 1094–1096. https://doi.org/10.1126/science.aao2998 arXiv:https://science.sciencemag.org/content/359/6380/1094.full.pdf

[29] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2020. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence* 2, 1 (2020), 2522–5839.

[30] 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 *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence* (New York, New York, USA) *(IJCAI'16)*. AAAI Press, 3818–3824.

[31] Jing Ma, Wei Gao, Zhongyu Wei, Yueming Lu, and Kam-Fai Wong. 2015. Detect Rumors Using Time Series of Social Context Information on Microblogging Websites. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management* (Melbourne, Australia) *(CIKM '15)*. Association for Computing Machinery, New York, NY, USA, 1751–1754. https://doi.org/10.1145/2806416.2806607

[32] Jing Ma, Wei Gao, and Kam-Fai Wong. 2017. Detect Rumors in Microblog Posts Using Propagation Structure via Kernel Learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 708–717.

[33] Hana Matatov, Adina Bechhofer, Lora Aroyo, Ofra Amir, and Mor Naaman. 2018. DejaVu: A System for Journalists to Collaboratively Address Visual Misinformation. In *Computation+ Journalism Symposium. Miami, FL.*

[34] Adam Mosseri. [n. d.]. Working to Stop Misinformation and False News. https://www.facebook.com/formedia/blog/working-to-stop-misinformation-and-false-news.

[35] Luiza Nacshon, Rami Puzis, and Amparo Sanmateho. 2016. Pinpoint Influential Posts and Authors. *arXiv preprint arXiv:1609.02945* (2016).

[36] Inc. NewsGuard Technologies. [n. d.]. Why Should You Trust Us? https://www.newsguardtech.com/about/why-should-you-trust-us/. Accessed: 2023-09-06.

[37] Gordon Pennycook and David G Rand. 2019. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences* 116, 7 (2019), 2521–2526.

[38] Huyen Trang Phan, Ngoc Thanh Nguyen, and Dosam Hwang. 2023. Fake news detection: A survey of graph neural network methods. *Applied Soft Computing* (2023), 110235.

[39] Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 2931–2937. https://doi.org/10.18653/v1/D17-1317

[40] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. https://arxiv.org/abs/1908.10084

[41] Nils Reimers and Iryna Gurevych. 2020. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for

Computational Linguistics. https://arxiv.org/abs/2004.09813
[42] Yuxiang Ren, Bo Wang, Jiawei Zhang, and Yi Chang. 2020. Adversarial Active Learning Based Heterogeneous Graph Neural Network for Fake News Detection. In *2020 IEEE International Conference on Data Mining (ICDM)*. 452–461. https://doi.org/10.1109/ICDM50108.2020.00054
[43] Paul Resnick, Aviv Ovadya, and Garlin Gilchrist. 2018. Iffy quotient: A platform health metric for misinformation. *School of Information Center for Social Media Responsability University of Michigan* 1, 10 (2018), 1–10.
[44] Maor Reuben, Aviad Elyashar, and Rami Puzis. 2022. Iterative query selection for opaque search engines with pseudo relevance feedback. *Expert Systems with Applications* 201 (2022), 117027.
[45] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2019. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE International Conference on Computer Vision*. 1–11.
[46] Burr Settles. 2009. *Active Learning Literature Survey*. Computer Sciences Technical Report 1648. University of Wisconsin–Madison.
[47] Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. 2018. The spread of low-credibility content by social bots. *Nature communications* 9, 1 (2018), 1–9.
[48] 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 (2017), 22–36.
[49] Kai Shu, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Mining Disinformation and Fake News: Concepts, Methods, and Recent Advancements. In *Disinformation, Misinformation, and Fake News in Social Media: Emerging Research Challenges and Opportunities*, Kai Shu, Suhang Wang, Dongwon Lee, and Huan Liu (Eds.). Springer International Publishing, Cham, 1–19. https://doi.org/10.1007/978-3-030-42699-6_1
[50] Craig Silverman. 2016. Here are 50 of the biggest fake news hits on Facebook from 2016. *Buzzfeed News* (2016), 1–12.
[51] Eugenio Tacchini, Gabriele Ballarin, Marco L Della Vedova, Stefano Moret, and Luca de Alfaro. 2017. Some like it hoax: Automated fake news detection in social networks. *arXiv preprint arXiv:1704.07506* (2017).
[52] Alaa Tharwat and Wolfram Schenck. 2023. A Survey on Active Learning: State-of-the-Art, Practical Challenges and Research Directions. *Mathematics* 11, 4 (2023). https://doi.org/10.3390/math11040820
[53] 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)*. 809–819.
[54] Ciprian-Octavian Truică and Elena-Simona Apostol. 2023. It's all in the embedding! fake news detection using document embeddings. *Mathematics* 11, 3 (2023), 508.
[55] Cristian Vaccari and Andrew Chadwick. 2020. Deepfakes and disinformation: exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media+ Society* 6, 1 (2020), 2056305120903408.
[56] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151. Publisher: American Association for the Advancement of Science.
[57] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151.
[58] 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)*. Association for Computational Linguistics, Vancouver, Canada, 422–426. https://doi.org/10.18653/v1/P17-2067
[59] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In *Advances in Neural Information Processing Systems*. 9054–9065.
[60] Melissa Zimdars. 2016. False, misleading, clickbait-y, and satirical "news" sources. *Google Docs* (2016).

A ROBUSTNESS CHECKS ON THE 1,006 SHARING-NETWORK SOURCES

This section presents additional results from offline and online experiments, highlighting the performance of sharing- and exposure-based models on the 1006 sources in the sharing network.

Table 4 shows the PR-AUC of the best-performing model (Random Forest) using different feature sets on the 1006 sources. Results for offline (columns 2-4) and online (columns 5-7) settings are included. We report PR-AUC separately for sharing and exposure features, and the relative improvement of exposure-based models over sharing-based models (columns four and seven). The last column compares the best-performing models in online and offline settings using the same features.

The results in Table 4 show superior performance of the exposure-based models compared to the sharing-based models across feature families except for baseline features. The PR-AUC of the exposure-based model outperforms the sharing-based model by 2.5% to 38.9% in offline settings and 19.6% to 22.6% in online settings. The only exception is the model based on the baseline features, which exhibits a 14.4% degradation in the offline settings and 0.1% in the online settings. Further investigation suggests that in offline settings the number of fake neighbors adds noise to many non-fake sources, possibly due to fake sources being co-consumed with a large number of non-fake sources. When using all available features, the exposure-based model outperforms the sharing-based model by 11.5% in offline settings and by 6.9% in online settings. Similar to results on the full set of sources, all individual feature families improve over the baseline features. Note that the inclusion of multiple features on the subset of 1006 sources leads to marginal overfitting compared to individual feature families.

B GENERALIZABILITY ON OTHER DATASETS: FAKE NEWS CASCADES

To evaluate the generalizability of our methodology, we applied it to Reuben et al.'s dataset [44], which includes social sharing cascades of 3,355 news items (771 fake, 2,584 non-fake) posted by 1,621 Twitter users and potentially exposed to 188,459 others.

While this dataset only has a partial exposure network, it offers an opportunity to test the portability of our methodology. In particular, we aimed to check if exposure features still outperform sharing features in this new dataset. If similar trends were found, it would support the generalizability and robustness of our main findings for related fake news detection tasks.

Table 5 shows the PR-AUC results of models trained with different feature sets. Network representations significantly enhance performance over the baseline for both sharing and exposure features. Notably, the Exposure-Network model outperforms the Sharing-Network model by 28.11%. Incorporating exposure information with text features also yields similar positive gains.

Therefore, despite having only a highly incomplete view of the exposure network, incorporating exposure signals still enhances model performance. While the original dataset used in the paper with its full exposure network and user demographic information remains ideal for a comprehensive evaluation, these supplemental findings affirm the generalizability of exposure networks for improving fake news source detection.

C ACTIVE LEARNING IN OFFLINE SETTINGS

In this section, we report the complementary results of the experiments conducted in §4.3, focusing on *offline* settings. We maintained a consistent setup with the offline settings, utilizing all available features while imposing the constraint of acquiring a fixed number of labels at a fixed number of labeling intervals. This approach ensured that all active learning strategies commenced with the same initial set of 100 sources. The test sources of each five-fold cross-validation were used as a held-out set. We experimented with

	Offline settings			Online settings			Online vs. Offline
Features	Sharing [14]	Exposure	%	Sharing [14]	Exposure	%	%
Baseline	.694	.595	-14.4	.541	.540	-0.1	-22.1
% Fake Neighbors	.667	.151	-77.4	.602	.455	-24.5	-9.7
Popularity	.376	.595	+58.1	.250	.510	+104.0	-14.3
Network	.706	.783	+10.9	.582	.696	+19.6	-11.2
UF-ISF	.680	.717	+5.4	.546	.701	+28.4	-2.2
Node2vec	.701	.705	+0.6	.680	.566	-16.7	-3.6
Audience	.698	.716	+2.5	.654	.805	+23.2	+12.5
Text (TF-IDF)	.629	.873	+38.9	.593	.728	+22.6	-16.7
Text (Embedding)	.775	.846	+9.1	.668	.767	14.8	-9.3
Baseline+Audience	.724	.721	-0.4	.670	.773	+15.3	+6.7
Baseline+Network	.739	.739	+0.1	.650	.695	+6.9	-6.0
Baseline+Audience+Network	.736	.737	+0.1	.679	.743	+9.4	+0.8
Baseline+Text	.709	.866	+22.1	.668	.738	+10.4	-14.8
All features	.762	.850	+11.5	.715	.765	+6.9	-10.1

Table 4: The PR-AUC of the best-performing classification model using different feature families on the 1006 source subset. Offline settings results were evaluated using five-fold cross-validation over sources (columns 2-4). Online settings results were evaluated on emerging sources in increments of two weeks with the model gradually having access to more labels (columns 5-7). In both settings, sharing- and exposure-networks are reported separately, with the percentage gain of exposure models over sharing models. The last column shows the percentage improvement of the best online model over the best offline model.

Feature	Sharing	Exposure	%
Baseline	0.363	0.368	+1.38
-%Fake Neighbors	0.323	0.321	-0.62
-Popularity	0.295	0.307	+4.07
Network	0.249	0.319	+28.11
-UF-ISF	0.250	0.260	+4.00
-Node2vec	0.299	0.378	+26.42
Text (TF-IDF)	0.362	0.396	+9.39
Baseline+Network	0.369	0.397	+7.59
Baseline+Text	0.435	0.450	+3.45
All Features	0.434	0.440	+1.38

Table 5: PR-AUC results of the best-performing classification model using different feature families on the fake news cascade dataset. Both the offline and online evaluation reports the results for the sharing- and exposure-networks separately with the improvement of the exposure-based model relative to the sharing-based model in percentages

different labeling "budgets" in 10, 20, 40, 50, 100 and found similar results. For brevity, we only report the findings with 40 new labels in each epoch as representative of this range and conducted 10 labeling intervals. The full description of the strategies used for selecting sources for labeling appears in §3.4.

The results of our active learning experiments are shown in Figure 3. The figure shows PR-AUC as a function of the number of labeled sources. Each line represents the PR-AUC of the different active learning strategies based on the number of unknown sources that were sent for labeling averaged over the five folds. The left panel shows the results obtained using the sharing network and the right panel shows the results of the exposure network.

The results in Figure 3 show that the Uncertainty Sampling strategy outperforms all other strategies both when using sharing- and exposure-network information ($p<0.05$). After a few iterations, Uncertainty Sampling achieves comparable performance to the Unlimited Budget strategy despite having fewer labels. For example, Uncertainty Sampling using the exposure-based model reached a PR-AUC of 0.803 after only five iterations (a total of 300 labels), which is a mere 28.5% of the labels used in the Unlimited Budget benchmark with a PR-AUC of 0.808. Similarly, Uncertainty Sampling using the sharing-based model achieved a PR-AUC of 0.582 with just 220 labels, equivalent to 27.3% of the Unlimited Budget benchmark with PR-AUC of 0.593. These results demonstrate that our methodology can achieve near-optimal performance with significantly fewer labels also for historical data labeling.

Similar to results in the active learning experiment on the online settings, Figure 3 shows that Random Sampling is a strong baseline that is superior or comparable to three other sampling strategies. When using the exposure network (right panel), the PR-AUC of the High Degree, Certainty Sampling, and Diversity Sampling strategies are significantly lower than Random Sampling. When using the sharing network (left panel), only Certainty Sampling is significantly lower, while the two other strategies are comparable to Random Sampling. This strengthens the indication that sampling strategies that focus on just one "type" of sources are not optimal for identifying fake news sources, and a different sampling strategy is needed to perform better than random.

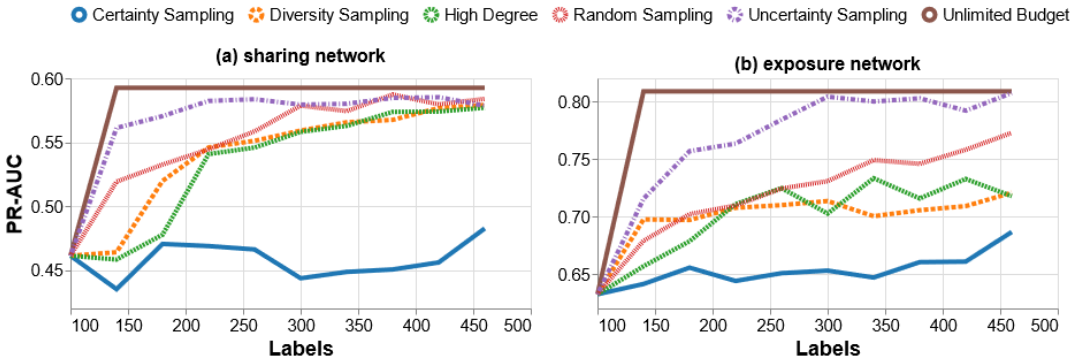


Figure 3: PR-AUC as a function of the number of labeled sources in offline settings based on sharing (left) and exposure (right) networks. Each line represents the PR-AUC of the different active learning strategies.

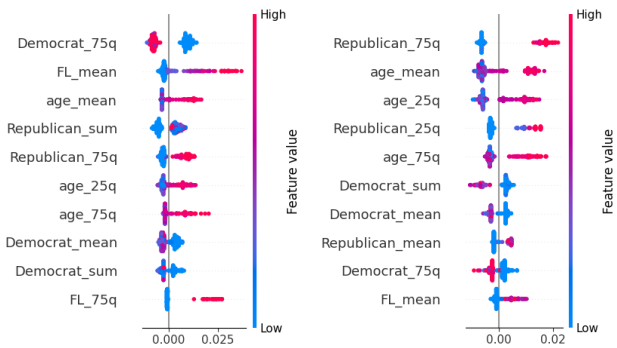
D FEATURE ANALYSIS ON SHARING-BASED MODELS

In addition to the feature analysis conducted in §4.2, we report the SHAP values of our best-performing models over the sharing network.

Figure 4 presents the SHAP values of both offline (left) and online (right) models computed over the sharing network with all available features. Features on the y-axis are sorted according to their importance from top to bottom. Each point represents a source and its position on the x-axis corresponds to the SHAP value of a particular feature. The color indicates the magnitude of the raw feature value. For example, the average age of people sharing a source (age_mean in panel a) shows that many fake news sources have a high average age and that is the most important feature. It is important to note that uninterpretable features were omitted from the figure (e.g., Node2vec embeddings) and we reported separately on text features below due to their plurality.

Similar to the feature analysis results on exposure-based models, age quantiles and mean appear high in both panels of Figure 4, indicating that older audiences were more likely to share or consume content from fake news sources. Moreover, party registration features were shown as important, strengthening the indication that fake news sources were less associated with Democrat audiences and more with Republicans. Finally, the U.S. state of Florida appears to have larger audiences of fake news sources consumers, but again it might be due to the small sample size in some states.

Next, we examined the top five text terms having the highest SHAP values. In offline settings, the top terms were bombshell, breaking, exposed, hillary, and WikiLeaks. In online settings, the top terms were hillary, video, WikiLeaks, soros, and breaking. The above terms represent public figures and topics that were involved and linked to fake news stories during the 2016 election. Clearly, these textual features are also tied to the specific time-span of the data.



(a) Offline **(b) Online**

Figure 4: SHAP values of the best-performing sharing model in the offline (left) and online (right) settings.