



# A data-driven approach to supporting fact-checking and mitigating misinformation and disinformation through domain quality evaluation

Kaveh Kadkhoda Mohammadmosaferi<sup>1,2\*</sup>, Anna Bertani<sup>1,3,4</sup>, Thomas Louf<sup>1,5</sup> and Riccardo Gallotti<sup>1</sup>

Handling Editor: Diogo Pacheco

\*Correspondence: [kkadkhodamohammad001@dundee.ac.uk](mailto:kkadkhodamohammad001@dundee.ac.uk)

<sup>1</sup>Fondazione Bruno Kessler, Trento, Italy

<sup>2</sup>University of Dundee, Dundee, UK  
Full list of author information is available at the end of the article

## Abstract

Misinformation and disinformation spread rapidly on social media, threatening public discourse, democratic processes, and social cohesion. One promising strategy to address these challenges is to evaluate the trustworthiness of entire domains (source websites) as a proxy for content credibility. This approach demands methods that are both scalable and data-driven. However, current solutions such as NewsGuard and Media Bias/Fact Check (MBFC) rely on expert assessments, cover only a limited number of domains, and some (e.g., NewsGuard) require paid subscriptions. These constraints limit their usefulness for large-scale research. This study introduces a machine-learning-based system designed to assess the quality and trustworthiness of websites. We propose a data-driven approach that leverages a large dataset of expert-rated domains to predict credibility scores for previously unseen domains using domain-level features. Our supervised regression model achieves moderate performance on test data and high performance on independent datasets, highlighting its ability to generalize to unseen domains. Using feature importance analysis, we found that PageRank-based features provided the greatest reduction in prediction error, suggesting that link-based indicators play a central role in domain trustworthiness. The solution's scalable design accommodates the continuously evolving nature of online content, ensuring that evaluations remain timely and relevant. The framework enables continuous assessment of thousands of domains with minimal manual effort. This capability allows stakeholders (social media platforms, media monitoring organizations, content moderators, and researchers) to allocate resources more efficiently, prioritize verification efforts, and reduce exposure to questionable sources.

**Keywords:** Domain trustworthiness assessment; Fact-checking algorithms; Misinformation and disinformation mitigation; Machine learning for credibility analysis

## 1 Introduction

The rapid spread of misinformation and disinformation on social media threatens public discourse, policymaking, and societal integrity [1]. Fact-checking individual posts is

© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

labor-intensive and unsustainable on a large scale. Therefore, assessing the quality of domains shared on social media offers a more practical and scalable alternative. Domains consistently producing reliable content differ substantially from those disseminating misinformation. Evaluating domain trustworthiness can help users recognize potential biases, which aids their interpretation of content. At the same time, this approach assists platforms and fact-checking organizations in prioritizing verification efforts, developing content-moderation strategies, and reducing exposure to low-quality information.

Current domain rating services, such as NewsGuard [2] and Media Bias/Fact Check (MBFC) [3], primarily depend on expert reviews and systematic analyses, which have proved valuable for scientific research [4–6]. However, these services cover a limited number of domains. As of September 2022, NewsGuard had evaluated only 8178 domains [7]. Additionally, their paid subscription models limit their applicability for large-scale research and automated content-moderation settings.

Despite these limitations, expert-based ratings have supported extensive academic research. NewsGuard data have been used to demonstrate increasing reliance on unreliable sources among U.S. politicians, highlighting substantial partisan differences [8]. Another study revealed that user engagement with unreliable sources primarily reflects personal biases rather than algorithmic issues [9]. Similarly, patterns of misinformation consumption across Europe indicate persistent engagement with unreliable domains [10]. During the Coronavirus Disease 2019 (COVID-19) pandemic, MBFC ratings helped identify highly polarized misinformation-sharing behaviors on social media, confirming the global relevance of domain-level trustworthiness evaluations [11].

These studies collectively emphasize the critical importance of scalable methods for domain quality assessment. An automated approach capable of evaluating and assigning reliability scores to a wide range of domains would considerably enhance misinformation research, policy interventions, and content moderation practices.

To address this need, we propose a scalable solution that leverages a comprehensive dataset of 11,520 domains compiled by Lin et al. [7]. This dataset merges expert ratings from multiple credible sources, including NewsGuard [2], MBFC [3], Ad Fontes Media [12], crowdsourced judgments of news source quality [13], the Iffy Index [14], and Lasser et al. [8]. Moreover, Lin et al. [7] applied advanced imputation techniques to address missing data, thereby improving the dataset's completeness.

Building upon this dataset, our method utilizes machine learning and domain analytics to dynamically evaluate website trustworthiness. We extract various informative domain features, such as PageRank, Domain Authority, and Spam Score, to characterize reliability. By employing a LightGBM (Light Gradient Boosting Machine) regression model, we achieved a mean absolute error of 0.11 when predicting expert-based trustworthiness ratings for previously unseen domains. This result demonstrates moderate accuracy and practical applicability. Furthermore, our feature analysis revealed that PageRank-based measures provided the largest reduction in prediction error, highlighting the central role of link-based indicators in domain trustworthiness. In particular, the metrics from Open PageRank [15], an initiative that indexes billions of web pages via the Common Crawl corpus, demonstrated strong predictive performance.

The rest of this paper is organized as follows: Sect. 2 provides a review of the existing research related to our work. Section 3 describes the methods we used, including how we collected data, extracted features, and built our models. Section 4 shows our results,

explaining the main findings and how well the solution performed. Lastly, Sect. 5 discusses these findings in more detail, addresses limitations, and suggests possible next steps.

## 2 Background

Early research on detecting online misinformation focused on analyzing individual news items or social media posts [16, 17]. Many studies used machine learning techniques to classify news articles or posts as true or false by examining textual attributes, writing styles, and patterns of social media sharing. Surveys of online misinformation have noted that misleading stories tend to use exaggerated or dramatic language, sensational or attention-grabbing headlines, and specific or unusual diffusion patterns. While these content-level approaches can effectively identify false content in small settings, examining each article or post separately is computationally expensive and time-consuming, especially with millions of posts appearing daily. As a result, there is growing interest in broader domain-level methods of assessing credibility.

A promising way to address scalability issues is to evaluate credibility at the domain level, based on the assumption that a news source's overall reliability reflects the trustworthiness of its content. Services like NewsGuard, Media Bias/Fact Check (MBFC), and Ad Fontes Media provide expert-based ratings for entire domains, focusing on accuracy and bias. While these expert-curated lists are useful, they cover only a small portion of active news sites. To expand coverage, some researchers combined several lists into a single "wisdom of experts" set [7]. However, even this unified list does not solve the fundamental challenge of labeling the many domains that remain unrated. This gap motivates automated methods that can estimate credibility at scale.

To tackle this gap, several researchers have developed techniques for evaluating the credibility of news domains. One approach uses factual consistency as a proxy for trustworthiness. For example, Dong et al. introduced a Knowledge-Based Trust method that evaluates a website by extracting numerous factual claims from its pages and checking them against a knowledge base [18]. They found that sites publishing fewer false statements receive higher trust scores, demonstrating that internal content signals can complement external link metrics. Another line of research trains machine learning models with diverse features to predict the reliability of news outlets. Baly et al. built a domain-level classifier that combines multiple signals: the writing style of articles, metadata from Wikipedia and Twitter, the domain's URL structure, and web-traffic statistics [19]. This model, trained on a large set of labeled news sites, achieved higher accuracy than methods using any single type of feature. The authors emphasized that profiling sources in this way can serve as a prior for fact-checking systems. Subsequent work has extended this multifaceted approach. For example, Panayotov et al. introduced GREENER, a graph neural network model that represents news outlets as nodes in a graph based on audience overlap. By learning from this audience network and combining it with text and social features, their model improved factuality and bias predictions beyond text-only approaches [20].

Building on domain-level evaluation, recent studies have explored specific contexts and emerging threats. Lepird et al. introduced a framework with a Non-Credibility Score (NCS) based on how often a site's content is shared on Facebook relative to mainstream news outlets [21]. This score ranges from 0 to 1 and helps detect politically biased "pink slime" sites that pose as local news outlets [21]. Their work applies network analysis of social media sharing patterns to discover new hyper-partisan sites. In another example,

Chen and Freire developed a system to proactively surface new fake news domains by mining real-time Twitter data [22]. Their method constructs a graph linking domains that are shared by similar groups of users, and then applies a topic-agnostic classifier to prioritize the most suspicious sites. This approach identifies emerging misinformation sources before they appear on existing watchlists, and it includes an interface to help fact-checkers investigate the flagged domains [22]. A related study by Sehgal et al. analyzed how misinformation sites hyperlink to each other [23]. By building domain-level hyperlink and social-media networks, they found that misinformation domains interlink with each other far more than with trustworthy sites. Their classifier, trained on these link patterns and Twitter engagement data, highlighted clusters of interconnected misinformation peddlers [23].

A separate direction focuses on technical attributes of websites rather than their content or social media signals. De Mendonça et al. presented a credibility assessment model for Brazilian news domains that purposely avoids analyzing article text [24]. Instead, it relies solely on domain and infrastructural signals such as domain age, Domain Name System (DNS) records, server locations, and Secure Sockets Layer (SSL)/Transport Layer Security (TLS) certificate details. By using these publicly available indicators, they demonstrated that non-content signals can help distinguish reliable sites from unreliable ones. However, they found that domain geolocation features were less useful: both credible and non-credible news sites often share the same hosting providers, so location data was not effective for determining reliability. Similarly, Hounsel et al. explored using infrastructure-level features like domain registration details, TLS/SSL certificate characteristics, and hosting configurations [25]. They argued that such signals are valuable because they are available even before any content is published, enabling early identification of suspicious domains. In a pilot deployment, Hounsel and colleagues' classifier successfully discovered previously unreported disinformation websites [25].

Researchers have also developed content-agnostic detection systems that look at web traffic patterns and site behaviors. Papadopoulos et al. introduced Fake News Detection-as-a-Service (FNDaaS), a system that uses features like DNS record changes, domain age, page structure, and page loading patterns without examining the actual news content [26]. Similarly, Chalkiadakis et al. analyzed the lifecycle and traffic patterns of 283 fake-news websites and built a content-agnostic classifier [27]. They observed that fake-news sites often have short life spans and synchronized periods of activity, insights that can inform detection strategies [27].

Recently, Pereira et al. proposed a graph-based detection framework that leverages real browser traffic data [28]. They observed that existing domain classifiers often perform well on benchmark datasets but lose precision by up to an order of magnitude when deployed on live web traffic. To address this, the authors constructed a graph of user navigational patterns between websites and extracted traffic-based features to train a classifier [28]. Notably, the authors describe their classifier as a first-stage filter that surfaces suspicious domains for human review, rather than a final verdict on misinformation.

Despite these advances, current methods still face practical barriers. Many approaches require rich data or complex features that are expensive to obtain and maintain. For example, Baly et al. relied on multi-source information, which is costly to gather [19]. Lepird et al. noted that their NCS method depends on large-scale social media data, which can be difficult to access due to platform restrictions [21]. Similarly, systems like Chen and

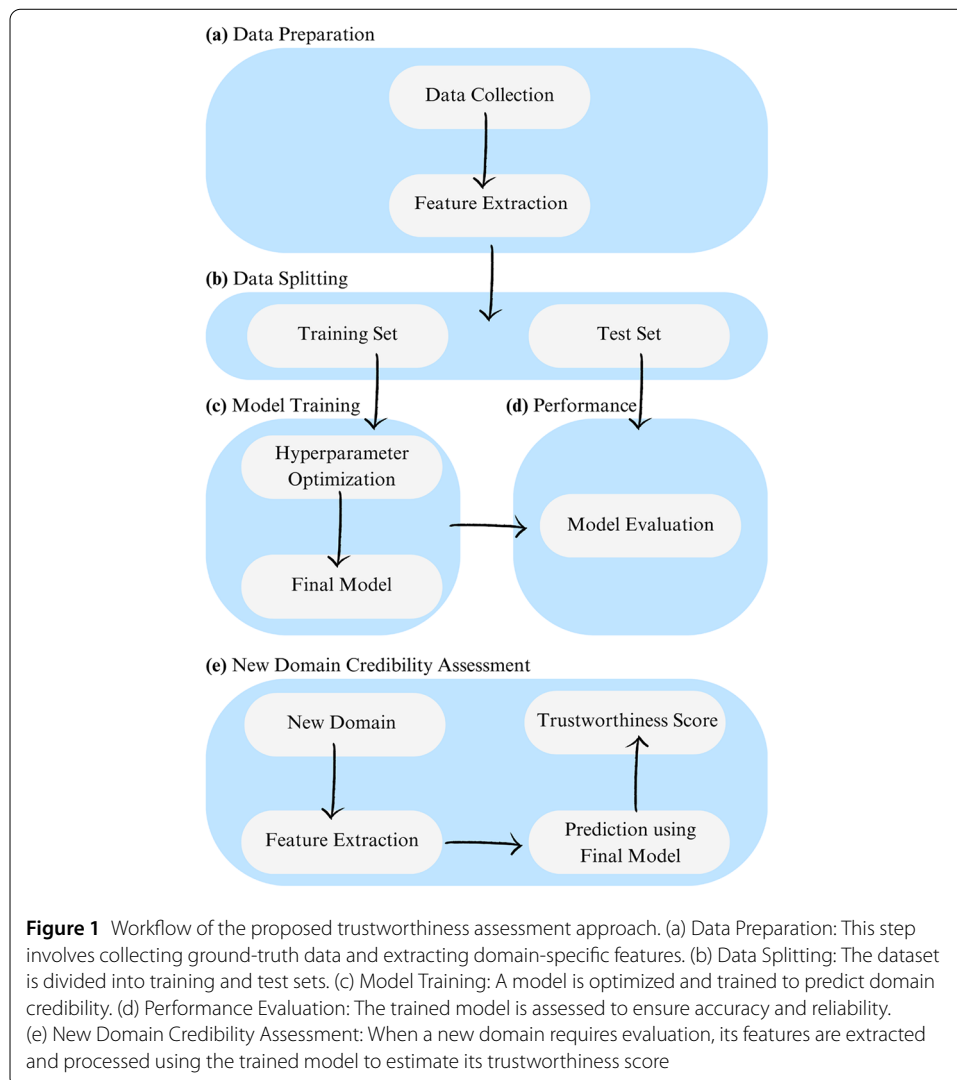
Freire's Twitter-based discovery tool [22] need real-time social media feeds that may not always be available. Some techniques rely on knowledge bases or technical metadata that do not exist for every site, for instance, a factual consistency approach like Dong et al.'s assumes a comprehensive database of truths [18]. Models built on very complex feature sets (combining content, network, and metadata signals) can also be resource-intensive to compute and update. Additionally, methods that use aggregated browser logs (such as Pereira et al.'s traffic-based classifier [28]) depend on data that may not be accessible to most researchers.

Another major challenge is that misinformation websites adapt quickly, which can undermine detection efforts. Many untrustworthy sites share infrastructure with legitimate ones; therefore, signals like domain geolocation provide little distinction [24]. These sites can also change their domain names, go dormant and then reappear, or frequently shift their hosting setups. As a result, classifiers trained on past data may struggle to recognize new or evolving misinformation sites. Researchers have documented the ephemeral nature of fake news domains [27] and noted that the performance of domain classifiers often drops sharply when moving from controlled test datasets to real-world web traffic [28]. This underscores the need for approaches that can continuously adapt to the fast-changing landscape of online misinformation.

Given these challenges, simple, scalable signals are needed that can flag suspicious domains quickly without relying on expensive or hard-to-obtain data. Our work aims to address this need by building a domain-level credibility model that uses only link-based metrics, such as PageRank, domain authority, and backlink counts, which are obtainable from public or low-cost sources. Because these metrics are readily available and inexpensive, and because practically all domains have link data, our system can efficiently evaluate tens of thousands of sites. Importantly, domain quality is not static: web signals and site behavior change over time. Unlike fixed expert lists, our approach supports continuous tracking of changes in a domain's reliability. We stress that our system is not intended to deliver final judgments about misinformation. Instead, it functions as a warning signal, a first-pass filter to help fact-checkers and platforms prioritize which domains to scrutinize more closely. This role is analogous to the "prior" described in the work of Baly et al. [19] and to the early-alert mechanism in Chen and Freire's system [22]. By narrowing the field to a manageable set of high-risk sites, our approach enables human reviewers to focus their efforts where they are most needed. Ultimately, any domain flagged by our model should undergo human verification and deeper content analysis before being labeled as a misinformation source. In this way, our method complements existing content-level and social network-level techniques, offering a scalable new tool to combat the rapidly evolving landscape of online misinformation.

### 3 Methodology

Evaluating the credibility of online domains requires a structured, data-driven workflow. Figure 1 illustrates our five-step process: (a) Data Preparation, (b) Data Splitting, (c) Model Training, (d) Performance Evaluation, and (e) New Domain Credibility Assessment. In (a) Data Preparation, ground-truth data are gathered and domain-specific features are extracted. This ensures the model is trained on reliable indicators of trustworthiness. Next, in (b) Data Splitting, the dataset is divided into training and test sets, allowing for robust validation. During (c) Model Training, hyperparameter optimization is performed



to maximize predictive accuracy. In (d) Performance Evaluation, the model's reliability is rigorously tested. Finally, in (e) New Domain Credibility Assessment, features from newly encountered domains are fed into the trained model to generate a trustworthiness score.

### 3.1 Data collection

We based our study on the dataset compiled by Lin et al. [7], which aggregates domain quality ratings from six expert sources. Because these sources vary in coverage and methodology, discrepancies exist among their domain ratings. To address missing values and ensure data consistency, Lin et al. [7] applied multiple imputation methods and then conducted a principal component analysis. The analysis showed that the first principal component explained about 68% of the variance in the dataset.

Originally, the dataset included 11,520 domains, each assigned a quality rating between 0 and 1. Scores closer to 1 indicate higher trustworthiness, whereas those near 0 indicate lower trustworthiness. To include only active, valid websites, we sent automated Hypertext Transfer Protocol (HTTP) requests to each domain and retained only those that returned an HTTP 200 status code. This filtering step excluded inactive or non-existent sites. After



**Table 1** Overview of input and target variables in the dataset. Domain names were anonymized using '###'

Domain	Input Variables				Target Variable
	PageRank Decimal	Domain Authority	Spam Score	...	Trustworthiness
fr###.net	5.47	75	1	...	0.72
th###.com	5.04	54	3	...	0.22
cl###.news	3.53	36	5	...	0.07
...	...	...	...	...	...

filtering, 9432 domains remained for analysis, each with a trustworthiness score from Lin et al. [7].

### 3.2 Feature extraction

We extracted domain-specific features from two external services. The first, the Moz Application Programming Interface (API), provides domain-level metrics including Domain Authority and Spam Score [29]. Domain Authority estimates how well a domain is likely to rank in search engine results based on its backlink profile, while Spam Score indicates the likelihood that a domain is associated with spammy links. These metrics offer insights into a domain's visibility, reliability, and potential influence. Supplementary Table S1 presents a comprehensive list of all 46 features obtained from Moz, along with their descriptions.

The second service, the Open PageRank API, supplied two additional reputation signals: PageRank Decimal and Rank [15]. Specifically, PageRank Decimal is a continuous score reflecting a domain's authority based on inbound links, while Rank denotes the domain's position among all indexed websites (indicating its relative popularity). In total, we gathered 48 features across both APIs.

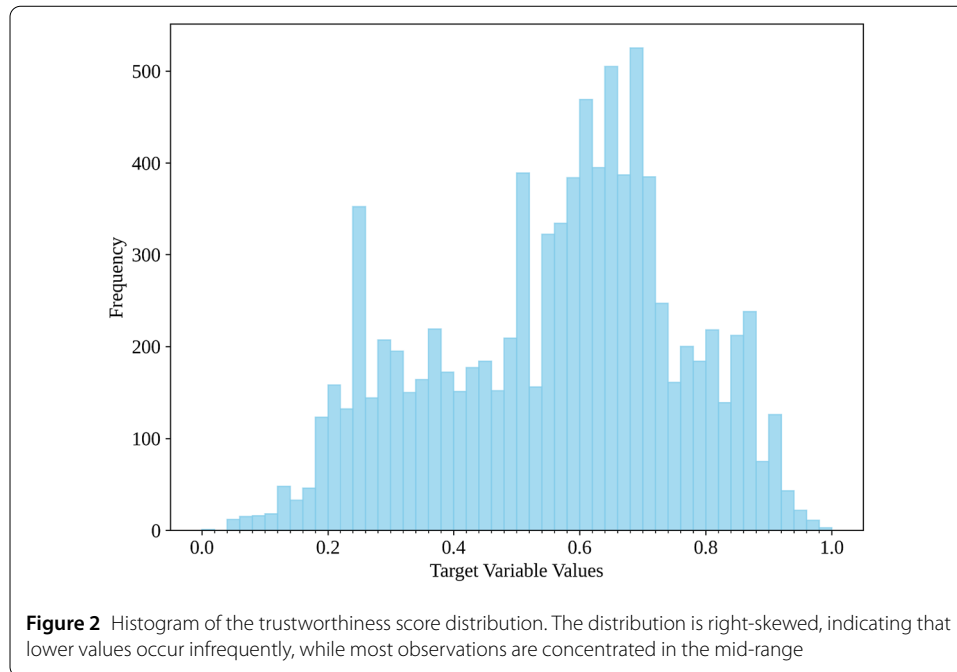
### 3.3 Data splitting

Table 1 summarizes the input variables (e.g., PageRank Decimal, Domain Authority, Spam Score) and the target variable (Trustworthiness), which ranges from 0 to 1. Figure 2 shows that the target variable's distribution is right-skewed: a small number of low scores are present, and most values cluster around the mid-range. This imbalance makes it essential to carefully manage the train/test split to ensure the model performs consistently across different reliability levels.

To address this, we used stratified data partitioning. Because the target variable is continuous, we binned its values for stratification, adjusting the number of bins according to the distribution of the data. This process helped maintain a similar overall distribution of the target variable in both the training and test sets. After binning, we allocated 90% of the data to training and 10% to testing. This ensured that the test set was representative, which allowed for an unbiased assessment of the model's generalization ability.

### 3.4 Model training

We selected Light Gradient Boosting Machine (LightGBM) [30] for its speed, efficiency, and ability to model complex feature interactions. We framed the prediction problem as a regression task, in which the model predicts a continuous trustworthiness score from the extracted domain features. To optimize performance, we employed Optuna [31], a hyperparameter optimization framework. Optuna tested different hyperparameter configurations over multiple trials, training a new LightGBM model in each trial and evaluating its performance. This process continued until the best hyperparameters were identified.

**Table 2** Performance of the Proposed Solution

Metric	Value
Mean Absolute Error (MAE)	0.11
Root Mean Squared Error (RMSE)	0.15
R <sup>2</sup> Score	0.46

### 3.5 Feature analysis

We analyzed the optimized LightGBM regression model to identify global feature importance using SHapley Additive exPlanations (SHAP) [32]. SHAP applies game theory to assign each feature a Shapley value, quantifying that feature's fair contribution to an individual prediction [33]. Collectively, the sum of these Shapley values equals the difference between a given instance's prediction and the model's average prediction. This sum offers a clear, additive explanation of how the model's output deviates from the norm [34]. Consequently, SHAP values illustrate precisely how each feature impacts a prediction relative to the dataset's average outcome.

## 4 Results

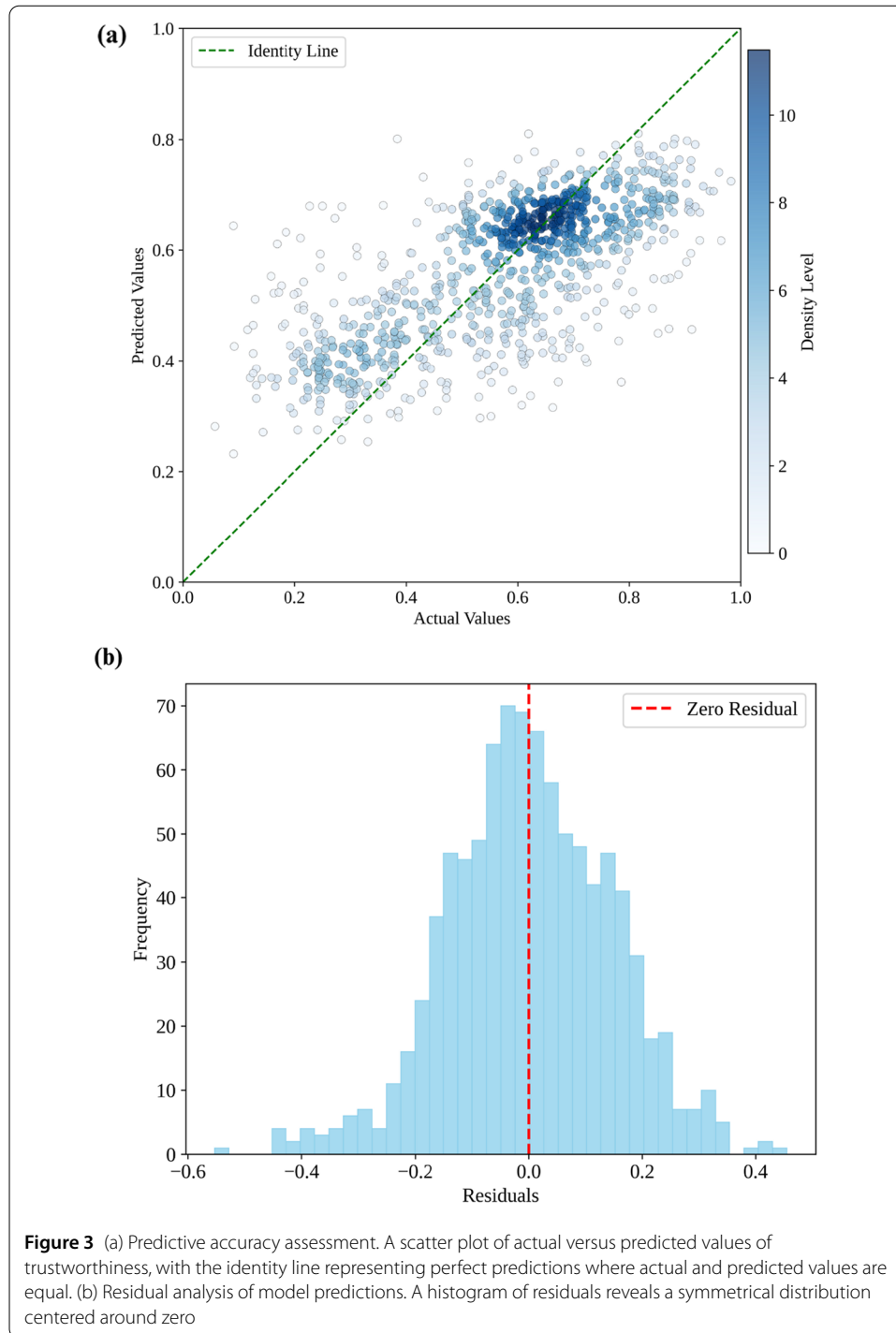
### 4.1 Performance

We evaluated the proposed solution on the test dataset to assess its ability to predict domain trustworthiness. As shown in Table 2, the model's performance metrics, including mean absolute error (MAE), root mean squared error (RMSE), and coefficient of determination ( $R^2$ ), indicate that the model explains part of the variance in trustworthiness and achieves moderate predictive accuracy.

### 4.2 Analysis

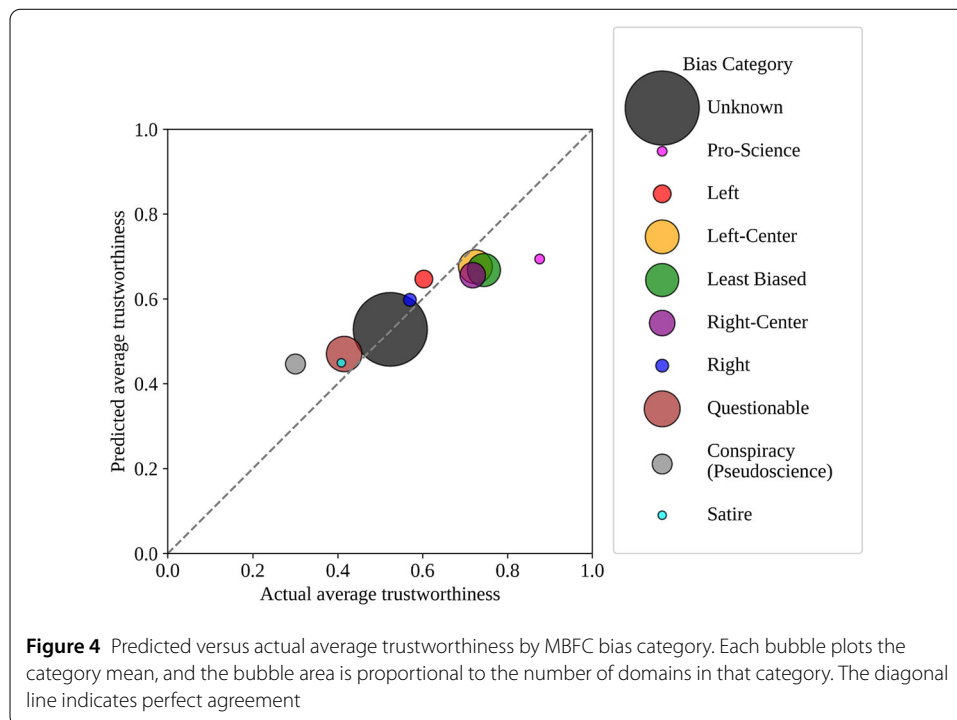
We further examined the model's accuracy on the test set. Figure 3a shows a scatter plot of each domain's actual trustworthiness score versus its predicted score. Ideally, all points





would lie on the 45° identity line, indicating perfect agreement between predictions and reality. Although most points do cluster near this line (indicating acceptable performance), the model tends to slightly underestimate highly trustworthy domains and slightly overestimate domains with low trustworthiness. Nonetheless, the overall Mean Absolute Error of 0.11 indicates a reasonable level of predictive accuracy.

Figure 3b shows a histogram of residuals, revealing a nearly symmetrical distribution around zero (mean residual  $\approx 0.002$ , skewness =  $-0.130$ ) with a standard deviation of



0.145. The residuals range from roughly  $-0.55$  to  $+0.45$ , suggesting that while underestimation or overestimation can occur, errors are not heavily skewed in one direction. The near-zero mean and balanced spread around zero imply that the model does not consistently favor overestimates or underestimates, which reinforces its overall reliability.

From a modeling perspective, achieving an  $R^2$  of 0.46 is scientifically meaningful. Most importantly, the features we selected (e.g., PageRank and Domain Authority) correlate with website trustworthiness, indicating that they help explain differences in trust levels in a meaningful way. While explaining 46% of the variance might seem modest, it is much better than random guessing and represents a meaningful step forward for a complex topic such as website credibility. This moderate explanatory power confirms that our chosen metrics are relevant. However, it also highlights that there is room to incorporate additional features in future research. Our results suggest that domain-level metrics capture trustworthiness only partially. This observation aligns with findings from other studies, which also typically report moderate explanatory power when predicting outcomes related to human judgment [35].

### 4.3 Bias-wise error analysis

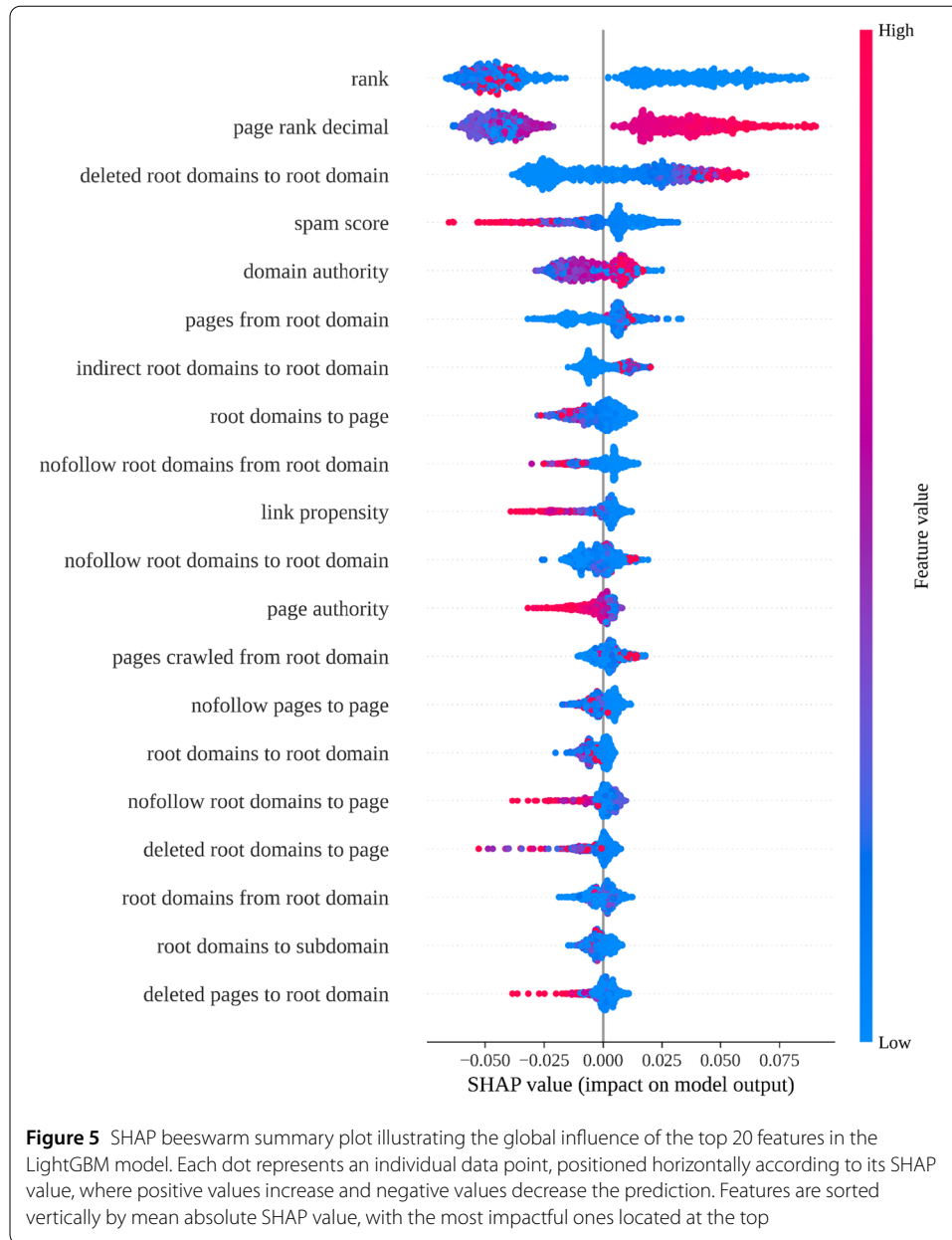
To investigate whether prediction errors were systematically related to content bias, we analyzed the model's performance across source categories defined by Media Bias/Fact Check (MBFC) [3]. MBFC classifies news sites by political leaning and factual reliability, assigning labels such as Left, Right, Left-Center, Right-Center, and Least Biased, as well as special designations like Pro-Science, Conspiracy/Pseudoscience, Questionable Source, and Satire. According to MBFC's framework, Least Biased and Pro-Science sources generally have the highest credibility, whereas Conspiracy/Pseudoscience and Questionable sources are among the least reliable. Figure 4 illustrates a bubble chart of the average predicted trustworthiness versus the actual average trustworthiness for each bias category. In

this chart, each bubble's position reflects the relationship between the model's prediction and the ground truth for that category, and the bubble size is proportional to the number of domains in the category. In an ideal scenario, all bubbles would lie on the 45° diagonal line (indicating no error for any bias group). Deviations from this parity line reveal where the model overestimates or underestimates trustworthiness for different bias groups.

Figure 4 shows that the model's largest underestimation occurs for the Pro-Science category. On average, Pro-Science domains (highly factual, science-based sources) have an actual trustworthiness of 0.875, but the model predicts only 0.694, a shortfall of about 0.181. In other words, the most trustworthy sites (as identified by MBFC's Pro-Science label) are underestimated by the model. Conversely, the model's most pronounced overestimation is for Conspiracy/Pseudoscience domains: the model predicts an average trust score of 0.446 for this group, compared to a much lower actual mean of 0.300 (an over-prediction by 0.146). This indicates that sources known for promoting conspiracies or pseudoscience are given too much credit by the model relative to their true trustworthiness. Notably, the model does recognize the overall credibility hierarchy to some extent, for example, it assigns higher scores to Pro-Science sites than to Conspiracy/Pseudoscience sites, aligning with the fact that the former are inherently more reliable.

Other bias categories exhibit smaller yet consistent errors that align with this trend. The model overestimates the trustworthiness of several biased or low-factuality groups by a modest amount. For instance, domains labeled Questionable (politically extreme and often poor in sourcing) have a mean actual score of 0.415, while the model predicts 0.470 (+0.055 higher than it should be). Similarly, the Satire category (deliberate humor or parody) is predicted at 0.449 vs. an actual 0.408 (+0.041), and Left-biased sources are predicted 0.647 vs. actual 0.603 (+0.044). The Right-biased category shows a smaller overestimate (0.598 vs. 0.570, about +0.028). In contrast, the model slightly underestimates several more moderate or highly factual groups. Least Biased news sites (centrist, with minimal bias) have an actual average trustworthiness of 0.744, but the model underrates them with an average prediction of 0.668 ( $\approx 0.076$  low). A similar underprediction appears for Left-Center (predicted 0.676 vs. actual 0.724,  $-0.048$ ) and Right-Center sources (0.656 vs. 0.718,  $-0.062$ ). These errors, while smaller in magnitude than those of Pro-Science or Conspiracy sites, consistently suggest the model leans toward deflating scores for more reliable sources and inflating scores for less reliable ones. Moreover, the Unknown group, domains with no MBFC bias classification, shows virtually zero error (the model's mean prediction is 0.528 vs. an actual mean of 0.524, a difference of only +0.004).

This bias-wise breakdown provides insight into the model's errors and complements the aggregate performance metrics like MAE and  $R^2$  reported earlier. The overall Mean Absolute Error of 0.11 and  $R^2$  of 0.46 summarize the model's average accuracy, but they do not reveal which sources of error are most significant. By examining errors by bias category, we uncover that a substantial portion of the remaining prediction error is not random noise but arises from under-valuation of certain high-credibility sources and over-valuation of low-credibility ones. In summary, the bias-wise error analysis (Fig. 4) adds an important layer to our evaluation by pinpointing biases in the predictions. This deeper diagnosis goes beyond the aggregate MAE and  $R^2$ , helping ensure that our model's credibility scoring is not only accurate on average but also helping assess whether errors differ systematically across MBFC bias categories and where additional calibration/features may be needed.



**Figure 5** SHAP beeswarm summary plot illustrating the global influence of the top 20 features in the LightGBM model. Each dot represents an individual data point, positioned horizontally according to its SHAP value, where positive values increase and negative values decrease the prediction. Features are sorted vertically by mean absolute SHAP value, with the most impactful ones located at the top

#### 4.4 Feature importance analysis

We assessed the global impact of features using a SHAP summary (beeswarm) plot. This plot ranks the top 20 features by their mean absolute SHAP value, which clearly illustrates each feature's relative importance (Fig. 5). Examining specific features in this plot reveals distinct patterns. For instance, the Rank feature shows a clear trend: low rank values (near the top of the ranking) correspond to positive SHAP contributions on the right side of the plot, whereas high rank values are associated with negative contributions on the left side. Therefore, the model predicts higher trustworthiness scores for domains with better (lower) rank positions and lower scores for domains with worse (higher) rank positions.

Additionally, the PageRank Decimal feature displays the opposite pattern. High PageRank Decimal values correspond to positive SHAP contributions, thereby boosting the

model's predictions. By contrast, low PageRank Decimal values are associated with negative SHAP contributions, leading to lower predicted scores. Consequently, a higher PageRank Decimal value positively influences the prediction, whereas a lower value has a negative influence.

Meanwhile, the Spam Score feature shows a straightforward negative association with trustworthiness. High Spam Score values are almost exclusively associated with negative SHAP values on the left side of the plot, indicating that a higher Spam Score consistently lowers the model's predicted trustworthiness score. In contrast, low Spam Score values correspond to positive SHAP values, which elevate the model's predictions.

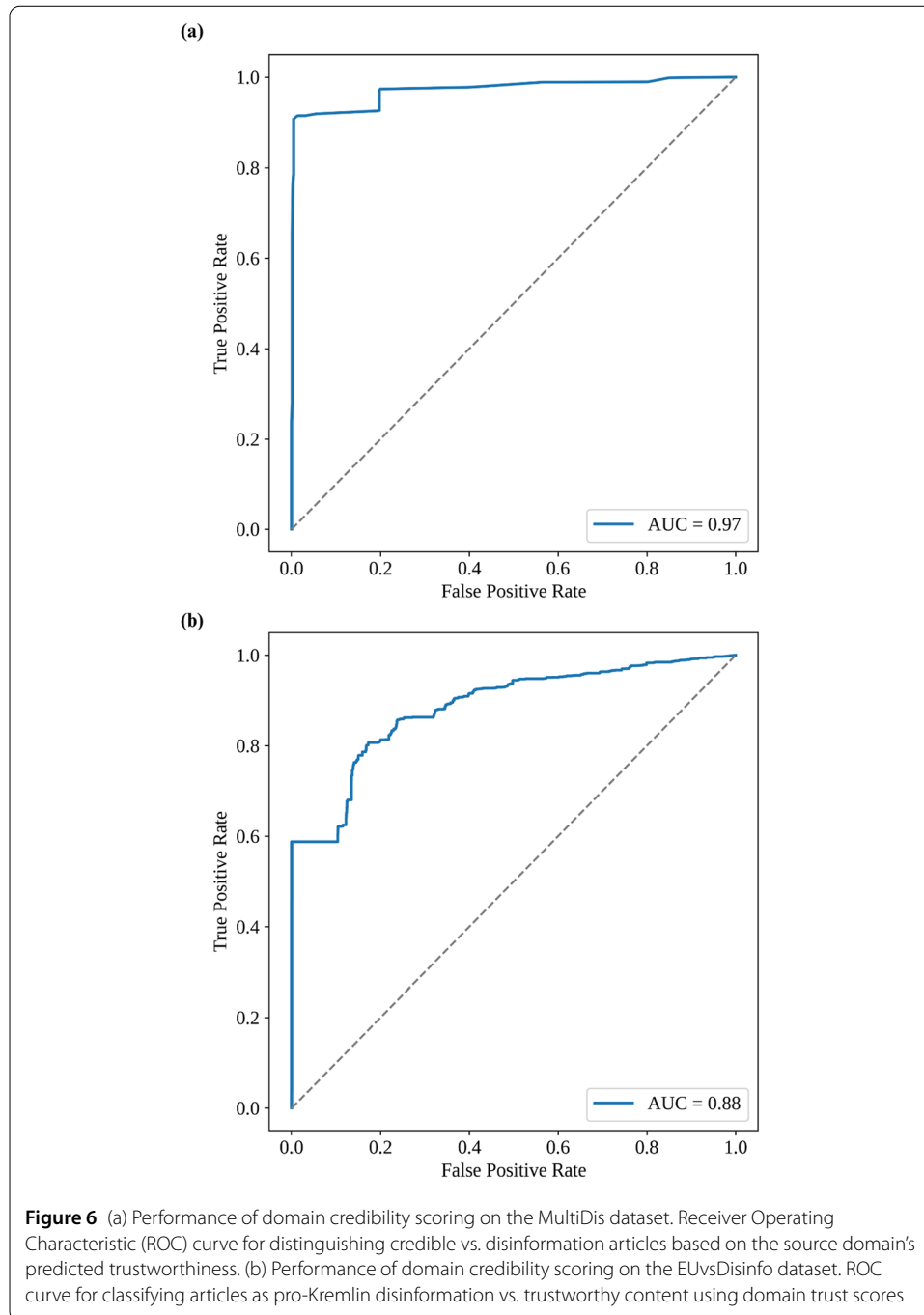
#### 4.5 Case study: external evaluation on MultiDis and EUvsDisinfo

To assess the generalizability of our domain-level model, we conducted an external evaluation on two independent datasets. The chosen datasets were MultiDis [36] and EUvsDisinfo [37], which provide ground-truth credibility labels for news content in different contexts. By comparing the model's predicted credibility scores to these ground-truth labels, we assess whether domains assigned high scores by the model correspond to credible information and whether those with low scores correspond to disinformation in each dataset. Because our model outputs a continuous credibility score rather than a fixed label, we assess its performance across the full range of classification thresholds using the Receiver Operating Characteristic (ROC) curve. We then summarize this curve using the Area Under the Curve (AUC), a single value between 0 and 1 that reflects how well the model separates trustworthy domains from disinformation.

After preprocessing MultiDis (URL-level aggregation and removal of tie cases), we retained 1903 unique articles, with 1260 labeled as credible information and 643 as disinformation. We applied our domain-level model to this dataset by assigning each article the predicted credibility score of its domain. This allowed us to test the model's generalizability beyond the original training data. When comparing these predictions to the known labels, the model achieved an AUC of 0.972 on MultiDis, indicating very strong discrimination between credible news and disinformation. As shown in Fig. 6a, the ROC curve for MultiDis rises steeply toward the upper-left corner of the plot, far above the diagonal chance line. In practical terms, this means the model can identify credible articles with a very high true positive rate while keeping the false positive rate extremely low.

The Spearman's rank correlation between the domain-predicted scores and the ground-truth labels is  $\rho = 0.775$ , which is a strong positive correlation. In other words, domains that our model scores highly generally correspond to credible content, whereas low-scoring domains typically host disinformation. These results demonstrate that the domain-based scores align well with actual content credibility in the MultiDis collection.

We further evaluated the model on the EUvsDisinfo dataset, which contains 18,249 entries (after preprocessing and label normalization), with 7567 labeled as trustworthy and 10,682 labeled as disinformation. On this larger and more challenging dataset, the model attained an AUC of 0.884 (Fig. 6b). Although this AUC is slightly lower than that for MultiDis, it is still very high, and the ROC curve lies well above the chance line, though not as close to the top-left corner as in the previous case. This indicates that the model continues to distinguish disinformation from truthful content with good accuracy, though with a bit more overlap between the two classes' score distributions. Consistently, the model assigned higher predicted credibility scores on average to EUvsDisinfo's trustworthy entries than to its disinformation entries, illustrating a clear separation between the classes.



The Spearman's  $\rho$  on EUvsDisinfo is 0.656, indicating a moderately strong monotonic relationship between the domain-based scores and the actual labels. In simpler terms, domains that our model evaluates as highly credible generally correspond to trustworthy information in this dataset, while those receiving low credibility scores tend to be sources of disinformation. This performance demonstrates that our domain-level model generalizes well to a different, real-world disinformation repository, successfully flagging likely unreliable sources even in a much larger and varied dataset.

**Table 3** Performance of various machine learning algorithms in predicting domain trustworthiness

Metric	LightGBM	Random Forest	Linear Regression	K-Nearest Neighbors Regression	Decision Tree Regressor
MAE	<b>0.114</b>	0.115	0.129	0.120	0.122
RMSE	<b>0.145</b>	<b>0.145</b>	0.164	0.153	0.154
R <sup>2</sup> Score	<b>0.463</b>	0.462	0.310	0.403	0.394

#### 4.6 Performance of various machine learning algorithms

To assess the impact of model choice, we evaluated a range of machine learning algorithms on the task of predicting domain trustworthiness. For a fair comparison, all models were trained on the same set of domain-level link-based features, ensuring that any performance differences reflect each algorithm's ability to exploit these signals. As shown in Table 3, ensemble methods such as LightGBM and Random Forest exhibited the best performance, achieving higher R<sup>2</sup> scores than the other algorithms and thus explaining more of the variance in trustworthiness. For example, a simple linear regression baseline attained a much lower R<sup>2</sup>, highlighting the advantage of these ensemble methods in capturing the complex patterns in the data.

### 5 Discussion and conclusions

One key insight from our work is that domain-level features are useful indicators of a news source's credibility. This approach helps combat online misinformation by focusing on the source of the news rather than examining each piece of content individually. Many studies suggest that the credibility of a source influences the reliability of its content [16, 38]. By identifying low-credibility websites in advance, fact-checkers and automated systems can concentrate on sources most likely to spread false or misleading information. This strategy can save time and resources compared to verifying individual articles or claims one by one. As summarized in Table 4, a variety of approaches have been proposed to detect misinformation, each with its own scope and required data, operational considerations, and benefits and limitations. This context highlights that while many sophisticated methods combine content analysis, social signals, or technical infrastructure cues, they often struggle to scale due to heavy data requirements. By contrast, our domain-level model relies on widely accessible link metrics, enabling efficient large-scale application.

Our results indicate that link-based metrics (such as those derived from PageRank) play an important role in identifying trustworthy domains. Highly referenced websites often provide more reliable content, whereas sites with fewer inbound links from reputable sources tend to be less credible. Using free services like Open PageRank makes it easier to conduct large-scale evaluations. This is helpful for research groups and social media platforms that need to evaluate many domains regularly. Additionally, data from paid services like Moz and other backlink tools can be important for maintaining acceptable performance. These platforms provide valuable metrics, such as Domain Authority and Spam Score, which improve accuracy and offer a clearer picture of a site's overall trustworthiness. By relying only on these link-based metrics, our framework avoids the extensive data requirements and social media platform dependencies that have hindered some prior approaches. Note that our current implementation obtains link metrics through the Moz and Open PageRank APIs, so it depends on the availability of these third-party services. If access to these APIs becomes restricted, the same feature extraction and model train-



**Table 4** Comparative Summary of Misinformation Detection Methods and Site Credibility Approaches

Approach / Method	Scope and Required Data	Operational Considerations	Benefits and Limitations
Content-level Machine learning detection [16, 17]	Operates at the level of individual news articles or claims. Requires the full text of each item (headline and body) with extracted linguistic features, along with large labeled datasets for model training. Some implementations also incorporate social engagement signals as additional inputs.	Processing is computationally intensive because each piece of content is analyzed separately. Scaling up to web-wide content volumes requires computational resources and efficiency optimizations.	Benefits: Enables fine-grained content analysis by capturing linguistic cues indicative of misinformation. Limitations: Often specific to certain domains or languages, and heavily dependent on the availability of labeled training data.
Expert-curated domain ratings [2, 3, 12]	Conducted at the domain (website) level. Expert reviewers evaluate sites based on editorial practices, ownership, transparency, and samples of content. Results are typically compiled into credibility lists or databases.	This approach is labor-intensive and does not scale easily. It often relies on proprietary or subscription-based services for access to the ratings. Coverage is limited to the domains that have been reviewed (on the order of only a few thousand sites).	Benefits: Provides highly accurate and context-rich evaluations with transparent justifications. Limitations: Cannot keep pace with the constant emergence of new websites, leaving many domains unrated. Some rating providers require paid access, which can restrict use in large-scale or automated settings.
Wisdom-of-experts [7]	A domain-level approach that aggregates multiple expert-generated lists of site credibility. It merges ratings or labels from different expert sources to produce a consensus score for each domain, covering the union of those sources.	Relies entirely on existing expert data and thus only includes sites present in the source lists. It is computationally lightweight to apply, but any update to the source lists requires recomputing the combined scores.	Benefits: Yields a unified credibility score that facilitates comparisons and generally extends coverage beyond any single list. Limitations: Provides no information for domains that are not in the input lists.
Knowledge-based trust estimation [18]	Domain-level method that extracts factual claims from website content and checks them against a reference knowledge base. Determining site credibility in this way requires large-scale web crawling to collect claims and a comprehensive knowledge database of known facts.	Deploying this approach is computationally demanding, given the need to crawl many pages and perform inference against the knowledge base. Moving from research prototypes to web-wide deployment would require substantial infrastructure investment.	Benefits: Assesses the correctness of content directly (rather than relying on popularity or other proxies), providing a truthfulness-oriented evaluation. Limitations: New or emerging topics often lack entries in the reference data, and the method's effectiveness depends on having an extensive ground-truth repository.

**Table 4** (Continued)

Approach / Method	Scope and Required Data	Operational Considerations	Benefits and Limitations
Multifactor news source classification [19]	Domain-level approach that utilizes a wide range of features from various sources. These include the writing style and linguistic attributes of content, metadata from Wikipedia, social media metrics (e.g., Twitter follower counts or engagement), the structure of the domain name, and site traffic statistics.	Requires integrating multiple data sources, which can be challenging if some information is missing (small or new sites might not have much data available). Applying this method involves scraping content and querying several APIs. If the requisite data can be obtained, the technique can be applied to thousands of sites, though data collection may become a bottleneck.	Benefits: By combining content-based features with network and metadata features, this model can improve prediction accuracy and serve as a useful initial credibility estimate. Limitations: Data gathering can be costly and time-consuming, and the approach depends on third-party services (for instance, Twitter or web analytics platforms) which may impose rate limits or change over time.
Graph-based news source profiling [20]	A domain-level approach that builds a graph-based representation of news sources. It incorporates audience overlap metrics, site performance indicators, textual content features, social media signals (from platforms like Twitter, YouTube, Facebook), and descriptions from sources such as Wikipedia.	This method depends on several data inputs, some of which are proprietary or no longer publicly available. The model typically uses graph neural networks, which are computationally intensive to train. In practice, overall coverage is constrained by the availability of all required data for each domain.	Benefits: Integrates audience network patterns with content and social features, capturing community-level relationships that pure content models might miss. This can outperform content-only models in identifying low-credibility sites. Limitations: Relies on data (such as historical web traffic statistics) that may be discontinued or limited in scope. Predictions for very small or new sites can be noisy, and the complex model architecture can make results difficult to interpret.
Social virality-based ranking [21]	Domain-level scoring derived from the performance of sites on social media, specifically Facebook. It uses metrics of virality from Facebook posts that contain the domain's URLs, such as the number of likes, the count of unique Facebook pages sharing the domain, and posting frequency.	Requires data access through platforms like CrowdTangle (or a similar tool) to gather Facebook engagement data, which can be subject to platform policy changes. Once the data have been obtained, calculating scores is straightforward. Coverage is inherently limited to domains that appear in the social media feed; however, popular new misinformation sites can be detected if they quickly gain traction on Facebook.	Benefits: Produces a quantitative ranking that can highlight previously unknown low-credibility (pink slime) news sites based on their unusual sharing patterns. Reflects community co-sharing behavior and has demonstrated strong performance in identifying problematic sources. Limitations: Only flags sites that achieve significant circulation on Facebook, so it may miss less widely shared domains. Moreover, obtaining large-scale, real-time data is difficult, especially as CrowdTangle is being phased out, making the approach harder to use operationally.

**Table 4** (Continued)

Approach / Method	Scope and Required Data	Operational Considerations	Benefits and Limitations
Real-time social feed discovery [22]	A domain-level method focused on proactively finding new misinformation websites through Twitter. It continuously monitors a stream of tweets for URLs, extracting page-level features from those linked websites (such as Hypertext Markup Language (HTML) layout and stylistic elements).	This approach requires sustained access to the Twitter streaming API and the ability to crawl the linked web pages. It can scale to cover events or topics by filtering the tweet stream with keywords, though its reach is limited by the chosen keywords and what is shared on Twitter.	Benefits: Enables early detection of emerging misinformation domains by leveraging real-time sharing patterns and page presentation features. It can surface new sites at their onset and provides an interface to support investigative analysis. Limitations: It is tied specifically to Twitter data, meaning it will miss sites that do not appear in the Twitter stream. The method requires selecting appropriate seed keywords and continuous access to Twitter's data, which may be subject to restrictions.
Mutual hyperlink network analysis [23]	Domain-level approach that investigates the network of hyperlinks among websites. Starting from known sets of misinformation and reputable domains, it constructs a graph of interlinking by crawling hyperlinks and using backlink APIs. The model also incorporates social engagement data (e.g., how often domains are shared on Twitter) and requires a set of ground-truth labels for training.	Building and analyzing the hyperlink network is data-intensive and requires extensive web crawling or access to large link index databases. The approach's effectiveness depends on the availability of sufficient link data and social signals for the domains in question.	Benefits: Reveals clusters of misinformation sites that frequently link to each other, a pattern which can be predictive of low credibility. Inclusion of Twitter sharing patterns further improves detection by highlighting commonly co-shared domains. Limitations: Depends on having a list of seed misinformation domains to start the crawl and on the availability of social media data for link relationships. It may not identify entirely new misinformation sites that lack connections to known ones, and it relies on Twitter-derived signals which may not be comprehensive.
Domain registration and geolocation signals [24]	A domain-level classification approach using website registration and hosting attributes. It looks at WHOIS (domain registration lookup service) and Domain Name System (DNS) records (such as registrar and Autonomous System Number (ASN) information), characteristics of domain names (e.g., use of unusual or newly created TLDs (Top-Level Domains)), TLS/SSL certificate details (issuer and validity period), and server geolocation. This method was demonstrated on a dataset of websites from Brazil, using those specific regional examples for evaluation.	Gathering these domain attributes is relatively quick on a per-domain basis, as it mainly involves lookups. However, the approach's scalability beyond the tested 222 sites remains unproven in larger settings.	Benefits: Can potentially flag suspicious domains even before any content is published, by spotting telltale patterns in a site's infrastructure and registration (for example, domains with very short lifespans or privacy-masked ownership). Such technical signals can differentiate generally reliable sources from likely misinformation sources. Limitations: The initial study was limited to a specific country (Brazil) and a small sample, so it is unclear how well the findings generalize elsewhere. Some signals like IP geolocation proved weak as indicators, and unusual hosting configurations might sometimes be used by legitimate sites, leading to false positives.

**Table 4** (Continued)

Approach / Method	Scope and Required Data	Operational Considerations	Benefits and Limitations
Infrastructure-based site vetting [25]	Domain-level approach that collects a broad array of technical infrastructure features to identify disinformation websites. It compiles WHOIS and domain registrar details, the use of uncommon top-level domains, domain age and privacy settings, DNS resolution info, TLS certificate attributes, hosting autonomous system (AS) and country, as well as website technologies like WordPress themes or plugins. The model uses known disinformation site lists for training labels and draws on various services (DomainTools, CertStream, crtsh, the Internet Archive, and social platforms like Twitter/Reddit) to discover and profile new domains.	Integrating data from multiple sources and tracking new domain appearances is complex and requires constant monitoring. Nonetheless, this content-agnostic approach can be run with moderate resources since it relies on querying external databases rather than processing site content directly.	Benefits: Uses indicators that are available even before a site gains an audience, allowing early flagging of potentially problematic domains. In a pilot study, this method uncovered several disinformation sites that were not yet listed by experts, demonstrating its ability to detect previously unreported domains. Limitations: The predictive patterns (such as certain registration or hosting traits) may evolve over time, necessitating continuous updates to what the model considers suspicious. Technical signals alone cannot determine the nature of misinformation on a site, and savvy adversaries might mimic benign characteristics to avoid detection.
Content-agnostic web profiling (FNDaaS) [26]	A domain-level detection technique that monitors web infrastructure behavior and changes over time, without analyzing page content. It gathers telemetry such as page load performance and resource usage, Document Object Model (DOM) structure and element statistics, historical DNS records and registration changes, and hosting/Internet Protocol (IP) address shifts. Data is sourced from specialized services (e.g., HostlerStats, Whoisology, SpyOnWeb, ViewDNS) to compile a technical fingerprint of each site.	Requires access to historical DNS and hosting data, as well as the capability to crawl pages for performance metrics. Resource demands are moderate: the method is scalable since it bypasses content analysis, focusing instead on automated scans of technical attributes across many sites.	Benefits: Has broad applicability because it does not depend on language or content, focusing on common technical patterns. It can detect signals like very rapid site setup, reuse of templated site designs, or abnormal load behaviors often observed in fake news sites. Limitations: May produce false positives by flagging benign sites that have uncommon technical setups. Conversely, misinformation sites that employ normal-looking infrastructure may evade detection. The approach benefits from temporal data to catch changes over time, so completely new sites with no history can be difficult to assess.

**Table 4** (Continued)

Approach / Method	Scope and Required Data	Operational Considerations	Benefits and Limitations
Lifecycle and traffic pattern analysis [27]	Domain-level analysis that examines how long sites persist and how they attract users over time. It uses indicators such as domain age, patterns of uptime and downtime (site appearance or disappearance), creation and expiration dates, as well as user traffic volumes, engagement metrics, and the presence of trackers or advertisements.	Requires assembling historical records of site availability and traffic from multiple sources. Analyzing long-term traffic trends can be computationally intensive. In tests, the approach was applied to a few hundred sites (around 283), indicating moderate scalability but not yet proven at a larger scale.	Benefits: Identifies distinctive behaviors of fake news sites, such as short operational lifespans or sudden coordinated surges in traffic, which are strong signals of inauthentic operations. Limitations: This method is inherently data-intensive and retrospective, relying on past observation of site behavior. It may not be well-suited for real-time detection, since it focuses on historical patterns and typically requires a data collection period before conclusions can be drawn.
Browsing traffic graph approach [28]	A domain-level approach that constructs a graph of user navigation flows to and from websites, based on large-scale browsing data. It requires access to massive, real-time logs of user traffic (clickstreams), which are often proprietary and raise privacy concerns. This graph-based model essentially maps how users move through the web and identifies domains that are central in the misinformation network.	Building and updating a comprehensive web navigation graph is computationally demanding. If such data were available, the approach could in principle cover the entire web, but maintaining a real-time graph of global browsing activity is a significant technical challenge.	Benefits: Achieves high precision in filtering out misinformation sites by using a content-agnostic, language-independent signal derived from aggregate user behavior. Graph-based filtering leverages the structure of web traffic to pinpoint suspect domains effectively. Limitations: Access to the necessary browsing data is the barrier, such data are usually proprietary and not publicly available. Even with access, there are privacy and legal considerations in using detailed user traffic logs.
Proposed solution	Domain-level prediction model introduced by the authors, which relies on link analysis metrics instead of content or social media features. Key inputs include web link-based authority measures such as Open PageRank, as well as domain reputation indicators like Moz's Domain Authority and Spam Score. All required data can be obtained through public or low-cost APIs, and the model is trained on domains with expert-provided credibility scores.	The method runs with low computational resources since it primarily involves querying APIs for each domain. Inference (scoring new domains) is fast, and the approach can be scaled to evaluate tens of thousands of domains.	Benefits: Simple and efficient; can quickly produce credibility estimates for a large number of sites without needing any site content or social data. This content-agnostic technique is easy to maintain and update. Limitations: Offers only moderate predictive power, as it captures a limited aspect of credibility (link-based popularity and network patterns). It may fail to detect quality issues that do not manifest in link structure, and thus should be complemented by other indicators for a comprehensive assessment.

ing workflow can be applied using link data from other providers or comparable metrics computed from open web crawls such as Common Crawl [39].

Our model achieved an  $R^2$  of 0.46, meaning it explains about 46% of the variation in website trustworthiness. Although this  $R^2$  value might seem low, it is actually reasonable. Trustworthiness is inherently difficult to measure precisely, as it is a psychometric construct involving many factors that cannot be directly quantified [40]. Accordingly, domain-level metrics such as PageRank or Domain Authority can only approximate one component of trustworthiness. Indeed, factors such as information accuracy, writing quality, and user perceptions influence trustworthiness, but they are not directly captured by the link-based metrics used here. Other researchers facing similar challenges have utilized richer datasets, although doing so typically involves trade-offs in scalability. For instance, K  kol et al. introduced the Content Credibility Corpus and developed a regression model trained on over 15,000 crowdsourced page-level evaluations. Their approach, however, depends primarily on human judgments and textual features, which must be gathered for each page. While highly valuable, this method presents practical challenges for continuously evaluating large numbers of new or evolving domains [41].

Furthermore, an  $R^2$  value around 0.4 is typical in studies involving human judgment or content evaluation. In fact,  $R^2$  values between 0.1 and 0.5 are often considered acceptable in such contexts [35]. Comparable studies support this point. For example, McKnight and Kacmar were able to explain only about 38–44% of the variance in how users judge website credibility, despite using detailed quality and reputation measures [42]. Thus, our model's 46% explanatory power aligns well with findings from previous research. However, these results also indicate that adding more informative features, such as content characteristics, linguistic patterns, or user engagement metrics, could lead to even better predictive accuracy. Future research could build upon these findings to further refine predictive models, thus contributing to more effective strategies for detecting misinformation.

Beyond these initial results, a key advantage of our solution is that it relies on a website's current state each time we evaluate its credibility. Since websites can change ownership, editorial policies, or journalistic standards over time, static lists that are updated only occasionally can quickly become outdated. Our model uses the latest signals (such as inbound links) each time it evaluates a domain's credibility. This approach helps address the limitations of expert-based lists, which may not cover a sufficient number of domains and can lose relevance over time. This dynamic design is especially important given how rapidly misinformation sites can change or reappear to evade detection. Although our system uses current link metrics when scoring a domain, the model is trained on a fixed snapshot of expert ratings and feature data. As domains and their link profiles evolve, the relationship learned during training may change; therefore, the model may need periodic retraining on updated data to maintain the performance reported here.

Our model is trained to predict the trustworthiness scores provided by Lin et al. [7], which aggregate expert ratings from several assessment sources. For this reason, the system should be seen as an amplifier of expert judgment rather than a replacement for it. It extends these expert evaluations to domains that have not been rated yet by learning how trustworthiness relates to domain-level signals. To keep the model accurate as the web changes, future updates will still depend on the continued availability of high-quality human-labeled data for retraining and recalibration.

However, it is important to acknowledge the limitations of a domain-centered strategy. While low-quality sources account for a large portion of false news, this does not mean that every piece of content from an unreliable source is false, nor does it mean that reputable websites never make mistakes. Source-level labels alone do not guarantee accuracy at the article level; even outlets generally considered reliable may publish inaccurate stories. Our model provides a warning signal, rather than a final judgment, on individual articles. Often, the preferred approach is a layered process that combines domain-level and content-level checks, using different signals to form a more complete picture of credibility [43]. From a practical standpoint, automatically scoring domains helps researchers and fact-checkers prioritize their efforts. For instance, they can focus on high-risk websites before examining each individual story in detail. Social media platforms and recommendation systems can also use these trust signals to label content from untrustworthy sources or reduce its visibility. Such actions could limit the spread of misinformation. Prior research suggests that highlighting or downgrading content from questionable domains could reduce user engagement with false information [13].

In summary, our domain-level scoring method is a useful tool for fact-checking and combating online misinformation. By scoring a site's credibility, we can identify likely misinformation at its source, thereby reducing the burden on claim-level verification. This proactive approach can be part of a broader multi-layered strategy to maintain information quality. Such a system is best used alongside human expertise and specialized content-checking tools [44]. Because the misinformation landscape is always evolving, scalable solutions like ours will remain essential for keeping online information healthy and accurate. We hypothesize that domain-based analysis, backed by up-to-date data, will become increasingly important for maintaining trust and accuracy in online spaces.

#### Abbreviations

API, Application Programming Interface; AS, Autonomous System; ASN, Autonomous System Number; AUC, Area Under the Curve; COVID-19, Coronavirus Disease 2019; DNS, Domain Name System; DOM, Document Object Model; FNDaaS, Fake News Detection-as-a-Service; HTML, Hypertext Markup Language; HTTP, Hypertext Transfer Protocol; IP, Internet Protocol; LightGBM, Light Gradient Boosting Machine; MAE, Mean Absolute Error; MBFC, Media Bias/Fact Check; NCS, Non-Credibility Score;  $R^2$ , Coefficient of Determination; RMSE, Root Mean Squared Error; ROC, Receiver Operating Characteristic; SHAP, SHapley Additive exPlanations; SSL, Secure Sockets Layer; TLD, Top-Level Domain; TLS, Transport Layer Security; URL, Uniform Resource Locator; WHOIS, Domain registration lookup protocol/service.

#### Supplementary information

**Supplementary information** accompanies this paper at <https://doi.org/10.1140/epjds/s13688-026-00628-3>.

**Additional file 1.** (DOCX 28 kB)

#### Acknowledgements

Not applicable.

#### Author contributions

Kaveh Kadkhoda Mohammadmosaferi (K.K.M.) supervised the study, managed the project, and contributed to conceptualization, methodology, software development, data curation, investigation, formal analysis, validation, and visualization. K.K.M. also wrote the original draft and participated in reviewing and editing. Anna Bertani contributed to conceptualization, validation, visualization, and writing (review and editing). Thomas Louf contributed to conceptualization, validation, visualization, and writing (review and editing). Riccardo Gallotti supervised the study, managed the project, and participated in writing (review and editing). All authors reviewed and approved the final manuscript.

#### Funding information

This study was funded by the European Union's Horizon Europe research and innovation program under grant agreement 101070190. R.G. acknowledges the support of the PNRR ICSC National Research Centre for High Performance Computing, Big Data and Quantum Computing (CN00000013), under the NRRP MUR program funded by NextGenerationEU.



**Data availability**

Data availability: The aggregated domain rating dataset from Lin et al. [7] is publicly available via the Open Science Framework (OSF) repository (<https://doi.org/10.17605/osf.io/9jwzs>). Additional feature data are available from the corresponding author upon request. Due to licensing restrictions (specifically Moz domain metrics), the full compiled dataset is available only upon reasonable request.

**Clinical trial number**

Not applicable.

**Code availability**

<https://github.com/KavehKadkhoda/domain-quality-evaluation/tree/main>

**Declarations****Ethics approval and consent to participate**

Not applicable.

**Consent for publication**

Not applicable.

**Competing interests**

The authors declare no competing interests.

**Author details**

<sup>1</sup>Fondazione Bruno Kessler, Trento, Italy. <sup>2</sup>University of Dundee, Dundee, UK. <sup>3</sup>University of Trento, Trento, Italy. <sup>4</sup>Centre for Sociology of Humans and Machines (SOHAM), Trinity College Dublin, 3-5 Foster Place, D02 YT92, Dublin, Ireland. <sup>5</sup>Universidad Carlos III de Madrid, Departamento de Matemáticas, Grupo Interdisciplinar de Sistemas Complejos, Leganés, Spain.

Received: 21 July 2025 Accepted: 4 February 2026 Published online: 09 March 2026

**References**

1. World Economic Forum (2025) Global risks report 2025. <https://web.archive.org/web/20250512234121/https://www.weforum.org/publications/global-risks-report-2025/>. Accessed 13 May 2025
2. NewsGuard (2025) Rating process and criteria. <https://web.archive.org/web/20250221152850/https://www.newsguardtech.com/ratings/rating-process-criteria/>. Accessed 21 Feb 2025
3. Media Bias/Fact Check (2025) Methodology. <https://web.archive.org/web/20250221152759/https://mediabiasfactcheck.com/methodology/>. Accessed 21 Feb 2025
4. Caldarelli G, De Nicola R, Petrocchi M, et al (2021) Flow of online misinformation during the peak of the COVID-19 pandemic in Italy. *EPJ Data Sci* 10:34. <https://doi.org/10.1140/epjds/s13688-021-00289-4>
5. Cinelli M, De Francisci Morales G, Galeazzi A, et al (2021) The echo chamber effect on social media. *Proc Natl Acad Sci USA* 118:e2023301118. <https://doi.org/10.1073/pnas.2023301118>
6. Gallotti R, Valle F, Castaldo N, et al (2020) Assessing the risks of 'infodemics' in response to COVID-19 epidemics. *Nat Hum Behav* 4:1285–1293. <https://doi.org/10.1038/s41562-020-00994-6>
7. Lin H, Lasser J, Lewandowsky S, et al (2023) High level of correspondence across different news domain quality rating sets. *PNAS Nexus* 2: pgad286. <https://doi.org/10.1093/pnasnexus/pgad286>
8. Lasser J, Aroyehun ST, Simchon A, et al (2022) Social media sharing of low-quality news sources by political elites. *PNAS Nexus* 1: pgac186. <https://doi.org/10.1093/pnasnexus/pgac186>
9. Greene KT, Pisharody N, Meyer LA, et al (2024) Current engagement with unreliable sites from web search driven by navigational search. *Sci Adv* 10: eadn3750. <https://doi.org/10.1126/sciadv.adn3750>
10. Baqir A, Galeazzi A, Zollo F (2024) News and misinformation consumption: a temporal comparison across European countries. *PLoS ONE* 19(5):e0302473. <https://doi.org/10.1371/journal.pone.0302473>
11. Bertani A, Mazzeo V, Gallotti R (2024) Decoding the news media diet of disinformation spreaders. *Entropy* 26(3):270. <https://doi.org/10.3390/e26030270>
12. Ad Fontes Media (2025) Methodology. <https://web.archive.org/web/20250207065513/https://adfontesmedia.com/methodology/>. Accessed 7 Feb 2025
13. Pennycook G, Rand DG (2019) Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proc Natl Acad Sci USA* 116:2521–2526. <https://doi.org/10.1073/pnas.1806781116>
14. Iffy Index (2025) Iffy News. <https://web.archive.org/web/20250213052109/https://iffy.news/index/>. Accessed 13 Feb 2025
15. DomCop (2025) What is Open PageRank? <https://web.archive.org/web/20250126101929/https://www.domcop.com/openpagerank/what-is-openpagerank/>. Accessed 26 Jan 2025
16. Shu K, Sliva A, Wang S, et al (2017) Fake news detection on social media: a data mining perspective. *ACM SIGKDD Explor Newsl* 19(1):22–36. <https://doi.org/10.1145/3137597.3137600>
17. Rashkin H, Choi E, Jang JY, et al (2017) Truth of varying shades: analyzing language in fake news and political fact-checking. In: *Proceedings of EMNLP 2017*, pp 2931–2937. <https://doi.org/10.18653/v1/D17-1317>
18. Dong XL, Gabrilovich E, Murphy K, et al (2015) Knowledge-based trust: estimating the trustworthiness of web sources. *Proc VLDB Endow* 8(9):938–949. <https://doi.org/10.14778/2777598.2777603>
19. Baly R, Karadzhov G, Alexandrov D, et al (2018) Predicting factuality of reporting and bias of news media sources. In: *Proceedings of EMNLP 2018*, pp 3528–3539. <https://doi.org/10.18653/v1/D18-1389>

20. Panayotov P, Shukla U, Sencar HT, et al (2022) GREENER: graph neural networks for news media profiling. In: Proceedings of EMNLP 2022, pp 7470–7480. <https://doi.org/10.18653/v1/2022.emnlp-main.506>
21. Lepird CS, Ng LHX, Carley KM (2024) Non-credibility scores: relative ranking of news sites shared on social media to identify new “pink slime” sites. First Monday 29(9). <https://doi.org/10.5210/fm.v29i9.13544>
22. Chen Z, Freire J (2020) Proactive discovery of fake news domains from real-time social media feeds. In: Companion proceedings of the web conference 2020, pp 584–592. <https://doi.org/10.1145/3366424.3385772>
23. Sehgal V, Peshin A, Afroz S, et al (2021) Mutual hyperlinking among misinformation peddlers. [arXiv:2104.11694](https://arxiv.org/abs/2104.11694)
24. de Mendonça MPC, Moraes IM, Mattos DMF (2025) Automatic inference of Brazilian websites’ reliability for combating fake news: domain and geolocation features. J Internet Serv Appl 16(1):e5035. <https://doi.org/10.5753/jisa.2025.5035>
25. Hounsel A, Holland J, Kaiser B, et al (2020) Identifying disinformation websites using infrastructure features. In: Proceedings of the 10th USENIX workshop on free and open communications on the Internet (FOCI 2020). <https://doi.org/10.48550/arXiv.2003.07684>
26. Papadopoulos P, Spithouris D, Markatos EP, et al (2023) FNDaaS: content-agnostic detection of websites distributing fake news. In: Proceedings of the 2023 IEEE international conference on big data (big data 2023), pp 1438–1449. <https://doi.org/10.1109/BigData59044.2023.10386830>
27. Chalkiadakis M, Kornilakis A, Papadopoulos P, et al (2021) The rise and fall of fake news sites: a traffic analysis. In: Proceedings of the 13th ACM web science conference (WebSci 2021), pp 168–177. <https://doi.org/10.1145/3447535.3462510>
28. Pereira M, Greene K, Pisharody N, et al (2023) Navigating the web of misinformation: a framework for misinformation domain detection using browser traffic. [arXiv:2307.13180](https://arxiv.org/abs/2307.13180)
29. Moz (2025) Moz API. <https://web.archive.org/web/20250217184919/https://moz.com/products/api>. Accessed 17 Feb 2025
30. Ke G, Meng Q, Finley T, et al (2017) LightGBM: a highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst 30:3149–3157
31. Akiba T, Sano S, Yanase T, et al (2019) Optuna: a next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, pp 2623–2631. <https://doi.org/10.1145/3292500.3330701>
32. SHAP (2025) SHAP documentation. <https://web.archive.org/web/20250505062925/https://shap.readthedocs.io/en/latest/index.html>. Accessed 5 May 2025
33. Lundberg SM, Lee S-I (2017) A unified approach to interpreting model predictions. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in neural information processing systems. NeurIPS 2017, vol 30, pp 4765–4774
34. Ponce-Bobadilla AV, Schmitt V, Maier CS, et al (2024) Practical guide to SHAP analysis: explaining supervised machine learning model predictions in drug development. Clin Transl Sci 17(11):e70056. <https://doi.org/10.1111/cts.70056>
35. Ozili PK (2023) The acceptable R-square in empirical modelling for social science research. In: Saliya CA (ed) Social research methodology and publishing results: a guide to non-native English speakers. IGI Global, Hershey, pp 134–143. <https://doi.org/10.4018/978-1-6684-6859-3.ch009>
36. Modzelewski A, Sosnowski W, Labruna T, Wierzbicki A, Da San Martino G (2025) PCoT: persuasion-augmented chain of thought for detecting fake news and social media disinformation. [arXiv:2506.06842](https://arxiv.org/abs/2506.06842)
37. Leite JA, Razuwayevskaya O, Bontcheva K, Scarton C (2024) EUvsDisinfo: a dataset for multilingual detection of pro-Kremlin disinformation in news articles. In: Proceedings of the 33rd ACM international conference on information and knowledge management (CIKM’24), Boise, Idaho, USA, 21–25 October 2024 Association for Computing Machinery, New York, pp 5380–5384. <https://doi.org/10.1145/3627673.3679167>
38. Bondielli A, Marcelloni F (2019) A survey on fake news and rumour detection techniques. Inf Sci 497:38–55. <https://doi.org/10.1016/j.ins.2019.05.035>
39. Common Crawl (2026) Web graphs. <https://web.archive.org/web/20260126140957/https://commoncrawl.org/web-graphs>. Accessed 28 Jan 2026
40. Singal H, Kohli S (2016) Trust necessitated through metrics: estimating the trustworthiness of websites. Proc Comput Sci 85:133–140. <https://doi.org/10.1016/j.procs.2016.05.199>
41. Kąkol M, Nielek R, Wierzbicki A (2017) Understanding and predicting web content credibility using the content credibility corpus. Inf Process Manag 53(5):1043–1061. <https://doi.org/10.1016/j.ipm.2017.04.003>
42. McKnight DH, Kacmar CJ (2007) Factors and effects of information credibility. In: Proceedings of the 9th international conference on electronic commerce (ICEC’07), pp 423–432
43. Aïmeur E, Amri S, Brassard G (2023) Fake news, disinformation and misinformation in social media: a review. Soc Netw Anal Min 13:30. <https://doi.org/10.1007/s13278-023-01028-5>
44. Guo Z, Schlichtkrull M, Vlachos A (2022) A survey on automated fact-checking. Trans Assoc Comput Linguist 10:178–206. [https://doi.org/10.1162/tacl\\_a\\_00454](https://doi.org/10.1162/tacl_a_00454)

## Publisher’s note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.