Reliability Estimation of News Media Sources: *Birds of a Feather Flock Together*

Sergio Burdisso[⋆,1], **Dairazalia Sánchez-Cortés**[1], **Esaú Villatoro-Tello**[1] and **Petr Motlicek**[1,2]
[1]Idiap Research Institute, Martigny, Switzerland
[2]Brno University of Technology, Brno, Czech Republic
{sergio.burdisso,dairazalia.sanchez-cortes,esau.villatoro,petr.motlicek}@idiap.ch

Abstract

Evaluating the reliability of news sources is a routine task for journalists and organizations committed to acquiring and disseminating accurate information. Recent research has shown that predicting sources' reliability represents an important first-prior step in addressing additional challenges such as fake news detection and fact-checking. In this paper, we introduce a novel approach for source reliability estimation that leverages reinforcement learning strategies for estimating the reliability degree of news sources. Contrary to previous research, our proposed approach models the problem as the estimation of a reliability degree, and not a reliability label, based on how all the news media sources interact with each other on the Web. We validated the effectiveness of our method on a news media reliability dataset that is an order of magnitude larger than comparable existing datasets. Results show that the estimated reliability degrees strongly correlates with journalists-provided scores (Spearman=0.80) and can effectively predict reliability labels (macro-avg. F_1 score=81.05). We release our implementation and dataset, aiming to provide a valuable resource for the NLP community working on information verification.

1 Introduction

As of 2023, the number of internet users is over 5.18 billion worldwide, meaning that around two-thirds of the global population is currently connected to the WWW (Petrosyan, 2023). The Web has democratized and radically changed how people consume and produce information by shifting the paradigm from a news-centred one to a user-centred one. Nowadays, any person on the Web can potentially be a "news medium" providing information either by creating websites, blogs and/or by making use of social media platforms.

Nevertheless, news media can no longer perform its role as "gatekeeper" deciding which stories to disseminate to the public or not (Munger, 2020) since most of the information on the Internet is unregulated by nature (Cuan-Baltazar et al., 2020). As a consequence, an enormous proliferation of misinformation has emerged leaving the public vulnerable to incorrect or misleading information about the state of the world which, among others, increased polarization and decreased trust in institutions and experts (Lewandowsky et al., 2017; Strömbäck et al., 2020). The World Health Organization (WHO) recently declared a worldwide "infodemic" characterized by an overabundance of misinformation (Van Der Linden, 2022). The best-known type of misinformation is *fake news* (Lazer et al., 2018) defined as "false information intentionally created to mislead and/or manipulate a public through the appearance of a news format with an opportunistic structure to attract the reader's attention" (Baptista and Gradim, 2022).

In an attempt to limit the impact of fake news, a large number of initiatives have been undertaken by media, journalists, governments, and international organizations to identify true and false information across the globe (Shaar et al., 2020). For instance, the Duke University's center for journalism research, the Reporters' Lab, lists a total of 419 fact-checking active sites online[1] from which FactCheck.org, Snopes, Full Fact and Politifact are the most well-known. These sites manually and systematically assess the validity of thousands of claims. However, human annotators will always be outnumbered by the claims that need to be verified, reducing the impact of such services in a large-scale scenario. Consequently, we have witnessed a growing interest in using different machine learning models, ranging from non-neural (Kwon et al., 2013; Popat et al., 2016; Nguyen et al., 2018) to deep learning-based ones (Ma et al., 2016; Wang,

[1] https://reporterslab.org/fact-checking/ (Oct. 2023)

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 6900–6918

2017; Popat et al., 2018b; Wang et al., 2018; Fajcik et al., 2023), to determine the validity of claims, news and online information. Nevertheless, these models' performance still has not reached confident accuracy values, limiting their applicability in real-world scenarios (Baly et al., 2018).

A more recent paradigm to fight fake news proposes to focus on the source rather than on the content (Baly et al., 2018, 2020), a task referred to as profiling news medium. The underlying hypothesis states that even though fake news spreads mainly through social media, they still need an initial website hosting the news. Hence, if information about websites is known in advance, identifying potentially fake news can be done by verifying the *reliability* of the source. In fact, this activity is also performed by journalists, who often consult rating services for news websites like NewsGuard[2] or MBFC[3]. Nonetheless, these services are not exhaustive and difficult to keep up-to-date as they rely on human evaluators, highlighting the need for scalable automatic solutions that can be applied in real-world scenarios.

Previous research has shown that predicting reliability is an important first-prior step for fact-checking systems (Nguyen et al., 2018; Popat et al., 2017; Mukherjee and Weikum, 2015) and also the most important aspect that journalists consider when manually verifying the trustworthiness of the information (Baly et al., 2018). Thus, in this paper, we focus on the task of *source reliability estimation*, i.e., automatically analyzing the source that produces a given piece of information and determining its reliability degree. Concretely, we address the posed task by investigating the following research question: *to what extent can we predict the reliability of a news media source solely based on its interactions with other sources?* Contrary to previous research, our proposed method represents a scalable and language-independent approach that can be further enriched via content-based features. Our performed experiments shed light on the immediate (positive) effects of profiling news mediums through its interactions with other sources and also in combination with traditional content-based attributes. Our research holds the potential to uncover deeper insights by incorporating more recent content-based technologies to further explore the nuanced dynamics of news websites, opening the door to a broader NLP research community.

Overall, the main contributions of this paper can be summarized as follows: *(i)* we propose a methodology capable of modeling the *source reliability estimation* problem in a real-world scale scenario that contrary to previous research, estimates the reliability degree (*i.e.* a continuous value) rather than a categorical value and does not depend on any third-party resources; *(ii)* we pioneer the introduction and evaluation of different algorithms to estimate the reliability score, exploring a spectrum from vanilla reinforcement learning strategies to task-specific variations; *(iii)* we build the largest news media reliability dataset available, orders of magnitude larger than existing datasets; *(iv)* we present empirical evidence that establishes the feasibility of predicting the reliability of a news media source solely through its interactions with other sources (which further improves when content-based features are incorporated); and *(v)* we release both the dataset and source code to the wider NLP research community.[4]

2 Related Work

The task of determining information veracity has been approached from different angles and perspectives, from micro to macro, depending on the object of study. For instance, fact-checking focuses on validating a single statement, *i.e.* the claim; fake news detection analyses a whole document, *i.e.* the content of the news article. In this work, we focus on the source that produces a given piece of information, also known as *source reliability estimation*.

Within social media, the sources are individual users creating the content, and previous work has focused on identifying different types of users such as *spammers* (Liubchenko et al., 2022; Stringhini et al., 2010), *bots* (Lei et al., 2023; Knauth, 2019), fake profiles (Roy and Chahar, 2020; Ramalingam and Chinnaiah, 2018), paid users (Mihaylov et al., 2015b), and *trolls* (Tomaiuolo et al., 2020; Miao et al., 2020; Mihaylov et al., 2015a), among others (Sansonetti et al., 2020; Burdisso et al., 2022). However, in the broader case of the WWW, sources are individual websites (Dong et al., 2015), and in our case, news media websites.

Previous studies have tangentially addressed news media source reliability as part of the study of automatic fact-checking systems, either as a *prior*

[2] https://www.newsguardtech.com/
[3] https://mediabiasfactcheck.com
[4] https://github.com/idiap/News-Media-Reliability

in probabilistic graphical models (Nguyen et al., 2018; Mukherjee and Weikum, 2015) or as features for stance classification models (Popat et al., 2018a, 2017, 2016). In these studies, reliability estimation relied on indirect measures since no gold labels were used. For instance, some works use the *AlexaRank*[5] and *PageRank* (Brin and Page, 1998) scores of the websites as proxies for their reliability (Baly et al., 2018; Popat et al., 2016) while others the proportion of articles that refute false claims and support true claims (Popat et al., 2018a, 2017). However, in the latter, authors rely on a fact-checking model and the selected true and false claims while, in the former, on scores that only capture the authority and popularity of the sources, not necessarily their trustworthiness — for instance, think of popular unreliable gossip websites[6] or satirical news websites, like *The Onion*,[7] highly popular, attracting huge web traffic.

Recently, Baly et al. (2020, 2019, 2018) addressed the source reliability estimation task on its own, modeling it as a classification task using source-level gold annotations. In particular, authors focused on predicting websites factual reporting and political bias using the values published by a news media rating service as ground truth. However, their proposed method relies on collecting and extracting information from multiple external and restricted sources (e.g. Twitter, Facebook, YouTube, etc.) for generating content-based, audience-based, and metadata-based features for the classification model, limiting its practical use on a large-scale scenario. In this paper, we also address the task using gold annotations, however, we adopt an easier-to-scale approach. Specifically, we model the problem as estimating a continuous value (i.e. the reliability *degree*) based simply on how all news media sources interact with each other on the World Wide Web.

3 Methodology

3.1 Problem Formulation

Let S be the set of all news media sources on the Web, and $G = \langle S, E, w \rangle$ be the weighted directed graph where there is an edge $(s, s') \in E$ if source s contains articles (hyper) linked to s' and where the weight $w(s, s') \in [0, 1]$ is the proportion of total hyperlinks in s linked to s'. Given two disjoint subsets $\hat{S}^+, \hat{S}^- \subset S$ containing, respectively, some known reliable and unreliable news sources, the goal is to estimate the *reliability degree* $\rho(s)$ *for all* $s \in S$. More precisely, a total function $\rho : S \mapsto \mathbb{R}$ such that:

1. $\rho(s) > 0$ if s is reliable
2. $\rho(s) \leq 0$ if s is unreliable
3. $\rho(s) < \rho(s')$ if s' is more reliable than s.

The underlying intuition behind using hyperlinks to build the graph is that the more frequently one source links to another (i.e., the higher $w(s, s')$), the higher the chances of a random reader to (click and) reach the reliable/unreliable source s' from s. Notably, hyperlinks also serve as a proxy for content-based interactions, as they are typically used to cite content from the referred article. Thus, a higher $w(s, s')$ also implies a stronger content-based relationship. Therefore, this simple weighted, hyperlink-based, and source-centered approach potentially captures both interaction types among news sources simultaneously, while being relatively easy to scale.

3.2 Reinforcement Learning Strategy

Our reinforcement learning reliability framework models reliability in terms of a Markov Decision Process (MDP). An MDP is defined by a 4-tuple $\langle \mathbb{S}, A, P_a, r_a \rangle$ where $\mathbb{S}$ is a set of states, A a set of actions, $P_a(s, s')$ is the probability that action a in state s will lead to state s', and $r_a(s, s')$ is the immediate reward perceived after taking action a in state s leading to state s' (Sutton and Barto, 2018; Puterman, 2014; Kaelbling et al., 1996).

Given an MDP, a decision process is then specified by defining a policy π that provides the probability of taking action a in state s. In turn, given a policy π we can estimate the value of each state, $V^\pi(s)$, in terms of how good it is to be in that state following the policy. In particular, the value $V^\pi(s)$ is given by the *Bellman equation* which is defined, for any state $s \in \mathbb{S}$, recursively as:

$$V^\pi(s) = \sum_{s' \in \mathbb{S}} P^\pi(s, s')[r(s') + \gamma V^\pi(s')] \quad (1)$$

where $\gamma \in [0, 1)$ is known as the *discount factor* and $r(s)$ is the immediate reward received when reaching s. Thus, we address the reliability estimation as an MDP $\langle \mathbb{S}, A, P, r \rangle$ such that: (a) The set

[5] https://www.alexa.com/

[6] http://www.ebizmba.com/articles/gossip-websites

[7] https://www.theonion.com/

Algorithm 1 *F-Reliability* strategy for $\rho(s)$.

Set $\forall s \in S, \rho(s) = 0$
repeat
  $\Delta = 0$
  for all $s \in S$ **do**
    $\rho'(s) = \sum_{s' \in S} P(s, s')[r(s') + \gamma\rho(s')]$
    $\Delta = max(\Delta, |\rho'(s) - \rho(s)|)$
  $\rho = \rho'$
until Δ *is small enough*

Algorithm 2 *P-Reliability* strategy for $\rho(s)$.

Set $\forall s \in S, \rho(s) = 0$
repeat
  $\Delta = 0$
  for all $s \in S$ **do**
    $\rho'(s) = r(s) + \gamma \sum_{s' \in S} P(s', s)\rho(s')$
    $\Delta = max(\Delta, |\rho'(s) - \rho(s)|)$
  $\rho = \rho'$
until Δ *is small enough*

of states $\mathbb{S}$ are all the news media websites on the Web —*i.e.* we have $\mathbb{S} = S$; (b) The set of actions A contains only one element, the *"move to a different news media website"* action; (c) The probability P of moving from s to s' will be given by the proportion of hyperlinks in s connecting to s' —*i.e.* we have $P(s, s') = w(s, s')$; and (d) The reward r of moving to a source is determined only by the source itself, and it will be positive or negative for known reliable or unreliable sources respectively —*i.e.* we have $r(s, s') = r(s')$ where $r : S \mapsto \mathbb{R}$ such that $r(s) = 1$ if $s \in \hat{S^+}$; $r(s) = -1$ if $s \in \hat{S^-}$; $r(s) = 0$ otherwise. In simple words, we can think of modeling the problem as if there was a "virtual user" browsing from one news media source to another with probability proportional to how strongly connected they are, and who will perceive a positive or negative signal (the reward) when arriving to known reliable or unreliable sources, respectively. Given this framework, now the challenge is how to estimate the *reliability scores* $\rho(s)$.

3.2.1 Perceived Future Reliability

Under this simple framework, our initial approach involves estimating reliability by "looking to the future". To be more precise, we will assume the reliability degree $\rho(s)$ is proportional to the *expected* perceived reliability (reward) by the virtual user. Consequently, a source is considered more reliable (or unreliable) if it is expected to guide the virtual user to well-known reliable (or unreliable) sources.

To achieve this, we can simply set $\rho(s) = V(s)$, as Equation 1 defines $V(s)$ as the discounted long-term future rewards received following a policy π. Note that, given that we only have one possible action in $\mathbb{A}$, the policy π is trivial and thus $P^\pi(s, s') = P(s, s')$. Therefore, a source s will have a higher (lower) $\rho(s)$ the more positive (negative) its total expected future reward $V(s)$. In other words, it reflects how much s is expected to guide us to known reliable (unreliable) sources, as intended. We will refer to this strategy as *"F-Reliability"*. In practice, the computation of $V(s)$ can be done using the *Value Iteration* algorithm (Sutton and Barto, 2018). Thus, we compute $\rho(s)$ for our specific MDP as shown in Alg. 1.

3.2.2 Accumulated Past Reliability

An alternative approach is to estimate reliability by "looking to the past" rather than the future. Specifically, we assume that the reliability degree $\rho(s)$ is proportional to the accumulated reliability (reward) perceived by the virtual user in reaching the current source s. Consequently, a source becomes more reliable (unreliable) as more known reliable (unreliable) sources lead to it.

To formalize the above intuition, we leverage the *reverse Bellman equation* introduced by Yao and Schuurmans (2013). This equation is recursively defined for any state $s \in \mathbb{S}$ as:

$$R^\pi(s) = r(s) + \gamma \sum_{s' \in \mathbb{S}} P^\pi(s', s) R^\pi(s') \qquad (2)$$

In contrast to Equation 1 that looks forward from a state to define its value, this equation looks backward to define it —note $P^\pi(s, s')$ is swapped to $P^\pi(s', s)$. More precisely, while $V(s)$ defines the value of a state based on the forward accumulated reward, $R(s)$ does so in terms of the historical accumulated reward. Therefore, by setting $\rho(s) = R(s)$, a source s will have a higher (lower) $\rho(s)$ the more positive (negative) is the accumulated reward $R(s)$ —*i.e.* the more known reliable (unreliable) sources lead to s, as intended. We will refer to this strategy as *"P-Reliability"*. In practice, we can again employ *Value Iteration* to compute $\rho(s)$ using $R(s)$ as shown in Algorithm 2.

3.2.3 Past and Future Perceived Reliability

Lastly, we can explore an approach that combines both "the future and the past". Intuitively, we can argue that the transfer of reliability between news

Algorithm 3 *I-Reliability* strategy for $\rho(s)$.

Set $\forall s \in S, \rho(s) = r(s)$
repeat
  for all $s \in S$ **do** ▷ Investment step
    $totalcredits(s) = \sum_{s' \in S} w(s', s)\rho(s')$
  for all $s \in S$ **do** ▷ Credit collection step
    $profit = \sum_{s' \in S} w(s, s')credits_s(s')$
    $\rho(s) = \rho(s) + profit$
until n *times*

media sources and their neighboring sources is asymmetric. Specifically, the impact on the reliability, $\rho(s)$, of a source s when referencing a reliable source is not equivalent to the effect of a reliable source referencing s.[8] Moreover, this asymmetry extends to both reliable and unreliable sources. That is, a reliable source referencing s carries a different weight than an unreliable one referencing s, and vice versa.[9] In a broader sense, we can think of a source s increasing its reliability $\rho(s)$ as more reliable sources link to it, while losing reliability as it links to more unreliable sources.

To formalize this asymmetric behavior, we can incorporate both $R(s)$ and $V(s)$ into our reliability model. More precisely, let $V^-(s)$ be $V(s)$ where only negative rewards $r(s)$ are allowed, and analogously, $R^+(s)$ with only positive rewards, then we can define $\rho(s)$ as:

$$\rho(s) = V^-(s) + R^+(s) \quad (3)$$

As a result, a source s will have a higher reliability $\rho(s)$ the more reliable sources link to it (*i.e.* the higher $R^+(s)$), and lower reliability the more it links to unreliable sources (*i.e.* the lower $V^-(s)$). We will refer to this strategy as *"FP-Reliability"*.

3.3 Reliability Investment Strategy

A well-established algorithm used in the field of *truth discovery* (Li et al., 2016) is the *Investment* algorithm (Pasternack and Roth, 2010). This algorithm is an iterative method in which two interdependent steps are repeated: (1) sources uniformly "invest" their trustworthiness among their claimed values; (2) sources collect credits back from the claimed values which update, in turn, their trustworthiness. Inspired by this "invest and collect" intuition, we now formulate an algorithm based on the same principle. Initially, each source will distribute its reliability $\rho(s)$ among neighboring sources in proportion to the strength of their links, *i.e.* $\propto w(s, s')$. In essence, during the investment step, the total credits invested in each source s is defined as follows:

$$totalcredits(s) = \sum_{s' \in S} w(s', s) \cdot \rho(s') \quad (4)$$

In the subsequent credit collection step, the total credits are distributed among investors, s', in proportion to their contribution to the source s:

$$credits_{s'}(s) = w_{s'}(s) \cdot totalcredits(s) \quad (5)$$

Here, $w_{s'}(s) \in [0, 1]$ represents the proportion of total *inbound* hyperlinks in s originating from s'. The reliability degree is then updated by collecting the credits back in proportion to the invested percentage:

$$\rho(s) = \rho(s) + \sum_{s' \in S} w(s, s') \cdot credits_s(s') \quad (6)$$

Finally, we repeat this process n times to update $\rho(s)$ considering values from up to n-hop-away sources in the graph, as illustrated in Algorithm 3.

CC-News snapshot	#articles	Graph #nodes	#edges
2019/08	17M	6,799	171,810
2020/08	23M	11,427	276,666
2021/08	28M	10,938	315,447
2022/08	35M	10,607	354,386
all above	103M	17,057	909,354

Table 1: CC-News snapshots and the obtained graphs. The last row corresponds to our final graph.

4 Data

4.1 A real-world scale news media graph

The *Common Crawl Foundation*[10] maintains the *Common Crawl News Dataset* (*CC-News*), the world's largest collection of news articles crawled from global news web sites since 2016. The data is updated daily and published as a series of snapshots organized by year and month.

We developed a Python CC-News processing pipeline that takes care of building the news media graph, G, from CC-News snapshots (details in Appendix C). Similar to the *CCNet* pipeline (Wenzek et al., 2020), our pipeline utilizes the language classifier from fastText (Joulin et al., 2017; Grave et al., 2018) to categorize news articles into 176 languages. Consequently, for a given CC-NEWS

[8] *e.g.* it is not the same for your reputation as a news media if The New York Times references you as you referencing it.

[9] *e.g.* The New York Times referencing you has not the same impact on your reputation as a fake news media referencing you, or as you referencing a fake news media.

[10] https://commoncrawl.org

snapshot URL, the pipeline generates one graph for each supported language showing how news sources relate to each other in that language. However, in this paper, we focus exclusively on the English graph due to the predominance of available ground truth data for experimentation in this language. Specifically, for experimentation, we will use the English graph obtained from joining four different *CC-News* snapshots corresponding to August over the past four years (2019 to 2022).[11] As indicated in Table 1, this process resulted in a unified graph containing around 17k English-speaking news media sources and nearly 1M connections —graph shown in Figure 2 (Appendix B).

4.2 Ground truth datasets

To facilitate a comparative analysis with previous studies on the source reliability classification task, we employ the largest dataset published earlier by Baly et al. (2018). This dataset encompasses 1066 annotated news media URL domains extracted from *Media Bias/Fact Check* (MBFC) —refer to the first row of Table 2 for details.[12] Furthermore, for a more comprehensive evaluation, we employ an extended dataset meticulously created by merging ground truth labels from various sources, as outlined below:

• **MBFC:** we followed a similar process as in Baly et al. (2018) but crawling the entire MBFC website to extract 4138 ground truth labels. Following Gruppi et al. (2020), we aggregated these labels into three classes: "reliable" for sources with high or very high factual reporting, "unreliable" for sources flagged as conspiracy, pseudoscience, or with low/very low factual reporting, and "mixed" for sources with mixed factual reporting.

• **Wikipedia's perennial sources:** the platform hosts a list of sources discussed by the community regarding their reliability and use on the platform.[13] We extracted 553 ground truth labels from this list applying the following policy: sources marked as *generally reliable* were labeled as "reliable"; sources marked as *generally unreliable*, *deprecated*, or *blacklisted* were labeled as "unreliable"; and sources marked as *no consensus*, *stale discussions* or *discussion in progress* as "mixed";

[11] By selecting the same month, we ensure a consistent 4-year time span while limiting the processed news articles to approximately 100M.

[12] Original factuality labels were transformed into reliability labels following Gruppi et al. (2020) strategy.

[13] https://en.wikipedia.org/wiki/Wikipedia:Reliable_sources/Perennial_sources

Dataset	Label distribution *unreliable*	*mixed*	*reliable*
Baly et al. (2018)	256	268	542
Our own	1425	1461	2446
MBFC	*546*	*1363*	*2229*
Wikipedia	*298*	*98*	*157*
Fake News	*556*	-	-
NewsGuard	*25*	-	*60*

Table 2: Datasets details. Bottom part shows individual contributions to our final dataset.

• **Fake news:** we manually collected a list of 556 unreliable sources from fake news websites, including the Wikipedia list of fake news websites[14] and a report from the Institute for Strategic Dialogue identifying active and inactive fake news domains (ISD, 2020).

• **NewsGuard:** a paid rating service similar to MBFC, provides both a verdict and a reliability score based on predefined journalistic criteria (details in Appendix D). Due to license limitations, we could only use the 85 ground truth values included in the NELA-GT-2018 dataset (Nørregaard et al., 2019). However, as detailed in Section 5.2, the inclusion of NewsGuard enables us to measure the correlation between the estimated reliability degrees $\rho(s)$ and the scores provided by journalists.

Hence, our final aggregated dataset comprises 5332 news URL domains, each annotated with 3-class reliability labels. As illustrated in Table 2, its scale surpasses that of the largest one to date (Baly et al., 2018), being an order of magnitude larger. For evaluation, it is crucial that the source s is present in the graph, as we want to assess how well the reliability degree $\rho(s)$ is computed from it. Therefore, we limit our experimentation to using the subset of the ground truth dataset corresponding to the nodes within our graph. This subset contains approximately 40% of the total ground truth sources. In particular, 400 sources from Baly et al. (2018) (294 "reliable," 85 "mixed," and 21 "unreliable") and 2117 sources from our own dataset (1630 "reliable," 321 "mixed," and 166 "unreliable"). Additionally, since our goal is to evaluate the ability of $\rho(s)$ to distinguish reliable from unreliable sources (see conditions 1 and 2 in Section 3.1), we merge "unreliable" and "mixed" labels to create the following three experimentation sets:

• **ExpsetA:** 294 *reliable* and 106 *unreliable* sources

[14] https://en.wikipedia.org/wiki/List_of_fake_news_websites

from Baly et al. (2018).

• **ExpsetB:** 1630 *reliable* and 487 *unreliable* sources from our dataset.

• **ExpsetB$^-$:** 1630 *reliable* and 166 *unreliable* sources. A simpler version of *ExpsetB* removing "mixed" from *unreliable* sources.

5 Experiments and Evaluation Results

For experimentation, we define the reward values based on ground truth labels as $r(s) = 1$ if the label is "reliable", $r(s) = -1$ if "unreliable", and $r(s) = 0$ otherwise. In addition, selecting appropriate hyperparameter values is crucial. For *I-Reliability*, n controls how far to look in the neighborhood for investments (how many nodes away). Similarly, in the reinforcement learning strategies, the discount factor γ controls the distance of looking back/forward; $\gamma \approx 0$ focuses mostly on present reward $r(s)$, while $\gamma \approx 1$ considers all history/future to compute $\rho(s)$. We performed a grid search over $n \in [1, 10]$ and $\gamma \in [0.05, 0.95]$ to determine the best hyperparameter values on each of the three experimental sets. The grid search was performed using 5-fold cross validation selecting the n and γ that obtained the best *macro avg.* F_1 on the reliability classification task, as described in Section 5.1. We observed that, independently of the dataset, *better reliability estimation is achieved when looking mostly at nearby sources*, as better performance was obtained with small n ($n \leq 2$) and γ ($\gamma < 0.5$) values —details in Appendix A.

5.1 Reliability Classification Results

In this section, we focus on evaluating the first two conditions for $\rho(s)$ given in Section 3.1. These conditions allow us to measure the ability of $\rho(s)$ to distinguish reliable from unreliable sources. For comparison, we follow the evaluation procedure from Baly et al. (2018) and report results for 5-fold cross-validation. More precisely, in each k-fold iteration, we only use ground truth rewards $r(s)$ from four folds to compute $\rho(s)$ for all 17k sources in the graph, and using conditions 1 and 2, all s in the hold-out fold are classified as *reliable* ($\rho(s) > 0$) or *unreliable* ($\rho(s) \leq 0$).

Table 3 shows the evaluation results obtained on the three experimentation sets along with two naive baselines for reference, random and majority class classifiers.[15] In addition, for *ExpsetA*, we also report the results obtained using the classification models introduced in previous works (Baly et al., 2018, 2020). These classifiers combine multiple content-based, audience-based, and metadata-based features about the sources. Authors released the pre-computed features values for the Baly et al. (2018) dataset and thus, using their source code,[16] we were able to train and evaluate their classifiers on the *ExpsetA* set. However, since building these features relies on multiple external sources (*e.g.* Twitter, Facebook, YouTube, etc.), we were not able to evaluate their method in our new dataset given current API restrictions to access them.

Data	Strategy	F_1 score			Acc.
		macro avg.	*reliable*	*unreliable*	
ExpsetA	M-BL	42.33	84.66	0.00	73.44
	R-BL	48.85	61.76	35.94	52.33
	Baly18	67.87	84.81	50.92	76.95
	Baly20	65.24	82.99	47.50	74.37
	F-R	61.52	87.62	35.42	79.26
	P-R	**72.67**	**90.05**	**55.29**	**83.79**
	FP-R	69.28	89.23	49.34	82.29
	I-R	**72.81**	**90.03**	**55.60**	**83.77**
ExpsetB	M-BL	43.50	87.00	0.00	77.00
	R-BL	47.48	62.17	32.80	51.63
	F-R	61.85	79.72	43.98	70.34
	P-R	**74.69**	**88.29**	**61.10**	**82.00**
	FP-R	55.95	68.08	43.82	59.38
	I-R	**75.51**	**89.30**	**61.72**	**83.28**
ExpsetB$^-$	M-BL	47.58	95.15	0.00	90.76
	R-BL	39.17	63.04	15.31	48.55
	F-R	62.18	90.23	34.12	83.02
	P-R	**78.90**	**95.83**	**61.97**	**92.48**
	FP-R	59.20	84.66	33.74	75.11
	I-R	**81.05**	**96.71**	**65.39**	**93.99**

Table 3: 5-fold cross-validation average results for reliability classification. The best-performing values are **underlined**, while the 2nd-best results appear in **bold** font. R-BL and M-BL refer to random and majority class baselines; Baly18 and Baly20 refer to Baly et al. (2018) and Baly et al. (2020); and *-R stands for *-Reliability.

Observing the performance across the different datasets, all four strategies outperform the random and majority class baselines by a statistically significant difference (paired t-test with $p\text{-}value < 0.02$). In addition, both *P-Reliability* and *I-Reliability* consistently outperform other strategies, including those presented in previously published works ($p\text{-}value < 0.03$), however, the difference between these two strategies is not statistically significant

[15] For a comprehensive view of additional metrics, such as precision, recall, and confidence intervals, refer to Table 7 in the Appendix.

[16] github.com/ramybaly/News-Media-Reliability.

Strategy	F_1 score			
	macro avg.	*reliable*	*unreliable*	**Acc.**
P-Reliability	*72.67*	**<u>90.05</u>**	*55.29*	***<u>83.79</u>***
+Baly18	**77.11**	87.75	**66.47**	82.11
+Baly20	74.36	86.02	62.70	79.69
I-Reliability	*72.81*	**90.03**	*55.60*	***83.77***
+Baly18	**<u>77.47</u>**	87.89	**<u>67.06</u>**	82.34
+Baly20	72.88	85.46	60.30	78.74

Table 4: Ensemble results for *P-Reliability* and *I-Reliability* strategies on *ExpsetA*. The best performance results are **<u>underlined</u>**, while the 2nd-best appear in **bold** font.

Rank	Domain	Score	$\hat{\rho}(s)$
1	bbc.co.uk	100.0	0.995
2	cnbc.com	95.0	0.995
3	dailysignal.com	92.5	0.830
4	thinkprogress.org	90.0	0.907
5	independent.co.uk	87.5	0.968
1	sputniknews.com	7.5	-0.992
2	truepundit.com	12.5	-0.957
3	dailymail.co.uk	15.0	-0.998
4	theduran.com	17.5	-0.954
5	thegatewaypundit.com	20.0	-0.994

Table 5: NewsGuard top-5 unique most scored (top part) and least scored sources (bottom part) along with the estimated $\rho(s)$ given by *P-Reliability*.

(p-*value* > 0.5). From the results we can also see that, regardless of the chosen strategy and dataset, the F_1 score for the *unreliable* class consistently remains lower when compared to the *reliable* class. This suggests that identifying unreliable sources is more challenging, likely due to the dataset imbalance favoring reliable sources —note that in both *ExpsetA* and *ExpsetB*, only approximately 25% of the sources are unreliable. This imbalance results in models having fewer negative signals (*i.e.* rewards $r(s) = -1$) to learn to identify unreliable sources effectively. Another contributing factor is the inclusion of "mixed" labels in the *unreliable* group, making the task more challenging by incorporating unreliable sources whose reliability is not clearly defined. This hypothesis is supported by the results from *ExpsetB*$^-$, where the removal of "mixed" labels results in improvements across all metrics —note that the highest *unreliable* F_1 score is achieved while the dataset is significantly more unbalanced (ten times fewer unreliable sources than reliable ones). Concerning the reinforcement learning strategies, *P-Reliability* significantly (p-*value* ≤ 0.02) outperformed *F-Reliability* suggesting that the reliability of a source *is more significantly influenced by its origins than by the destinations it reaches*. On the other hand, *FP-Reliability* shows poor performance, mainly due to the different nature of $V^-(s)$ and $R^+(s)$ in Equation 3 —note that $V(s)$ is defined as an expectation, whereas $R(s)$ is not.[17]

Finally, to assess the complementarity of our graph-based strategies with content-based ones, we performed an additional experiment: a simple voting ensemble between our strategies and Baly's models. Specifically, sources were classified as reliable only when both models agreed on the classification. Table 4 presents the obtained results for our two best-performing models, *P-Reliability* and *I-Reliability*, on *ExpsetA*.[18] The ensemble approach enabled the models to further improve their performance, particularly in detecting unreliable sources, achieving the highest *macro avg.* F_1 score on this dataset (77.47).

5.2 Correlation with human judgment

In this section, we focus on evaluating the final condition in the definition of $\rho(s)$ in Section 3.1. This condition enables $\rho(s)$ to assess the reliability of s *relative* to other sources, allowing the ranking of sources based on their reliability degrees. For evaluation, we use the *NewsGuard* dataset introduced in Section 4.2 containing the 85 ground truth reliability scores provided by trained journalists. The score ranges from 0 to 100 and is obtained by answering 9 questions that address different journalistic criteria —details in Appendix D. Table 5 shows examples of NewsGuard scores and their estimated reliability degree.[19]

We measure the correlation between these human-provided scores and their estimated reliability degree by computing the *Pearson correlation coefficient* (PCC) and the *Spearman's rank correlation coefficient* (SRCC). PCC measures the linear relationship between two variables, whereas SRCC assesses the monotonic relationship, making

[17] Transition probabilities are not normalized in both directions, only in the forward direction. Consequently, there is no inherent mathematical symmetry that ensures $V^-(s)$ and $R^+(s)$ will equilibrate to zero ($\rho(s) = 0$) when they correspond to an equivalent number of unreliable and reliable sources

[18] Full results included in Table 7 (Appendix).

[19] For ease of comparison, in this table, $\rho(s)$ is normalized in the range $[-1, 1]$ by dividing it by the maximum (when $\rho(s) \geq 0$) and minimum value (when $\rho(s) < 0$).

Strategy	PCC	*p-value*	SRCC	*p-value*
Random baseline	0.058	*0.6*	0.066	*0.6*
PageRank baseline	0.313	*0.008*	0.544	*1e-06*
F-Reliability$^{\diamondsuit}$	0.556	*5e-07*	0.295	*1e-02*
P-Reliability$^{\diamondsuit}$	**0.647**	*1e-09*	0.668	*2e-10*
FP-Reliability$^{\diamondsuit}$	0.636	*3e-09*	**0.677**	*3e-09*
I-Reliability$^{\diamondsuit}$	0.589	*7e-08*	0.657	*5e-10*
F-Reliability$^{\clubsuit}$	0.927	*1e-30*	0.544	*1e-06*
P-Reliability$^{\clubsuit}$	0.912	*9e-33*	**0.801**	*6e-17*
FP-Reliability$^{\clubsuit}$	**0.929**	*8e-32*	0.775	*7e-15*
I-Reliability$^{\clubsuit}$	0.757	*2e-19*	0.792	*8e-12*

Table 6: Correlation between $\rho(s)$ and journalist-provided reliability scores. ♣: w/ rewards; ♢: w/o rewards.

it *more suitable for evaluating the relative reliability* of sources, as it captures ranked associations regardless of the exact numerical values (condition 3 for ρ). In particular, we perform the evaluation under two scenarios, when sources are known to be (un)reliable and, more challenging, when their reliability is not known in advance.[20] In other words, we perform the evaluation following, respectively, two experimental settings: (♣) we use all the ground truth rewards from the largest experimental set, *ExpsetB*, to learn the reliability degree $\rho(s)$ of all 17k sources in the graph; and (♢) we repeat the same process but removing the rewards for all the 85 domains used for evaluation.

Table 6 shows the obtained results along with two baselines for reference, random and *PageRank* algorithm (Brin and Page, 1998),[21] scatter plots in Appendix E. We can observe that, as expected, correlations are weaker under the hardest scenario without rewards, specially in terms of PCC which is more sensitive to the bias introduced by the rewards.[22] Nevertheless, in both scenarios, obtained correlation coefficients are statistically significant (*p-value* $\leq$ 5e-07 for PCC, *p-value* $\leq$ 1e-06 for SRCC) and higher than the baselines, except for *F-Reliability*. In general, the strategies that correlate more strongly with the journalist-provided scores are *P-Reliability* and *FP-Reliability*, showing both a strong linear (PCC) and ranking-based (SRCC) relationship independently of whether rewards were used or not.[23] *FP-Reliability* results suggest that combining *F-Reliability* and *P-Reliability* strategies could be advantageous for estimating *relative* reliability.[24] Overall, results are inspiring considering that the learning process for all $\rho(s)$ values in the graph leverages only a subset of binary ground truth rewards ($r(s) = -1$ or 1), *without any explicit notion of ground truth score or degree*. In contrast to reliability scores derived from various qualitative journalistic criteria, $\rho(s)$ approximates the reliability degree solely based on the propagation of these initial rewards throughout the network's structure.[25]

6 Conclusion and Future Work

In this study, we introduced an approach for assessing the reliability of news media through their network interactions. This approach diverges from previous models that depend on content, audience feedback, and/or metadata. Moreover, unlike in previous works, our method estimates a *reliability degree* rather than a reliability label. We assessed the quality of the estimated values in terms of reliability classification and correlation with journalists-provided scores. We found that a source's origins is more indicative of its reliability than its reach and show evidence that it is feasible to predict the reliability of news media using only their network interactions, providing an easier-to-scale approach than prior methods. As future work, we plan to expand the study by building a larger graph and designing more sophisticated strategies that leverage content-based features. Additionally, we aim to explore the estimation of other news source properties, such as political bias, using the same approach. Finally, we intend to investigate the use of the estimated reliability values in downstream tasks like fact-checking and fake news detection.

[20] For instance, is $\rho(cnbc.com) < \rho(bbc.co.uk)$? that is, is "bbc.co.uk" more reliable than the "cnbc.com"? knowing in advance that both are reliable ($r(cnbc.com) = r(bbc.co.uk) = 1$) vs. not knowing it ($r(cnbc.com) = r(bbc.co.uk) = 0$).

[21] Note that PageRank is unsuitable for classification experiments as its non-negative scores always predicted the positive class, reducing it to a majority-class classifier. Hence, its exclusion from Table 3.

[22] Note that sources with $r(s) = 1$ will naturally tend to have a final $\rho(s)$ close to 1 while sources with $r(s) = -1$ close to -1, this bias is heavily reduced when instead of using the actual $\rho(s)$ value we use its ranking (as in SRCC).

[23] In fact, we performed an additional experiment in which we set $\rho(s) = \frac{\rho_p(s)+\rho_{fp}(s)}{2}$, i.e., the reliability degree was defined as the average of the values obtained by the best performing strategies, *P-Reliability* and *FP-Reliability*, obtaining the strongest correlation values (PCC=0.933, SRCC=0.803 and PCC=0.715, SRCC=0.697 with and without rewards, respectively).

[24] In contrast to reliability classification, where $\rho(s) = 0$ is the fixed threshold separating reliable from unreliable sources.

[25] We are releasing the list of estimations for all 17k sources along with this paper.

7 Ethical Considerations

The work presented in this paper has been done in the scope of the CRiTERIA project[26] that follows the H2020 ethical standards and guidelines. The Consortium Agreement includes the partners' commitment to FAIR (findable, accessible, interoperable and re-usable) data management practices and responsible research practices. The framework of the research questions and preliminary results were reviewed by the Project Ethics Check and Audit committee in the form of an on-going work deliverable.

In the present research paper there is no gender bias to be investigated or addressed. The data comes from hyperlinks and URL domain names that can not be associated to a gender. As described in Section 4, there is no intentional collection of personal data in any form, and as a consequence, there is no need for data anonymisation or pseudonymisation. Similarly, there is no need of informed and singed consents since there is no direct human participation in the construction of the graphs to calculate the reliability values.

Regarding data and processing security, a snapshot of CC-News data is transferred, held in the local servers and processed for the reliability estimation. The data can be deleted at any time and easily downloaded from the original public CC-News sources (as described in Section 4). Any further data processing carried out, is going to be publicly available, along with the reference to this paper.

Regarding other sources, only data collected by other sources (under Apache or open source licenses) with well described data collection methodology, validated with published results and that is publicly available was used. The overview of the datasets is described in Section 4.2, the description includes the annotations distribution in Table 2, and the original sources are properly acknowledge along the paper.

The estimated reliability values were obtained based on initial ground truth labels. This labels capture mainly the factuality of reporting and do not consider other aspects like, for instance, political bias or press freedom rating. Therefore, computed reliability values should not be considered as *de facto* values.

From a societal perspective, this paper brings a positive impact, improving the situational awareness of decision makers, including fact-checkers. The mathematically defined algorithms are robust to content-related biases since they are both language and content independent (political, religious, racial, etc.).

[26] https://www.project-criteria.eu/

8 Limitations

8.1 Methodology

The main constraint of the proposed methodology lies in the requirement for news media sources to be included in the graph for their reliability to be calculated. This limitation may arise due to temporal or size constraints in the data used to construct the graph. For instance, a recently emerged news source might not be referenced by others until some time has passed. To address this limitation to a certain degree, assigning $\rho(s) = 0$ to such sources can be considered. This implies that their reliability is indeterminate, indicating an unknown or undetermined status, meaning they are neither reliable nor unreliable.

8.2 Experimentation

The main three limitations of the present work regarding the experimentation and evaluation of the proposed approach are:

1. **Only English-speaking news sources:** the proposed methodology is content- and language-independent. However, we focused exclusively on the English-speaking news sources due to the predominance of available ground truth data for experimentation in this language. Further studies need to be done with ground truth for non-English-speaking sources to assess the robustness of the methodology across languages.

2. **Restricted graph size:** the graph we used was built processing around 100M news articles from 4 months spanning a 4-year time window. This imposes not only a temporal limitation but also restricts the number of sources in the graph. However, along with this paper, we release and open source, under Apache 2.0 license, the Python CC-News processing pipeline and the dataset for the community to reproduce the proposed methodology in larger scale.

3. **Corpus used to build the graph:** although CC-News is continuously growing on a daily

basis, the crawling of news articles started in 2016. Therefore, CC-News does not contain articles prior to 2016 and news sources that existed before 2016 but not after, will not be reachable. Consequently, for these sources to be included, their articles need to be crawled from a different corpus or manually from the Web.

9 Acknowledgments

This work was supported by CRiTERIA, EU project funded under the Horizon 2020 program, grant agreement number 101021866. We would also like to express our sincere gratitude to the anonymous reviewers for their valuable feedback, which has helped us enhance the quality of this work.

References

Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018. Predicting factuality of reporting and bias of news media sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3528–3539, Brussels, Belgium. Association for Computational Linguistics.

Ramy Baly, Georgi Karadzhov, Jisun An, Haewoon Kwak, Yoan Dinkov, Ahmed Ali, James Glass, and Preslav Nakov. 2020. What was written vs. who read it: News media profiling using text analysis and social media context. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3364–3374, Online. Association for Computational Linguistics.

Ramy Baly, Georgi Karadzhov, Abdelrhman Saleh, James Glass, and Preslav Nakov. 2019. Multi-task ordinal regression for jointly predicting the trustworthiness and the leading political ideology of news media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2109–2116, Minneapolis, Minnesota. Association for Computational Linguistics.

João Pedro Baptista and Anabela Gradim. 2022. A working definition of fake news. *Encyclopedia*, 2(1).

Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1):107–117. Proceedings of the Seventh International World Wide Web Conference.

Sergio Burdisso, Juan Pablo Zuluaga-Gomez, Esaú Villatoro-Tello, Martin Fajcik, Muskaan Singh, Pavel Smrz, and Petr Motlicek. 2022. IDIAPers @ causal news corpus 2022: Efficient causal relation identification through a prompt-based few-shot approach. In *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, pages 61–69, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Jose Yunam Cuan-Baltazar, Maria José Muñoz-Perez, Carolina Robledo-Vega, Maria Fernanda Pérez-Zepeda, and Elena Soto-Vega. 2020. Misinformation of covid-19 on the internet: infodemiology study. *JMIR public health and surveillance*, 6(2):e18444.

Xin Luna Dong, Evgeniy Gabrilovich, Kevin Murphy, Van Dang, Wilko Horn, Camillo Lugaresi, Shaohua Sun, and Wei Zhang. 2015. Knowledge-based trust: Estimating the trustworthiness of web sources. *Proc. VLDB Endow.*, 8(9):938–949.

Martin Fajcik, Petr Motlicek, and Pavel Smrz. 2023. Claim-dissector: An interpretable fact-checking system with joint re-ranking and veracity prediction. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10184–10205, Toronto, Canada. Association for Computational Linguistics.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Maurício Gruppi, Benjamin D. Horne, and Sibel Adali. 2020. NELA-GT-2019: A large multi-labelled news dataset for the study of misinformation in news articles. *CoRR*, abs/2003.08444.

ISD. 2020. *Anatomy of a Disinformation Empire: Investigating NaturalNews*. Institute for Strategic Dialogue.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. 1996. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285.

Jürgen Knauth. 2019. Language-agnostic Twitter-bot detection. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 550–558, Varna, Bulgaria. INCOMA Ltd.

Sejeong Kwon, Meeyoung Cha, Kyomin Jung, Wei Chen, and Yajun Wang. 2013. Aspects of rumor

spreading on a microblog network. In *Social Informatics: 5th International Conference, SocInfo 2013, Kyoto, Japan, November 25-27, 2013, Proceedings 5*, pages 299–308. Springer.

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

Zhenyu Lei, Herun Wan, Wenqian Zhang, Shangbin Feng, Zilong Chen, Jundong Li, Qinghua Zheng, and Minnan Luo. 2023. BIC: Twitter bot detection with text-graph interaction and semantic consistency. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10326–10340, Toronto, Canada. Association for Computational Linguistics.

Stephan Lewandowsky, Ullrich KH Ecker, and John Cook. 2017. Beyond misinformation: Understanding and coping with the "post-truth" era. *Journal of applied research in memory and cognition*, 6(4):353–369.

Yaliang Li, Jing Gao, Chuishi Meng, Qi Li, Lu Su, Bo Zhao, Wei Fan, and Jiawei Han. 2016. A survey on truth discovery. *SIGKDD Explor. Newsl.*, 17(2):1–16.

Nataliia Liubchenko, Andrii Podorozhniak, and Vasyl Oliinyk. 2022. Research application of the spam filtering and spammer detection algorithms on social media. In *CEUR Workshop Proceedings*, volume 3171, pages 116–126.

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*, IJCAI'16, page 3818–3824. AAAI Press.

Lin Miao, Mark Last, and Marina Litvak. 2020. Detecting troll tweets in a bilingual corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6247–6254, Marseille, France. European Language Resources Association.

Todor Mihaylov, Georgi Georgiev, and Preslav Nakov. 2015a. Finding opinion manipulation trolls in news community forums. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 310–314, Beijing, China. Association for Computational Linguistics.

Todor Mihaylov, Ivan Koychev, Georgi Georgiev, and Preslav Nakov. 2015b. Exposing paid opinion manipulation trolls. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 443–450, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

Subhabrata Mukherjee and Gerhard Weikum. 2015. Leveraging joint interactions for credibility analysis in news communities. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, page 353–362, New York, NY, USA. Association for Computing Machinery.

Kevin Munger. 2020. All the news that's fit to click: The economics of clickbait media. *Political Communication*, 37(3):376–397.

An Nguyen, Aditya Kharosekar, Matthew Lease, and Byron Wallace. 2018. An interpretable joint graphical model for fact-checking from crowds. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Jeppe Nørregaard, Benjamin D Horne, and Sibel Adalı. 2019. Nela-gt-2018: A large multi-labelled news dataset for the study of misinformation in news articles. In *Proceedings of the international AAAI conference on web and social media*, volume 13, pages 630–638.

Jeff Pasternack and Dan Roth. 2010. Knowing what to believe (when you already know something). In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 877–885, Beijing, China. Coling 2010 Organizing Committee.

Ani Petrosyan. 2023. Internet usage worldwide - statistics & facts. *Statista*.

Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2016. Credibility assessment of textual claims on the web. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, page 2173–2178, New York, NY, USA. Association for Computing Machinery.

Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2017. Where the truth lies: Explaining the credibility of emerging claims on the web and social media. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, page 1003–1012, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2018a. Credeye: A credibility lens for analyzing and explaining misinformation. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 155–158, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018b. DeClarE: Debunking fake news and false claims using evidence-aware deep learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 22–32, Brussels, Belgium. Association for Computational Linguistics.

Martin L Puterman. 2014. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

Devakunchari Ramalingam and Valliyammai Chinnaiah. 2018. Fake profile detection techniques in large-scale online social networks: A comprehensive review. *Computers & Electrical Engineering*, 65:165–177.

Pradeep Kumar Roy and Shivam Chahar. 2020. Fake profile detection on social networking websites: A comprehensive review. *IEEE Transactions on Artificial Intelligence*, 1(3):271–285.

Giuseppe Sansonetti, Fabio Gasparetti, Giuseppe D'aniello, and Alessandro Micarelli. 2020. Unreliable users detection in social media: Deep learning techniques for automatic detection. *IEEE Access*, 8:213154–213167.

Shaden Shaar, Nikolay Babulkov, Giovanni Da San Martino, and Preslav Nakov. 2020. That is a known lie: Detecting previously fact-checked claims. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3607–3618, Online. Association for Computational Linguistics.

Gianluca Stringhini, Christopher Kruegel, and Giovanni Vigna. 2010. Detecting spammers on social networks. In *Proceedings of the 26th annual computer security applications conference*, pages 1–9.

Jesper Strömbäck, Yariv Tsfati, Hajo Boomgaarden, Alyt Damstra, Elina Lindgren, Rens Vliegenthart, and Torun Lindholm. 2020. News media trust and its impact on media use: Toward a framework for future research. *Annals of the International Communication Association*, 44(2):139–156.

Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.

Michele Tomaiuolo, Gianfranco Lombardo, Monica Mordonini, Stefano Cagnoni, and Agostino Poggi. 2020. A survey on troll detection. *Future Internet*, 12(2).

Sander Van Der Linden. 2022. Misinformation: susceptibility, spread, and interventions to immunize the public. *Nature Medicine*, 28(3):460–467.

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

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*, pages 849–857.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Hengshuai Yao and Dale Schuurmans. 2013. Reinforcement ranking. *arXiv preprint arXiv:1303.5988*.

A Hyperparameter optimization

We performed a grid search to determine the best hyperparameter values on each of the three experimental sets. More precisely, as in Section 5.1, the evaluation was performed using 5-fold cross validation on the reliability classification task. For reinforcement learning strategies, we evaluated γ from 0.05 to 0.95 in increments of 0.05 (*i.e.,* $\gamma \in \{0.05, 0.1, 0.15, \ldots, 0.95\}$). For *I-Reliability*, n from 1 to 10 (*i.e.,* $n \in \{1, 2, 3, \ldots, 10\}$). Finally, the hyperparameter values obtaining the best *macro avg. F_1 score* were the one selected on each experimental set. Namely, the selected values for each strategy were:

- **F-Reliability:** $\gamma = 0.05$ for *ExpsetA* and *ExpsetB*$^-$, $\gamma = 0.5$ for *ExpsetB*.
- **P-Reliability:** $\gamma = 0.15$ for *ExpsetA*, $\gamma = 0.3$ for *ExpsetB*, and $\gamma = 0.2$ for *ExpsetB*$^-$.
- **FP-Reliability:** $\gamma = 0.1$ for *ExpsetA*, $\gamma = 0.05$ for *ExpsetB* and *ExpsetB*$^-$.
- **I-Reliability:** $n = 1$ for *ExpsetA* and *ExpsetB*, and $n = 2$ for *ExpsetB*$^-$.

As shown in Figure 1, we can observe that *better reliability estimation is achieved when looking mostly at nearby sources*, as better performance is obtained with small n and γ values, namely $n \leq 2$ and $\gamma < 0.5$. Furthermore, *P-Reliability* (orange line) outperforms the other reinforcement learning strategies, consistently, while being more robust to the choice of γ, except when $\gamma > 0.7$ from which performance starts to decrease.

Finally, for the Baly et al. (2018) and Baly et al. (2020) classifiers in Table 3, we follow the same process described by the authors to tune the SVM hyperparameters, i.e., the cost C, the kernel type, and the kernel width γ using the 5-fold cross validation maximizing the *F_1 score* as with our methods.[27]

B Temporal Ablation Analysis

To evaluate the robustness of our proposed approach concerning both the graph size and the temporal span used in its construction, a temporal ablation study was conducted. In addition to the graph used for experimentation, illustrated in Figure 2, we generated four different graphs, each corresponding to one of the four CC-News snapshots (refer to Table 1 in Section 4.1 for detailed information on each graph). Subsequently, employing each of these four graphs, we replicated the evaluation procedure described in Section 5.1. These evaluations allowed us to measure how the performance of the proposed strategies changed, when changing the graph, compared to the reported values in Table 3.

Figure 3 shows, without loss of generality, the results obtained with the two best-performing strategies reported in Table 3, *P-Reliability* and *I-Reliability*, on the largest experimental set *ExpsetB*. We observed that, independently of the strategy and the dataset, the best results were always obtained with the largest (in size and time) graph joining all the snapshots. We also observed that not all strategies exhibit equal robustness to changes in the graph. For instance, we can see in Figure 3 how *P-Reliability* is more sensitive than *I-Reliability* to the choice of snapshot for graph construction. Despite this variation, both strategies demonstrate improvement and achieve their best results with reduced uncertainty when considering all snapshots.

C Graph Construction Steps

The *Common Crawl News Dataset* (*CC-News*)[28] is published as *WARC* files[29] grouped by year and month, called snapshots.[30] To construct the news media graph, $G = \langle S, E, w \rangle$, from CC-News snapshots, we follow to the subsequent steps:

1. Download each WARC file and parse each news article in it to extract its URL and all the hyperlinks in its body. At the end of this step, we have a set of news article URLs, U, and a set of hyperlinks L_u for each article $u \in U$.
2. Generate the graph nodes S from U simply as $S = \{domain(u) : u \in U\}$ which contains the URL domain names (*e.g.* "nytimes.com", "cnn.com", etc.) of all processed news articles.
3. For each domain $s \in S$ create the list of all its hyperlinks, L_s, by aggregating the hyperlinks of all its articles, *i.e.* $L_s = \bigcup_{u \in U : domain(u) = s} L_u$.

[27] We used the author's source code containing the hyperparameter search in it (github.com/ramybaly/News-Media-Reliability).

[28] https://commoncrawl.org/blog/news-dataset-available

[29] A file format that resembles the raw HTTP request and response of each crawled web page.

[30] https://data.commoncrawl.org/crawl-data/CC-NEWS/index.html

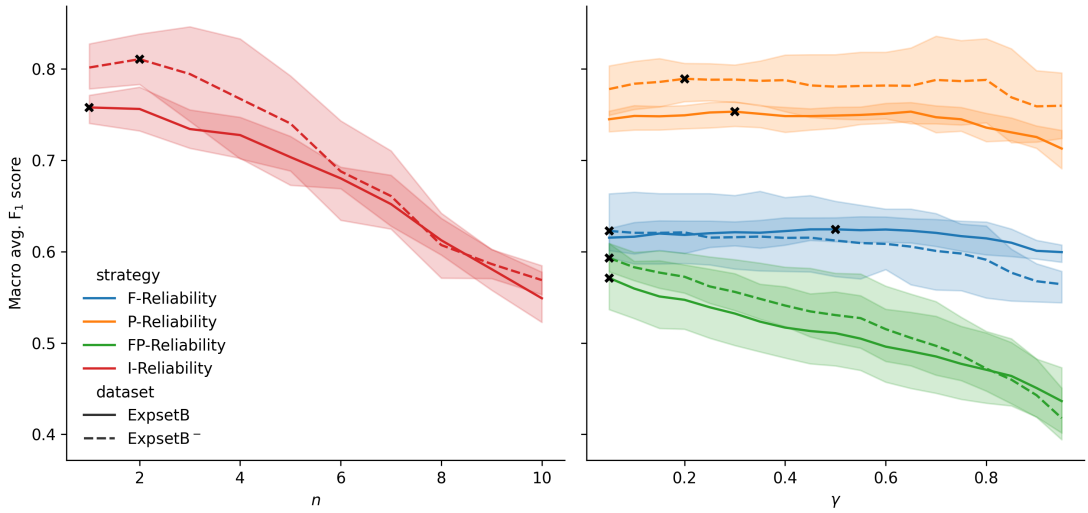

Figure 1: Performance variation across searched values of n (left side) and γ (right side) on the *ExpsetB* (solid line) and *ExpsetB*$^-$ (dashed line) datasets. The lines represent the mean values across the 5 folds, and 95% confidence intervals are depicted. Markers highlight selected hyperparameter values.

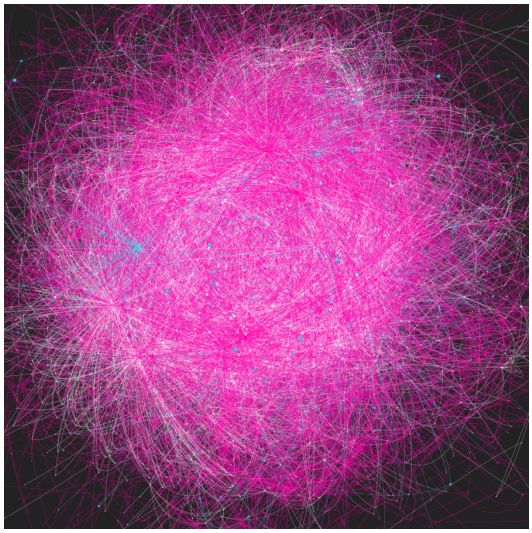

Figure 2: News media graph built from all four CC-News snapshot (only English-speaking sources) and used for experimentation.

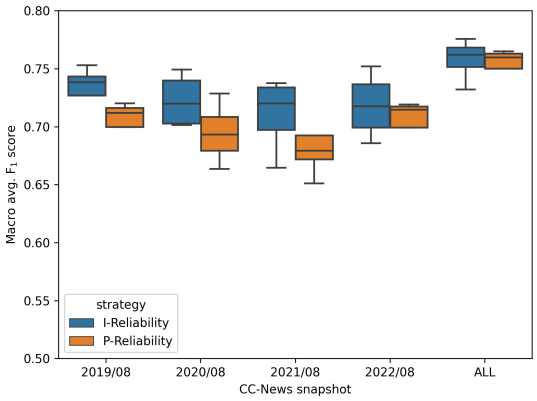

Figure 3: 5-fold cross-validation results obtained on the *ExpsetB* dataset with the two best strategies, *P-Reliability* and *I-Reliability*, using different graphs. The x-axis represents the CC-News snapshot used to build the graph, and the y-axis the *macro averaged* F_1 *score*.

4. Finally, generate the graph edges $(s, s') \in E$ for each $s \in S$ by creating an edge to each unique domain s' in its hyperlinks L_s weighted by the proportion of links whose domain is s', *i.e.* $w(s, s') = |\{l \in L_s : domain(l) = s'\}|/|L_s|$.

D NewsGuard Score Details

In Section 5.2, we used the scores from the *NewsGuard* dataset introduced in Section 4.2 to measure the correlation with estimated $\rho(s)$ values. NewsGuard employs a team of journalists and experienced editors to produce these reliability scores for news and information websites. The score ranges from 0 to 100 and NewsGuard is transparent about the methodology used to compute it. Namely, they compute this reliability score based on the following nine apolitical criteria, each is worth the indicated number of points, based on importance:

1. ***Does not repeatedly publish false content?*** *(22 points)*

2. ***Gathers and presents information responsibly?*** *(18 points)*

3. ***Regularly corrects or clarifies errors?*** *(12.5 points)*

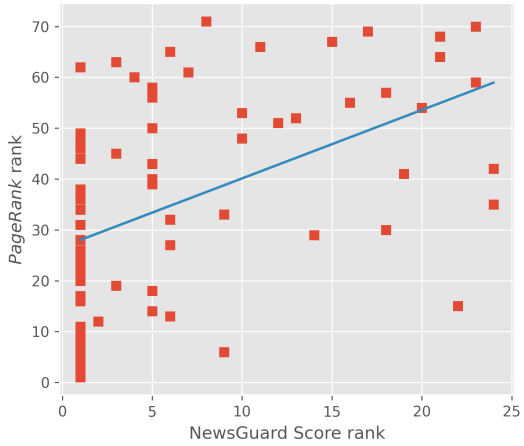

Figure 4: Scatter plot showing the correlation between the rankings obtained by PageRank values (y-axis) and News Guard scores (x-axis).

4. ***Handles the difference between news and opinion responsibly?*** *(12.5 points)*
5. ***Avoids deceptive headlines?*** *(10 points)*
6. ***Website discloses ownership and financing?*** *(7.5 points)*
7. ***Clearly labels advertising?*** *(7.5 points)*
8. ***Reveals who's in charge, including any possible conflicts of interest?*** *(5 points)*
9. ***Provides information about content creators?*** *(5 points)*

More details can be found in the "Rating Process and Criteria" section of their website.[31]

E Detailed Results

Table 7 shows the detailed evaluation results obtained on the three experimentation sets along with two naive baselines for reference, random and majority class classifiers. Furthermore, for *ExpsetA*, we also report the results obtained using the classification models introduced in previous works (Baly et al., 2018, 2020) along with the ensemble results. We can see that on *ExpsetA*, unlike on the other experimental sets, our strategies have 100% precision for the *unreliable* sources but its recall on the same sources is quite low (from 22.20% to 38.72%) showing the models are detecting only a small portion of unreliable sources but with high precision (probably only the easiest cases). By performing the ensemble with Baly models, the recall for the *unreliable* group increases allowing the graph-based strategies to identify more unreliable sources, in turn improving the macro average F_1 scores.

Finally, in Figure 4 is shown the scatter plot showing the correlation between the ranking obtained by PageRank values and the one obtained by the News Guard scores. Likewise, Figures 5, 6, 7, and 8 show the scatter plots for *F-Reliability*, *P-Reliability*, *FP-Reliability*, *I-Reliability*, respectivelly.

[31] https://www.newsguardtech.com/ratings/rating-process-criteria/

Data	Strategy	Precision			Recall			F_1 score			Acc.
		macro avg.	*reliable*	*unreliable*	*macro avg.*	*reliable*	*unreliable*	*macro avg.*	*reliable*	*unreliable*	
ExpsetA	*M-BL*	*36.72±1.29*	*73.44±2.57*	*0.00±0.00*	*50.00±0.00*	***100.00***±*0.00*	*0.00±0.00*	*42.33±0.85*	*84.66±1.70*	*0.00±0.00*	*73.44±2.57*
	R-BL	*51.37±4.21*	*74.63±3.49*	*28.11±6.08*	*51.56±5.30*	*52.87±6.72*	*50.26±7.86*	*48.85±5.29*	*61.76±5.63*	*35.94±6.94*	*52.33±5.52*
	Baly18	*70.11±5.74*	***82.48***±*5.02*	*57.74±8.72*	*67.68±7.52*	*87.65±4.31*	*47.70±16.75*	*67.87±6.89*	*84.81±2.65*	*50.92±11.97*	*76.95±4.11*
	Baly20	*71.03±9.32*	*78.42±3.14*	*63.64±20.83*	*64.64±3.62*	*88.74±7.19*	*40.55±8.33*	*65.24±4.03*	*82.99±2.06*	*47.50±6.75*	*74.37±2.84*
	F-R	89.02±1.75	78.03±3.51	**100.00**±0.00	61.10±4.70	**100.00**±0.00	22.20±9.40	61.52±6.83	87.62±2.19	35.42±11.92	79.26±3.59
	P-R	**90.96**±0.88	**81.91**±1.75	**100.00**±0.00	**69.30**±3.53	**100.00**±0.00	**38.59**±7.05	**72.67**±4.10	**90.05**±1.05	**55.29**±7.87	**83.79**±1.59
	FP-R	90.29±1.07	80.57±2.14	**100.00**±0.00	66.53±3.31	**100.00**±0.00	33.06±6.61	69.28±3.91	89.23±1.32	49.34±7.13	82.29±2.07
	I-R	**90.95**±1.02	**81.89**±2.04	**100.00**±0.00	**69.36**±2.75	**100.00**±0.00	**38.72**±5.51	**72.81**±3.18	**90.03**±1.23	**55.60**±5.86	**83.77**±1.81
	F-R+Baly18	76.54±3.69	87.12±2.64	65.95±7.81	75.85±3.02	87.65±4.31	64.06±7.47	75.84±3.01	87.29±2.02	64.39±4.59	81.32±2.67
	F-R+Baly20	**77.80**±7.68	84.57±2.43	**71.04**±17.10	74.79±2.98	**88.74**±7.19	60.84±3.90	75.34±4.18	86.38±2.69	64.30±5.75	80.30±3.67
	P-R+Baly18	77.47±3.96	**87.99**±1.21	66.95±8.47	**77.23**±1.76	87.65±4.31	**66.82**±3.31	**77.11**±2.95	**87.75**±2.06	**66.47**±4.27	**82.11**±2.72
	P-R+Baly20	77.04±7.79	83.89±2.25	**70.19**±17.23	73.70±2.95	**88.74**±7.19	58.66±4.34	74.36±4.10	86.02±2.57	62.70±5.84	79.69±3.50
	FP-R+Baly18	74.25±4.78	85.08±3.73	63.42±8.54	72.67±4.33	87.65±4.31	57.68±8.86	73.01±4.32	86.24±2.66	59.79±6.25	79.54±3.72
	FP-R+Baly20	76.01±7.86	82.96±3.47	69.06±17.57	72.10±4.12	**88.74**±7.19	55.46±9.83	72.67±4.28	85.46±2.11	59.88±7.00	78.72±2.99
	I-R+Baly18	**77.79**±3.45	**88.29**±1.75	67.30±7.86	**77.73**±1.47	87.65±4.31	**67.82**±4.62	**77.47**±2.40	**87.89**±1.89	**67.06**±3.40	**82.34**±2.40
	I-R+Baly20	76.12±7.77	82.85±2.25	69.38±17.58	72.10±2.28	**88.74**±7.19	55.46±3.69	72.88±3.46	85.46±2.36	60.30±4.77	78.74±3.13
ExpsetB	*M-BL*	*38.50±0.04*	*77.00±0.09*	*0.00±0.00*	*50.00±0.00*	***100.00***±*0.00*	*0.00±0.00*	*43.50±0.03*	*87.00±0.06*	*0.00±0.00*	*77.00±0.09*
	R-BL	*51.09±1.39*	*78.05±1.30*	*24.13±1.49*	*51.53±1.96*	*51.72±3.06*	*51.34±3.70*	*47.48±1.83*	*62.17±2.45*	*32.80±1.93*	*51.63±2.21*
	F-R	61.74±3.19	83.63±0.71	39.85±5.85	63.24±2.10	76.38±6.35	50.11±2.82	61.85±3.12	79.72±3.69	43.98±2.61	70.34±4.36
	P-R	**74.66**±2.51	**88.44**±1.09	**60.89**±4.20	**74.78**±2.26	**88.16**±1.62	**61.41**±3.72	**74.69**±2.31	**88.29**±1.15	**61.10**±3.52	**82.00**±1.73
	FP-R	59.01±1.95	85.73±1.42	32.29±2.67	62.60±2.66	56.63±5.39	**68.57**±3.22	55.95±3.21	68.08±4.09	43.82±2.55	59.38±3.98
	I-R	**76.68**±3.08	**87.98**±1.10	**65.39**±5.31	**74.60**±2.40	**90.67**±1.68	58.53±3.78	**75.51**±2.61	**89.30**±1.22	**61.72**±4.02	**83.28**±1.87
ExpsetB$^-$	*M-BL*	*45.38±0.05*	*90.76±0.10*	*0.00±0.00*	*50.00±0.00*	***100.00***±*0.00*	*0.00±0.00*	*47.58±0.03*	*95.15±0.06*	*0.00±0.00*	*90.76±0.10*
	R-BL	*49.82±2.41*	*90.62±2.36*	*9.02±2.47*	*49.43±7.19*	*48.34±0.88*	*50.52±14.11*	*39.17±2.54*	*63.04±1.09*	*15.31±4.20*	*48.55±1.71*
	F-R	60.60±2.73	94.14±0.58	27.07±5.03	66.84±2.97	86.69±3.07	46.99±5.59	62.18±3.15	90.23±1.75	34.12±4.90	83.02±2.77
	P-R	**77.75**±5.20	**96.46**±0.83	**59.04**±9.83	**80.48**±4.25	**95.21**±1.42	**65.74**±7.79	**78.90**±4.50	**95.83**±0.98	**61.97**±8.04	**92.48**±1.74
	FP-R	59.20±0.39	95.98±0.63	22.41±0.49	72.21±1.94	75.77±2.32	68.66±5.99	59.20±0.54	84.66±1.23	33.74±0.83	75.11±1.60
	I-R	**83.22**±4.09	**96.13**±0.72	**70.32**±7.83	**79.41**±3.58	**97.30**±0.85	61.52±7.02	**81.05**±3.40	**96.71**±0.60	**65.39**±6.22	**93.99**±1.09

Table 7: 5-fold cross-validation detailed results for reliability classification. Mean and standard deviation is shown for each metric. **Bold** indicates 2nd-best-performing values, while **underlined** the best-performing values in each dataset. R-BL and M-BL refer to random and majority class baselines; Baly18 and Baly20 refer to Baly et al. (2018) and Baly et al. (2020); and *-R stands for *-Reliability.

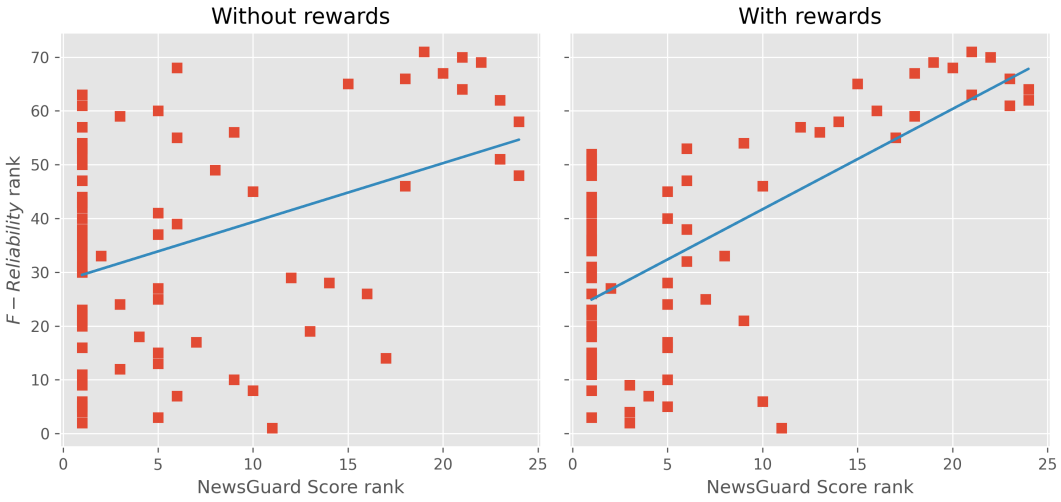

Figure 5: Scatter plot showing the correlation between the rankings obtained by *F-Reliability* values (y-axis) and News Guard scores (x-axis). Left side without rewards and right side with rewards.

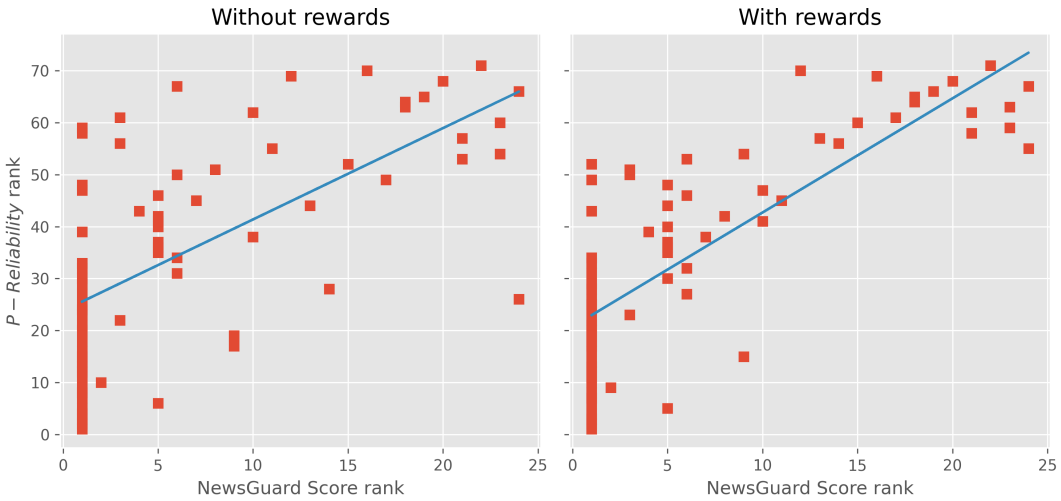

Figure 6: Scatter plot showing the correlation between the rankings obtained by *P-Reliability* values (y-axis) and News Guard scores (x-axis). Left side without rewards and right side with rewards.

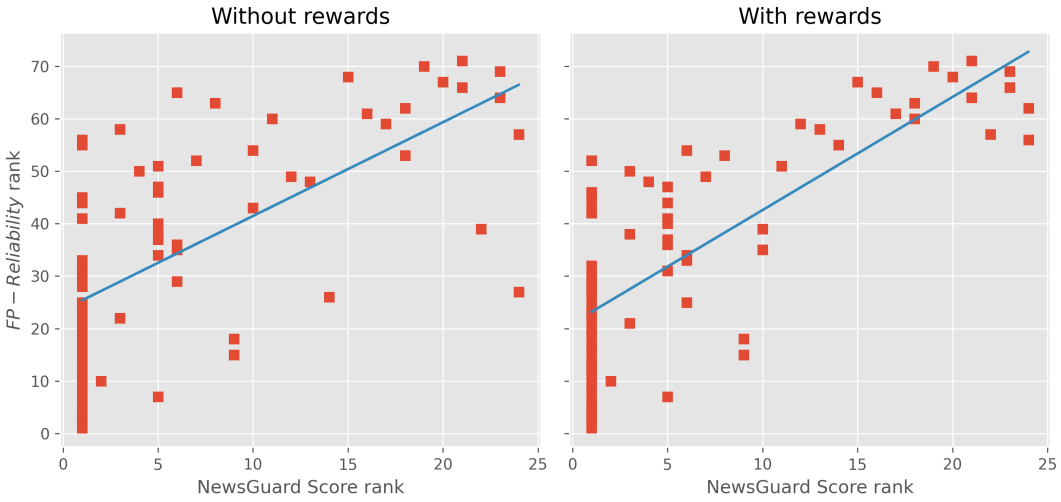

Figure 7: Scatter plot showing the correlation between the rankings obtained by *FP-Reliability* values (y-axis) and News Guard scores (x-axis). Left side without rewards and right side with rewards.

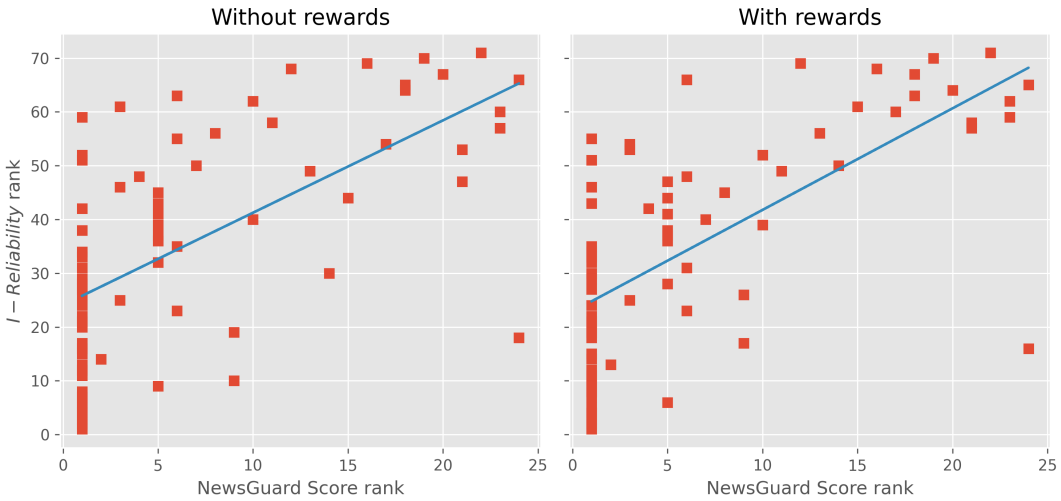

Figure 8: Scatter plot showing the correlation between the rankings obtained by *I-Reliability* values (y-axis) and News Guard scores (x-axis). Left side without rewards and right side with rewards.