Evaluating Trustworthiness of Online News Publishers via Article Classification

John Bianchi*
IMT School for Advanced Studies Lucca
Italy
john.bianchi@imtlucca.it

Manuel Pratelli
IMT School for Advanced Studies Lucca
IIT-CNR
Italy
manuel.pratelli@imtlucca.it

Marinella Petrocchi
IIT-CNR
IMT School for Advanced Studies Lucca
Italy
marinella.petrocchi@iit.cnr.it

Fabio Pinelli
IMT School for Advanced Studies Lucca
Italy
fabio.pinelli@imtlucca.it

ABSTRACT

The proliferation of low-quality online information in today's era has underscored the need for robust and automatic mechanisms to evaluate the trustworthiness of online news publishers. In this paper, we analyse the trustworthiness of online news media outlets by leveraging a dataset of 4033 news stories from 40 different sources. We aim to infer the trustworthiness level of the source based on the classification of individual articles' content. The trust labels are obtained from NewsGuard, a journalistic organization that evaluates news sources using well-established editorial and publishing criteria. The results indicate that the classification model is highly effective in classifying the trustworthiness levels of the news articles. This research has practical applications in alerting readers to potentially untrustworthy news sources, assisting journalistic organizations in evaluating new or unfamiliar media outlets and supporting the selection of articles for their trustworthiness assessment.

CCS CONCEPTS

• **Computing methodologies** → *Natural language processing*; *Machine learning*; • **Information systems** → *Decision support systems*; *Content ranking*.

KEYWORDS

Online News, Transparency and Reputability of Online News Sources, Multiclass Classification, Data Science for Social Good

*Corresponding Author

SAC'24, April 8–April 12, 2024, Avila, Spain

ACM ISBN 979-8-4007-0243-3/24/04…$15.00
https://doi.org/https://doi.org/10.1145/3605098.3636044

ACM Reference Format:
John Bianchi, Manuel Pratelli, Marinella Petrocchi, and Fabio Pinelli. 2024. Evaluating Trustworthiness of Online News Publishers via Article Classification. In *Proceedings of ACM SAC Conference (SAC'24).* ACM, New York, NY, USA, Article 4, 8 pages. https://doi.org/https://doi.org/10.1145/3605098.3636044

1 INTRODUCTION

Disintermediation, or the phenomenon of reducing intermediate flows, is a term coined back in 1983, when author Paul Hawken called by this name the set of processes by which consumers could directly manage financial investments in securities, rather than leaving their money in savings accounts [12].

Over time, many industries have experienced disintermediation. In tourism, the Internet provides users with access to a wealth of information, allowing them to seamlessly assemble various tourism services and create unique travel experiences on their own. Similarly, the growing trend of self-publishing places increasing responsibility on authors to oversee the entire process of producing and distributing their work [10]). Journalism has also experienced changes that reflect the evolving landscape of direct access to information and news dissemination. These shifts indicate a broader societal trend toward greater autonomy and control in various industries. Particularly with regard to journalism, the emergence of new web technologies and social networks has diminished the essential role of traditional journalists as the prevalence of participatory journalism facilitated by blogs and social networks continues to grow [3, 14]. In this regard, a recent UNESCO report[1] on the existential threat posed by social media to traditional news claims that online 'news outlets often struggle to get the clicks from readers that determine advertising revenue' and job cuts in journalism have resulted in a noticeable void in the information landscape.

The erosion of the mainstream journalism system in recent years, coupled with challenges such as understaffing and the pressure to publish attention-grabbing news to re-engage readers, has raised

[1]https://news.un.org/en/story/2022/03/1113702. All of the URLs in this document were last accessed on December 22, 2023.

concerns about the quality of information provided by online media. Various journalism organizations and indices, including NewsGuard[2], the MediaBias Fact Check[3], the Iffy index of unreliable sources[4], the Global Disinformation Index[5], the Ad Fontes Media[6] that conduct studies on the transparency and trustworthiness of online news sources, including their tendency to produce propagandistic and/or politically biased content.

Although different organizations use different criteria to determine the trustworthiness of an online media outlet, recent work has found excellent convergence in the labels each assigns to individual media outlets, confirming the degree of trustworthiness of the judgments [18, 27].

Unfortunately, the process of evaluating each news outlet is very cumbersome, especially in terms of time. For example, the procedure followed by the Global Disinformation Index is to select annotators who are experts in the online information system of a particular country. After training, the annotators select a group of newspapers that accurately reflect the country's information landscape. They then manually analyze these newspapers to find information on aspects such as ownership and funding sources. This is followed by a manual content analysis of a sample of articles per media outlet to check for unreliable, sensational and/or propagandistic content. The GDI then processes the results of the study, which are summarized in a score between 0 and 100 that indicates the risk that the media outlet is misinforming its readers[7].

In this paper, we aim to label the trustworthiness level of a news source from the classification of the news itself. Operationally, we start with a dataset of 4033 news stories from 40 online news outlets, which we have collected and to which we have attached labels regarding both the main topic and the trustworthiness score of the news outlet. The labels are collected by NewsGuard, which is licensed to the authors of this paper. Through qualified journalists, NewsGuard rates all news sources, which account for 95 percent of online engagement[8]. Each site is analyzed according to nine accepted journalistic criteria. Based on these nine criteria, the site receives a trustworthiness score from 0 to 100. The trustworthiness levels are 5, ranging from 'high credibility,' the best rating, to 'proceed with extreme caution,' indicating a site with a very low level of transparency and credibility.

On the one hand, tagging the articles in our dataset with NewsGuard labels relieves us of the tedious task of annotating the data and gives us a solid ground truth based on the work of specialized journalists. On the other hand, before moving on to the main goal of the work, which is to derive the level of trustworthiness of the news source from the analysis of individual articles, we will test the goodness of the dataset by classifying articles by topic and making sure that the predicted topic matches the topic assigned in the label.

We combine 3 standard topics, i.e., *Sports*, *Political News* and *Health*, with an escalating one, in the age of the internet and pandemics, vaccines and wars, namely *Conspiracy Theories* [4].

Results. Our models have proven to be highly effective in both classification tasks. The results are summarized as follows:

- **Trustworthiness Detection Task:** This is a multi-class classification task at the article level. Our model successfully predicts the level of trustworthiness of news sources based on the article text alone. Specifically, we obtain an average F1-macro of 0.843 and an average F1-micro of 0.882 for this task (see section 4.2).
- **Topic Detection Task:** This is a multi-class classification task at the article level. Our model achieves an average F1-macro of 0.925 and an average F1-micro of 0.929. This gives us an additional level of confidence in the original NewsGuard labeling of the data set. Wrong predictions arise when distinguishing between *'Conspiracy'* and (*'Health"* or *'Political'*), leading to some misclassifications (see Section 4.1).

Applications. The ability to predict a publisher's level of trustworthiness from the classification of individual articles suggests at least three possible applications, one to assist the user, the others to assist journalistic organizations (e.g., NewsGuard and GDI):

(1) at the user level, the classifier can be used as a tool to alert the reader by displaying a meaningful visual signal, such as the classic red flag, while the reader is viewing an article from an unfamiliar news outlet. The red flag could say something like 'the article you are reading is similar to those produced by untrustworthy newspapers. Supplement your reading with other readings of articles produced by trustworthy newspapers'.
(2) at the organizational level, let the reader assume that the media outlet is new or completely unknown to the evaluator (which is very common these days given the constant proliferation of alternative online media outlets [31]). An initial idea of its level of trustworthiness can be obtained by collecting a number of articles and applying the trustworthiness ranking model to them. At a later date, if the evaluator deems it necessary, a more comprehensive investigation can be conducted using traditional journalistic analysis.
(3) still at the organizational level, the selection of articles to analyze to assess the publisher's trustworthiness are typically the most shared articles on social media and/or articles containing a set of representative keywords.[9]. Unfortunately, this method may not be sufficient to select a sample of articles that is truly representative of the publisher. For example, if we relied on the most shared articles on social media, we might select a sample consisting only of straight news stories (e.g., traffic accidents, robberies, etc.). Therefore, we argue that the models presented in this paper can be used to process a selection of articles from the target media for a more balanced sample that can be manually analyzed. This approach ensures a balanced assessment in terms of both trustworthiness levels and topics.

[2]NewsGuard: https://www.newsguardtech.com
[3]MediaBias Fact Check: https://mediabiasfactcheck.com/
[4]Iffy Index: https://iffy.news/index/
[5]The Global Disinformation Index: https://www.disinformationindex.org/
[6]Ad Fontes Media: https://adfontesmedia.com/
[7]The second and third authors are familiar with the procedure, having participated as annotators in the GDI country study on the Italian online media market [25].
[8]https://www.newsguardtech.com/solutions/newsguard/
[9]https://www.disinformationindex.org/country-studies/2023-06-08-disinformation-risk-assessment-the-online-news-market-in-thailand/

2 PROBLEM DEFINITION

In this section, we formalize the main problem addressed in this paper, called *Trustworthiness Level* Detection, and present the performance metrics used to evaluate the resulting models. The *Topic Detection* task is also formalized, since we use the results of this classifier to experimentally evaluate the quality of the labeling procedure obtained from NewsGuard.

Let be A a set of articles. This set represents a collection of articles characterized by their textual content. Formally, $A = \{a_1, a_2, \ldots, a_n\}$, where n is the total number of articles. Each article a has attributes $text_a$ representing the textual content of the article and $newspaper_a$, the newspaper from which the article originates.

The set of Newspapers N comprises various newspapers, each associated with a specific level of trustworthiness. Formally, $N = \{n_1, n_2, \ldots, n_m\}$, where m is the total number of newspapers. Each newspaper n_i has $trust_{n_i}$ that represents the level of trustworthiness associated with it.

Notice that different levels of trustworthiness associated with newspapers belong to the set L. Formally, $L = \{l_1, l_2, \ldots, l_p\}$, where p is the total number of trustworthiness levels. The level of trustworthiness is typically a continuous value, but we prefer to aggregate values in bins that correspond to our levels of trustworthiness L. We provide more details about the bins in Section 3.

Notice that each article a_i is associated with one and only one level of trustworthiness ($trust_{a_i}$) from the set of trustworthiness levels (L) inherited by the newspaper it originated from.

Therefore, given the sets A, N, and L, and the constraints mentioned above, we formalize the problem as follows: we aim to develop a text classifier C_{trust} that associates the level of trustworthiness $trust_a \in L$ for each article a based on its textual content $text_a$. The Classifier C_{trust} is a machine learning model that addresses the *Trustworthiness Level* Detection task, and it relies on the NewsGuard annotation process described in Section 1. Given this, we can reasonably assume that the level of trustworthiness associated with each newspaper is reliable, a notion further supported by considering the primary topic associated with each newspaper.

To this end, we define the *Topic Detection* task as follows. Each newspaper n_i has an additional attribute associated with it: $topic_{n_i}$ that indicates the topic it primarily covers. This information is gathered by NewsGuard. The set T defines the possible categories or topics into which articles can be classified. Formally, $T = \{t_1, t_2, \ldots, t_k\}$, where k is the total number of topics. As for the previous classification task, each article a_i is associated with one and only one Topic ($topic_{a_i}$) from the set of Topics (T) inherited by the newspaper it originated from. Thus, given the sets A, N, and T, and the constraints mentioned above, we formalize the problem as follows: we aim to develop a text classifier C_{topic} that associates a $topic_a \in T$ for each article a based on its textual content $text_a$. The Classifier C_{topic} is a machine learning model for *Topic Detection* task. As in [26], which addressed a similar challenge to ours, the performance of the classifiers is evaluated using the F1 Micro and F1 Macro metrics, defined as follows:

$$\text{F1 Micro} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{1}$$

$$\text{F1 Macro} = \frac{1}{|M|} \sum_{i=1}^{|M|} \frac{2 \cdot \text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \tag{2}$$

Using these two metrics allows us to apply a consistent measurement approach, as established by [26], and thereby illuminate two different facets of classification performance. The F1 Micro (1) considers all predictions comprehensively, while the F1 Macro (2) treats each class independently by calculating a weighted average of the F1 scores for each class. In Equation 2, we denote a general set of classes M. In our tasks, the classes are defined by the set L representing levels of trustworthiness for the *Trustworthiness Level* Detection and the set T representing topics for the *Topic* Detection.

3 DATASET

This section outlines the steps leading to the final dataset of articles used in our experiments: (i) selecting a representative list of online media, i.e., the N set of newspapers introduced above, (ii) retrieving the textual content of the articles published by the selected sources, i.e., the A set of articles, and (iii) data cleaning.

3.1 Online Media Outlets Selection

The goal of this selection process is to create a fair and representative list of online media that accurately reflects real-world conditions and allows for effective testing of our models. To achieve this goal, we started with the NewsGuard dataset of tagged online news sources, which is available to authors under the NewsGuard license. This dataset contains news sources that have been rigorously evaluated by expert journalists. Notably, the news sources within this dataset have significant impact, collectively contributing to 95% of online engagement. Our focus within this dataset is on specific source-level attributes, namely, 'topics' and 'trustworthiness scores'. We have identified four key topics for the analysis: *Political news or commentary*, *Conspiracy theories or hoaxes*, *Sports and athletics*, and *Health or medical information*, i.e., the set T introduced in Section 2. Starting with the original dataset, we first selected the sources associated with one and only one topic.

Now we want to obtain a list of news sources, 10 per topic, that preserves the distribution of NewsGuard's trust scores for each selected topic. In other words, we want to obtain a set of news sources that represent the true distribution of trust scores in the original dataset. We use a stratified sampling approach based on the trustworthiness levels, i.e., the set L of Section 2, defined by NewsGuard (see Table 1).

For example, let the reader consider the online sources related to the topic *Sports and athletics*. The distribution of trustworthiness scores for these sources is as follows: 30% of them have a trustworthiness score of 100; 50% fall within the range 75-99, and 20% fall within the range 60-74. So, to keep the original distribution, we randomly take 3 sources with l=100, 5 sources with l=75-99, and 2 sources with l=60-74.

Assuming that the original NewsGuard dataset accurately represents the actual online media landscape, our stratified sampling approach allows us to obtain 40 sources that closely match and mirror the original NewsGuard dataset in terms of the distribution

of trustworthiness across the four specified topics. Finally, we exclude sources with a paywall and those that are not in English. This represents our set N of newspapers.

Table 1: Newsguard Trustworthiness Levels

Score	Description
100	High Credibility
75-99	Generally Credible
60-74	Credible with Exceptions
40-59	Proceed with Caution
0-39	Proceed with Maximum Caution

3.2 Articles collection

The next stage is to collect a sample of articles for each selected source. Specifically, we collect the textual content of the most recently published articles on the sources' websites. The goal is to generate the set A of articles presented in 2.

To accomplish this task, we follow a multi-step process. We compile a list of URLs corresponding to the published articles. From the homepage addresses of the media outlets reported in the NewsGuard dataset, we manually identify the web page that contains the list of published articles (the so-called 'news history' page).

We use Selenium library[10] to develop a script that automatically scrolls through these web pages to collect the necessary URLs. For example, if we look at the website of *The Sun*[11], the news history page is available at: *https://www.the-sun.com/news/us-news/**page/1**/*. The developed application systematically jumps to the next pages by increasing the number in the URL. In the example, it becomes *https://www.the-sun.com/news/us-news/**page/2**/*, and for each of these pages, it collects all the URLs related to individual news articles. Once we have the URL lists for each source's articles, we retrieve the HTML content of each article page referenced in the URL list. Specifically, we utilize the GNU Wget[12] command to retrieve and save the complete HTML content of these pages in WARC-format archives. This ensures a readily accessible offline copy of the HTML pages, facilitating subsequent textual content extraction. The final step involves the creation of custom text extractors tailored to each news source. These extractors utilize XPATH-based information to effectively extract the embedded text from each article's web page. We do not consider news articles with less than 200 words; we acquire at least 40 articles per news outlet, with a maximum of 294. Thus, out of 5006 articles, we collected the texts of 4033 of them. The resulting dataset A, as depicted in Figure 1 (top and bottom) and Table 2, comprises articles extracted from the selected media outlets N during the period spanning from May 4th to May 15th, 2023. The figures provide an overview of the distribution of l per topic for the 40 news outlets. The *Conspiracy* topic is entirely dominated by articles falling within the lowest score range. For *Health or medical information* and *Political news or commentary*, the data show a more evenly distributed pattern, with a prevalence in the l=75-99 score range for the number of sources

[10] https://www.selenium.dev/
[11] https://www.the-sun.com/
[12] https://www.gnu.org/software/wget/

Table 2: Number of Articles per trustworthiness level, broken down by topic

Levels	Political	Conspiracy	Sports	Health
0 - 39	204	900	0	183
40 - 59	162	0	0	79
60 - 74	51	0	83	146
75 - 99	579	0	849	269
100	59	0	314	155
Total	**1055**	**900**	**1246**	**832**

in the latter topic. *Sports and athletics*, which includes the largest number of articles, 1246, presents a high concentration of elements in the l=75-99 interval.

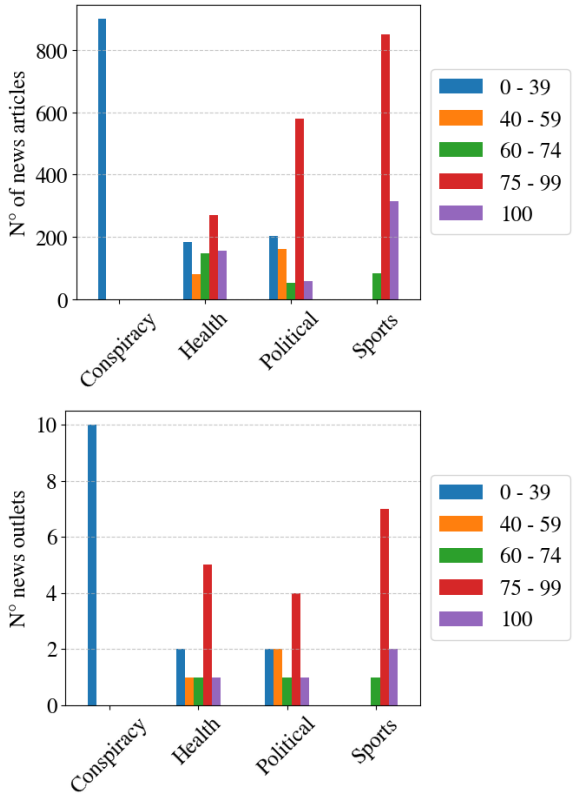

Figure 1: Number of articles (top) and news outlets (bottom) per trustworthiness level, broken down by topic.

3.3 Data cleaning

While our XPATH-based extraction method (see Subsection 3.2) provides flexibility by adapting to how each source organizes article content on its site, it often results in the inclusion of extraneous text fragments. These can be divided into two groups: (i) repetitive phrases (e.g., thank you messages, signatures, slogans) and (ii) other miscellaneous linguistic elements (e.g. dates).

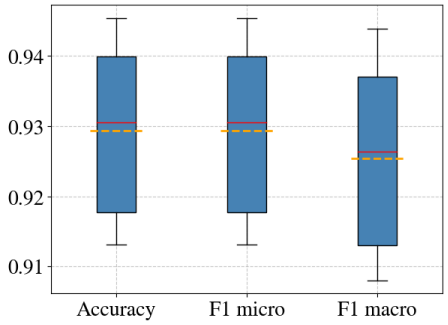

Figure 2: Evaluation results for topic detection. The yellow dotted line represents the average, while the red line represents the median.

We therefore clean the data by applying both *Spacy*[13], specifically by utilizing the "en_core_web_lg" model, and manual verification. Statistics about the final dataset are in Table 2. This dataset represents the set A of articles of the problem definition (see Section 2).

4 RESULTS AND DISCUSSION

We present the experimental setup and results for the tasks defined in Section 2, specifically, *topic* and *trustworthiness* detection. We use BERT [7], the well-known state-of-the-art pre-trained language model. We use the *transformers*[14] library in Python to deploy and fine-tune BERT, as well as to compute performance metrics. In particular, we adopt *BertForSequenceClassification* because it combines the capabilities of a highly trained language model with the adaptability to address specific tasks. For the validation, we adopt the 10-fold stratified cross-validation implemented by *scikit-learn*[15], thus using 10% of the dataset as test and 90% as training in each step. This choice guarantees an even distribution of the target class across each fold, which ensures a more robust evaluation.

4.1 Topic detection

Before we continue with the multi-class classification task, we show the results of the topic detection task. We remind the reader that the execution of this task and the evaluation of its results should not be considered as the main result of this article. The task is valuable for understanding the quality of the dataset annotations, which we did not provide ourselves.

Figure 2 shows the evaluation results on the 10 stratified folds. We achieve an average F1-macro of 0.925 (min = 0.908 and max = 0.943) and an average F1-micro of 0.929 (min = 0.913 and max = 0.945). For completeness, we also report the accuracy, precision and recall values in Figure 6 in the appendix, which are quite high for each topic.

Figure 3 (top) shows the confusion matrix for the fold with the worst F1 macro score. The errors made by the classifier are mainly due to the misclassification of *Political news or commentary* articles as *Conspiracy theories or hoaxes*. This result is not surprising, since

[13]https://spacy.io/
[14]https://github.com/huggingface/transformers
[15]https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html

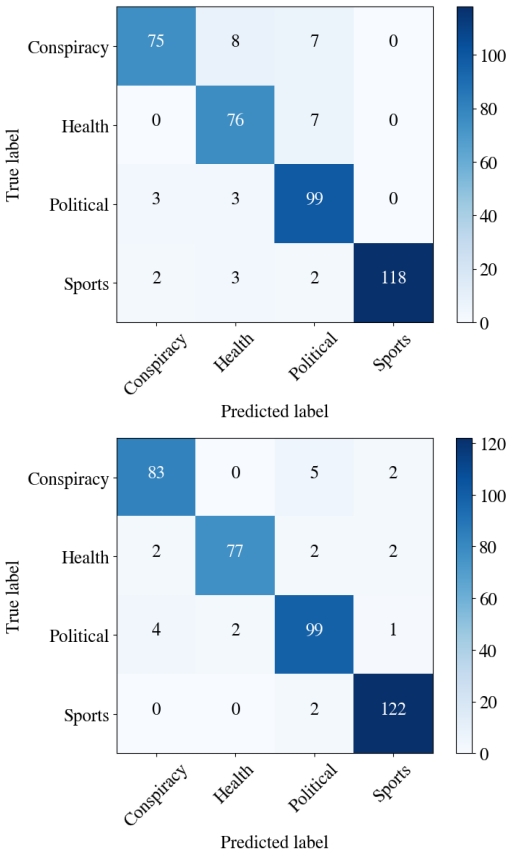

Figure 3: Topic: Confusion matrix for the fold with the lowest (top) and the highest (bottom) F1 macro

the lines between legitimate political news and conspiracy theories can be blurred by information manipulation strategies [8]. Also, some conspiracy articles are mistakenly categorized as related to health or medical information. This is not surprising since conspiracy theories often touch on topics related to public health, a phenomenon that has become more pronounced during and after the COVID-19 global pandemic [29]. This level of misclassification is not observed when examining the best-performing fold (Figure 3 bottom), where errors still exist, albeit to a lesser extent. Despite some inaccuracies, we argue that these results are satisfactory in that there is a very good match between articles and assigned topics.

4.2 Trustworthiness Detection

This section presents the results of detecting articles' trustworthiness level as defined in Section 2. The experiments are performed on the final dataset described in Section 3, following the methods described at the beginning of this section.

The primary goal of this task is to develop a classifier capable of assigning a level of trustworthiness (trust_a) to each article (a) based on its textual content (text_a). These trustworthiness levels, which include five different categories identified and assigned by NewsGuard, provide a nuanced characterization of publisher-level

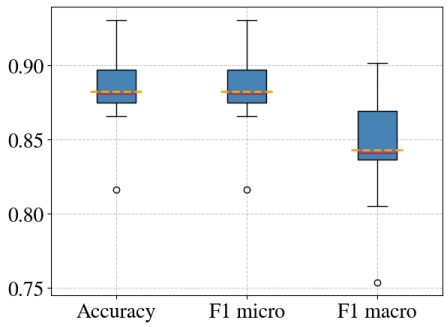

Figure 4: Evaluation results for trustworthiness level detection. The yellow dotted line represents the average, while the red line represents the median.

trustworthiness (see Table 1). This is a multi-class classification task at the article level.

Our results, shown in Figure 4, demonstrate the strong capability of the model to accurately associate the article to one of the five trustworthiness levels. The model achieves an average F1-macro of 0.843 (min = 0.753 and max = 0.901) and an average F1-micro of 0.882 (min = 0.816 and max = 0.930).

We analyze the confusion matrices associated with the best and worst Macro F1 scores to gain deeper insight into the results. Figure 5 (top) shows the confusion matrix for the fold associated with the lowest Macro F1 score. Here we can observe two errors: 49 items assigned to neighboring classes and 25 items assigned to classes significantly different from the true ones. In the best case, as shown in Figure 5 (bottom), the situation is characterized by lower values, with 17 items assigned to adjacent classes and 11 to distant classes.

The importance of the two types of errors can vary depending on how we want to use the model. As mentioned earlier, the trustworthiness levels represent a nuanced characterization of trustworthiness. There may also be situations where a coarser classification is desired. For example, we might consider redefining NewsGuard's thresholds to produce only two levels of trustworthiness: the $0-59$ and $60-100$ ranges to identify untrusted and trusted publishers. Redefining the thresholds to create coarser levels of trustworthiness can improve the performance of our model, thereby increasing its practical utility in real-world scenarios.

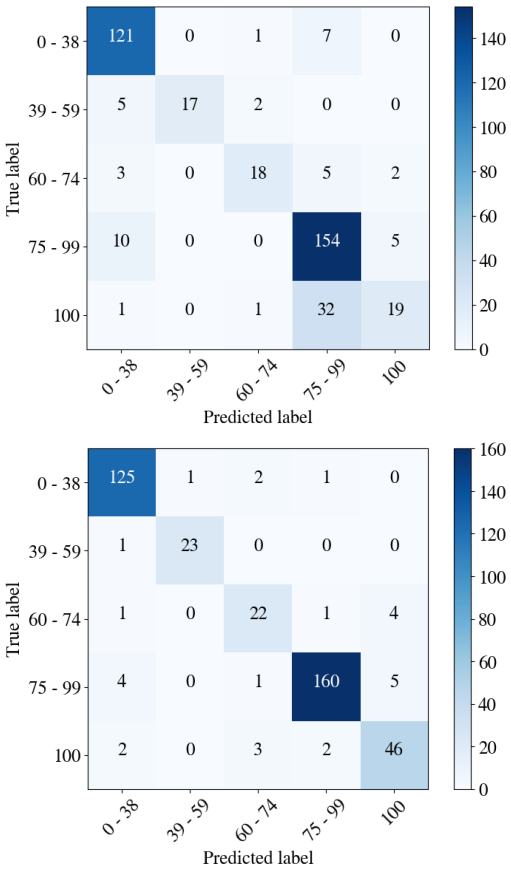

Figure 5: Trustworthiness level: Confusion matrix for the fold with the lowest (top) and highest (bottom) F1 macro

5 RELATED WORK

Online news consumption is now vital for insights into societal and cultural trends. Automated analysis, particularly through machine/deep learning, proves crucial in categorizing web-based news. A recent survey [23] identified 51 studies (2000-2019) using supervised and unsupervised learning and generative models.

Early work was primarily aimed at testing the ability of traditional supervised machine learning tools, such as Support Vector Machines, to recognize news topics. The analysis was conducted on texts in English as well as in other languages, such as German [30], and the features used were mostly n-grams or whole words extracted via the BoW method [11].

Among the works using 'traditional' approaches, an interesting work is [6], where the authors aim not so much to identify the topic of the news item as to identify its structure. In publishing, a news story is defined, for example, as having an 'Inverted Pyramid' structure when there is a so-called *lede* that introduces the story and a body section that, in an expository style, sets out other facts of the news. The paper considers both literal features and the writing style of the news to classify them according to their structure, e.g., the Inverted Pyramid, the Martini Glass, and the Kabob structures [13].

As we approach today, research focuses not only on news topics and structure recognition but also on detecting propagandistic, biased or untruthful content. This is the case with various challenges launched over the years, such as, for example, the Semantic Evaluation series, *Semeval* [20], and CheckThat!@CLEF, which, for example, in the 2022 edition dealt with the infodemic phenomenon developed during the Covid-19 pandemic, launching tasks on recognizing tweets worth checking based on their veracity, and useful for fact checking others [22]. Task 3 of the 2023 edition of Semeval 2023 [26], entitled 'SemEval-2023 Task 3: Detecting the Category, the Framing, and the Persuasion Techniques in Online News in a Multi-lingual Setup', extended the search for persuasion techniques in the news to include two additional types of categorization, *category* (namely, whether the text read represents opinion, objective

news or satire) and *framing*[16]. Although it is not the focus of our paper, the topic detection task we performed in Section 4.1 to assess the good match of topics to articles in our dataset can be considered, to some extent, a framing detection task. A conspiracy theory could be presented with a political, sports, or health framing. Conversely, news about politics, sports, and health might be written in a 'conspiracy' style and be associated with conspiracies. Nevertheless, SemEval-2023 Task 3 is by no means comparable with our work, the former being much more complex, with 14 possible frames, with articles associable with more than one frame, and multilingual.

On the Evaluation of News Publisher's Trustworthiness . The publisher's trustworthiness is a feature most often used to discern between true and fake news [1, 15]. Assessing this value is crucial for studying the spread of misinformation online [17]: trustworthiness values can be used by social platforms to limit reader exposure to content from untrustworthy sources[17] and, conversely, highlight credible sources of information [16]. Since the procedure of assessing a publisher's trustworthiness is time-consuming and traditionally requires experienced annotators [24], in this paper, we tried to approach the problem through the automatic classification of articles. Similar work to ours is in [2], where articles are annotated by hand (while we rely on the external source NewsGuard), and Bert is used for classification, amongst other language models. Like our work, this also considers multiclass classification at the article level. The articles were indeed associated with 4 trustworthiness intervals. Unfortunately, a fair comparison between our results and their performance is impossible because the articles in [2] are in Czech (the best classification performance in F1 is 0.52). The work in [28] considers a labelled news dataset in much the same way as ours: the tag of the individual news item is inherited from the trustworthiness of its source. Specifically, untrustworthy sources are collected from Politifact[18] while trustworthy sources are extracted from a study conducted by the Pew Research Center in [21]. The classification task uses multiple models, including BERT, and the accuracy achieved in testing is 0.99 for articles whose source has already been known by the model under training. The strong difference between [28] and the current work lies in the fact that in the former, the classification is binary, trustworthy vs. untrustworthy, whereas we consider a multiclass classification task.

On a very inspiring final note, a recent paper tested Chat-GPT's API to rank the reputability of an online news publisher, and the model obtains an Area Under the Curve of 0.89 in a binary ranking scenario (trusted/not trusted) [32]. Of course, the judgments are based on sources known to the model during training, but the first results suggest a LLM can also help analysts with the task of ranking the reputability of an online media outlet.

6 CONCLUSIONS

This paper examines the quality of the online news landscape, with particular reference to the level of trustworthiness of the news source. Many organizations, often formed by journalists and communication experts, have been trying for years to guide readers to read online media more trustworthyly by assigning trustworthiness ratings to various online newspapers. Of course, this process requires experienced annotators and is time-consuming. In this paper, we have tried to speed up the work of these organizations by evaluating the quality of an automatic ranking of an article's trustworthiness. The results are very promising when compared to the few existing related works.

Our approach is not intended to replace the careful procedures of journalistic organizations that invest much time and manpower in ranking online news media. Instead, as introduced at the beginning of the paper, we believe that the proposed article ranking can provide such organizations with initial guidance, both in selecting articles for human annotators to analyze and gaining insight into completely unfamiliar media outlets. In addition, our model can suggest to users the similarity - or otherwise - of the news they are reading to news from less reputable sources.

Our work contributes to real-world applications to combat the spread and impact of low-credibility content. However, the proposed work can be extended in three main directions: i) enriching the dataset with more sources, including non-English ones, and adding more articles per source; ii) exploring models other than BERT; and iii) integrating eXplainable Artificial Intelligence (XAI) techniques to understand textual differences between articles with different levels of trustworthiness and whether there is a specific reason why some sources are not classified correctly.

[16] In communication research, 'to frame' means 'to select some aspects of a perceived reality and make them more salient in a communicating text' [5, 9, 19].
[17] https://transparency.fb.com/it-it/policies/improving/timeline/
[18] https://www.politifact.com

ACKNOWLEDGMENTS

Work partially supported by project SERICS (PE00000014) under the NRRP MUR program funded by the EU - NGEU; by the Integrated Activity Project TOFFEe (TOols for Fighting FakEs) https://toffee.imtlucca.it/; by the IIT-CNR funded Project re-DESIRE (DissEmination of ScIentific REsults 2.0).

REFERENCES

[1] Parisa Bazmi, Masoud Asadpour, and Azadeh Shakery. 2023. Multi-view co-attention network for fake news detection by modeling topic-specific user and news source credibility. *Information Processing & Management* 60, 1 (2023).
[2] Matyáš Boháček, Michal Bravanskỳ, Filip Trhlík, and Václav Moravec. 2022. Fine-grained Czech News Article Dataset: An Interdisciplinary Approach to Trustworthiness Analysis. *arXiv preprint arXiv:2212.08550* (2022).
[3] S. Bowman and C Willis. 2003. We Media: How Audiences are Shaping the Future of News and Information. The Media Center at the American Press Institute..
[4] Michael Butter and Peter Knight. 2021. *Routledge Handbook of Conspiracy Theories.* Routledge.
[5] Dallas Card et al. 2015. The Media Frames Corpus: Annotations of Frames Across Issues. In *53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers).* Association for Computational Linguistics, 438–444. https://doi.org/10.3115/v1/P15-2072
[6] Zeyu Dai, Himanshu Taneja, and Ruihong Huang. 2018. Fine-grained Structure-based News Genre Categorization. In *Events and Stories in the News.* Association for Computational Linguistics, 61–67. https://aclanthology.org/W18-4308
[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
[8] Karen M Douglas, Joseph E Uscinski, Robbie M Sutton, Aleksandra Cichocka, Turkay Nefes, Chee Siang Ang, and Farzin Deravi. 2019. Understanding conspiracy theories. *Political psychology* 40 (2019), 3–35.
[9] Robert M. Entman. 1993. Framing: Toward Clarification of a Fractured Paradigm. *Journal of Communication* 43, 4 (1993), 51–58. https://doi.org/10.1111/j.1460-2466.1993.tb01304.x
[10] Forbes Business Council. 2022. Self-Publishing Versus Traditional Publishing: Pros And Cons For Leaders To Consider. https://www.forbes.com/sites/forbesbusinesscouncil/2022/08/15/self-publishing-versus-traditional-publishing-pros-and-cons-for-leaders-to-consider/. Accessed: 2023-12-22.

[11] Yoav Goldberg and Graeme Hirst. 2017. *Neural Network Methods in Natural Language Processing*. Morgan & Claypool Publishers.
[12] Paul Hawken. 1983. *The Next Economy*. Nenry Holt & co., New York, NY.
[13] Bahareh Heravi. 2022. Storytelling Structures in Data Journalism: Introducing the Water Tower structure. In *Computation+ Journalism 2022*.
[14] Dariusz Jemielniak and Aleksandra Przegalinska. 2020. *Collaborative Society*. MIT Press.
[15] Antino Kim, Patricia L. Moravec, and Alan R. Dennis. 2019. Combating Fake News on Social Media with Source Ratings: The Effects of User and Expert Reputation Ratings. *Journal of Management Information Systems* 36, 3 (2019), 931–968. https://doi.org/10.1080/07421222.2019.1628921
[16] Nadarevic L., R. Reber, A.J. Helmecke, et al. 2020. Perceived truth of statements and simulated social media postings: an experimental investigation of source credibility, repeated exposure, and presentation format. *Cogn. Research* 5, 56 (2020). https://doi.org/10.1186/s41235-020-00251-4
[17] David M. J. Lazer et al. 2018. The science of fake news. *Science* 359, 6380 (2018), 1094–1096. https://doi.org/10.1126/science.aao2998
[18] Hause Lin, Jana Lasser, Stephan Lewandowsky, Rocky Cole, Andrew Gully, David G Rand, and Gordon Pennycook. 2023. High level of correspondence across different news domain quality rating sets. *PNAS Nexus* 2, 9 (2023).
[19] Siyi Liu et al. 2019. Detecting Frames in News Headlines and Its Application to Analyzing News Framing Trends Surrounding U.S. Gun Violence. In *Computational Natural Language Learning (CoNLL)*. Association for Computational Linguistics, 504–514. https://doi.org/10.18653/v1/K19-1047
[20] Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles. In *Semantic Evaluation, SemEval@COLING*. International Committee for Computational Linguistics, 1377–1414. https://doi.org/10.18653/v1/2020.semeval-1.186
[21] Amy Mitchell, Jeffrey Gottfried, Jocelyn Kiley, and Katerina Eva Matsa. 2014. Political polarization & media habits. (2014).
[22] Preslav Nakov et al. 2022. Overview of the CLEF–2022 CheckThat! Lab on Fighting the COVID-19 Infodemic and Fake News Detection. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. Springer, 495–520.
[23] Mauricio Pandolfi-Gonzalez, Christian Quesada-Lopez, Alexandra Martinez, and Marcelo Jenkins. 2020. Automatic Classification of Web News: A Systematic Mapping Study. In *IntelliSys 2020: Intelligent Systems and Applications*. 558–574.
[24] Gordon Pennycook and David G. Rand. 2019. Fighting misinformation on social media using crowdsourced judgments of news source quality. *National Academy of Sciences* 116, 7 (2019), 2521–2526. https://doi.org/10.1073/pnas.1806781116
[25] M Petrocchi and A Spognardi. 2022. The Online News Market in Italy. Online: https://www.disinformationindex.org/country-studies/2022-1-31-the-online-news-market-in-italy/.
[26] Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023. SemEval-2023 Task 3: Detecting the Category, the Framing, and the Persuasion Techniques in Online News in a Multi-lingual Setup. In *Semantic Evaluation (SemEval-2023)*. Association for Computational Linguistics, 2343–2361. https://doi.org/10.18653/v1/2023.semeval-1.317
[27] Manuel Pratelli and Marinella Petrocchi. 2022. A Structured Analysis of Journalistic Evaluations for News Source Reliability. In *Workshop Proceedings of the 16th International AAAI Conference on Web and Social Media*. https://doi.org/10.36190/2022.51
[28] Piotr Przybyla. 2020. Capturing the Style of Fake News. In *Conference on Artificial Intelligence*. AAAI Press, 490–497. https://doi.org/10.1609/aaai.v34i01.5386
[29] Lotte Pummerer, Robert Böhm, Lau Lilleholt, Kevin Winter, Ingo Zettler, and Kai Sassenberg. 2021. Societal effects of COVID-19 conspiracy theories. *Social Psychological and Personality Science* (2021).
[30] M. Scharkow. 2013. Thematic content analysis using supervised machine learning: An empirical evaluation using German online news. *Quality and Quantity* 47 (2013), 761–773. https://doi.org/10.1007/s11135-011-9545-7
[31] Galen Stockinh et al. 2022. *The Role of Alternative Social Media in the News and Information Environment*. Pew Research Center. Retrieved October 10, 2023 from https://www.pewresearch.org/journalism/2022/10/06/the-role-of-alternative-social-media-in-the-news-and-information-environment/
[32] Kai-Cheng Yang and Filippo Menczer. 2023. Large language models can rate news outlet credibility. *CoRR* abs/2304.00228 (2023). https://doi.org/10.48550/arXiv.2304.00228

7 APPENDIX

The trustworthiness classification model is available here.

For the convenience of the reader, Table 3 lists the domains of the news outlets used to create our dataset. Topic and source trustworthiness tags cannot be released because they are proprietary to NewsGuard and licensed to the paper's authors. We would be happy to agree with NewsGuard to release some of the tags. Figure 6 and Figure 7 show the performances obtained for the Topic Classification Task and the Trustworthiness Classification Task regarding Accuracy, Precision, Recall and F1.

Table 3: List of news domains considered in the final dataset

Domain names		
12up.com	aclu.org	americanprogress.org
bizpacreview.com	californiahealthline.org	cbssports.com
celebritiesdeaths.com	clutchpoints.com	nih.gov
news-front.info	ewg.org	famadillo.com
flagandcross.com	harmonyhustle.com	historyfact.in
labourlist.org	nationalrighttolifenews.org	now8news.com
nowtheendbegins.com	on3.com	outsideonline.com
politichome.com	powerofpositivity.com	psypost.org
pulsetoday.co.uk	realrawnews.com	sbnation.com
scoopearth.com	skepticalraptor.com	sportscasting.com
theamericanmirror.com	thecovidblog.com	thegrayzone.com
thelibertytimes.com	thepatriotjournal.com	theplayerstribune.com
theringer.com	thetentacle.com	tnewsnetwork.com
trendingpoliticsnews.com	truthdig.com	unz.com
wavefunction.info	consciousreminder.com	countylocalnews.com
deadspin.com	dreddymd.com	drjockers.com

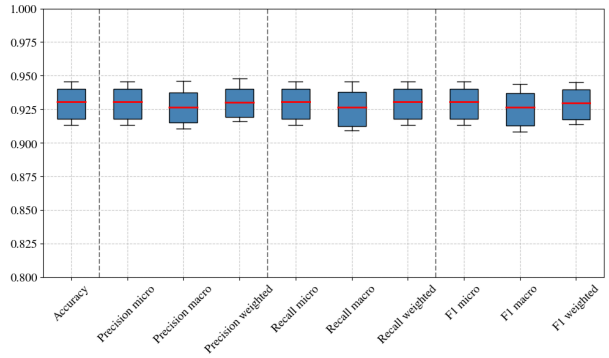

Figure 6: Complete evaluation results for Topic Detection

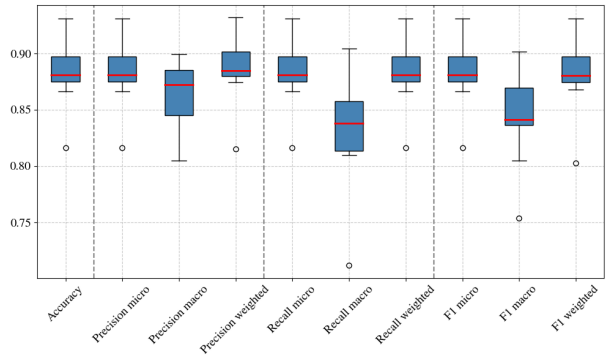

Figure 7: Complete evaluation results for Trustworthiness Detection