Yesterday's News: Benchmarking Multi-Dimensional Out-of-Distribution Generalization of Misinformation Detection Models

Ivo Verhoeven[1*], Pushkar Mishra[2], and Ekaterina Shutova[1]

[1]ILLC, University of Amsterdam
i.o.verhoeven@uva.nl
[2]Meta AI, London

This article introduces `misinfo-general`*, a benchmark dataset for evaluating misinformation models' ability to perform out-of-distribution generalization. Misinformation changes rapidly, much more quickly than moderators can annotate at scale, resulting in a shift between the training and inference data distributions. As a result, misinformation detectors need to be able to perform out-of-distribution generalization, an attribute they currently lack. Our benchmark uses distant labeling to enable simulating covariate shifts in misinformation content. We identify time, event, topic, publisher, political bias, and misinformation type as important axes for generalization, and we evaluate a common class of baseline models on each. Using article metadata, we show how this model fails desiderata, which is not necessarily obvious from classification metrics. Finally, we analyze properties of the data to ensure limited presence of modelling shortcuts. We make the dataset and accompanying code publicly available.*[1]

1. Introduction

The field of misinformation detection aims to develop classification models that reliably moderate online content. Despite burgeoning academic interest (Wu et al. 2019; Zhou and Zafarani 2020) in a multitude of fields (Kruijver et al. 2025), and impressive classification results on existing datasets, mis- and disinformation content continues to propagate online and cause significant societal harm.

The rapid evolution of online content—which significantly outpaces model development cycles—partially explains this. News, and more generally all online content,

* Corresponding author.

[1] `https://github.com/ioverho/misinfo-general`.

Action Editor: Wei Gao. Submission received: 26 May 2025; revised version received: 14 November 2025; accepted for publication: 23 November 2025.

https://doi.org/10.1162/COLI.a.585

is valued primarily for its novelty. It will often contain unseen entities, events, and entity-event relationships. Further exacerbating issues is the fact that what does or does not constitute misinformation is constantly changing, and highly dependent on the perspective of the labeler (Yee 2025). Verifying content manually using domain experts is prohibitively expensive, and typically requires context not available when news is emerging.

The misinformation datasets used for misinformation detector training, however, are collections of yesterday's news. These datasets typically have a narrow focus on particular events or misinformation forms, and are collected well after the fact. Consequently, state-of-the-art moderation systems lag behind the news landscape, and will encounter inference-time data distributions that have shifted away from the distribution of the training data. For misinformation detection to be successful at mitigating harms during deployment, and especially in situations with limited availability of social or historical context (i.e., when news is emerging), models will need to be robust to many forms of distribution shifts.

Currently, this property of Out-of-Distribution (OoD) generalization is lacking in many state-of-the-art (SoTA) NLP models, and especially misinformation detection models. Performance significantly degrades when evaluated on unseen:

- time periods (Bozarth and Budak 2020; Horne, Nørregaard, and Adali 2020; Kochkina et al. 2023; Stepanova and Ross 2023),
- publishers (Rashkin et al. 2017; Zhou et al. 2021),
- events (Lee et al. 2021; Cheng, Nazarian, and Bogdan 2020; Ding et al. 2022; Wu and Hooi 2022),
- topics (Przybyla 2020),
- domains (Hoy and Koulouri 2022; Kochkina et al. 2023; Verhoeven et al. 2024), and
- cultures or languages (Horne, Gruppi, and Adalı 2020; Chu, Xie, and Wang 2021; Ozcelik et al. 2023).

We primarily attribute this to the present state of misinformation datasets. While plentiful, these are often small, collected over a short time span, centered around specific events, biased towards popular content, or contain a too homogenous set of publishers. These properties are generally believed to be detrimental to the generalization capabilities of modern NLP models, which require large, diverse pre-training datasets, especially when text or labels are noisy.

Creating a dataset that *does* enforce OoD generalization, however, is not easy. Given the expense involved in collecting these datasets, prior attempts at doing so have invariably had to make trade-offs in size, diversity, or label fidelity (see Section 3). As a result, these datasets are not representative of the misinformation landscape, and evaluation with such datasets will overestimate model performance during deployment (Aïmeur, Amri, and Brassard 2023; Xiao and Mayer 2024; Kuntur et al. 2024). Generating high-quality misinformation labels for a realistically sized, naturalistic dataset remains intractable due to the cost of domain experts and the inherent subjectivity present in online content. This article will not solve this problem. Instead, we focus on more accurately estimating a model's robustness to expected distributional shifts.

Specifically, we present `misinfo-general`, a dataset meant for testing the generalization performance of automated misinformation detectors holistically. We do so by processing a distantly labeled series of corpora intended for publisher reliability labeling. While this introduces noise into the labels, we argue that the scale and diversity of the data make it useful for *generalizability* evaluation. To mitigate said noise, we perform extensive pre-processing of the data (Section 4), and post-hoc testing of dataset properties (Section 8). This ensures a balance of article quantity and label quality, providing one with rich metadata for a heterogeneous set of publishers, across a long time-span, covering a multitude of events and topics.

To showcase the utility of such a benchmark, we identify and operationalize six generalization axes—(1) time, (2) specific events, (3) topics, (4) publisher style, (5) political bias, and (6) misinformation type. We then train a simple yet representative baseline model. We find that generalization to different classes of publishers is particularly challenging, whereas within-publisher variation across years is smaller than expected. Using the metadata available to us, we provide additional analysis of publisher-level determinants of performance, and find some undesirable model behaviors not discussed in prior literature: Less-frequent publishers see degraded performance, and models treat different political biases differently. Juxtaposed to these results, we also find some initial evidence that the scale and diversity of this dataset can benefit model generalization ability when trained on.

2. Related Work

2.1 Generalizable Misinformation Detection

Generalization abilities of misinformation classifiers have been tested in many settings, at smaller scales. Horne, Nørregaard, and Adali (2020) found that performance degrades quickly when evaluating on future events, which Bozarth and Budak (2020) corroborate and extend to changes in domain. The same issues have also been reported in misinformation detection in other modalities (Stepanova and Ross 2023; Verhoeven et al. 2024). Zhou et al. (2021) find that models tend to overfit to publisher idiosyncrasies more than article content, especially in publisher-level annotated datasets.

Results on existing benchmark datasets are generally not indicative of downstream performance. Kochkina et al. (2023) found that performance within one dataset vastly overestimates performance on other datasets or time spans. Even when controlling for the time period or topic, Hoy and Koulouri (2022) found that models overfit to the training dataset and perform worse on similar but unseen datasets. In a recent systematic review of the literature, Xiao and Mayer (2024) come to the conclusion that:

> . . . detection tasks are often meaningfully distinct from the challenges that online services actually face. Datasets and model evaluation are often non-representative of real-world contexts, and evaluation frequently is not independent of model training. (p. 1)

This sentiment matches the earlier discussion in Aïmeur, Amri, and Brassard (2023); current misinformation benchmarks and evaluation setups can yield deceptively high performance scores.

Despite this paucity in benchmarks and labels, there has been some interest in developing generalizable or adaptive misinformation detection techniques. This has been attempted through weak supervision (Shu et al. 2020), multitask training (Lee

et al. 2021), utilizing external agents (Ding et al. 2022; Mosallanezhad et al. 2022), data resampling or active learning (Hu et al. 2023), adversarial learning (Lin et al. 2022), or gradient-based meta-learning (Zhang et al. 2021a; Yue et al. 2022). While these research directions are promising, their utility for out-of-distribution misinformation detection has not been sufficiently tested on large, diverse benchmark data.

2.1.1 Synthetic Distribution Shifts. This article focuses on exploring the robustness of automated misinformation detectors to natural distribution shifts, i.e., those one might expect to occur during transfer from training to deployment-time inference.

A related strand of research is the analysis of model performance under *synthetic* distribution shifts. These techniques can avoid the cost of collecting and extracting misinformation data, while elucidating model behavior under covariate shifts and adversarial attacks.

In general, misinformation classifiers have been found to be fragile against adversarial attacks (Zhou et al. 2019; Koenders et al. 2021). Przybyła, Shvets, and Saggion (2024) found that larger LMs were more fragile to data augmentation techniques that minimize semantic distance, while maximizing performance degradation. Despite this, those same LMs were successful in generating adversarial examples (Przybyła, McGill, and Saggion 2025). On the other hand, several recent works find that incorporating adversarial data augmentation techniques during training (Smith et al. 2021; Ahmed et al. 2024) can boost robustness. In extreme cases, LLM-generated misinformation is used as a proxy for sampled misinformation data in misinformation detector evaluation (Lucas et al. 2023), which theoretically should allow for fine-grained control over distribution shifts being tested.

2.2 Publisher Reliability Estimation

A related field to misinformation classification, especially when utilizing publisher-level labels, is publisher reliability estimation. Instead of yielding article level moderation decisions, a publisher reliability model uses the content of one or many articles from one publisher to yield a reliability estimate of the publisher as a whole. This is a relatively well-studied problem. At present, this is usually achieved through a mix of content-based (Rashkin et al. 2017; Bianchi et al. 2024) and metadata features (Baly et al. 2018a, b, 2019, 2020a, b; Nakov et al. 2024).

Relative to article-level misinformation classifiers, publisher-level classification can greatly reduce the computational cost needed for classification (Burdisso et al. 2024). However, this typically involves incorporating additional historical context, world-knowledge (Yang and Menczer 2025), or social context (Pratelli, Saracco, and Petrocchi 2024). This can make publisher reliability models *transductive* instead of *inductive* learners—moderation decisions come from specific prior experience rather than general rules.

This mimics how moderators or users might analyze the reliability of a publisher, potentially before ingesting the contents of a specific article. However, such approaches might fail in cases of where publishers are unknown, ambiguous, or evolving. In those situations, moderation decisions at the article-level is necessary. While `misinfo-general` is suited for either approach, we focus on testing the generalization of inductive article-level classifiers. These models naturally provide classification in cases where limited context or prior experience is available, and are required to utilize (non-spurious) general rules.

3. Biases in Misinformation Datasets

At risk of repetition: Misinformation models' performance degrades quickly under covariate distribution shifts expected to occur during model deployment, an observation whose cause we attribute to the datasets they were trained on. Due to the exorbitant cost of acquiring high-fidelity misinformation labels, misinformation datasets tend not to reflect the true variance in online (misinformation) content.

To illustrate this, we analyze common properties of datasets specifically constructed for the development of misinformation detectors, by ways of an inexhaustive, yet representative survey of existing misinformation datasets. We provide an overview of these datasets in Appendix A Table A.1. Furthermore, in this section, we (1) broadly categorize datasets into different labeling methods; (2) provide specific examples of how misinformation is collected and labeled; (3) discuss how these operationalizations can lead to biases in the datasets; and finally, (4) provide a discussion on the merits and demerits of publisher-level labeled datasets for the purposes of model generalization.

3.1 Dataset Labeling Granularity

Generally speaking, one can classify misinformation datasets into 3 annotation schemes. Listed from most fine-grained to most coarse-grained:

1. **Claim**: Experts fact-check individual (but complete) statements in isolation. Claims are usually small spans sourced from larger documents or utterances.
2. **Article**: Experts label the *overall* veracity of entire documents. These can contain many claims, whose factuality need not be consistent with each other.
3. **Publisher**: Experts label publishers for their propensity for factual reporting, based on historical records and prescribed authorial intent. These labels are often used as a proxy for finer-grained labels. The articles produced by publishers do not necessarily have the same label as the publisher.

The more fine-grained annotation methods yield high-quality labels, but can be prohibitively expensive to procure, or evaluate texts without the context those texts would naturally have. Furthermore, these labeling methods are typically forced to exclude unverifiable texts (e.g., highly subjective texts or opinions), despite these being prevalent in online discourse. On the other hand, the more coarse-grained annotation methods run the risk of introducing noise into the labels, by assuming consistency between finer-grained labels. For example, an article may contain many factual statements, but a single blatant lie. Since there are increasingly fewer units at each level, however, labels are far easier to procure.

3.2 Survey of Misinformation Datasets

In Appendix A Table A.1 we present various misinformation datasets, their labeling granularity, their size, and a description of how their data was sampled. In this subsection, we briefly expand on some common trends on how misinformation data and labels were sourced.

Claim-level annotations represent some of the oldest (LIE DETECTOR [Mihalcea and Strapparava 2009]) and largest (CREDBANK [Mitra and Gilbert 2015]) collections of misinformation text. The claims can be sourced from directly sampling social media (CREDBANK [Mitra and Gilbert 2015]) or sampling specific utterances flagged for review (LIAR [Wang 2017]; POLITIFACT-OSLO [Poldvere, Uddin, and Thomas 2023]).

While labels sourced from domain experts are dominant, using lay people as a method of crowdsourcing for either data or label collection has also proven popular. For the former, as an example, articles are collected only if these were flagged by (trusted) users of social media sites (WEIBO15 [Ma et al. 2016], WEIBO17 [Jin et al. 2017], WECHAT [Wang et al. 2020]). In some cases, lay volunteers were even used in the production of misinformation (LIE DETECTOR [Mihalcea and Strapparava 2009], FAKENEWSAMT [Pérez-Rosas et al. 2018]).

The benefit of crowdsourcing is clear; especially at the article level, datasets that use expert annotations (BUZZFEED-WEBIS [Potthast et al. 2018]; ALLCOTT & GENTZKOW [Allcott and Gentzkow 2017]; FAKENEWSCORPUS [Pathak and Srihari 2019]) tend to be much smaller than those leveraging crowdsourcing. A common strategy to combat this is to blend the Article and Publisher level labeling schemes (FAKENEWSNET/GOSSIPCOP [Shu et al. 2019]; FAKENEWSCORPUS [Pathak and Srihari 2019]; MM-COVID [Li et al. 2020]). Either factual or misinformation articles are manually verified, and the complementary class is sampled from a set of publishers commonly associated with misinformation or factual articles, respectively.

A similar strategy is to blend the Article and Claim level labeling schemes. A claim made in an article is annotated for veracity in annotation, and its label is propagated to the entirety of the article (FAKENEWSNET/POLITIFACT [Shu et al. 2019]; COAID [Cui and Lee 2020]; POLITIFACT-OSLO [Poldvere, Uddin, and Thomas 2023]).

The most consistent method for generating large, diverse corpora, however, proves to be using Publisher-level labeling (TSHP-17 [Rashkin et al. 2017]; KAGGLE FAKE NEWS [Risdal 2016]; SOME LIKE IT HOAX [Tacchini et al. 2017]; FAKE VS SATIRE [Golbeck et al. 2018]; QPROP [(Barrón-Cedeño et al. 2019)]), as discussed above.

Typically, the topics covered in the corpus are not further analyzed by dataset authors, although some datasets specifically focus on articles from various perspectives on the same events (MEDIAEVAL15 [Boididou et al. 2015]; PHEME [Zubiaga, Liakata, and Procter 2017]; BUZZFEED-WEBIS [Potthast et al. 2018]). In some cases, these instead influence the features used in automated misinformation classification (PRZYBYŁA CREDIBILITY [Przybyla 2020]).

Similarly, while most datasets are fairly general, some focus on specific domains. Very common are those focusing on social media or microblogging texts (CREDBANK [Mitra and Gilbert 2015]; MEDIAEVAL15 [Boididou et al. 2015]; WEIBO15 [Ma et al. 2016]; WEIBO17 [Jin et al. 2017]; WECHAT [Wang et al. 2020]). Another common domain involves celebrity rumors, typically annotated for verification rather than veracity, and also commonly sourced from social media posts (WEB DATASET CELEBRITY [Pérez-Rosas et al. 2018]; FAKENEWSNET/GOSSIPCOP [Shu et al. 2019]). During the COVID-19 pandemic, various health-related datasets were introduced (FAKEHEALTH [Dai, Sun, and Wang 2020]; MM-COVID [Li et al. 2020]; FAKECOVID [Shahi and Nandini 2020]; COAID [Cui and Lee 2020]).

3.3 Sources of Dataset Bias

In this subsection, we discuss how specific operationalizations can introduce bias in the dataset, adversely affecting model generalization performance.

Differing Definitions. Even among domain experts, there exists substantial disagreement on what does and does not constitute misinformation (Altay et al. 2023), with disagreements as to which degree content, medium, or intent is relevant to defining misinformation (Gelfert 2018; Yee 2025). Recent systematic reviews have found that this disagreement has carried over to the computer sciences (see for example Wu et al. [2019]; Oshikawa, Qian, and Wang [2020]; Zhou and Zafarani [2020]; Aïmeur, Amri, and Brassard [2023]; Bodaghi et al. [2024]; Xiao and Mayer [2024]). Indeed, the surveyed definitions of misinformation in Table A.1 seem to agree on basic properties of misinformation, but disagree on the specific forms. As a result, the forms of misinformation which are included can vary considerably. For example, misinformation forms like Satire and Propaganda are either explicitly included or excluded, proving to be especially divisive.

Inconsistent Label Sourcing. Another source of between-dataset variation is the source of misinformation labels. While most datasets rely on domain experts, some use lay volunteers to verify content, either explicitly (CREDBANK [Mitra and Gilbert 2015]; WEIBO15 [Ma et al. 2016]; WEIBO17 [Jin et al. 2017]) or implicitly (SOME LIKE IT HOAX [Tacchini et al. 2017]).

Recently, datasets have started using many misinformation sources (FAKECOVID [Shahi and Nandini 2020]; MUMIN [Nielsen et al. 2022]; MCFEND [Li et al. 2024]). These can come from different countries and cultures, some of which are likely to disagree on their misinformation definitions. Furthermore, this requires aggregating the different misinformation labeling formats.

Most misinformation definitions require specific authorial intent to deceive. However, in some datasets this is missing in the original content (LIE DETECTOR [Mihalcea and Strapparava 2009]; FAKENEWSAMT [Pérez-Rosas et al. 2018]), or ambiguous due to misinformation being defined as a lack of credible information (FAKEHEALTH [Dai, Sun, and Wang 2020]; PHEME [Zubiaga, Liakata, and Procter 2017]; FAKENEWSNET/GOSSIPCOP [Shu et al. 2019]).

Few Publishers. Many datasets limit the number of publishers in either class. In some cases, this is due to deliberate scoping of the dataset (BUZZFEED-WEBIS [Potthast et al. 2018]; FAKENEWSCORPUS [Pathak and Srihari 2019]); however in most cases this is due to publisher scarcity. Misinformation annotators, like Snopes, Politifact, GossipCop, etc., understandably tend to focus on verifiable misinformation pieces. As a result, datasets sampling annotations from these sources incur a large positive bias. A common strategy to counteract this is by including samples from a few mainstream publishers (TSHP-17 [Rashkin et al. 2017]; MM-COVID [Li et al. 2020]; COAID [Cui and Lee 2020]; FAKENEWSNET [Shu et al. 2019]).

An unwanted side effect of having a small, homogenous publisher set is the introduction of a modelling shortcut; misinformation classifiers no longer need to analyze the veracity or intent of input content, but rather simply discriminate between a few publishers with unique idiosyncrasies. Similarly, in datasets where misinformation is constructed by editing factual information (LIE DETECTOR [Mihalcea and Strapparava 2009]; FAKENEWSAMT [Pérez-Rosas et al. 2018]), the labels can be inferred by discriminating between the stylistic preferences of the original texts' authors and those of the editors.

Few Events or Topics. Similarly, many datasets sample content from a narrow time-span, or from a small set of events or topics. This can reduce the cost of generating labels, but will likely induce overfit in automated moderation systems trained on these corpora.

Focus on Obvious or Popular Misinformation. In several of the discussed datasets, misinformation texts are collected based on user reports, or from third-party fact-checkers. These run the risk of introducing a selection bias, resulting in a dataset that is not representative of all produced misinformation.

A secondary effect of this is that unverifiable content (e.g., those relying purely on opinion and speculation) are implicitly excluded. Some datasets explicitly exclude unverifiable content (FAKENEWSNET [Shu et al. 2019]), whereas others include this as a specific category (BUZZFEED-WEBIS [Potthast et al. 2018]). Most datasets do not discuss unverifiable cases, despite these forming a sizable part of produced online content (see Section 8.3).

Conclusion. In short, we find that the realities of misinformation data collection results in many datasets making a trade-off between label quality and corpus size. As a result, these datasets introduce some bias, which we suggest as a primary reason for the reported brittleness of misinformation detectors under covariate shift. Given that these covariate shifts are practically guaranteed in online content or news, testing misinformation detectors before deployment for generalizability is crucial. Doing so, however, requires large, diverse datasets, which we have established is difficult to procure without bias. A related task, publisher reliability estimation, might provide an alternative.

3.4 Publisher Reliability Datasets

Related to the task of misinformation detection is publisher reliability estimation (see Section 2.2). Given an article, or a set of articles, from some publisher, a reliability estimator has to predict the overall publisher reliability.

Publisher reliability is a broader concept than factuality, and considers many aspects of a publisher, which are not necessarily clear when analyzing articles or claims from a publisher in isolation. These aspects include framing, publisher political or editorial bias, intended audience, sourcing practices, funding, etc. All of these factors are analyzed on a large collection of a publishers works, and used to provide an indication of the trustworthiness of past and future releases. Ultimately, however, the factuality of produced articles is an important dimension of publisher reliability.

Much like the publisher-level misinformation labeling scheme discussed above, it does not preclude less reliable publishers producing reliable content, or vice versa. It merely suggests that this is less likely to occur. Reliable publishers often produce sensationalist or subjective content to draw in readership, whereas unreliable publishers might intersperse their less reliable articles with more reliable ones to boost their perceived trustworthiness.

Implicitly, by using publisher-level labels as a proxy for article-level reliability, we (as well as many Publisher-level datasets) make the assumption that the article-level factuality of an article from a reliable publisher is stochastically higher than that of an article from a less reliable publisher.

3.4.1 Measuring Generalization with Publisher Reliability Labels. Relative to misinformation datasets, for generalizability aspects, publisher-reliability datasets are far easier to produce at scale, and given publisher-level metadata, can be built specifically to enforce diversity in both publishers and text. Furthermore, articles can be collected across much longer time-spans, which naturally includes shifts in article topics or events.

Perhaps most importantly, however, if a large enough set of publishers is collected, the resulting dataset becomes a naturalistic view of published online content and (mis)information. Instead of a dataset including only verified misinformation, which are typically the least ambiguous or popular cases due to the selection bias of third-party fact-checkers, the dataset is more aligned with online content as it would appear post deployment (Section 8). As a result, statistics about model evaluation are more representative, and model developers can derive stronger conclusions.

In this article, we propose using a publisher-level reliability estimation dataset for the purpose of evaluating the generalizability of article-level misinformation detectors. While this runs the risk of tarring all articles from a publisher with the same brush, we believe the size and diversity of the dataset, along with access to publisher-level metadata, can offset the induced bias and still allow for conclusive inferences about model behavior under distribution shift.

Specifically, we assume that the effect of covariate distributional shifts on the predictive quality of a model is positively correlated between the two labeling approaches. In other words, we assume that model performance degrades under the same distributional shift in both labeling set-ups. Thus, implicitly, we assume that the level of robustness to distributional shifts on a dataset like `misinfo-general` serves as a good indicator for robustness in article-level misinformation detection.

4. The `misinfo-general` Dataset

Here we introduce `misinfo-general`, a benchmark for testing the generalization capacity of misinformation detection models, built on top of a series of noisy publisher-level datasets. While best suited for publisher reliability estimation models, we instead use the publisher labels as a proxy for article labels.

Based on the prior discussion, we foresee two sources of bias: (1) labels might not be accurate at the article level, and (2) models will learn to infer the article's publisher and its label instead of inferring the label from the article. We take the following steps to mitigate these biases as much as possible:

1. relabeling existing articles (Section 4.2)
2. masking or removing publisher identifiable text in articles (Section 4.3)
3. removing any article- and sentence-level duplicates (Section 4.3)
4. masking self-references, along with other private or identifiable information (Section 4.3)

In Section 8.1 we show that these pre-processing steps have made publisher identification from articles alone difficult.

In this section, we describe how we gather the dataset content and labels, and generate any additional metadata. Later sections make use of article-level metadata, and we specifically test for model overfit to publisher style (Section 5 and Section 6); we show that including a diverse set of publishers is beneficial to generalization performance (Section 7.2), and we try to find publishers with high degrees of mislabeling by assessing the necessity of model memorization (Section 8.2).

4.1 Article Provenance

All raw articles come from the various **Ne**ws **La**ndscape (NELA) corpora produced by the MELA lab[2] (Horne, Khedr, and Adali 2018; Nørregaard, Horne, and Adalı 2019; Gruppi, Horne, and Adalı 2020, 2021, 2022, 2023). The corpora cover 2017–2022 (6 iterations) almost continuously, with articles from a diverse group of publishers. In their original form, the 6 iterations together consist of 7.2 million long-form articles.

The original authors' goal was to study the dynamic behavior of news and news publishers. They deemed existing corpora inadequate for their goals, because of (1) a small, relatively homogenous collection of articles or publishers, (2) too narrow a focus on specific events, (3) bias towards popular publishers, and (4) limited ground truth labeling (Horne, Khedr, and Adali 2018; Nørregaard, Horne, and Adalı 2019).

4.2 Publisher Labeling

From the 2018 iteration onwards, the NELA datasets come with publisher-level labels. However, due to inconsistencies across dataset iterations and the frequency of labeling errors, we chose to relabel the dataset completely.

Similar to the initial NELA corpora labels, we scraped Media Bias/Fact Check[3] (MBFC). MBFC is a curated database of news publishers, with thorough analyses of publisher origins, bias, and credibility. Despite being run by lay volunteers, MBFC labels correlate well with professional fact-checking sources (Kiesel et al. 2019; Broniatowski et al. 2022; Pratelli and Petrocchi 2022). MBFC labels have been used in many earlier works (Rashkin et al. 2017; Baly et al. 2018a, 2020a; Burdisso et al. 2024; Casavantes et al. 2024; Szwoch et al. 2024). We use the metadata available as of October 2024, well after the final publication dates of articles in the corpus.

Using the URL domain of the scraped articles, we first mapped all articles to a consistent set of publishers before removing any publishers known to be news aggregators or social media sites. This gives an article-publisher mapping that is consistent across dataset iterations, and removes cases of where articles were republished on different sites. Each publisher was linked to a publisher in the scraped MBFC database. We provide further detail in Appendix B.1. Ultimately, we identified 488 distinct publishers, many of which were falsely attributed in NELA's original set of publishers. The metadata available for each publisher is provided in Appendix B.6.

The MBFC database is dynamic, and it does happen that the publisher label or metadata annotations change.[4] Usually, this presents as a relatively minor change in political bias. During the data collection and processing period (January 2017–October 2024), we found 20 instances where the change was substantive (see Appendix B.5 Table B.6). In the majority of cases (12/20), this resulted in a previously reliable$^+$ publisher failing too many fact checks, resulting in their rating being downgraded to unreliable$^-$. Ultimately, all these cases are due to additional information about the publishers' editorial practices coming to light, rather than those practices changing. In 5 cases publishers either corrected articles with failed fact checks or shifted their editorial policies, resulting in a label shift from unreliable$^-$ to reliable$^+$.

2 https://melalab.github.io
3 https://mediabiasfactcheck.com/.
4 See https://mediabiasfactcheck.com/changes-corrections/.

4.3 Data Processing

Beyond errors in the article-publisher and publisher-label mappings, the texts themselves frequently contain duplicates or scraping errors. Of the 6.7M re-labeled articles, roughly $\approx$ 22% or 1.5M articles were duplicates. Many of the remaining unique articles were deemed malformed or semantically void. These contain either very little text, substantial amounts of markup, or include too many special tokens to be human-readable. We filter these using a few simple rules (see Appendices B.2 and B.3). Altogether, we remove approximately $\approx$ 43% of all downloaded articles. The final dataset contains 4.2 million cleaned articles.

In the remaining texts, we mask various forms of private or identifiable information (PII), both to enhance safety and reduce the number of available classification "shortcuts". We furthermore standardize the copyright masking procedure introduced in Gruppi, Horne, and Adalı (2020). This introduces 4 new special tokens: `<copyright>` replacing NELA's repeated `@` tokens, `<twitter>`, `<url>`, and `<selfref>` for any self-references.

Despite our efforts, the datasets retain a level of "noise" customary to data sourced from the internet. For example, articles from the same publisher tend to contain unique by-lines, attribution messages, or donation requests. Further cleaning efforts might reduce the realism of the benchmark.

4.4 Topic Clustering

One of our aims is to test model generalization across different events and topics. To discover these, we used a modified variant of `BERTopic` (Grootendorst 2022) with a `gte-large`[5] (Li et al. 2023) backbone. This produced thousands of event clusters for every dataset iteration, each with a TF-IDF representation vector. We aggregate these events into overarching topics by applying spectral clustering to the adjacency matrix induced by the inter-event cosine similarity of the TF-IDF matrix. We arbitrarily limit the number of topics to 10, each with varying numbers of events in them. This process is further described in Appendix B.4.

This largely mimics the process used in Przybyla (2020), and extends the work of Litterer, Jurgens, and Card (2023) on identifying "news storms" in the NELA corpora to a larger time-span, and a larger set of publishers.

5. Generalization Taxonomy

In this section, we describe various dimensions along which we believe covariate shifts likely to occur, and which are feasible to simulate using `misinfo-general`. We consider a total of 6 specific generalization axes.

Time-based generalization measures the extent to which changes in publisher style affect a model's predictions. The publishers considered in each split should be held constant to avoid confounding with different publishers.

Evolution of article content will also impact performance. We focus on two specific forms of such change: (1) due to spontaneous **events**, which we define as news-worthy happenings with a definite and narrow time-span, or (2) due to evolving **topics**, which

5 `https://huggingface.co/thenlper/gte-large`.

Table 1
A schematic overview of the generalization taxonomy. The left columns provide relevant generalization category and axis, whereas the right columns provide examples of in domain and out-of-distribution article sets.

Generalization	Axis	In Distribution			Out-of-Distribution		
Time	Time	CNN 2018	AP 2018	Vox 2018	CNN 2017	AP 2020	Vox 2019
Content	Event	not COVID-19 events			COVID-19 events		
	Topic	Crime, Sports			Elections		
Publisher	Publisher	CNN	MSNBC	OANN	Reuters	AP	True Activist
	Political Bias	AP Centre	Reuters Centre	Fox News Right	Vox Left	Daily Beast Left	True Activist Left
	Misinfo Type	Vox Reliable	NYT Reliable	OANN Questionable	MSNBC Reliable	911Truth Conspiracy	Age of Autism Pseudosci.

we define as large, overarching collections of events that remain relatively static over a long period. Across these events and topics, we expect markedly different language.

The distribution of publishers is also expected to change between training and inference time. All *publishers* exhibit some form of editorial bias or style, which can be memorized by classification models. While models should use style to inform moderation decisions, they should also not overfit to stylistic idiosyncrasies. One related, usually implicit, expectation of misinformation detectors is a robustness to different *political biases* or *misinformation types*. Predictions ought to be based on a publisher's intent, not their norms and values. By excluding these from training, we can test a model's ability to generalize to different classes of publishers.

5.1 Data Splits

To operationalize these generalization axes, we build 6 (+1 baseline) train/test splits of the dataset using the publisher-level metadata available to us. Each split is meant to simulate one of the above described covariate shift scenarios, while ensuring minimal cross-scenario confounding.

Throughout, we approximate the same 70%/10%/20% article proportions per training/validation/test split, respectively. The validation split, used for early stopping, is sampled independently and identically distributed (i.i.d) from the training set. For all scenarios, we repeat each split independently for each dataset year, for a total of 6 times. The only exception is the Event axis, for which we combine all years into a single dataset (Table 1).

Briefly, we construct splits (schematically displayed in Table 2) for the scenarios as follows:

0. **Uniform**: Standard stratified random splitting of articles into disjoint article sets. No article meta-data is used.
1. **Time**: The training set consists of a single dataset year, while the test set contains articles from publishers seen during training in all other dataset years. This tests within publisher variation.
2. **Event**: The dataset has been annotated for several thousands of events, but we focus on a singular one: the COVID-19 pandemic. We reserve all

Table 2
Article-level classification performance comparing performance on the ID and OoD evaluation sets. The top row uses uniform splitting for both (OoD = ID), serving as a baseline value. Time based splitting has strongly varying class proportions, making F1 values inappropriate.

Generalization Form	MCC			F1 Reliable			F1 Unreliable		
	ID	*OoD*	Δ	*ID*	*OoD*	Δ	*ID*	*OoD*	Δ
Uniform	0.46	0.46	0.00	0.86	0.86	0.00	0.57	0.57	0.00
Time	0.46	0.33	−0.13			N/A			
Event	0.43	0.46	0.03	0.87	0.86	−0.01	0.52	0.55	0.03
Topic	0.46	0.38	−0.08	0.87	0.84	−0.03	0.56	0.50	−0.06
Publisher	0.48	0.37	−0.10	0.87	0.84	−0.03	0.58	0.53	−0.05
Political Bias									
Left	0.49	0.30	−0.19	0.85	0.87	0.02	0.61	0.38	−0.23
Right	0.56	0.19	−0.37	0.95	0.60	−0.34	0.58	0.26	−0.32
Misinformation Type									
Consp.-PSci.	0.43	0.42	−0.01	0.87	0.82	−0.05	0.53	0.53	0.01
Questionable	0.43	0.23	−0.20	0.94	0.62	−0.33	0.41	0.25	−0.16

articles containing any related keywords for testing, and we train on all non-COVID articles.

3. **Topic**: We reserve the k smallest topic clusters for the test set, such that these contain roughly 20% of all articles, and we train on the remaining articles.
4. **Publisher**: Similarly, we reserve the k least frequent publishers for the test set, such that these contain roughly 20% of all articles, and we train on the remaining articles.
5. **Political Bias**: We reserve all articles from either all Left- or Right-biased publishers for testing, and train on articles from the opposite political bias, along with any Center-biased publishers.
6. **Misinformation Type**: Similarly, we reserve all articles from either all Questionable Source or Conspiracy-Pseudoscience publishers for testing, and train on articles from the other misinformation class. We use an i.i.d. split of reliable articles to ensure a similar class distribution in all splits.

We include a substantially expanded description of each split's construction in Appendix C. The Topic and Publisher splits were constructed by sampling from the smallest topics and publishers. This was to ensure that all splits have approximately the same size while simultaneously maximizing the diversity of the held-out test sets. This could introduce a bias towards the more prolific publishers and topics; however, (1) this bias is already present in the training data (see Appendix E.3, Tables E.2 and E.3, parameter train count); and (2) we do not believe this had an undue amount of influence on the quality of the training models (see Section 7.2).

It is important to note that from the model's perspective, each scenario seems identical. The same labels are present in each split, with roughly the same article counts

in the same class proportions. Without additional context, one should expect similar performance across these splits.

6. Experiments and Results

To showcase the utility of `misinfo-general` for model training and evaluation, we use a simple yet powerful baseline model. Specifically, we fine-tune an instance of `DeBERTa-v3`[6] (He, Gao, and Chen 2022) where we reset the pooler and classification weights but freeze the model's remaining weights. To enable using dataset-specific tokens, we allow the token embedding layer to train with a very low learning rate. The model's pre-training data included a closed-source news dataset (CC-News), dated between September 2016 and February 2019 (Liu et al. 2020), and thus should be easily adapted to `misinfo-general`. Similar architectures have shown surprisingly adequate performance on other benchmark datasets, including various NELA versions (Pelrine, Danovitch, and Rabbany 2021; Zhou et al. 2021; Raza and Ding 2022).

We fine-tune the models on the different splits outlined in Section 5. We keep the hyperparameters and compute budget constant (which were tuned on the validation sets of the Uniform splits), which we outline in Appendix D. Training occurs at the article level, using publisher-level labels. We binarize the article publisher's MBFC label for training labels: all Questionable Source, Conspiracy-Pseudoscience, and Satire publishers were deemed unreliable$^-$, and all others reliable$^+$. Other publisher-label mappings have been used in other works, and is deserving of future research for this dataset.

To assess model performance at the article level, we employ the F1-score computed independently for each class, along with the Matthews Correlation Coefficient (MCC). The F1-score's interpretation is largely dependent on the class proportion (Flach and Kull 2015), making it less suited to comparison across experiments, whereas MCC is more robust to this (Chicco and Jurman 2020, 2022). MCC is 0 for random performance, and 1 only for perfect classification.

6.1 OoD Generalization

Table 2 displays the article level classification results for the various generalization splits outlined in Section 5. The larger the deviation between the in distribution (ID) articles in the validation set and the out-of-distribution (OoD) articles in the test set, the worse we consider the model's generalization performance.

Firstly, we note that classification performance falls short of desired. While the F1-score for the reliable$^+$ class tends to be high (in the range of 0.85–0.95 at a $\sim$ 60% class proportion), classifying unreliable$^-$ articles is considerably more difficult—a trend that holds consistently across generalization forms. This is largely due to low recall scores for the unreliable$^-$ class. This is especially surprising given the high accuracy scores reported for similar models on other misinformation datasets.

We see no degradation in performance when applying the model to articles from an unseen event (here, the COVID-19 pandemic). Despite the introduction of many unseen terms to the articles' vocabulary, it appears the manner in which established publishers discuss this new event deviates little from preceding articles.

6 `https://huggingface.co/microsoft/deberta-v3-base`.

Both Publisher and Topic splitting show moderate decreases in MCC scores, carried primarily by a decrease in the F1-scores for the unreliable$^-$ class. Generalization to completely unseen publishers or topics, cases where one would expect distinctly different linguistic style or vocabulary, is more challenging. The magnitude of this performance degradation, however, is smaller than we initially expected. We attribute this to two effects:

1. More mainstream, prolific publishers are obscuring performance on publishers with fewer articles (see Appendix B.8). We correct for this effect by including a publisher-level analysis in Section 7.1.
2. The training data is heterogeneous enough for the models to learn generalization across publishers. We test for this in Section 7.2.

Since it is conceivable that different publishers prefer particular topics, we compute a correlation between the produced test sets. While we find a small but consistent overlap between the Publisher and Topic test sets, we do not believe this alone accounts for the similarity in performance (see Appendix C.2).

The final two generalization axes exclude a particular misinformation type or political bias from the training set. For the former, we can see little to no effect when removing the Conspiracy-Pseudoscience class of articles, but a drastic one if removing Questionable Source articles. We posit this is due to the Questionable Source being the class of articles written with the explicit purpose of mimicking reliable$^+$ publishers, whereas Conspiracy-Pseudoscience tends to discuss completely separate topics. In other words, the conspiracy or pseudo-scientific articles tend to be easier to identify as unreliable$^-$.

For the Political Bias generalization axis, we see an inability to generalize to opposing political biases. Training on center and right biased articles sees a **0.19** drop in MCC, whereas training on center and left yields a drastic **0.37** drop. While this is a form of publisher splitting, in both cases the magnitude of the degradation is substantially larger. Especially for transfer to right-biased articles, there exists a drop for both reliable$^+$ and unreliable$^-$ classification, indicating that it is more challenging for the model to determine article reliability.

6.2 Generalization Across Time

When applying the models to unseen years, we find the models to be surprisingly robust, as shown in Table 3. At the article level, despite consistent degradation in

Table 3
Article level MCC scores of models trained with uniform splitting on different years of the dataset.

		Eval					
		2017	2018	2019	2020	2021	2022
Train	2017	0.50	0.43	0.41	0.40	0.40	0.38
	2018	0.29	0.42	0.43	0.39	0.41	0.37
	2019	0.26	0.38	0.44	0.40	0.41	0.40
	2020	0.34	0.39	0.47	0.47	0.47	0.45
	2021	0.31	0.37	0.46	0.46	0.47	0.45
	2022	0.33	0.38	0.46	0.45	0.46	0.46

performance, proximal years tend to achieve similar scores. Only in very distant years does performance degrade dramatically.

We speculate that these differences are due to differences in the various dataset iterations, while publisher style or idiosyncrasies are relatively static. For example, all models not trained on the 2017 iteration perform poorly on the 2017 iteration (between 0.26 and 0.34 MCC), whereas the 2020–2022 editions perform reasonably well on each other's years. Indeed, visually, Table 3 correlates strongly with Appendix B.8 Table B.9, showing the amount of overlap in publishers across dataset years.

6.3 LLM Performance

We compare the performance of the fine-tuned models to that of `llama-3-8b-instruct`[7] (Llama Team 2024), prompted to determine reliability of an article in a 0-shot setting with 512 token context (see Appendix D.2).

Despite the LLM's parameter count and the recency of its pre-training data, we find `llama-3-8b-instruct` to be inferior to the fine-tuned models for the purpose of article-level reliability classification. It manages an MCC of 0.25, compared to our fine-tuning models achieving 0.46 on ID years, and 0.33 on OoD years.

The use of this modestly-sized LLM already incurs a computational cost far greater than that of the fine-tuning models. In our experiments, using a single A100 GPU, the LLM took ∼70 hours to yield a prediction for all articles in the corpus, whereas each of the fine-tuning models were trained and evaluated in ∼12.5 hours. For deployment scenarios where the amount of compute necessary for inference far exceeds that of training, this difference will likely be more pronounced, and platforms with a large influx of text (i.e., social media networks) will need to balance the substantial computational overhead of model inference with classification performance.

While it is conceivable that larger, more recent language models might achieve strong misinformation detection performance, due to the size of the corpus and length of the articles, we did not experiment further with such models. Additionally, recent experiments with in-context learning misinformation detection have resulted in subpar performance (Yang and Menczer 2025).

6.3.1 Reasoning LLMs. We additionally experiment with some reasoning models (DeepSeek-AI 2025; Gemini-2.5-team 2025). Unlike standard LLMs, these models are post-trained for reasoning tasks, and produce long thoughts before answering a question.

We compare these models against the fine-tuned models on a small, stratified subset of the entire dataset[8] that contains two articles per publisher-topic combination for a maximum of 120 articles per publisher, totalling 28k articles. Similar to the above LLM and fine-tuning experiments, we only provided 512 tokens of context.

Table 4 shows the MCC and F1 scores these models achieve. We find that these models can achieve significantly better performance on this publisher-topic stratified subset, primarily through higher unreliable$^-$ precision scores.

Again, it should be noted that the reasoning models have orders of magnitude more parameters and pre- and post-training data. As a result, it is plausible that these models have some knowledge about the articles and the events they depict, and are capable

7 `https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct`.
8 Due to budget constraints.

Table 4
Performance of Uniform fine-tuned decoder-only models, compared to several SoTA reasoning LLMs (via API), on a small, stratified subset of the dataset. The left column provides the name of the model, and the rows below the thinking budget provided to the model.

Model	MCC	F1 Reliable	F1 Unreliable
Fine-Tuned	0.41	0.78	0.58
Gemini 2.5 flash lite			
	0.46	0.70	0.75
DeepSeek Reasoner			
	0.52	0.77	0.76

of placing individual articles in substantially more context than the fine-tuned models can. This becomes readily apparent when analyzing the reasoning models' "thoughts". These contain frequent references to quoted publishers and entities, whose reliability is known a priori, and the models seem to have a keen understanding of how these interact with reliable journalistic practices.

As a result, the results are likely not directly comparable to the purely inductive, fine-tuned models, with the reasoning models being able to apply a mixture of generalizable rules and external knowledge. This likely means that their generalization performance is overestimated, and one might expect the same set of issues identified in Sections 1 and 2 to occur: The models are not being evaluated for their performance on OoD data.

7. Analysis of Model Generalization

7.1 Determinants of Performance

The analysis of our results, thus far, has been constrained to article-level classification. While this reflects how misinformation classifiers interact with articles, it does not match how we annotate the dataset and can obscure performance on smaller, less mainstream publishers. Ideally, as highlighted by Baly et al. (2018b) and Burdisso et al. (2024), classification performance is also evaluated at the publisher-level, testing which publisher properties aid or interfere with misinformation detection.

To that end, we use a binomial logistic regression on the average publisher-level accuracy[9] to assess which aspects of a publisher determine the achieved accuracy score (for details and full model specification, see Appendix E.3). Unlike standard logistic regression, the dependent variable is modelled as a ratio.

In Figure 1 we show coefficient magnitudes for several important determinants, expressed as effect sizes (Chinn 2000; Lampinen et al. 2022). A positive effect size indicates that the variable increases the odds of accurate classification, *ceteris paribus*.

The size of the training set has a large positive effect, with a 1.91 multiplicative increase in the odds for each 10-fold increase in training samples. Thus, a publisher with

9 Defined as the ratio of correctly predicted articles to all articles for a publisher, i.e., true positive rate, recall.

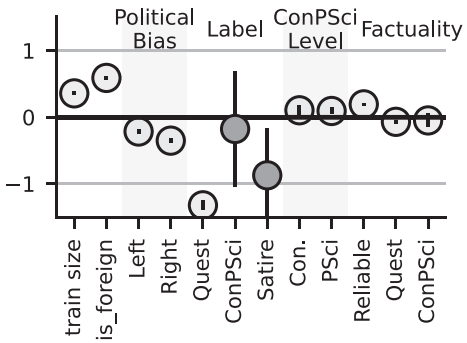

Figure 1
Coefficients of the determinants model, expressed as effect sizes. Circles are centered toward the effect size, with lines giving the 95% confidence interval.

Table 5
Median predicted publisher-level accuracies averaged over combinations of the MBFC label (rows) and political bias (columns). The row headers correspond to (R) reliable, (Q) Questionable Source, (C) Conspiracy Pseudoscience, and (S) Satire.

	Left	Center	Right
R	86.22%	92.90%	79.03%
Q	46.13%	46.11%	30.92%
C	32.47%	79.41%	31.20%
S	7.10%	–	14.11%

1,000 articles in the training set is 3.82 times more likely to have correct classification in the test set than a publisher with only 10 articles. Foreign publishers also prove easier to identify.

Relative to center-biased publishers, publishers on the left or right sides of the political spectrum see slightly degraded performance, even after controlling for the publisher label. Moreover, classification on the unreliable$^-$ classes suffers more than on reliable$^+$ publishers. Because the odds ratios or effect sizes are difficult to interpret, we also provide the estimated marginal mean publisher-level accuracy for those combinations in Table 5.

Despite right biased unreliable$^-$ sources being far more prevalent in the training data, for both the Questionable Source and Conspiracy-Pseudoscience classes, model performance is noticeably worse than on left and center biased sources. This somewhat confirms the results in the Political Bias rows of Table 2: models struggle disproportionately to discriminate between reliable and unreliable$^-$ right-biased articles.

We see a slight positive correlation with the MBFC Conspiracy-Pseudoscience level. The higher the value, the further the publishers' articles tend to deviate from convention. As a result, strongly conspirational or pseudo-scientific publishers are 4.84 and 4.72 times more easily identified than publishers where this effect is weaker.

Finally, when looking at the factuality level (the propensity for a publisher to publish factual articles) we find a positive interaction for reliable articles, and weakly negative interactions for unreliable articles. The more a publisher goes against the expectation (factual for reliable for publishers, false for unreliable), the more difficult it becomes to disambiguate the source.

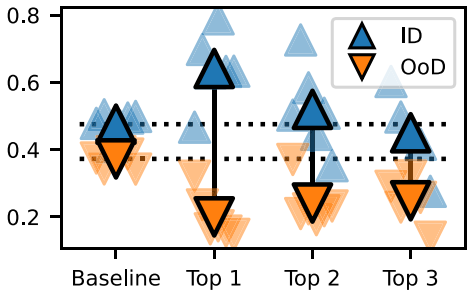

Figure 2
MCC scores for different Publisher split test sets with only the top 1, 2, or 3 most prolific publishers retained per publisher class. The Baseline column corresponds to the standard publisher splitting described in Section 5. The lightly shaded shapes provide a value for each year of the dataset; the solid shapes their average. The blue upward triangles are ID publishers; the orange downward triangles instead represent OoD publishers.

7.2 Effect of Publisher Diversity

Here we test to what extent the diversity of the publishers present in the training data has an effect on the generalization capacity of the models. We re-run the Publisher split experiment with smaller training sets, constrained to only the most prolific publishers. Specifically, for each MBFC label, we only include the top-n most frequent publishers in the training data, while leaving the test set untouched. While this reduces the amount of variation in publishers considerably, it minimally affects the total amount of data present. Each training set still consists of hundreds of thousands of articles.

Figure 2 displays the generalization gap (in terms of MCC) induced when increasing publisher homogeneity. Especially when limiting performance to the top 1 most common publishers, the models show increased overfit to the training set. Where the Publisher split saw a 0.1 MCC delta, this increased to an average degradation of 0.5. As the number of included publishers increases, the generalization gap decreases and starts to converge to the previously seen Publisher values. Notably, the variance in values is also substantially higher in the limited publisher settings.

From this, we conclude that (1) the splitting described in Section 5 has minimally altered the heterogeneity present in the dataset, and (2) the models improve with publisher heterogeneity. The former finding suggests that the underestimation of the generalization gap will be especially egregious in datasets with a small pool of publishers (e.g., those that sample from a single reliable source to boost label balance). The latter, instead, provides some initial evidence for the utility of using large, diverse publisher-level datasets for pre-training article-level misinformation detectors; while fine-tuning on high-fidelity labels is likely necessary, using distantly supervised datasets might encourage more robust models before fine-tuning.

8. Analysis of Publisher-Level Labeling

8.1 Publisher Identifiability

The use of publisher-level labels as a form of weak supervision, especially in misinformation detection, can lead to models overfitting to publisher styles instead of article veracity. This was shown to be a serious concern by Zhou et al. (2021), and efforts

to mitigate this effect at the data level were discussed in Section 4. Despite this, in Sections 6.1 and 7.1 we still found models to overfit to specific publishers and publisher classes, and in Section 7.2 we found a negative correlation between publisher diversity and the magnitude of the generalization gap.

As such, here we directly test the identifiability of the publisher from article content by replacing the misinformation labels with a unique identifier for each publisher. In other words, instead of classifying into the set $\{\text{reliable}^+, \text{unreliable}^-\}$, the model classifies into the set of all possible publishers.

Using the same learning algorithm, we find this to be a substantially more difficult task. While models exhibit above-random article-level performance, with an average MCC score of 0.18, this is much lower than scores achieved with misinformation labels. Furthermore, when aggregating F1-scores across classes proportionally according to publisher frequency (micro) we get 0.14, whereas with a flat average (macro) we obtain a mere 0.04 F1. In short, while it is possible to predict the publisher from an article with above random performance, this is only really possible for the most prolific publishers, and this cannot entirely explain performance in misinformation classification.

8.2 Publisher Memorization

In this subsection, we analyze to which extent models need to memorize specific publishers. If there exists a lot of disagreement between the label of an article and the label assigned to its publisher, it is likely impossible to generalize to the unseen publisher from seen publishers; the labels of similar publishers clash. In this case, for classification to be successful, it is necessary for the misinformation detector to memorize publisher idiosyncrasies.

Inspired by the works of Pleiss et al. (2020) and Jenkins, Talafha, and Goodwin (2023) in automated mislabeling detection, and Swayamdipta et al. (2020) on diagnosing dataset issues using "dataset cartography", to estimate the necessity of memorization of specific publishers, we run an experiment comparing the average article confidence (mean logit assigned to the correct class) and disagreement (variance of logit assigned to the correct class) of publishers when included or excluded from the dataset. If there is a large shift in either the confidence or disagreement of a publisher when in- or excluded, this might indicate that the publishers' articles' labels are not aligned with those of similar publishers.

We rerun the 2021 uniform split experiment 15 times. We exclude each publisher in 5 runs, at random, while taking care to minimize the number of exclusion set co-occurences and stratifying the exclusion across fine-grained MBFC publisher classes. As such, there should always be similar publishers available to excluded ones.

Figure 3 shows the effect of exclusion on the average confidence and disagreement scores for all publishers. Overall, and unsurprisingly, including a publisher during training increases average article confidence and decreases disagreement. However, for most publishers, the shift between training or exclusion is small, and likely attributable to the inherent stochasticity of mini-batch training. We assume that these publishers' labels can largely be learned from similar publishers, and that these align well with each other.

While there are significant shifts, these mostly occur for the largest, typically well-performing reliable^+ publishers, and take the form of significantly increased disagreement and decreased confidence when excluded. These include sources which MBFC believes to produce typically reliable, but sensationalist, subjective news (e.g., The Sun, The Daily Mirror), or anti-US propaganda sources (e.g., Pravda Report, Asia Pacific

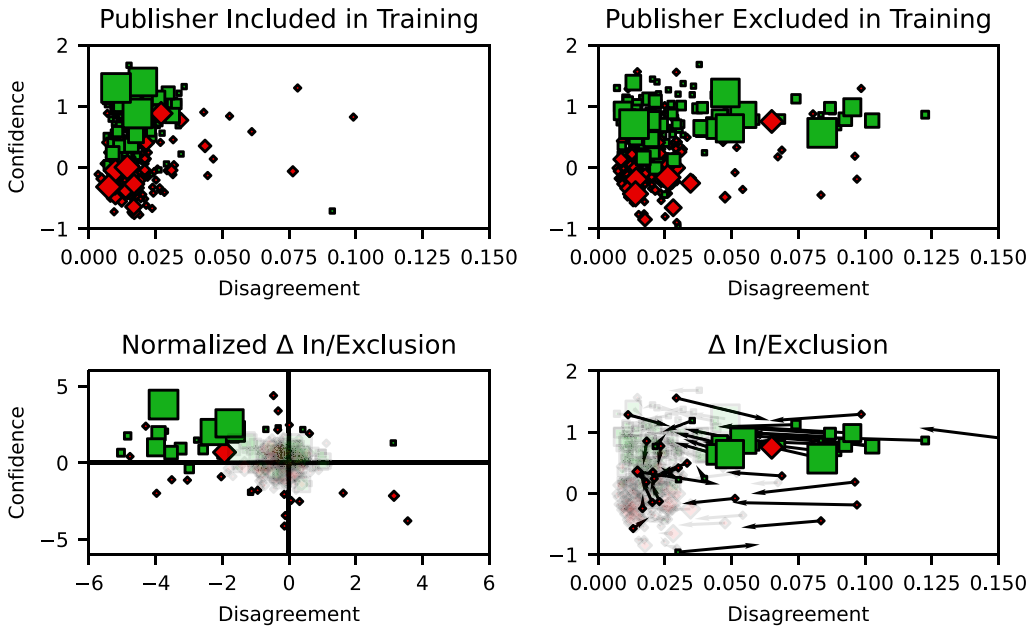

Figure 3
The average article-level confidence and disagreement for different publishers. The top left panel shows scores when publishers are included in training, whereas the top right panel shows scores when publishers were excluded. The bottom left panel shows their difference, normalized. Only publishers with a shift magnitude above 2 standard deviations (i.e., significantly far from origin) are shown with low opacity. The bottom right panel shows the direction of differences for significantly shifted publishers. In all panels, the size of the circles are proportional to a publisher's size in the dataset. Green colored squares represent reliable^{+} publishers, and red colored diamonds represent unreliable^{-} ones.

Research), as well as highly reputable sources (e.g., CBS News, BBC).[10] Comparatively, most large unreliable^{-} publishers see far smaller shifts.

There are substantially fewer significant shifts in the other quadrants (Figure 3, bottom left panel), and those that do show up tend to be for much smaller publishers. Publishers that see a significant *reduction* in confidence when included in training do exist, although these constitute a small minority with typically very few articles. Looking more closely at such publishers, these include cases where site ownership changed during article collection (Viral News Network, Infinite Unknown), or whose articles are extremely noisy (Alternative Media TV), which would serve as good candidates for removal. We also find particularly difficult cases here, like neutral, objectively written articles promoting climate change denial (Climate Etc).[11]

All in all, while there appears to be some ambiguity in article labels, largely due to publisher editorial biases, we find no evidence of mislabeling beyond a level expected for a corpus scraped from the internet. Despite having missed some publishers in the data cleaning phase (see Appendix B), these represent a small minority of all publishers, and collectively contain a small minority of all included articles.

10 These categorizations originate from MBFC, and do not reflect the authors' opinions.
11 Idem.

8.3 Article Properties

In this subsection, we automatically analyze various properties of our dataset at the article level, both to assess their presence for different classes of publishers, and their correlation with news reliability labels.

Subjectivity Analysis. The first property we annotate for is subjectivity. Objective news presents facts in a neutral, unbiased manner, and is commonly considered the antithesis of hyperpartisan or misinformation news, which is written specifically to incite an emotional response from readers, thereby inducing sharing (Bojic, Prodanovic, and Samala 2024). Subjectivity has shown some promise as a feature in discriminating reliable and unreliable news (Jeronimo et al. 2019). Despite this, reliable publishers also produce subjective text, likely to drive engagement. This can make articles unverifiable, hampering article-level labeling.

To assess the degree of objectivity, we ask `ChatGPT-4o-mini` to provide a rating for an article using a 5-point Likert scale, ranging from entirely objective to entirely subjective. While by no means SoTA, similar setups have shown reasonable performance in prior work (Galassi et al. 2023; Struß et al. 2024; Shokri et al. 2024).

Figure 4 shows the estimated proportions of each subjectivity level. Despite reliable news being in the majority, most (∼73%) articles have a subjectivity level of Mixed or higher. In fact, Mostly Subjective articles seem to be most common.

Upon inspection of the dataset, these annotations seem to match our findings. While unreliable news is substantially less likely to present itself as objective, reliable publishers still publish a plethora of discussion and opinion pieces. This is especially true for publishers with a more pronounced political bias.

We repeat the binomial logistic regression analysis used in Section 7.1 to determine what publisher-level properties correlate with subjectivity. Unlike earlier, where the ease of classification correlated strongly with the form of misinformation, article subjectivity tends to correlate strongly with publisher political bias. Where an unreliable publisher reduces the odds of an objective article by between 0.34 and 0.68, moving to a left or right political bias does so by between 0.24 and 0.25. In other words, both

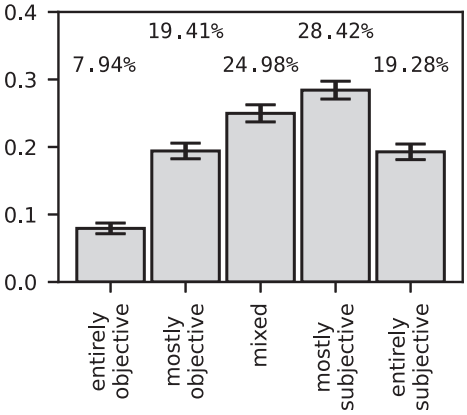

Figure 4
The estimated proportions for each subjectivity level in `misinfo-general`. Errors bars give the 95% Agresti-Coull binomial proportion confidence interval (Agresti and Coull 1998).

reliable and unreliable publishers produce subjective, potentially unverifiable articles, especially when publishing from a biased standpoint. The full model, including estimated marginal medians, and prompt specification, can be found in Appendix E. We also compare the agreement with `ChatGPT-4o`, which seems to lean towards an even greater subjectivity propensity.

Manual Annotation. To complement the automated subjectivity analysis, and to verify the alignment of article- and publisher-level labels, we manually annotated a small subset of articles. Specifically, we took 362 articles sampled from the subjectivity annotated subset, stratified over publisher and subjectivity level. Then we check whether the article—in isolation—violates common journalistic norms and practices.

For unreliable$^-$ publishers, 43.50% (36.62%–50.37%)[12] of articles were clear cases of non-credible news. The proportion of non-credible articles differs substantially between different publishers, with some unreliable$^-$ publishers mixing innocuous articles or clear opinion pieces on general topics with misinformation on specific ones. In reliable$^+$ publishers, we deem 8.20% (4.07%–12.32%) of articles to be non-credible. Practically all these cases come from hyper-partisan articles, rather than instances of clear misinformation. Overall, we find that the odds of a non-credible article being published by an unreliable publisher are 8.62 times higher than for a reliable publisher.

Emotion Analysis. Another property we annotate for is emotion. Affective language in journalistic texts has long been understudied, despite emotion and its effect playing an increasingly important role in the modern media landscape (Koivunen et al. 2021). It is especially prevalent in unreliable news, and is used to both persuade readers and incite sharing (Alba-Juez and Mackenzie 2019). The persuasiveness of emotional language in fake news is a matter of debate (Martel, Pennycook, and Rand 2020; Phillips et al. 2024), with prior work showing that high affective state in people after ingesting misinformation is associated with both increased susceptibility (Martel, Pennycook, and Rand 2020; Bago et al. 2022) and skepticism (Horner et al. 2021; Lühring et al. 2024). From a computational perspective, however, combining emotion detection with misinformation detection has shown some promise (Ghanem, Rosso, and Rangel 2020; Zhang et al. 2021b). Especially low valence emotions like anger, sadness, anxiety, surprise, and fear are believed to be prevalent in misinformation texts (Liu et al. 2024).

We use a similar setup as above to annotate articles with one of eight Plutchik basic emotions (Plutchik 1980) and neutrality, a common emotion model used for annotation in NLP (Bostan and Klinger 2018).

Figure 5 shows the association of different emotions with different publisher classes. Visually, these results largely mirror the objectivity analysis; both reliable and unreliable publishers use emotional language, with political bias being a more important determinant of affect than reliability. The most notable difference in association is with Neutral and Center-Reliable publishers; relative to other publisher categories, neutral writing occurs relatively often for this publisher class. Low valence emotions like Anger, Disgust, and Sadness are prevalent throughout, but are especially associated with more politically biased, less reliable publishers. Especially Satire publishers seem to be characterized by relatively high amounts of Joy and Surprise. Anticipation is highly prevalent

12 The brackets represent the 95% Agresti-Coull binomial proportion confidence intervals (Agresti and Coull 1998).

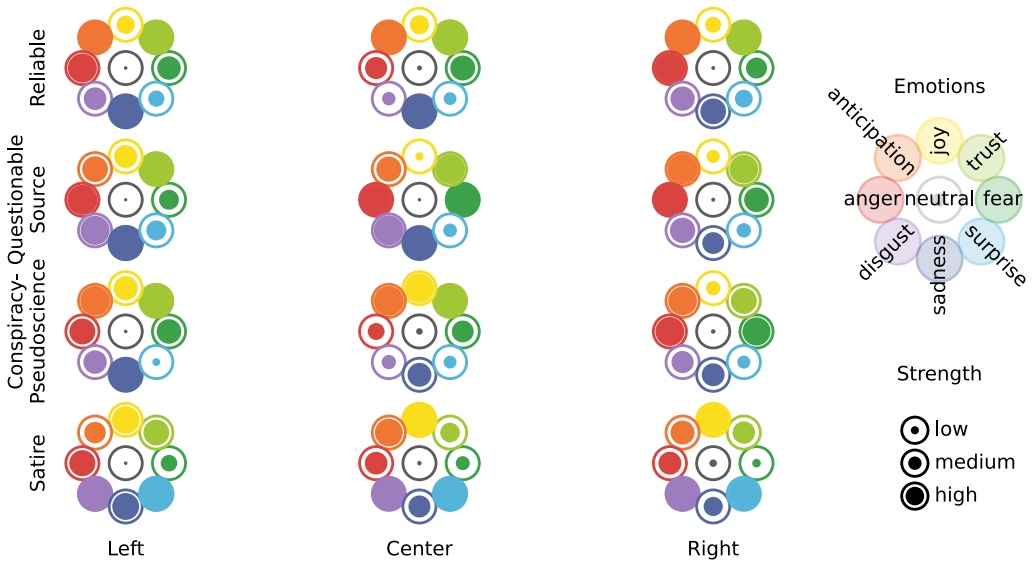

Figure 5
The association of emotions in articles with different publisher categories, as measured using pointwise mutual information. Each circle represents one of Plutchik's 8 emotions, along with neutrality in their center. The area of a circle represents the strength of the association of the emotion with that publisher class (as measured using PPMI), relative to the maximal association found. The legend provides each emotion's color and location in the color wheel.

in articles discussing expected future events, which appears to occur regularly, across all publisher classes.

Overall, however, much like subjectivity, there exists a good balance of emotions across the different publisher classes. While there is some correlation between emotion or subjectivity with publisher reliability, these appear to be insufficient to function as a shortcut for misinformation prediction.

While we partially corroborate the finding that different emotions are present (or at least, in different proportions) across publisher classes, ultimately there exists a good balance of emotions across the different publisher classes. Much like subjectivity, simply using the emotions present in an article to determine whether it comes from a reliable or unreliable publisher is likely insufficient.

As argued in Section 1, to reliably estimate the generalization gap of misinformation detectors, it is crucial to have access to a naturalistic corpus of misinformation, which is representative in terms of the diversity contained within. With this analysis, we have shown that for the properties of emotion and subjectivity, this diversity is present, with subjective and emotional language being present in articles from both reliable and unreliable publishers.

9. Conclusion

In the present state of the art, automated misinformation detectors cannot be safely or reliably deployed. While impressive performance is often reported, time and again, papers show that in more realistic settings, where out-of-distribution generalization is required, these models fail. In part, this is inherent to the problem of misinformation

detection; as Yee (2025) argues, the informational norms in any community is continually evolving, and any equilibrium is transient. The nonstationary nature of misinformation is, and likely always will be, difficult to model.

One source of this brittleness to distributional shifts, as substantial empirical evidence has shown, is a failure of misinformation datasets to adequately simulate misinformation detection scenarios. Due to the prohibitively expensive cost of procuring misinformation labels, practically all surveyed misinformation datasets have had to make unrealistic assumptions, which introduce undesirable biases.

To accommodate recommended misinformation evaluation practices, and thereby enable the development of *generalizable* misinformation models, this article introduces `misinfo-general`. It consists of a benchmark built on top of a cleaned, weakly supervised corpus of online articles, which have rich article- and publisher-level metadata, and an operationalization of various generalization axes. We have shown that this dataset is challenging for a common class of misinformation detection models, and especially so when generalization to unseen forms of content or publishers is required.

The metadata annotations enable us to further analyze the determinants of model performance. We find large discrepancies across political biases and misinformation forms, but have also shown that increased diversity has a positive effect on generalization.

While publisher-level labeling introduces noise, we believe the increased scale, diversity, and affluence of metadata make up for this. The first two properties can enable OoD generalization or robustness, whereas the latter enables evaluation, analysis, and potentially generalization-aware training.

We make the dataset publicly available, and hope it will serve as a resource for OoD generalization-focused training and evaluation. While the implementation of trustworthy automated misinformation detectors remains out of reach, we hope that this dataset at least makes evaluating and diagnosing generalization capacities of misinformation detection models easy enough for widespread academic adoption.

9.1 Extending `misinfo-general`

One of the central claims in this article is that online content and misinformation tends to change, rapidly and comprehensively. While "misinfo-general" represents a large and diverse corpus, useful for pre-training misinformation detectors and evaluating generalization abilities of trained models, it is inevitable that this collection will become misaligned with future forms of misinformation. For example, with the latest articles coming from the end of 2022, it likely misses the newly emerging category of LLM-generated (mis)information. Another reason for updating the dataset is to avoid data leakage; new NLP models will likely be trained on collections that overlap with `misinfo-general`, and can include hyper-textual reference to `misinfo-general` content (e.g., fact-checking articles). However, we are confident that the collection can be relatively easily maintained and updated.

With the accompanying metadata database, it is trivial to extract information on the publishers already in the dataset, allowing scraping of future content, and these are all linked to the MBFC front-end, which can be used to track labels. Incorporating article-level label providers, which the surveyed datasets in Appendix A Table A.1 show are becoming more accessible, could allow for blending the noisy, but pragmatic distant labeling with high-quality, high-cost article-level labels.

10. Limitations and Ethical Considerations

While the dataset includes a diverse set of publishers, events, and topics, ultimately the publisher metadata comes from a single source. The information provided by MBFC assumes a narrow, US-centric world-view. This is especially prominent when discussing foreign publishers from nations with geopolitical ambitions at odds with the United States. As such, the metadata provides limited nuance, providing only a single perspective of an inherently subjective assessment. It is expected that across cultural backgrounds, the publisher information is bound to change.

In general, the news included in the dataset is US-centric, with the vast majority of publishers being American, producing articles for an American audience. This is exacerbated by us excluding all non-English articles. This means cross-lingual or cross-cultural generalization cannot be evaluated.

That said, we do discuss one weak form of cross-cultural transfer. Besides prioritizing different events, the primary distinction between the political ideologies discussed in Section 5 are their differing norms and values. As such, the poor political bias generalization bodes ill for the more general cross-cultural generalization tasks.

While much previous work has shown incorporating non-text modalities in the classification pipeline benefits classification (Alam et al. 2022; Xiao and Mayer 2024), in its current form, `misinfo-general` does not include non-text modalities.

In the case of social media context, all references to such content was removed. We consider such content inherently PII, and their use in misinformation detection is fraught with ethical problems (Mishra, Yannakoudakis, and Shutova 2021).

Embedded images and videos were also not included. This data was not available in the progenitor datasets. This excludes a large and important class of misinformation detectors, which can leverage the interplay of text and non-text context. Incorporating these contexts would make for valuable future work, allowing for models that more closely emulate the decision process used by human misinformation annotators. At present, however, this lies outside the scope of this project.

In general, the deployment of misinformation detectors comes with legal and ethical issues. Pre-emptive moderating of communication, which is typically the implicit goal of automated misinformation classifiers, is in essence a prior restraint on speech, regardless of the accuracy or OoD robustness of the model (Llansó 2020). While OoD robustness mitigates the propensity of such human rights violations (Tobi 2024), it cannot remove it entirely.

10.1 Dataset Access and Licensing

We aim to make `misinfo-general` as easy to use as possible, but have had to make some restrictions. The dataset contains texts that are toxic, hateful, or otherwise harmful to society if disseminated. The dataset itself or any derivative formats of it, like language models, should not be released for non-research purposes.

The NELA corpora were initially released under a CC0 1.0 license,[13] essentially being released to the public domain. From 1 January, 2024, the NELA authors have de-accessioned their repository. Upon request, the authors note their desire to restrict usage to non-commercial research.

13 `https://creativecommons.org/publicdomain/zero/1.0/deed.en`.

Given the potentially harmful content, and our colleagues' wishes, we (re-)release our dataset under a more restrictive CC BY-NC-SA 4.0[14] license. This allows for redistribution and adaption as necessary for academic research, while preventing commercial use-cases and requiring adaptions to maintain these restrictions. To circumvent copyright of the original texts, we have extended the effort made by the original NELA authors, and have "poisoned" all texts with special tokens.

We have released `misinfo-general` through two media that allow for restricted access. Specifically, we use Harvard's Dataverse (which implements per-file access restrictions) and HuggingFace's Dataset Hub (which implements repository-level gating). We plan to review access applications manually, limiting use-cases to academic research only.

14 `https://creativecommons.org/licenses/by-nc-sa/4.0/deed.en`.

Appendix A. Misinformation Dataset Survey

Table A.1
A long table with an (inexhaustive) sampling of misinformation datasets. Each row provides a single dataset, with name and citation, along with the labeling granularity (see Section 3), the dataset size (where units "k" and "M" denote thousands and millions, respectively), and a short description of how the dataset authors generated misinformation labels.

Dataset	Label	Size	Description
Lie Detector Mihalcea and Strapparava (2009)	Claim	0.3k	MTurkers[a] produce short arguments that align and oppose their stance on various topics
CREDBANK Mitra and Gilbert (2015)	Claim	60M	Many MTurkers[a] annotate tweets for veracity and verifiability, with the majority annotation becoming the label
Weibo15 Ma et al. (2016)	Article	5k	The authors scraped user nominated misinformation articles from the Sina Weibo Community Management center[b]. Unannoted posts were included as factual posts
MediaEval15 Boididou et al. (2015)	Claim	12k	The authors generated a list of events which were verified as true or false, and a collection of tweets discussing these events. The tweets were manually verified
BuzzFeed-Webis Potthast et al. (2018)	Article	1.6k	Articles from a small set of sources were manually rated between mostly true or mostly false by expert journalists from BuzzFeed[c]
TSHP-17 Rashkin et al. (2017)	Publisher	70k	Trusted news articles were sampled from the Gigaword News corpus, whereas unreliable news was sampled from specific publishers.
Kaggle Fake News Risdal (2016)	Publisher	13k	The authors scraped articles from unreliable sources using a third-party tool. No reliable articles were included
Allcott & Gentzkow Allcott and Gentzkow (2017)	Article	0.2k	Verified fake news articles were scraped from Snopes[d], PolitiFact,[e] and BuzzFeed[c]. No reliable articles were included
PHEME Zubiaga, Liakata, and Procter (2017)	Claim	5.8k	Tweets related to 5 mainstream events were manually annotated as unverified rumor or verified
Liar Wang (2017)	Claim	13k	Short snippets from famous politicians scraped from the PolitiFact[e] API
Weibo17 Jin et al. (2017)	Article, Publisher	10k	They take posts reported as false from trusted users, and take articles from mainstream publishers for their factual class
Some Like it Hoax Tacchini et al. (2017)	Publisher	15.5k	Articles were scraped from Facebook groups dedicated to sharing scientific or pseudo-scientific articles
Fake vs Satire Golbeck et al. (2018)	Publisher	0.5k	Articles were sampled from identified satire or fake news sites. The authors constrained the number of articles per publisher to ensure a diverse publisher set. All ambiguous cases were removed
FakeNewsAMT Pérez-Rosas et al. (2018)	Article	0.5k	A small set of manually verified articles were taken from mainstream publishers, and minimally edited by MTurkers[a] to produce misinformation
Web Dataset Celebrity Pérez-Rosas et al. (2018)	Article	0.5k	To complement FakeNewsAMT, the authors collect articles from rumor and tabloid publications, and manually verify articles using sites like GossipCop[f]
FakeNewsNet PolitiFact Shu et al. (2019)	Claim, Article	23k	The authors label an article based on a claim made within, where the claim is labeled by PolitiFact[e]
FakeNewsNet GossipCop Shu et al. (2019)	Article, Publisher	23k	Unverified rumor articles were taken from GossipCop[f], with verified rumors coming from a few mainstream publishers
FakeNewsCorpus Pathak and Srihari (2019)	Article, Publisher	0.7k	~700 articles were taken from questionable source publishers, and used as misinformation, and 26 expert labeled factual news articles. Satire and unverifiable news were explicitly excluded
QProp Barrón-Cedeño et al. (2019)	Publisher	51k	Uses MBFC to assign articles the label of their publisher. They manage to sample from 104 different sources, although only include 10 propagandistic sources
Przybyła Credibility Przybyła (2020)	Publisher	100k	Scrapes articles from websites classified as non-credible by PolitiFact[e]. The authors specifically evaluate publishers as credible or non-credible, as opposed to fake or factual news
FakeHealth Dai, Sun, and Wang (2020)	Article	2.3k	Both variants of the dataset (HealthStory and HealthRelease) include text manually verified by experts from HealthNewsReview.org[g] on the *credibility* of the information provided
MM-COVID Li et al. (2020)	Article, Publisher	4.2k	Articles with manual labels were collected from Snopes[d] and Poynter[h], and to complement reliable articles, they sample from mainstream media sources

Table A.1
Continued.

Dataset	Label	Size	Description
FakeCovid Shahi and Nandini (2020)	Article	5.2k	Specifically COVID articles with labels from Snopes[d] and Poynter[h] were collected. The authors make sure to include labels from 92 separate organizations across 105 countries
WeChat Wang et al. (2020)	Article	4k	The authors collected articles flagged by WeChat users. A small subset was annotated by experts, while a larger subset was unannotated, meant for unsupervised training
CoAID Cui and Lee (2020)	Claim, Article, Publisher	3.7k	Misinformation articles about the COVID19 pandemic were scraped directly from various fact-checking sources. Factual articles were scraped from 9 reliable publishers. Claims were scraped from official government sites
MuMIN Nielsen et al. (2022)	Claim	13k	The authors collected a set of 115 fact checking organizations from the Google Fact Check Tool[i] API, and then collected all fact-checked claims from these organizations. They use a separate classifier to collate different labeling schemas
PolitiFact-Oslo Poldvere, Uddin, and Thomas (2023)	Claim, Article	2.7k	Claims were extracted from PolitiFact[e], and the post or article from which the claim originated was manually extracted. The authors specifically highlight the importance of publisher-level metadata
MCFEND Li et al. (2024)	Article	24K	Articles annotated by various fact-checking organizations around the world were collected, and manually mapped to a single annotation schema.

[a] MechanicalTurk: crowdsourced lay volunteers.

[b] Weibo Community Management Center: credible Weibo users can report posts.

[c] Buzzfeed: a digital media company.

[d] Snopes: expert journalist website for debunking misinformation.

[e] PolitiFact: expert journalist website for fact checking politicians.

[f] GossipCop: a defunct website dedicated to fact-checking celebrity rumors.

[g] HealthNewsReviews: a defunct website dedicated to reviewing medical claims.

[h] Poynter: a global, non-profit organization with annotations from partnered organizations.

[i] Google Fact Check Tool: a unified API for fact-checking annotations.

Appendix B. `misinfo-general` Processing and Statistics

We downloaded the initial NELA corpora from the Harvard Dataverse, under a CC0 1.0 license.[15] The corpora have since been de-accessioned, and can no longer be downloaded. We expand on this in Section 10.1.

Appendix B.1 Re-Labeling

As a first processing step, we relabeled all publishers. This was done to (1) attribute articles to their original publisher (where possible), (2) ensure publisher information was up-to-date (MBFC had expanded their catalogue considerably), and (3) mitigate the effects of publishers that might interfere with the learning signal.

An important class of publishers that belong under that third point are *aggregation sites*. Such sites either do not produce original content, or intersperse articles from (usually more reputable) other sources through their content. While the collection of articles as a whole might express some editorial bias, for the most part, these sorts of publishers introduce noise into an already noisy labeling scheme.

15 `https://creativecommons.org/publicdomain/zero/1.0/deed.en`.

Table B.1
Size of the datasets, in terms of millions of articles, after each step of cleaning, per year. The lower percentage gives the step-to-step reduction in size. The Total column computes reduction relative to Original.

Year	2017	2018	2019	2020	2021	2022	Total
Original	0.14	0.70	1.12	1.78	1.86	1.78	7.24
De-aggregation & Labeling	0.13 −1%	0.61 −9%	0.96 −12%	1.62 −5%	1.71 −3%	1.66 −3%	6.69 −8%
Exact Deduplication	0.12 −6%	0.55 −11%	0.86 −12%	1.34 −18%	1.39 −20%	1.32 −21%	5.58 −23%
Cleaning	0.12 −3%	0.53 −4%	0.71 −17%	1.12 −16%	1.16 −17%	1.11 −16%	4.74 −35%
Near Deduplication	0.11 −11%	0.47 −12%	0.65 −8%	1.03 −8%	1.06 −9%	1.01 −9%	4.33 −40%
Language Detection	0.11 0%	0.47 −1%	0.64 −1%	1.02 −1%	1.05 −1%	0.99 −1%	4.16 −43%

Table B.2
Publishers removed for being news aggregation sites, with article counts post de-aggregation.

Publisher	# Articles
drudgereport	92,186
bonginoreport	17,895
whatreallyhappened	57,431
thelibertydaily	6,346
yahoonews	52,797
theduran	10,722
Total	237,377

We manually re-mapped all URL domains to a set of publishers consistent across years, excluding all news aggregation platforms and social media sites (see Tables B.2 and B.3). It should be noted that the 2018 edition of NELA did not contain URLs, making relabeling in this manner impossible.

Table B.9 shows the amount of publisher overlap that exists between different dataset years.

Appendix B.2 Deduplication

Models trained on this dataset have to (implicitly) learn a mapping from an article in its publisher's class. As a result, any instances of duplicate content unique to a publisher or publisher class likely induces overfit. The model need only memorize those cases to make a prediction, independent of the article content.

Table B.3
Publishers removed for being social media sites, with article counts post de-aggregation.

Domain	# Articles
soundcloud.com	106
youtu.be	4,009
apps.apple.com	414
amazon.com	112
facebook	108
amazon	113
youtube	2,902
play.google.com	421
twitter	274
instagram	15
reddit	4
dnyuz.com	1,900
Total	10,378

Table B.4
Frequent duplicated articles.

Type	Sample Text
Banners	Click for more article by Guest …
Previews	To read the full blog, please check out the complete post …
Prayers	Our Father, who art in heaven, hallowed be thy Name …
Podcast Descriptions	Don't forget to tune in to …
Error Messages	403 Forbidden nginx

To minimize the effect of these duplicates, we apply several stages of deduplication.

1. We only keep the first (as determined by publication date) instance of exact duplicates for each publisher. Besides plagiarism, the most common duplicates are due to errors in scraping or parsing, resulting in artifacts (see Table B.4).
2. We remove all articles with duplicate titles or URLs, across the entire dataset. This primarily removes articles that are updated at a later stage (e.g., live-blogs or summaries), or result from URL re-directions.
3. We remove all articles if more than 5% of its sentences are duplicates from the same publisher.

The former two (Exact Deduplication in Table B.1) can handle document-level duplicates, like plagiarized articles or updated blog-posts, but miss near duplicates. The latter should filter out near duplicates, like lightly edited documents (Near Deduplication in Table B.1).

An added benefit of these deduplication steps is the identification of common scraping errors. In fact, we found this to be the most frequent form of duplication. We list some common texts in Table B.4. Since the majority of such errors were automated responses unique to individual publishers, we could filter these out with relative ease using the described Exact Deduplication approaches.

Ultimately, the separate deduplication steps together resulted in removing 3.1M articles, comprising roughly ≈22% of the dataset after relabeling.

Appendix B.3 Cleaning and Filtering

Many of the included articles either include too many non-semantically relevant tokens or are otherwise malformed. This is aggravated by the NELA authors "poisoning" the dataset (from 2019 onwards) with repeated @ tokens to avoid copyright infringement. We do our best to clean the article texts, and remove those with no discernible semantic information.

We normalize all punctuation and remove any embedded URLs, HTML, or Markdown markup. SpaCy's (Montani et al. 2023) `en_core_web_sm` was used to annotate all tokens in the corpora. We identify self-references as named entities with a large longest common substring relative to the publisher's name. We also mask any sentences which SpaCy flags as being part of an email, URL, or Twitter handle. Finally, we standardize the NELA copyright poisoning, applying it to all dataset years equally.

All-in-all, this introduces 4 new special tokens: `<copyright>` replacing NELA's repeated `@` tokens, `<twitter>` for X (formerly known as Twitter) handles, `<url>` for any URLs, and `<selfref>` for any self-references.

To ensure all articles contained enough grammatical text to reasonably classify, we removed any article that did not abide by the rules delineated in Table B.5. As a final step, we use Lingua (Stahl 2024) to filter out any non-English texts.

Despite removing almost half of the articles, the dataset retains a level of "noise" customary to data sourced from the internet. This is inherent to the domain, and further cleaning might negatively affect the realism of the benchmark. One type of noise that could interfere with learning is the presence of stylistically unique substrings identifying publishers. For example, articles from the same publisher tend to contain similar by-lines, attribution messages, or donation requests.

Table B.5
Filtering rules.

Rule	VALUE
Articles must contain at least ${VALUE} tokens	16
Articles must not contain more than ${VALUE} tokens	4,096
The article must have a title	–
The article must be at least ${VALUE} times longer than its title	3
The article must have a mean token length greater than ${VALUE}	2
The article must have a mean token length less than ${VALUE}	10
The article must have at most ${VALUE}% copyright tokens	20

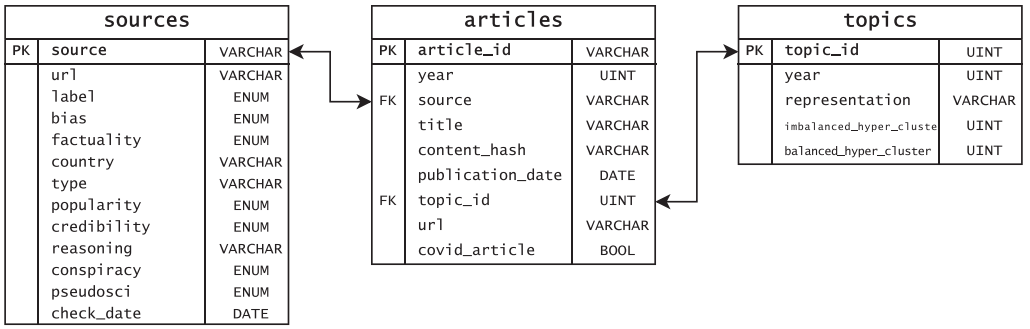

Figure B.1
Schema for the generated metadata database. PK indicates primary key, FK denotes a foreign key relationship.

Appendix B.4 Event and Topic Clustering

To generate the metadata necessary for the Topic generalization form (see Section 5 or Appendix C), we opted for a bottom-up approach. This involved first clustering the dataset into thousands of fine-grained events, before clustering the event clusters into overarching topics.

To achieve this, we used a heavily modified variant of `BERTopic` (Grootendorst 2022) with a `gte-large`[16] (Li et al. 2023) backbone.

After embedding an entire year of the dataset, we first reduce their dimensionality using mini-batched principle component analysis (PCA), and whiten the data. We construct a vocabulary over all the documents, by aggregating mini-batches of vocabularies. We then apply `UMAP` dimensionality reduction, using the PCA solution as initialization, and the cosine distance as the distance metric. Clustering was performed via `HDBScan` to mini-batches of the embeddings, assigning all articles to their nearest event cluster. Using the dataset-wide vocabulary, we generate a single TF-IDF representation matrix.

After mini-batching, we generate a hierarchical clustering on top of the event-based TF-IDF matrix. We merge any events with a distance below 0.8 to mitigate the effect of mini-batching, and re-construct the TF-IDF representations. We deem these the final event representations. Table B.7 provides the total number of events and the average number of articles contained within each.

Finally, we apply a round of spectral clustering to these representations, compressing these events into 10 groups, which we deem topics. The numbers of events in each topic varies (see Table B.7).

Appendix B.5 Metadata

Article content is stored in `arrow` files, accessed using HuggingFace's `datasets` library (HuggingFace 2021). We store all generated metadata in a `duckdb` database (Mühleisen and Raasveldt 2024). Figure B.1 depicts the metadata schema. Table Sources provides information sourced from MBFC on different publishers, whereas tables Articles and

16 `https://huggingface.co/thenlper/gte-large`.

Topics provide information on the produced articles and events/topics, respectively. The articles are all given unique article identifiers. This ensures we can quickly generate relevant splits of the dataset and link these to article-level predictions.

Appendix B.6 Publisher-level Metadata

All publisher metadata is stored in the Sources table see Figure B.1. Each source is identified by a unique name, linked in a 1-to-many relationship to articles in the Articles table. The `url` column provides a link to the MBFC page with the source metadata. The `label` column provides the MBFC label, also shown in the lower rows of Table B.7. In column `bias`, the political bias (one of Extreme Left, Left, Left-Center, Least Biased, Pro-Science, Right-Center, Right or Extreme Right) of the publisher is provided. The `factuality` column provides an ordinal value for the propensity of a publisher to report factual news. Columns `country` and `type` provide information about the country of origin of the publisher, and their primary form of publication (e.g., TV, blog posts), respectively. In column `popularity`, the average number of visits to the publisher's main Web site is provided as an ordinal categorical value. `credibility` provides an overall assessment of the publishers' credibility, aggregating all other variables into a single score. Finally, the `conspiracy` and `pseudosci` columns provide the "strength" of the conspiracy or pseudo-science source. The larger this value is, the more these sources deviate from the public or scientific consensus. This is only provided for publishers labeled as Conspiracy-Pseudoscience.

The MBFC label and the metadata of publishers is subject to change, as additional information is made available or corrections are processed. We list the 20 publishers whose labels changed substantively during the data collection period in Table B.6. In

Table B.6
Publishers whose MBFC label changed substantively, from initial data collection (2017) to the publication of `misinfo-general` (Oct. 2024).

Publisher Name	Previous Label	New Label	Year	Reason
Big League Politics	Right	Questionable	2019	Failed fact checks
Daily Wire	Questionable	Right	2021	Correcting fact checks
Ecowatch	Pseudoscience	Left	2020	Editorial improvements
End Time Headlines	Conspiracy	Right-Center	2022	Improved publishing
FoxNews	Right	Questionable	2021	Failed fact checks, state propaganda
Just the News	Right	Questionable	2021	Failed fact checks
Newsmax	Right	Questionable	2020	Failed fact checks
One America News	Right	Questionable	2020	Failed fact checks
Pravda Report	Right	Questionable	2020	
Russia Insider	Right-Center	Questionable	2020	
Strategic Culture Foundation	Right-Center	Questionable	2019	Propaganda, conspiracy theories
The Drudge Report	Questionable	Right-Center	2020	Improved standards
The Political Insider	Right	Questionable	2019	
Townhall	Right	Questionable	2020	Failed fact checks
Truth Theory	Pseudoscience	Left	2020	Change in direction
Turning Point USA	Right	Questionable	2019	Failed fact checks
Washington Times	Right-Center	Questionable	2020	Failed fact checks
Western Journal	Right	Questionable	2020	Failed fact checks
WhatFinger	Right	Questionable	2022	Conspiracy theories
Wings over Scotland	Left-Center	Questionable	2021	Conspiracies, hate speech

most instances, the label changed from reliable$^+$ to unreliable$^-$ due to additional fact check failures being incorporated into the MBFC database (12/20). The inverse also occurs, where publishers improve their editorial practices or correct articles that failed fact checks (5/20). Most metadata changes correspond to relatively minor changes in political bias (e.g., Left-Center to Right-Center), and do not alter the analyses presented in this article.

Appendix B.7 Article Statistics

We present various statistics at the article level for each iteration of the final `misinfo-general` dataset in Table B.7. This includes the number of articles, the average number of tokens per article (post truncation), the number of events and the size of topics, and finally the label proportions, both in aggregated form as reliable$^+$ or unreliable$^-$,

Table B.7
Various statistics on the dataset and labels. ARTICLE COUNTS provides the total number of articles in each dataset year. LABEL PROPORTION provides the relative occurrence of reliable vs. unreliable articles, whereas PUBLISHERS provides the number of such publishers present. Section TRUNCATED TOKEN COUNTS provides the mean and the 25, 50, 75th quantiles of the number of tokens per article. Note that is after truncation at 512; the raw articles tend to be much longer. EVENT CLUSTERING provides the number and average size (in articles) of events. We also provide the number of events belonging to the smallest and largest topics. Sections RELIABLE LABELS and UNRELIABLE LABELS provides the proportion of each MBFC label in the dataset, split into publisher categories.

Statistic		2017	2018	2019	2020	2021	2022
Article Counts		0.10M	0.46M	0.59M	1.02M	1.04M	0.99M
Publishers	Reliable	43	99	134	189	183	184
	Unreliable	49	60	100	249	219	201
Truncated Token Counts	Mean	398.53	383.59	430.22	432.87	428.15	435.90
	Q25	299	266	358	369	366	383
	Q50	488	448	512	512	512	512
	Q75	512	512	512	512	512	512
Event Clustering	# Events	2,674	6,288	7,718	7,931	8,758	8,336
	Mean Event Size	38.67	73.38	76.87	128.49	118.24	119.29
	Smallest Topic Size	2.7K	20.5K	24.6K	27.7K	51.3K	48.4K
	Largest Topic Size	38.3K	94.1K	113.1K	192.3K	211.4K	209.4K
Label Proportion	reliable$^+$	61.39%	72.82%	75.91%	70.73%	71.36%	72.00%
	unreliable$^-$	38.61%	27.18%	24.09%	29.27%	28.64%	28.00%
reliable$^+$	Left	17.91%	13.32%	11.13%	9.01%	8.76%	9.88%
	Left-Center	24.09%	29.62%	35.19%	36.50%	34.53%	35.01%
	Least Biased	0.93%	2.18%	4.38%	4.91%	5.13%	3.97%
	Right-Center	4.34%	12.66%	9.81%	8.70%	10.38%	9.52%
	Right	14.12%	15.05%	15.41%	11.40%	12.03%	13.08%
	Pro-Science				0.21%	0.53%	0.55%
unreliable$^-$	Satire	1.70%	0.99%	0.82%	0.66%	0.48%	0.42%
	Questionable Source	27.83%	20.14%	17.34%	19.41%	19.00%	18.45%
	Consp.-Pseudosci.	9.09%	6.05%	5.93%	9.20%	9.17%	9.13%

Table B.8
An example of the topics present in the 2020 split of `misinfo-general`. As described in Section 5, to form the OoD set, the k smallest topics are chosen, such that collectively they (approximately) comprise 20% of all articles.

#	Description	# Articles		Labels		ID/OoD
				reliable$^+$	unreliable$^-$	
7	US Federal Elections	192k	18.8%	64.11%	35.89%	ID
9	International Affairs	161k	15.8%	75.80%	24.20%	ID
8	Health & COVID	149k	14.6%	70.85%	29.16%	ID
4	Entertainment & Sports	112k	11.0%	86.30%	13.70%	ID
5	Economy & Social Issues	112k	10.9%	63.50%	36.50%	ID
6	Science & Technology	89k	8.7%	71.68%	28.31%	ID
3	US Local Politics	66k	6.5%	68.85%	31.15%	OoD
2	Crime & Justice	64k	6.2%	68.93%	31.07%	OoD
0	Conflict	45k	4.4%	69.37%	30.63%	OoD
1	Biden Administration	28k	2.7%	60.43%	39.57%	OoD

but also per publisher category. In total, the dataset comprises some 2B tokens once truncated. Without truncation, this is likely substantially higher.

In Table B.8 we present descriptions of the various topics in the 2020 split of `misinfo-general`. These topics are latent, and automatically generated from the inter-event distance matrix. The space in which the articles were clustered into events is high-dimensional, and each cluster was assumed to be non-convex. As a result, it is difficult to find an "average" representation of any event, and even more difficult to find one for a topic (i.e., a cluster of clusters). Instead, we derived topic descriptions by sampling events from each topic, and using the `BERTopic` generated event representation.

Appendix B.8 Publisher Statistics

Across all dataset iterations, the vast majority of articles are written by a small minority of prolific publishers. This is shown in Figure B.2. The most prolific publishers have authored between 40% and 60% of articles within a year. The next deciles manage only 20%, then 10–15%, etc., with the smallest publishers authoring only a few articles. This effect seems to become more drastic for the later dataset years.

As discussed in Section 6 this can obscure poor performance when using article-level evaluation metrics. In such cases, the average performance score will tend to be dominated by the performance of the largest publishers, instead of the body of publishers as a whole. As shown in Figure 1 and Tables E.2 and E.3, performance on these less mainstream publishers tends to be substantially worse. As a result, the OoD performance metrics in Table 2 likely underestimate the generalization gap.

This dataset property means that publisher-level analyses are necessary, in addition to the far more common article-level performance measures. That said, prominent misinformation datasets tend to sample from a far smaller and homogenous set of publishers, with typically more balanced authorship proportions.

Another noteworthy statistic is the amount of overlap between publishers across the dataset years. While the dataset as a whole contains 488 distinct publishers, as can

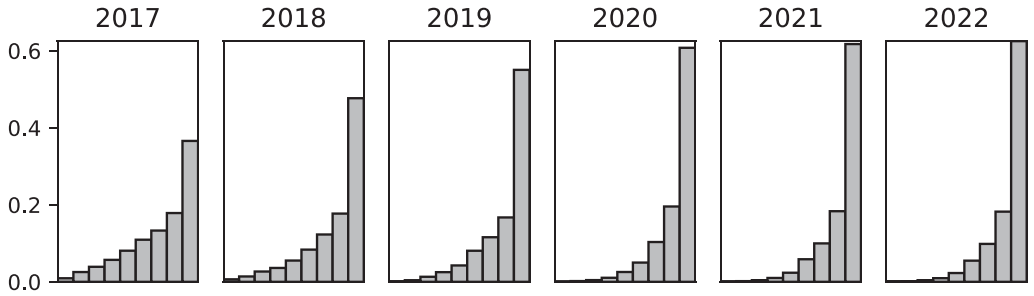

Figure B.2
The authorship of articles in the dataset, aggregated into deciles. The rightmost column of each panel represents the top 10% most prolific publishers, the leftmost the bottom 90%.

Table B.9
Overlap of publishers between years.

	2017	2018	2019	2020	2021	2022
2017	1.00	0.43	0.31	0.17	0.16	0.16
2018	0.43	1.00	0.62	0.34	0.33	0.32
2019	0.31	0.62	1.00	0.52	0.50	0.49
2020	0.17	0.34	0.52	1.00	0.86	0.80
2021	0.16	0.33	0.50	0.86	1.00	0.89
2022	0.16	0.32	0.49	0.80	0.89	1.00

be seen in Table B.7, each dataset year contains fewer total publishers. Furthermore, the distribution of those publishers (in terms of article counts) shifts across the years. This is partially due to a changing collection methodology used by the NELA authors.

As a result, the different dataset years have different degrees of overlap between each other. Table B.9 shows the publisher intersection-over-union overlaps between the 6 different dataset years. The largest block of overlap is between the years 2020 and 2022, with overlaps across datasets being well above 0.80. The IoU scores to other dataset years are far lower, with 2017 having particularly low overlaps.

Appendix C. Additional Information on Generalization Axes and Splits

Uniform. For this generalization axis, we use the industry standard method of stratified and shuffled train/validation/test splits. We maintain the ratios 80/10/20, respectively, throughout. This is meant to form the baseline set of values, both in terms of ID–OoD article performance, but also the expected delta between those.

Time. Misinformation changes quickly, but models have been found to be brittle to articles outside the time-span of the training data (Bozarth and Budak 2020; Horne, Nørregaard, and Adali 2020). We train models on a dataset from a single year, and evaluate performance on all other years. To avoid conflating Time and Publisher results, we only evaluate on articles from publishers seen in the training set. The proportion of publishers—and as a result, the reliable/unreliable proportions—does change drastically across dataset years.

Event. We define events as occurrences that span news articles with a definite time-span and sudden occurrence. Novel events spawn articles with a markedly different vocabulary, introducing words, names, or terms not yet available to the model. Prior work has shown models suffer when evaluated on unseen events (Lee et al. 2021; Ding et al. 2022). The dataset contains a multitude of such events, but we focus specifically on the COVID-19 pandemic; an event that is particularly notable for its rapid onset and the volume of misinformation it spawned (Bradd 2024). Models are trained on a subset of articles *not* containing COVID-19 keywords, and evaluated on ones that do. The article discovery-process is outlined in more detail in Appendix C.1.

Topic. We define topics to be relatively static collections of events, wherein the style or publishers' opinions change little across years or decades. We discover such topics automatically, bottom-up. Specifically, we apply a heavily modified variant of the `BERTopic` (Grootendorst 2022) algorithm to discover the many thousands of events that occur in each year of the dataset. Each event is textually represented by a TF-IDF vector. To group events into topics, we apply a second round of spectral clustering on the adjacency matrix induced by the inter-event cosine distance matrix. The number of articles contained by a topic differs substantially (see Table B.7). We reserve the n smallest topics, which collectively contain roughly 20% of all articles for testing, and train on articles from all other topics.

Publisher. All scraped articles are annotated with the news Web site, outlet, or publisher that produced the article (see Figure B.1). In general, news publishers have some editorial bias or stance, making their corpus of articles distinctly different from that of other publishers. Prior work has found misinformation classifiers to be particularly sensitive to changes in publisher (Barrón-Cedeño et al. 2019; Bozarth and Budak 2020). We specifically test for this by reserving the n smallest publishers, which collectively contain roughly 20% of all articles for the test set, and train on articles from all other publishers. We stratify this splitting across political bias, to ensure each political bias is represented equally in both the train and test sets.

Political Bias. Separate from editorial bias, publishers tend to exhibit a political bias as well. MBFC has annotated all publishers on a left-right political bias continuum. We map all Extreme Left and Left publishers to a coarser Left group, and vice versa for the Right-side of the political spectrum. All other political biases (Center-Left, Center-Right, Least Biased, and Pro-Science) were mapped to Center bias. We reserve either the Left or Right group for evaluation, and train on the collection of articles from the opposite political bias, along with all Center biased publishers.

Misinformation Type. Misinformation presents in various forms. MBFC divides unreliable publishers in three categories: Questionable-Source (publishers that exhibit extreme bias or propaganda), Conspiracy-Pseudoscience (publishers aligning with known conspiracies or pseudo-scientific practices), and Satire (publishers whose content is purposefully false for comedic effect). To test this form of generalization, we reserve either Questionable-Source or Conspiracy-Pseudoscience labels for evaluation, and train on all other articles. The test set includes a stratified sample of reliable sources as well.

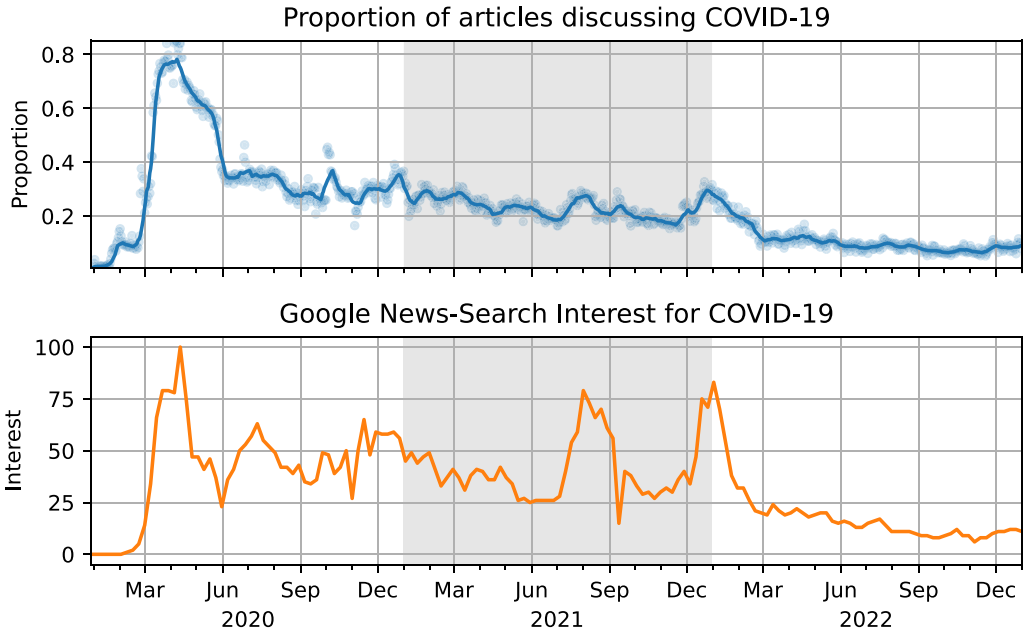

Figure C.1
The top figure shows the proportion of articles published that day that discuss the COVID-19 pandemic. The solid line provides a LOESS smoothed trend line. The bottom figure provides a Google search interest for topics that fall under Coronavirus disease 2019. The different years (2020, 2021, 2022) are displayed as banded vertical columns.

Appendix C.1 Identifying COVID-19 Articles

For Event splitting, we focus on a single event present over the latter 3 dataset years: the COVID-19 pandemic. We first derive a set of COVID-19 keywords (e.g., "sars-cov-2", "lockdown", "mask"), and combine it with the set of keywords defined by Gruppi, Horne, and Adalı (2021). We include all articles from 2020–2022 that include any of these terms in the held-out test set. While this is bound to induce a large number of false positives, it ensures no COVID-19 related terms contaminate the models' learned vocabulary. In Figure C.1 we display the correlation between the found number of articles and the amount of Google search volume.

Appendix C.2 Test Set Correlations

Publishers have a tendency to prioritize different topics. This is especially prevalent for less mainstream publishers, which have far fewer resources to spend on the breadth of their reporting. It is therefore plausible that the Publisher and Topic splits have some overlap.

We test for this by comparing the amount of article level overlap in the held-out test sets of the Publisher and Topic to random draws of the same size from uncorrelated uniform distributions. The larger this overlap, the more correlated the two test sets are. We display the results of our test in Table C.1. While we find more overlap than random draws, the effect is not large. Even in the most egregious case, there is only a

Table C.1
The amount of `article_id` overlap between the Publisher and Topic test sets. Column IoU gives the intersection over the union between the two held-out test sets. Column Random IoU compares this to the achieved value for 10,000 simulated draws from a uniform distribution of the same sizes as the actual test set sizes (given in the last two columns).

Year	IoU	Random IoU	Excess IoU	Test Set Sizes	
				Publisher	*Topic*
2017	0.1183	0.1084	9.07%	19,764	20,709
2018	0.1227	0.1162	5.62%	109,043	85,832
2019	0.1210	0.1133	6.80%	133,465	110,168
2020	0.1386	0.1116	24.24%	203,224	205,873
2021	0.1317	0.1055	24.82%	192,315	203,311
2022	0.1318	0.0992	32.93%	172,530	186,986

0.04 difference in absolute IoU. As such, we believe the two different test sets measure distinct generalization aspects.

Appendix D. Training Details

While `DeBERTa-v3` models can theoretically handle infinite sequences, to constrain memory requirements we truncate tokenization after the first 512 tokens. This allowed for a batch size of 64 during training, and 512 during validation. We used `Adam` with weight decay (Kingma and Ba 2015; Loshchilov and Hutter 2019) as the optimizer.

Some other important training hyperparameters are listed in Table D.1. We set a high learning rate for the classifier and pooler, but a very low value for the embeddings

Table D.1
An overview of important hyperparameters and their values. These were tuned on the validation set of the Uniform split dataset.

Hyperparameter	Value
`seed`	942
`max_length`	512
`val_prop`	0.1
`batchsize`	64
`pooler_dropout`	0.1
`lr`	
`embeddings`	5.0e-7
`pooler`	5.0e-5
`classifier`	1.0e-3
`weight_decay`	0.1
`lr_scheduler`	
`type`	polynomial
`power`	3.163
`max_steps`	3.0e+6
`warmup_ratio`	0.0213
`eval_prop`	0.05
`patience`	2

layer to avoid catastrophic forgetting. We utilized a polynomial decaying learning rate scheduler, which first linearly warms up, before decaying according to

$$\eta_t = \left(\frac{t}{\texttt{max\_steps}}\right)^{\texttt{power}} \eta_{t-1}$$

For each experiment, regardless of the size of the dataset, we allow for a maximum of $3.0e+6$ update steps, evaluating on the validation set every 5%. We use early stopping with patience of 2, through which training usually concludes well before reaching the maximum budget.

Appendix D.1 Additional Training Details

All experiments were conducted on SNELLIUS, a Linux SLURM-based supercomputer. Nodes consist of an Intel Xeon Platinum 8360Y CPU with 18 cores, 2.4 GHz speed, and a single NVIDIA A100 GPU accelerator (40 GiB of HMB2 memory), and 128 GiB of DDR4 memory.

We use Python 3.11 and PyTorch 2.2.2 built with CUDA 11.8. For training, we make extensive use of utilities implemented in HuggingFace's `transformers` library: `transformers` 4.37.2, `datasets` 2.19.0, and `accelerate` 0.30.1. All experiments were conducted under random seed 942. For local development, we use Ubuntu 20.04.6 LTS (GNU/Linux 5.15.90.1-microsoft-standard-WSL2×86_64).

Appendix D.2 LLM Inference

To conduct LLM inference, we use Meta's `llama-3-8B-Instruct` using the following prompt:

> <<SYS>>
> You are a content moderator working with journalistic articles. Your task is to identify articles from unreliable publishers.
> <</SYS>>
>
> Does the following text come from a reliable news publisher?
> Respond with 'yes' or 'no'.
> Article: ${ARTICLE}
> Does this article come from a reliable news publisher? 'yes' or 'no': [/INST]

Here, ${ARTICLE} is replaced by the first 512 tokens of each article.

We use the `pmi_dc` decision rule, introduced by Holtzman et al. (2021). We use the same prompt without the article tokens as the domain conditional text. Empirically, this performs slightly better across the entire corpus.

Due to the (very) long articles prevalent in this dataset, using few-shot exemplars was deemed intractable.

Appendix D.3 Annotating Article Properties

Subjectivity. To annotate articles for subjectivity, we used `ChatGPT4o-mini`, and `ChatGPT4o`. The former was used to annotate the bulk of articles, and the latter to verify the overall quality of the annotations. We used the following prompt:

> <<SYS>>
> You are a helpful assistant, helping analyse the properties of news articles. Before a final answer, make sure to explain your thinking.
> <</SYS>>
>
> Please classify how objective the following article is.
> Objective articles take a neutral stance on topics, and focus on reporting factual news.
> Subjective articles instead focus on opinions, which are more difficult to verify and can take specific stances for or against topics.
> The title and body are provided. After you provide your reasoning, respond with one of entirely objective, mostly objective, mixed, mostly subjective, entirely subjective, and nothing else.
> Article: `${ARTICLE}`

Table D.2 gives the classification correspondence between the two models. Overall, we find that the mini model aligns well with the larger model, and that the larger model tends towards *more* subjective annotations, rather than less. As such, it is plausible that the subjectivity within the dataset remains underreported. We include some examples of articles and their subjectivity annotations in `https://github.com/ioverho/misinfo-general/tree/main/assets/subjectivity_annotations_examples`.

Emotion. We perform a similar analysis, but now for emotions present in the articles. We use the following prompt, and only use `ChatGPT4o-mini`:

> <<SYS>>
> You are a helpful assistant, analyzing the properties of news articles. Before a final answer, make sure to explain your thinking.
> <</SYS>>

Table D.2
A confusion matrix between the subjectivity assessments of `ChatGPT4o` (rows) and `ChatGPT4o-mini` (columns).

	E. Obj.	M. Obj.	Mixed	M. Subj.	E. Subj.	
E. Obj.	5.71%	5.13%	0.39%	0.01%		
M. Obj.	1.80%	12.30%	3.76%	0.97%		
Mixed		3.37%	15.19%	7.62%	0.49%	
M. Subj.	0.05%	0.39%	4.35%	14.70%	4.00%	
E. Subj.			0.83%	5.08%	13.77%	

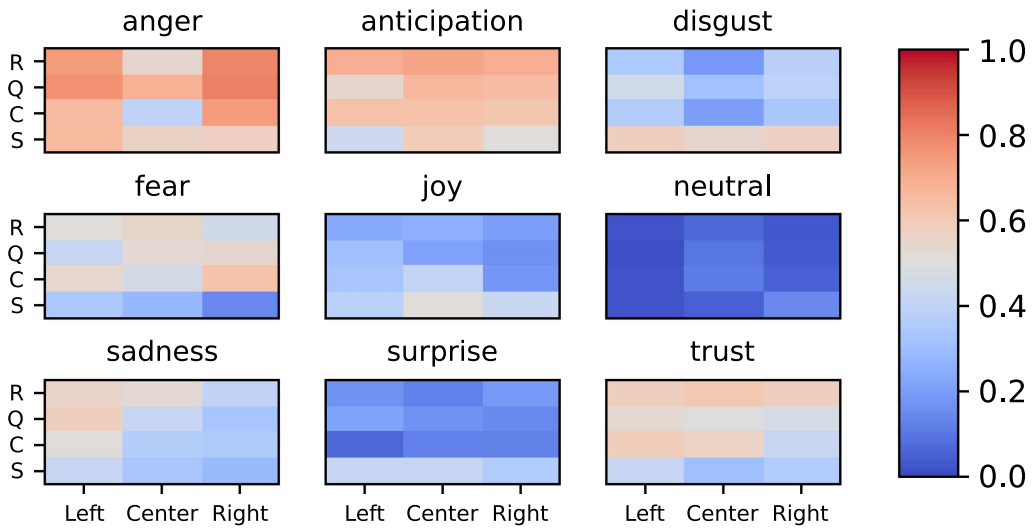

Figure D.1
Emotion propensity, like Figure 5, but now in absolute values, and presented for each emotion in isolation. More intense, red colors signify a higher presence of an emotion, and cooler, blue colors the opposite.

Please identify the dominant emotions in the following article.
We will be using Plutchik's 8 basic emotions, along with a label for neutral:

1. Joy: a feeling of happiness, pleasure, or contentment.
2. Sadness: a feeling of loss, disappointment, or grief
3. Trust: a sense of safety, security, and connection with others
4. Disgust: a strong aversion to something unpleasant, often related to taste, smell, or moral judgments
5. Fear: a response to perceived danger, leading to caution or escape behaviors
6. Anger: a reaction to perceived threats or injustice
7. Surprise: a reaction to something unexpected
8. Anticipation: looking forward to or expecting something, which can bring excitement or anxiety
9. Neutral: emotional balance, no strong positive or negative emotions are present

To respond, please list the emotions present in the article. When labeling the article for the 'Neutral' emotion, please make sure no other emotion is present.
The article's title and body are provided. After you provide your reasoning, respond with a list of joy, sadness, trust, disgust, fear, anger, surprise, anticipation, neutral, and nothing else.
Article: `${ARTICLE}`

Figure D.1 presents the same data as in Figure 5, but now transposed (all publisher class combinations per emotion), and in absolute values. We find emotion overall to be present in all forms of articles and publishers.

Appendix E. Additional Results

Appendix E.1 Political Bias

To further assess the model's interaction with political bias, we compute the complement of the expected publisher-level accuracy score for political bias and publisher label (reliable or unreliable), and divide it by the marginal expected score:

$$\frac{1 - \mathbb{E}[\text{ACC}_p|\text{bias}(p), y_p]}{1 - \mathbb{E}[\text{ACC}_p|y_p]} - 1$$

essentially giving the shifted exponentiated pointwise mutual information (PMI) of making an error for a particular combination of publisher political bias and label,

$$\frac{p(\hat{y}_p \neq y|\text{bias}(p), y)}{p(\hat{y}_p \neq y|y)} - 1 = \exp\left(\text{pmi}\left(\hat{y}_p \neq y; \text{bias}(p)|y\right)\right) - 1$$

The higher this value is, the greater the association between making an error for a particular political bias deviates from the expected value. If an error is made for a reliable publisher, this corresponds to a False Positive (FP), whereas an error for an unreliable publisher would correspond to a False Negative (FN). We introduce the shifting to ensure no bias corresponds with 0, whereas positive and negative bias correspond to positive and negative values, respectively.

In Figure E.1 we display various such biases for each iteration of the corpus. As seen in Table E.2, performance degrades for the more extreme political biases. There is, however, a notable difference between left-biased and right-biased publishers, and especially for the reliable right-biased publishers. The probability of the model generating a FP is substantially higher (sometimes near twice as likely) than for reliable leftist- or center-biased publishers. We conclude that the model tends to confuse articles from reliable right-biased publishers with unreliable ones, far beyond what would be considered "due to chance".

On the other hand, the model tends to be overly optimistic for center-biased publishers, with lower FP probabilities and a tendency for slightly higher FN probabilities. An unreliable publisher could therefore escape moderation by limiting its explicit political views. While this is a relatively unlikely example, as most unreliable publishers are unreliable *because* of overt extreme political views, it does showcase an important blind-spot for these models.

Given the high-stakes nature of misinformation moderation decisions, this indicates the presence of undesirable behavior. In their current state, our misinformation detection models discriminate against publishers of a particular political group.

Appendix E.2 Publisher Heterogeneity

In Section 7.2 we tested for the effect of publisher heterogeneity on generalizing to unseen publishers. We operationalized this by re-running the Publisher split experiment with only the top-n most prolific publishers for each MBFC label, while leaving the test set untouched.

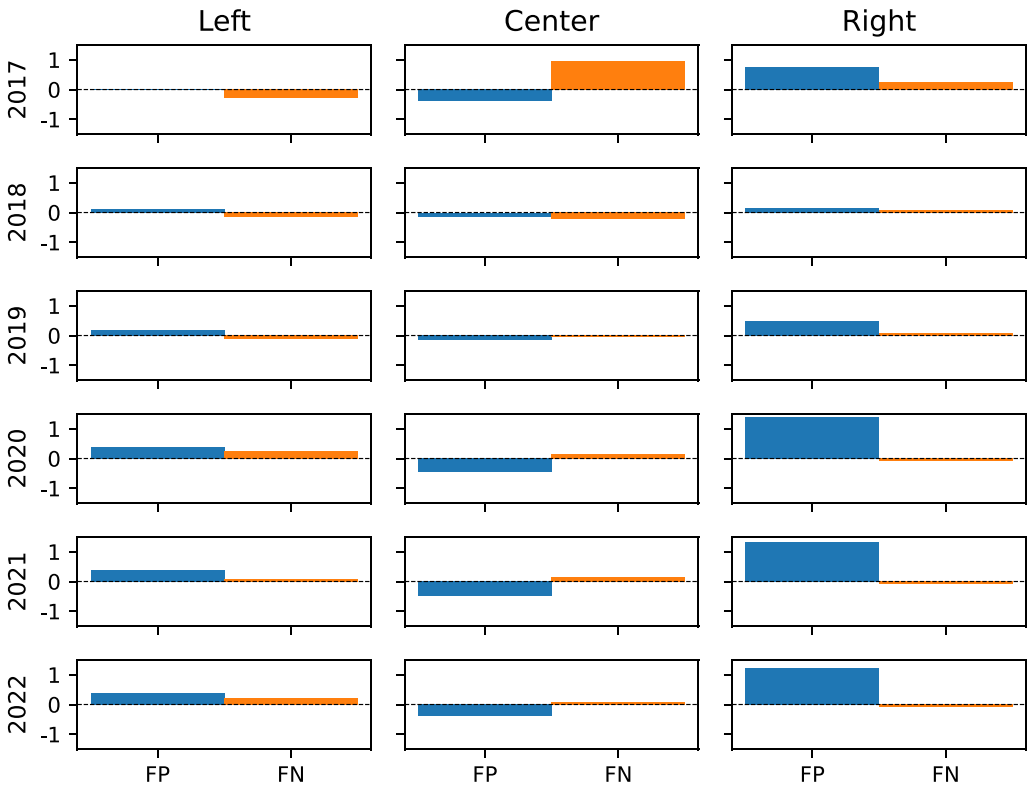

Figure E.1
Probabilities of making an error for publishers of a particular label and political bias, relative to the probability of making a mistake for publishers of the same label. FP columns denote the probability of error for reliable publishers, and FN vice versa.

Table E.1
Mean number of articles and publishers retained in the training set after limiting to the top-n most prolific publishers. Note that the number of articles and publishers deviates from the distributions depicted in Figure B.2. In this case, we additionally condition on the publisher's MBFC label.

Quantity	Top 1	Top 2	Top 3
Articles	28%	43%	55%
Publisher	11%	20%	30%

This dramatically reduces the amount of variation in publishers, while minimally affecting the total amount of data present. This can be seen in Table E.1. Note that these are expressed as percentages, but in absolute terms constitute hundreds of thousands of articles.

Appendix E.3 Determinants of Publisher-Level Accuracy

To estimate which factors impact accuracy at the publisher level, we first define publisher level accuracy for a publisher p as

$$\text{ACC}_p = \frac{1}{|\mathcal{D}^{(p)}|} \sum_{d \in \mathcal{D}^{(p)}} \mathbb{1}(\hat{y}^{(d)} = y^{(p)})$$

or put otherwise, the expected rate of correct classification for all documents belonging to that publisher.

To determine which factors contribute to these scores, we use a multinomial logistic regression model. Specifically, we use `statsmodels` (Seabold and Perktold 2010) to estimate the dependent variables as

$$\sum_{d \in \mathcal{D}^{(p)}} \mathbb{1}(\hat{y}^{(d)} = y^{(p)}) + \mathbb{1}(\hat{y}^{(d)} \neq y^{(p)}) = \sigma(\beta_0 + \sum_{i=1}^{K} \beta_i x_i)$$

We use $\mu^2 \cdot (1 - \mu)^2$ as the variance function, Pearson's χ^2 as the scale value, and the logit as the link function σ.

Uniform Generalization. The fully specified model is displayed in Table E.2, with some important variables' coefficients depicted graphically in Figure 1. The leftmost column provides natural groupings of the different variables. The interpretation of the coefficient magnitudes should be adjusted to consider the range of possible values for the corresponding variable.

In the *Generic* category, we include the logarithm of the number of articles present in the training data, and whether the publisher is foreign or not. Both variables are assigned large coefficients. Every 10x increase of labeled articles increases the odds of correct classification by a factor of 1.91. Foreign publishers tend to be 2.91 times as easy to classify relative to U.S.-based publishers.

The *Year* category of variables includes a dummy variable for each iteration of the dataset. Here we find relatively little difference, despite statistical significance, with coefficients falling between 0.85 and 1.15.

We further include dummy variables for the different political biases and MBFC labels (assumes Center-Reliable to be the default group). We furthermore include an interaction effect between all political biases and MBFC labels. We find that unreliable publishers are substantially more difficult to classify, with all unreliable labels having large negative coefficients. This is especially true for the Questionable Source category of articles, with log-odds shifts of -1.25 and -1.84 for left- and right-biased publishers, respectively. Furthermore, corroborating the analysis in Appendix E.1, the models perform much more poorly for publishers on the right side of the political spectrum.

In the *Strength* block of variables, we include MBFC's strength of conspiracy or pseudo-science. All Conspiracy-Pseudoscience sources are rated on a scale, with larger values indicating a further deviation from the societal norm. Reliable sources are given a default score of 0. We find that there exists a slight, but consistent, positive correlation with the *strength* of the pseudo-scientific or conspiratorial claims. In other words, the more those publishers deviate from the status quo, the more easily these are identified.

Similarly, for the *Factuality Score* block, we include the MBFC score for a publisher's propensity for factual reporting as interaction has an affect on each MBFC label (excluding Satire). Again, we find a slight positive correlation for reliable sources, meaning that the more often such publishers report factual information, the better classification performance. For the unreliable publishers, however, the opposite holds: the less likely to present factual news, the easier the classification. In either case, an "average" level of factuality tends to correspond to more ambiguous cases, lying closing to the model's decision boundary.

Topic Distance. We repeat this analysis, but this time aggregating entities at the event/label level, using the Topic split results. The model specification is left as above, with the introduction of a single additional variable: the minimum cosine distance to any event in the training set. The coefficients are provided in Table E.3.

While comparison across Tables E.2 and E.3 is not directly possible, as each estimates using different entities and different models, it is encouraging to see similar magnitudes for most variables' coefficients.

Surprisingly, we observe a very small but positive effect: $\beta_{\text{min\_dist}} = 0.15$. This implies that a topic that is as far away as possible from all topics in the training set sees a 1.16 increased odds of correct classification. While this is a weak effect, it does suggest that events more typical to a certain topic are easier to classify than those near the topic cluster borders.

Table E.2
Coefficients of the publisher-level determinants model, using the uniformly split models. Range indicates the possible values each variable can take, with .. indicating all natural numbers between the extremes (inclusive). β provides the logistic parameter, and exp(β) its exponent (i.e., the log-odds ratio). Column Std. Err. provides the standard error, p the corresponding *p*-value, and 95% CI the confidence interval. In the parameter names, the colon indicates an interaction term.

Group	Parameter	Range	β	exp(β)	Std. Err.	p	95% CI	
Intercept	Intercept	1	−0.85	0.43	0.03	0.00	−0.91	−0.79
Generic	$\log_{10}$(train count)	$[0, \infty]$	0.65	1.91	0.01	0.00	0.64	0.66
	Foreign	0..1	1.07	2.91	0.01	0.00	1.06	1.08
Year	2018	0..1	0.07	1.08	0.02	0.00	0.04	0.11
	2019	0..1	0.14	1.15	0.02	0.00	0.10	0.17
	2020	0..1	−0.16	0.85	0.02	0.00	−0.19	−0.13
	2021	0..1	−0.06	0.94	0.02	0.00	−0.09	−0.03
	2022	0..1	0.00	1.00	0.02	0.85	−0.03	0.04
Political Bias	Left	0..1	−0.39	0.68	0.01	0.00	−0.41	−0.37
	Right	0..1	−0.64	0.53	0.01	0.00	−0.66	−0.62
Label	Conspiracy-PseudoScience	0..1	−0.32	0.73	0.22	0.15	−0.76	0.11
	Questionable Source	0..1	−2.40	0.09	0.02	0.00	−2.44	−2.37
	Satire	0..1	−1.58	0.21	0.18	0.00	−1.94	−1.23
Interactions	Left: Conspiracy-PseudoScience	0..1	−1.65	0.19	0.19	0.00	−2.02	−1.28
	Left: Questionable Source	0..1	1.54	4.69	0.03	0.00	1.48	1.61
	Left: Satire	0..1	−2.15	0.12	0.20	0.00	−2.54	−1.75
	Right: Conspiracy-PseudoScience	0..1	−1.70	0.18	0.19	0.00	−2.06	−1.33
	Right: Questionable Source	0..1	1.20	3.31	0.02	0.00	1.17	1.23
	Right: Satire	0..1	0.56	1.75	0.35	0.11	−0.12	1.25
Strength	Conspiracy	1..5	0.19	1.21	0.02	0.00	0.15	0.22
	PseudoScience	1..5	0.17	1.18	0.01	0.00	0.14	0.19
Factuality Score	Factuality: Reliable	1..5	0.35	1.42	0.00	0.00	0.34	0.36
	Factuality: Conspiracy-PseudoScience	1..5	−0.09	0.92	0.03	0.00	−0.14	−0.04
	Factuality: Questionable Source	1..5	−0.13	0.88	0.01	0.00	−0.14	−0.11

Table E.3
Idem, but now for the publisher-topic aggregated binomial logistic model. Since the aggregation level differs from that of the model presented in Table E.2, the coefficients are not directly comparable.

Group	Parameter	Range	β	exp(β)	Std. Err.	p	95% CI	
Intercept	Intercept	1	−0.49	0.61	0.05	0.00	−0.60	−0.38
Distance	Topic Distance	[0, 1]	0.15	1.16	0.03	0.00	0.09	0.21
Generic	$\log_{10}$(train count)	[0, ∞]	0.56	1.75	0.01	0.00	0.54	0.58
	Foreign	0..1	0.88	2.42	0.01	0.00	0.86	0.90
Year	2018	0..1	0.23	1.26	0.03	0.00	0.17	0.29
	2019	0..1	0.30	1.35	0.03	0.00	0.24	0.36
	2020	0..1	−0.18	0.83	0.03	0.00	−0.24	−0.13
	2021	0..1	−0.47	0.62	0.03	0.00	−0.53	−0.41
	2022	0..1	−0.27	0.76	0.03	0.00	−0.33	−0.21
Political Bias	Left	0..1	−2.81	0.06	0.47	0.00	−3.74	−1.88
	Right	0..1	−2.58	0.08	0.03	0.00	−2.64	−2.52
Label	Conspiracy-PseudoScience	0..1	−1.51	0.22	0.25	0.00	−1.99	−1.03
	Questionable Source	0..1	−0.54	0.58	0.01	0.00	−0.57	−0.51
	Satire	0..1	−0.73	0.48	0.02	0.00	−0.76	−0.70
Interactions	Left: Conspiracy-PseudoScience	0..1	−0.38	0.69	0.41	0.36	−1.18	0.43
	Left: Questionable Source	0..1	2.11	8.27	0.05	0.00	2.02	2.21
	Left: Satire	0..1	−1.83	0.16	0.31	0.00	−2.44	−1.22
	Right: Conspiracy-PseudoScience	0..1	−0.73	0.48	0.41	0.07	−1.53	0.06
	Right: Questionable Source	0..1	1.92	6.85	0.03	0.00	1.87	1.98
	Right: Satire	0..1	0.32	1.37	0.45	0.48	−0.56	1.19
Strength	Conspiracy	1..5	0.42	1.52	0.04	0.00	0.35	0.49
	PseudoScience	1..5	0.29	1.33	0.03	0.00	0.24	0.34
Factuality Score	Factuality: Reliable	1..5	0.38	1.47	0.01	0.00	0.37	0.40
	Factuality: Conspiracy-PseudoScience	1..5	0.42	1.53	0.05	0.00	0.33	0.52
	Factuality: Questionable Source	1..5	−0.44	0.65	0.01	0.00	−0.46	−0.41

Acknowledgments

This work was supported by Meta through a Facebook Research Individual Research Project grant, through the project "Modelling Emotion-inducing Communication Strategies to Detect Misinformation". We would like to thank the anonymous reviewers for their time, effort, and insights. Their work has helped improve this article substantially. We additionally would like to thank Mauricio Gruppi and Benjamin Horne, along with their colleagues, for making the NELA datasets available and their rapid response to questions.

References

Agresti, Alan and Brent A. Coull. 1998. Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, 52(2):119–126. https://doi.org/10.1080/00031305.1998.10480550

Ahmed, Hadeer, Issa Traore, Sherif Saad, and Mohammad Mamun. 2024. Effect of text augmentation and adversarial training on fake news detection. *IEEE Transactions on Computational Social Systems*, 11(4):4775–4789. https://doi.org/10.1109/TCSS.2023.3344597

Aïmeur, Esma, Sabrine Amri, and Gilles Brassard. 2023. Fake news, disinformation and misinformation in social media: A review. *Social Network Analysis and Mining*, 13(1):30. https://doi.org/10.1007/s13278-023-01028-5, PubMed: 36789378

Alam, Firoj, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimiter Dimitrov, Giovanni Da San Martino, Shaden Shaar, Hamed Firooz, and Preslav Nakov. 2022. A survey on multimodal disinformation detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6625–6643.

Alba-Juez, Laura and J Lachlan Mackenzie. 2019. Emotion, lies, and "bullshit" in journalistic discourse: The case of fake news. *Iberica*, (39).

Allcott, Hunt and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–236. https://doi.org/10.1257/jep.31.2.211

Altay, Sacha, Manon Berriche, Hendrik Heuer, Johan Farkas, and Steven Rathje. 2023. A survey of expert views on misinformation: Definitions, determinants, solutions, and future of the field. *Harvard Kennedy School Misinformation Review*. https://doi.org/10.37016/mr-2020-119

Bago, Bence, Leah R. Rosenzweig, Adam J. Berinsky, and David G. Rand. 2022. Emotion may predict susceptibility to fake news but emotion regulation does not seem to help. *Cognition & Emotion*, 36(6):1166–1180. https://doi.org/10.1080/02699931.2022.2090318, PubMed: 35749076

Baly, Ramy, Giovanni Da San Martino, James Glass, and Preslav Nakov. 2020a. We can detect your bias: Predicting the political ideology of news articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 4982–4991. https://doi.org/10.18653/v1/2020.emnlp-main.404

Baly, Ramy, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018a. 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. https://doi.org/10.18653/v1/D18-1389

Baly, Ramy, Georgi Karadzhov, Jisun An, Haewoon Kwak, Yoan Dinkov, Ahmed Ali, James Glass, and Preslav Nakov. 2020b. 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. https://doi.org/10.18653/v1/2020.acl-main.308

Baly, Ramy, 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. https://doi.org/10.18653/v1/N19-1216

Baly, Ramy, Mitra Mohtarami, James Glass, Lluís Màrquez, Alessandro Moschitti, and Preslav Nakov. 2018b. Integrating stance detection and fact checking in a unified corpus. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 21–27. https://doi.org/10.18653/v1/N18-2004

Barrón-Cedeño, Alberto, Israa Jaradat, Giovanni Da San Martino, and Preslav Nakov. 2019. Proppy: Organizing the news based on their propagandistic content. *Information Processing & Management*, 56(5):1849–1864. https://doi.org/10.1016/j.ipm.2019.03.005

Bianchi, John, Manuel Pratelli, Marinella Petrocchi, and Fabio Pinelli. 2024. Evaluating trustworthiness of online news publishers via article classification. In *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing*, SAC '24, pages 671–678. https://doi.org/10.1145/3605098.3636044

Bodaghi, Arezo, Ketra A. Schmitt, Pierre Watine, and Benjamin C. M. Fung. 2024. A literature review on detecting, verifying, and mitigating online misinformation. *IEEE Transactions on Computational Social Systems*, 11(4):5119–5145. https://doi.org/10.1109/TCSS.2023.3289031

Boididou, Christina, Katerina Andreadou, Symeon Papadopoulos, Duc-Tien Dang-Nguyen, Giulia Boato, Michael Riegler, and Yiannis Kompatsiaris. 2015. Verifying multimedia use at MediaEval 2015. In *Multimedia Benchmark Workshop*, volume 1436.

Bojic, Ljubisa, Nikola Prodanovic, and Agariadne Dwinggo Samala. 2024. Maintaining journalistic integrity in the digital age: A comprehensive NLP framework for evaluating online news content. *International Journal for Quality Research*, 20(1):201–218. https://doi.org/10.24874/IJQR20.01-13

Bostan, Laura Ana Maria and Roman Klinger. 2018. An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119.

Bozarth, Lia and Ceren Budak. 2020. Toward a better performance evaluation framework for fake news classification. *Proceedings of the International AAAI Conference on Web and Social Media*, 14:60–71. https://doi.org/10.1609/icwsm.v14i1.7279

Bradd, Sam. 2024. Infodemic. https://www.who.int/health-topics/infodemic

Broniatowski, David A., Daniel Kerchner, Fouzia Farooq, Xiaolei Huang, Amelia M. Jamison, Mark Dredze, Sandra Crouse Quinn, and John W. Ayers. 2022. Twitter and Facebook posts about COVID-19 are less likely to spread misinformation compared to other health topics. *PLoS ONE*, 17(1):e0261768. https://doi.org/10.1371/journal.pone.0261768, PubMed: 35020727

Burdisso, Sergio, Dairazalia Sanchez-cortes, Esaú Villatoro-tello, and Petr Motlicek. 2024. Reliability estimation of news media sources: Birds of a feather flock together. In *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. https://doi.org/10.18653/v1/2024.naacl-long.383

Casavantes, Marco, Manuel Montes-y-Gómez, Luis Carlos González, and Alberto Barróñ-Cedeno. 2024. Propitter: A Twitter corpus for computational propaganda detection. In *Advances in Soft Computing*, pages 16–27. https://doi.org/10.1007/978-3-031-47640-2_2

Cheng, Mingxi, Shahin Nazarian, and Paul Bogdan. 2020. VRoC: Variational autoencoder-aided multi-task rumor classifier based on text. In *Proceedings of The Web Conference 2020*, WWW '20, pages 2892–2898. https://doi.org/10.1145/3366423.3380054

Chicco, Davide and Giuseppe Jurman. 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(1):6. https://doi.org/10.1186/s12864-019-6413-7, PubMed: 31898477

Chicco, Davide and Giuseppe Jurman. 2022. An invitation to greater use of Matthews correlation coefficient in robotics and artificial intelligence. *Frontiers in Robotics and AI*, 9:876814. https://doi.org/10.3389/frobt.2022.876814, PubMed: 35402520

Chinn, Susan. 2000. A simple method for converting an odds ratio to effect size for use in meta-analysis. *Statistics in Medicine*, 19(22):3127–3131. https://doi.org/10.1002/1097-0258(20001130)19:22<3127::AID-SIM784>3.0.CO;2-M, PubMed: 11113947

Chu, Samuel Kai Wah, Runbin Xie, and Yanshu Wang. 2021. Cross-language fake news detection. *Data and Information Management*, 5(1):100–109. https://doi.org/10.2478/dim-2020-0025

Cui, Limeng and Dongwon Lee. 2020. CoAID: COVID-19 healthcare misinformation dataset. *arXiv preprint arXiv:2006.00885*. https://doi.org/10.48550/arXiv.2006.00885

Dai, Enyan, Yiwei Sun, and Suhang Wang. 2020. Ginger cannot cure cancer: Battling fake health news with a comprehensive data repository. *Proceedings of the International AAAI Conference on Web and Social Media*, 14:853–862. https://doi.org/10.1609/icwsm.v14i1.7350

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*. https://doi.org/10.48550/arXiv.2501.12948

Ding, Yasan, Bin Guo, Yan Liu, Yunji Liang, Haocheng Shen, and Zhiwen Yu. 2022. MetaDetector: Meta event knowledge transfer for fake news detection. *ACM Transactions on Intelligent Systems and Technology*. https://doi.org/10.1145/3532851

Flach, Peter and Meelis Kull. 2015. Precision-Recall-Gain curves: PR analysis done right. In *Advances in Neural Information Processing Systems*, volume 28.

Galassi, Andrea, Federico Ruggeri, Alberto Barrón-Cedeño, Firoj Alam, Tommaso Caselli, Mucahid Kutlu, Julia Maria Struß, Francesco Antici, Maram Hasanain, Juliane Köhler, et al. 2023. Notebook for the CheckThat! Lab at CLEF 2023. In *Conference and Labs of the Evaluation Forum*, volume 3497.

Gelfert, Axel. 2018. Fake news: A definition. *Informal Logic*, 38(1):84–117. https://doi.org/10.22329/il.v38i1.5068

Gemini-2.5-team. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*. https://doi.org/10.48550/arXiv.2507.06261

Ghanem, Bilal, Paolo Rosso, and Francisco Rangel. 2020. An emotional analysis of false information in social media and news articles. *ACM Transactions on Internet Technology*, 20(2):1–18. https://doi.org/10.1145/3381750

Golbeck, Jennifer, Matthew Mauriello, Brooke Auxier, Keval H. Bhanushali, Christopher Bonk, Mohamed Amine Bouzaghrane, Cody Buntain, Riya Chanduka, Paul Cheakalos, Jennine B. Everett, et al. 2018. Fake news vs satire: A dataset and analysis. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '18, pages 17–21. https://doi.org/10.1145/3201064.3201100

Grootendorst, Maarten. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*. https://doi.org/10.48550/arXiv.2203.05794

Gruppi, Maurício, Benjamin D. Horne, and Sibel Adalı. 2020. NELA-GT-2019: A large multi-labelled news dataset for the study of misinformation in news articles. *arXiv preprint arXiv:2003.08444*. https://doi.org/10.48550/arXiv.2003.08444

Gruppi, Maurício, Benjamin D. Horne, and Sibel Adalı. 2021. NELA-GT-2020: A large multi-labelled news dataset for the study of misinformation in news articles. *arXiv preprint arXiv:2102.04567*. https://doi.org/10.48550/arXiv.2102.04567

Gruppi, Maurício, Benjamin D. Horne, and Sibel Adalı. 2022. NELA-GT-2021: A large multi-labelled news dataset for the study of misinformation in news articles. *arXiv preprint arXiv:2203.05659*. https://doi.org/10.48550/arXiv.2203.05659

Gruppi, Maurício, Benjamin D. Horne, and Sibel Adalı. 2023. NELA-GT-2022: A large multi-labelled news dataset for the study of misinformation in news articles. *arXiv preprint arXiv:2203.05659*. https://doi.org/10.48550/arXiv.2203.05659

He, Pengcheng, Jianfeng Gao, and Weizhu Chen. 2022. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *International Conference on Learning Representations*.

Holtzman, Ari, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. 2021. Surface form competition: Why the highest probability answer isn't always right. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7038–7051. https://doi.org/10.18653/v1/2021.emnlp-main.564

Horne, Benjamin, Sara Khedr, and Sibel Adali. 2018. Sampling the news producers: A large news and feature data set for the study of the complex media landscape. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1). https://doi.org/10.1609/icwsm.v12i1.14982

Horne, Benjamin D., Maurício Gruppi, and Sibel Adalı. 2020. Do all good actors look the same? Exploring news veracity detection across the U.S. and the U.K. *arXiv preprint arXiv:2006.01211*. https://doi.org/10.48550/arXiv.2006.01211

Horne, Benjamin D., Jeppe Nørregaard, and Sibel Adali. 2020. Robust fake news detection over time and attack. *ACM Transactions on Intelligent Systems and*

Technology, 11(1):1–23. https://doi.org/10.1145/3363818

Horner, Christy Galletta, Dennis Galletta, Jennifer Crawford, and Abhijeet Shirsat. 2021. Emotions: The unexplored fuel of fake news on social media. *Journal of Management Information Systems*, 38(4):1039–1066. https://doi.org/10.1080/07421222.2021.1990610

Hoy, Nathaniel and Theodora Koulouri. 2022. Exploring the generalisability of fake news detection models. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 5731–5740. https://doi.org/10.1109/BigData55660.2022.10020583

Hu, Beizhe, Qiang Sheng, Juan Cao, Yongchun Zhu, Danding Wang, Zhengjia Wang, and Zhiwei Jin. 2023. Learn over past, evolve for future: Forecasting temporal trends for fake news detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 116–125. https://doi.org/10.18653/v1/2023.acl-industry.13

HuggingFace. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184.

Jenkins, Thomas, Sameerah Talafha, and Adam Goodwin. 2023. An ordered sample consensus (ORSAC) method for data cleaning inspired by RANSAC: Identifying probable mislabeled data. https://doi.org/10.36227/techrxiv.23511453.v1

Jeronimo, Caio Libanio Melo, Leandro Balby Marinho, Claudio E. C. Campelo, Adriano Veloso, and Allan Sales Da Costa Melo. 2019. Fake news classification based on subjective language. In *Proceedings of the 21st International Conference on Information Integration and Web-based Applications & Services*, pages 15–24. https://doi.org/10.1145/3366030.3366039

Jin, Zhiwei, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM International Conference on Multimedia*, MM '17, pages 795–816. https://doi.org/10.1145/3123266.3123454

Kiesel, Johannes, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. 2019. SemEval-2019 Task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 829–839. https://doi.org/10.18653/v1/S19-2145

Kingma, Diederik P. and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference for Learning Representations*.

Kochkina, Elena, Tamanna Hossain, Robert L. Logan, Miguel Arana-Catania, Rob Procter, Arkaitz Zubiaga, Sameer Singh, Yulan He, and Maria Liakata. 2023. Evaluating the generalisability of neural rumour verification models. *Information Processing & Management*, 60(1):103116. https://doi.org/10.1016/j.ipm.2022.103116

Koenders, Camille, Johannes Filla, Nicolai Schneider, and Vinicius Woloszyn. 2021. How vulnerable are automatic fake news detection methods to adversarial attacks? *arXiv preprint arXiv:2107.07970*. https://doi.org/10.48550/arXiv.2107.07970

Koivunen, Anu, Antti Kanner, Maciej Janicki, Auli Harju, Julius Hokkanen, and Eetu Mäkelä. 2021. Emotive, evaluative, epistemic: A linguistic analysis of affectivity in news journalism. *Journalism*, 22(5):1190–1206. https://doi.org/10.1177/1464884920985724

Kruijver, Kimberley, Neill Bo Finlayson, Beatrice Cadet, and Sico Van Der Meer. 2025. The disinformation lifecycle: An integrated understanding of its creation, spread and effects. *Discover Global Society*, 3(1):58. https://doi.org/10.1007/s44282-025-00194-5

Kuntur, Soveatin, Anna Wróblewska, Marcin Paprzycki, and Maria Ganzha. 2024. Fake news detection: It's all in the data! *arXiv preprint arXiv:2407.02122*. https://doi.org/10.48550/arXiv.2407.02122

Lampinen, Andrew, Ishita Dasgupta, Stephanie Chan, Kory Mathewson, Mh Tessler, Antonia Creswell, James McClelland, Jane Wang, and Felix Hill. 2022. Can language models learn from explanations in context? In *Findings 2022*, pages 537–563. https://doi.org/10.18653/v1/2022.findings-emnlp.38

Lee, Nayeon, Belinda Z. Li, Sinong Wang, Pascale Fung, Hao Ma, Wen-tau Yih, and Madian Khabsa. 2021. On unifying misinformation detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational*

Linguistics: Human Language Technologies, pages 5479–5485. https://doi.org/10.18653/v1/2021.naacl-main.432

Li, Yupeng, Haorui He, Jin Bai, and Dacheng Wen. 2024. MCFEND: A multi-source benchmark dataset for Chinese fake news detection. In *Proceedings of the ACM Web Conference 2024*, WWW '24, pages 4018–4027. https://doi.org/10.1145/3589334.3645385

Li, Yichuan, Bohan Jiang, Kai Shu, and Huan Liu. 2020. MM-COVID: A multilingual and multimodal data repository for combating COVID-19 disinformation. *arXiv preprint arXiv:2011.04088*. https://doi.org/10.48550/arXiv.2011.04088

Li, Zehan, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*. https://doi.org/10.48550/arXiv.2308.03281

Lin, Hongzhan, Jing Ma, Liangliang Chen, Zhiwei Yang, Mingfei Cheng, and Chen Guang. 2022. Detect rumors in microblog posts for low-resource domains via adversarial contrastive learning. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2543–2556. https://doi.org/10.18653/v1/2022.findings-naacl.194

Litterer, Benjamin, David Jurgens, and Dallas Card. 2023. When it rains, it pours: Modeling media storms and the news ecosystem. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6346–6361. https://doi.org/10.18653/v1/2023.findings-emnlp.420

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. RoBERTa: A robustly optimized BERT pretraining approach. In *International Conference on Learning Representations*.

Liu, Zhiwei, Tianlin Zhang, Kailai Yang, Paul Thompson, Zeping Yu, and Sophia Ananiadou. 2024. Emotion detection for misinformation: A review. *Information Fusion*, 107:102300. https://doi.org/10.1016/j.inffus.2024.102300

Llama Team, AI @ Meta. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*. https://doi.org/10.48550/arXiv.2407.21783

Llansó, Emma J. 2020. No amount of "AI" in content moderation will solve filtering's prior-restraint problem. *Big Data & Society*, 7(1):205395172092068. https://doi.org/10.1177/2053951720920686

Loshchilov, Ilya and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Lucas, Jason, Adaku Uchendu, Michiharu Yamashita, Jooyoung Lee, Shaurya Rohatgi, and Dongwon Lee. 2023. Fighting fire with fire: The dual role of LLMs in crafting and detecting elusive disinformation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14279–14305. https://doi.org/10.18653/v1/2023.emnlp-main.883

Lühring, Jula, Apeksha Shetty, Corinna Koschmieder, David Garcia, Annie Waldherr, and Hannah Metzler. 2024. Emotions in misinformation studies: Distinguishing affective state from emotional response and misinformation recognition from acceptance. *Cognitive Research: Principles and Implications*, 9(1):82. https://doi.org/10.1186/s41235-024-00607-0, PubMed: 39692779

Ma, Jing, 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, pages 3818–3824.

Martel, Cameron, Gordon Pennycook, and David G. Rand. 2020. Reliance on emotion promotes belief in fake news. *Cognitive Research: Principles and Implications*, 5(1):47. https://doi.org/10.1186/s41235-020-00252-3, PubMed: 33026546

Mihalcea, Rada and Carlo Strapparava. 2009. The lie detector: Explorations in the automatic recognition of deceptive language. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers on - ACL-IJCNLP '09*, pages 309. https://doi.org/10.3115/1667583.1667679

Mishra, Pushkar, Helen Yannakoudakis, and Ekaterina Shutova. 2021. Modeling users and online communities for abuse detection: A position on ethics and explainability. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3374–3385. https://doi.org/10.18653/v1/2021.findings-emnlp.287

Mitra, Tanushree and Eric Gilbert. 2015. CREDBANK: A large-scale social media corpus with associated credibility

annotations. *Proceedings of the International AAAI Conference on Web and Social Media*, 9(1):258–267. https://doi.org/10.1609/icwsm.v9i1.14625

Montani, Ines, Matthew Honnibal, Matthew Honnibal, Adriane Boyd, Sofie Van Landeghem, and Henning Peters. 2023. Explosion/spaCy: V3.7.2: Fixes for APIs and requirements. Zenodo. https://doi.org/10.5281/zenodo.10009823

Mosallanezhad, Ahmadreza, Mansooreh Karami, Kai Shu, Michelle V. Mancenido, and Huan Liu. 2022. Domain adaptive fake news detection via reinforcement learning. In *Proceedings of the ACM Web Conference 2022*, pages 3632–3640. https://doi.org/10.1145/3485447.3512258

Mühleisen, Hannes and Mark Raasveldt. 2024. duckdb: DBI Package for the DuckDB Database Management System. R package version 1.0.0.9000. https://github.com/duckdb/duckdb-r

Nakov, Preslav, Jisun An, Haewoon Kwak, Muhammad Arslan Manzoor, Zain Muhammad Mujahid, and Husrev Taha Sencar. 2024. A survey on predicting the factuality and the bias of news media. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15947–15962. https://doi.org/10.18653/v1/2024.findings-acl.944

Nielsen, Dan S., View Profile, Ryan McConville, and View Profile. 2022. MuMiN: A large-scale multilingual multimodal fact-checked misinformation social network dataset. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM Conferences, pages 3141–3153. https://doi.org/10.1145/3477495.3531744

Nørregaard, Jeppe, Benjamin D. Horne, and Sibel Adalı. 2019. NELA-GT-2018: A large multi-labelled news dataset for the study of misinformation in news articles. *Proceedings of the International AAAI Conference on Web and Social Media*, 13:630–638. https://doi.org/10.1609/icwsm.v13i01.3261

Oshikawa, Ray, Jing Qian, and William Yang Wang. 2020. A survey on natural language processing for fake news detection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6086–6093.

Ozcelik, Oguzhan, Arda Sarp Yenicesu, Onur Yildirim, Dilruba Sultan Haliloglu, Erdem Ege Eroglu, and Fazli Can. 2023. Cross-lingual transfer learning for misinformation detection: investigating performance across multiple languages. In *Proceedings of the 4th Conference on Language, Data and Knowledge*, pages 549–558.

Pathak, Archita and Rohini Srihari. 2019. BREAKING! Presenting fake news corpus for automated fact checking. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 357–362. https://doi.org/10.18653/v1/P19-2050

Pelrine, Kellin, Jacob Danovitch, and Reihaneh Rabbany. 2021. The surprising performance of simple baselines for misinformation detection. In *Proceedings of the Web Conference 2021*, pages 3432–3441. https://doi.org/10.1145/3442381.3450111

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

Phillips, Samantha Cavanaugh, Sze Yuh Nina Wang, Kathleen M. Carley, David Gertler Rand, and Gordon Pennycook. 2024. Emotional language reduces belief in false claims. *Judgment and Decision Making*. 20:e43. https://doi.org/10.1017/jdm.2025.10019

Pleiss, Geoff, Tianyi Zhang, Kilian Q Weinberger, and Ethan Elenberg. 2020. Identifying mislabeled data using the area under the margin ranking. In *34th Conference on Neural Information Processing System*, volume 33, pages 17044–17056.

Plutchik, Robert. 1980, Chapter 1 - A general psychoevolutionary theory of emotion. In Robert Plutchik and Henry Kellerman, editors, *Theories of Emotion*. Academic Press, pages 3–33. https://doi.org/10.1016/B978-0-12-558701-3.50007-7

Poldvere, Nele, Zia Uddin, and Aleena Thomas. 2023. The PolitiFact-Oslo corpus: A new dataset for fake news analysis and detection. *Information*, 14(12). https://doi.org/10.3390/info14120627

Potthast, Martin, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 231–240.

https://doi.org/10.18653/v1/P18-1022
Pratelli, Manuel and Marinella Petrocchi. 2022. A structured analysis of journalistic evaluations for news source reliability. *arXiv preprint arXiv:2205.02736*. https://doi.org/10.48550/arXiv.2205.02736
Pratelli, Manuel, Fabio Saracco, and Marinella Petrocchi. 2024. Unveiling news publishers trustworthiness through social interactions. In *ACM Web Science Conference*, pages 139–148. https://doi.org/10.1145/3614419.3644015
Przybyla, Piotr. 2020. Capturing the style of fake news. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):490–497. https://doi.org/10.1609/aaai.v34i01.5386
Przybyła, Piotr, Euan McGill, and Horacio Saggion. 2025. Attacking misinformation detection using adversarial examples generated by language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27614–27630. https://doi.org/10.18653/v1/2025.emnlp-main.1405
Przybyła, Piotr, Alexander Shvets, and Horacio Saggion. 2024. Verifying the robustness of automatic credibility assessment. *Natural Language Processing*, pages 1–29. https://doi.org/10.1017/nlp.2024.54
Rashkin, Hannah, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937. https://doi.org/10.18653/v1/D17-1317
Raza, Shaina and Chen Ding. 2022. Fake news detection based on news content and social contexts: A transformer-based approach. *International Journal of Data Science and Analytics*, 13(4):335–362. https://doi.org/10.1007/s41060-021-00302-z, PubMed: 35128038
Risdal, Meg. 2016. Getting real about fake news. https://doi.org/10.34740/kaggle/dsv/911, https://www.kaggle.com/datasets/mrisdal/fake-news
Seabold, Skipper and Josef Perktold. 2010. Statsmodels: Econometric and statistical modeling with Python. In *9th Python in Science Conference*. https://doi.org/10.25080/Majora-92bf1922-011
Shahi, Gautam Kishore and Durgesh Nandini. 2020. FakeCovid – A multilingual cross-domain fact check news dataset for COVID-19. *arXiv preprint arXiv:2006.11343*. https://doi.org/10.48550/arXiv.2006.11343
Shokri, Mohammad, Vivek Sharma, Elena Filatova, Shweta Jain, and Sarah Levitan. 2024. Subjectivity detection in English news using large language models. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 215–226. https://doi.org/10.18653/v1/2024.wassa-1.17
Shu, Kai, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. FakeNewsNet: A data repository with news content, social context and spatialtemporal information for studying fake news on social media. *arXiv preprint arXiv:1809.01286*. https://doi.org/10.1089/big.2020.0062, PubMed: 32491943
Shu, Kai, Guoqing Zheng, Yichuan Li, Subhabrata Mukherjee, Ahmed Hassan Awadallah, Scott Ruston, and Huan Liu. 2020. Leveraging multi-source weak social supervision for early detection of fake news. *arXiv preprint arXiv:2004.01732*. https://doi.org/10.48550/arXiv.2004.01732
Smith, Marcellus, Brandon Brown, Gerry Dozier, and Michael King. 2021. Mitigating attacks on fake news detection systems using genetic-based adversarial training. In *2021 IEEE Congress on Evolutionary Computation (CEC)*, pages 1265–1271. https://doi.org/10.1109/CEC45853.2021.9504723
Stahl, Peter M. 2024. Pemistahl/lingua-py.
Stepanova, Nataliya and Björn Ross. 2023. Temporal generalizability in multimodal misinformation detection. In *Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP*, pages 76–88. https://doi.org/10.18653/v1/2023.genbench-1.6
Struß, Julia Maria, Federico Ruggeri, Dimitar Dimitrov, Andrea Galassi, Georgi Pachov, Ivan Koychev, Preslav Nakov, Melanie Siegel, Michael Wiegand, Maram Hasanain, Reem Suwaileh, and Wajdi Zaghouani. 2024. Notebook for the CheckThat! Lab at CLEF 2024. In *Conference and Labs of the Evaluation Forum*, volume 3740.
Swayamdipta, Swabha, Roy Schwartz, Nicholas Lourie, Yizhong Wang,

Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293. https://doi.org/10.18653/v1/2020.emnlp-main.746

Szwoch, Joanna, Mateusz Staszkow, Rafal Rzepka, and Kenji Araki. 2024. Limitations of large language models in propaganda detection task. *Applied Sciences*, 14(10):4330. https://doi.org/10.3390/app14104330

Tacchini, Eugenio, Gabriele Ballarin, Marco L. Della Vedova, Stefano Moret, and Luca de Alfaro. 2017. Some like it Hoax: Automated fake news detection in social networks. In *Data Science for Social Good*, volume 1960.

Tobi, Abraham. 2024. Towards an epistemic compass for online content moderation. *Philosophy & Technology*, 37(3):109. https://doi.org/10.1007/s13347-024-00791-3

Verhoeven, Ivo, Pushkar Mishra, Rahel Beloch, Helen Yannakoudakis, and Ekaterina Shutova. 2024. A (more) realistic evaluation setup for generalisation of community models on malicious content detection. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 437–463. https://doi.org/10.18653/v1/2024.findings-naacl.30

Wang, William Yang. 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. https://doi.org/10.18653/v1/P17-2067

Wang, Yaqing, Weifeng Yang, Fenglong Ma, Jin Xu, Bin Zhong, Qiang Deng, and Jing Gao. 2020. Weak supervision for fake news detection via reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 516–523. https://doi.org/10.1609/aaai.v34i01.5389

Wu, Jiaying and Bryan Hooi. 2022. Probing spurious correlations in popular event-based rumor detection benchmarks. *arXiv preprint arXiv:2209.08799*. https://doi.org/10.48550/arXiv.2209.08799

Wu, Liang, Fred Morstatter, Kathleen M. Carley, and Huan Liu. 2019. Misinformation in social media: Definition, manipulation, and detection. *ACM SIGKDD Explorations Newsletter*, 21(2):80–90. https://doi.org/10.1145/3373464.3373475

Xiao, Madelyne and Jonathan Mayer. 2024. The challenges of machine learning for trust and safety: A case study on misinformation detection. https://doi.org/10.48550/arXiv.2308.12215

Yang, Kai Cheng and Filippo Menczer. 2025. Accuracy and political bias of news source credibility ratings by large language models. In *Proceedings of the 17th ACM Web Science Conference 2025*, Websci '25, pages 127–137. https://doi.org/10.1145/3717867.3717903

Yee, Adrian. 2025. The limits of machine learning models of misinformation. *AI & Society*. https://doi.org/10.1007/s00146-025-02324-8

Yue, Zhenrui, Huimin Zeng, Ziyi Kou, Lanyu Shang, and Dong Wang. 2022. Contrastive domain adaptation for early misinformation detection: A case study on COVID-19. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 2423–2433. https://doi.org/10.1145/3511808.3557263

Zhang, Qiang, Hongbin Huang, Shangsong Liang, Zaiqiao Meng, and Emine Yilmaz. 2021a. Learning to detect few-shot-few-clue misinformation. *arXiv preprint arXiv:2108.03805*. https://doi.org/10.48550/arXiv.2108.03805

Zhang, Xueyao, Juan Cao, Xirong Li, Qiang Sheng, Lei Zhong, and Kai Shu. 2021b. Mining Dual emotion for fake news detection. In *Proceedings of the Web Conference 2021*, pages 3465–3476. https://doi.org/10.1145/3442381.3450004

Zhou, Xiang, Heba Elfardy, Christos Christodoulopoulos, Thomas Butler, and Mohit Bansal. 2021. Hidden biases in unreliable news detection datasets. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2482–2492. https://doi.org/10.18653/v1/2021.eacl-main.211

Zhou, Xinyi and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys*, 53(5):1–40. https://doi.org/10.1145/3395046

Zhou, Zhixuan, Huankang Guan, Meghana Bhat, and Justin Hsu. 2019. Fake news

detection via NLP is vulnerable to adversarial attacks. In *Proceedings of the 11th International Conference on Agents and Artificial Intelligence*, pages 794–800. https://doi.org/10.5220/0007566307940800

Zubiaga, Arkaitz, Maria Liakata, and Rob Procter. 2017. Exploiting context for rumour detection in social media. In *Social Informatics*, pages 109–123. https://doi.org/10.1007/978-3-319-67217-5_8