SoK: Machine Learning for Misinformation Detection

Madelyne Xiao
Princeton University

Jonathan Mayer
Princeton University

Abstract

We examine the disconnect between scholarship and practice in applying machine learning to trust and safety problems, using misinformation detection as a case study. We survey literature on automated detection of misinformation across a corpus of 248 well-cited papers in the field. We then examine subsets of papers for data and code availability, design missteps, reproducibility, and generalizability. Our paper corpus includes published work in security, natural language processing, and computational social science. Across these disparate disciplines, we identify common errors in dataset and method design. In general, 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. We demonstrate the limitations of current detection methods in a series of three representative replication studies. Based on the results of these analyses and our literature survey, we conclude that the current state-of-the-art in fully-automated misinformation detection has limited efficacy in detecting human-generated misinformation. We offer recommendations for evaluating applications of machine learning to trust and safety problems and recommend future directions for research.

1 INTRODUCTION

Online services face a daunting task: There is an unceasing deluge of user-generated content, on the order of hundreds of thousands of posts per minute on popular social media platforms [1]. Some of that content is false, hateful, harassing, extremist, or otherwise problematic. How can platforms reliably and proactively identify these "trust and safety" issues?

Machine learning has proved an attractive approach in the academic literature, leading to large bodies of scholarship on misinformation detection [2], toxic speech classification [3], and other core trust and safety challenges (e.g., [4]). The conceptual appeal of machine learning is that it could address the massive scale of user-generated content on large platforms *and* the capacity constraints of small platforms. Recent work on ML-based approaches to automated misinformation detection claims impressive performance statistics: In the literature review that we conduct for this work, among publications that report performance metrics, about 70% of papers report over 80% accuracy on at least one detection task; some of these works report near-perfect accuracy [5, 6].

In recent years, disclosures from major tech companies have tempered these expectations. In 2023, Twitter's former head of trust and safety stated that large-scale automated detection of misinformation remains a hard problem, and that no generalizable automated solutions are currently available [7]. Automated misinformation detection and fact-checking tools are in development at Google and OpenAI; both companies have disclaimed that these tools still require human intervention in deployment [8]. These disclosures accord with our observation that, in practice, trust and safety functions at online services are driven by manual interventions such as user reports and human moderators.

In this work, we investigate the disconnect between scholarship and practice in applications of machine learning to trust and safety problems. Our project is inspired by recent research that has identified shortcomings in machine learning applications for many problem domains, including information security, social science, and medicine [9–12]. We use misinformation detection as a case study for trust and safety problems because the topic has recently generated a rich literature with diverse methods and claimed successes. Misinformation detection has substantive complexities that are common for trust and safety problems: linguistic and cultural nuance, sensitivity to context, and rapidly evolving circumstances.

We seek to answer four discrete research questions, which collectively shed light on the research-practice gap in automated misinformation detection.

RQ1. How well-suited are misinformation detection methods in the academic literature to the needs of online services, specifically social media platforms that host user-generated content?

RQ2. How do ML-based detection methods select targets, curate datasets, and evaluate performance? Are there identifiable best practices and missteps at each of these stages? Are there identifiable missteps related to target selection, dataset curation, feature selection, and evaluations of method performance in ML-based misinformation detection studies?

RQ3. How reproducible are published ML-based misinformation detection methods?

RQ4. How generalizable are published ML-based misinformation detection methods to out-of-domain data (i.e., to data types, time periods, and topics not present in training data)?

We address these research questions in three ways. First, we conduct a broad literature review and synthesis of the full paper corpus (248 papers), with a focus on detection targets and evaluation. We provide an in-depth review of a subset (87 papers) of the full paper set, with a focus on methods: dataset design, feature engineering, and model selection. Second, we attempt to obtain code and data for a subset of prior work. Third, we test several representative approaches for replication and generalizability. We arrive at the following results by applying these methods.

1. Detection tasks in scholarship are often steps removed from the misinformation content moderation challenges that platforms face. Detection targets may be of limited consequence, or may be more readily and accurately identified through manual means.
2. Methods frequently target *proxies* for the presence of misleading content; these approaches are easy to evade. Datasets used in publications are often non-representative of real-world contexts. Model evaluation often lacks independence from training and rarely involves close emulation of a real-world deployment.
3. Data and code availability problems pervade the literature, inhibiting replication. Where these are available, we are generally able to replicate prior work.
4. Prior work has poor generalizability when classifying content beyond what was included in training data.

We organize this work as follows: In **Section 2**, we situate this project within existing security literature and academic literature that critiques machine learning-driven methods. We motivate our choice of automated misinformation detection as a representative case study and highlight the particular relevance of this problem to the security community. In **Section 3**, we discuss detection goals and definitional inconsistencies and present the information taxonomies developed from our inductive coding of papers. We discuss findings from our reading and coding of papers in **Section 4**. Our systematization of literature progresses along two axes: 1) from unimodal to increasingly multi-modal methods, and 2) tracking design steps. Specifically, we discuss issues pertaining to method-target fit, dataset curation and feature selection, model selection, and method evaluation (RQ1, RQ2). In **Section 5**, we illustrate the issues identified in our literature review with a series of replication studies, with a focus on reproducibility and generalizability of results (RQ3, RQ4). In **Section 6**, we conclude with a discussion of findings from our literature review and replication studies and recommendations for evaluating ML-based interventions for trust and safety. Based on these observations and analyses, we propose directions for future work that uses ML models to address trust and safety concerns.

Through this work, we hope to underscore 1) the prevalence and severity of reproducibility and ML-based method design issues in the existing misinformation detection literature, 2) the need for careful and preemptive evaluation of ML-based methods at the point of problem formulation, and 3) the importance of method explainability and data accessibility. **We contribute taxonomies, our annotated corpus of 248 research papers, recent datasets, and recommendations for evaluating ML-based trust and safety tasks.**

2 Motivation

Platforms have been motivated to move toward fully-automated or hybrid content moderation systems in response to the burgeoning volume and diversity (in terms of language, topic, and medium) of social media content [13]. These systems must be able to process large volumes of content and be robust to distributional shifts in data, including those incurred by real-time events (e.g., natural disasters, mass casualty events). These systems should also be resistant to circumvention and attempts to "game" moderation [14]. Taken together, we summarize these desiderata as *generalizability*, *versatility in real-time*, and *robustness to evasion and attacks*.

The detection of misinformation and "influence operations" (IOs) is a topic of convergent interest in security, social science, and AI/ML; researchers in these areas and trust & safety practitioners in industry are the intended audience for this work. Methods developed in recent years for the detection of influence operations (e.g., astroturfing and coordinated misinformation campaigns) resemble techniques previously used by security researchers to detect botnets, advanced persistent threats (APTs), and malware [15–17]. These security-oriented approaches target anomalous network traffic, code semantics, misleading and inauthentic content, and evidence of coordinated activity. In general, security research requires adaptive responses to rapidly changing, possibly adversarial circumstances. As such, the security community is well-positioned to shepherd the responsible development of misinformation detection methods. In addition, misinformation research stands to benefit from the formal structure of security research methodology: human fact-checkers have expressed interest in adopting threat modeling practices similar to those found in security literature [18].

3 Preliminaries

In this section, we present terminology and taxonomies used in this work and describe our criteria for evaluating detection goals and performance, as reported by paper authors.

Definitions. Definitions of *misinformation, disinformation, malinformation,* and *influence operations* vary across the literature; as such, for each paper we review, we consider the paper authors' working definitions of these terms (if such definitions are provided) and evaluate model performance with respect to the definitions set forth [19–23]. We limit our analysis to text-based English-language content to facilitate independent verification of classification results (English is the primary language of both paper authors). We note that the nature of misinformation narratives and misinformation spread varies with language and geography. We additionally note that *disinformation* is commonly used to refer to incorrect information written with the intent to deceive, while *misinformation* refers to incorrect information in general; for completeness and concision, we refer to intentionally *and* unintentionally false information as *misinformation* unless otherwise qualified. We occasionally use *false rumors, false news,* and *false information* interchangeably, but avoid *fake news*, a politically charged term [24]. We use "influence operations" (IOs) or "coordinated campaigns" to discuss collaborative attempts to disseminate misinformation across networks [25].

Detection goals. We evaluate the experimental methods in our paper corpus for *generalizability* across topics and time; performance in *real-time* settings; and *robustness to evasion and distributional shifts*. These desiderata reflect real-world platform needs, as discussed in Section 2. Methods that exhibit "good" performance on out-of-distribution, real-time, and adversarial detection tasks are "well-suited" to the needs of actual platforms. We consider method performance with respect to the metrics, datasets, and detection settings reported by authors. These metrics include accuracy, precision and recall, AUC, and F1 scores. *Though we provide guidelines for evaluating these metrics in the next subsection, we emphasize that, with few exceptions, differences in dataset and target choice inhibit direct comparisons between papers.* We discuss target misalignment and metrics in Limitations (6.2).

Evaluation metrics. Precise thresholds for "good" performance on these metrics are hard to determine. Performance on common evaluation metrics must be weighed against the needs of the deployment setting. Class distributions and the semantic content of datasets should influence interpretation of quantitative measurements, and authors should clearly state any thresholds for acceptable performance and their rationale for these thresholds. For instance, if the detection setting concerns sensitive information for which a false negative is potentially high-risk (e.g., COVID symptoms), performance on recall will be more informative than overall accuracy. If the outcome of a false positive result is high-risk (e.g., flagging of spam bots or messages), precision scores should be reported.

While higher reported scores on all aforementioned metrics are generally preferable, readers should remain mindful of data overfitting and imbalanced class distributions, which might yield deceptively high scores on certain metrics. (We include information about the contents and class distributions of common datasets used for misinformative detection tasks in Appendix B.) As such, it is inadvisable to compare performance on different metrics within the same work, and without consideration of training and testing dataset contents: these numbers quantify different aspects of classifier performance on different subsets of data. Between papers, these comparisons are further hampered by differences in target formulation and dataset choice, and we critique these misalignments in our literature review. Ultimately, we use these metrics to demonstrate that, given assumptions about *generalizability* of performance and *dataset choice*, these numbers suggest that many algorithms perform suitably well for real world uses. We challenge these assumptions in this work.

Taxonomies. Through iterative inductive coding of papers, we develop a taxonomy of five "information scopes." As we read each paper in our corpus, we noted the smallest semantic or organizational unit of information that the method attempted to classify as true or false and sorted each paper into emergent categories. We ultimately identify five information scopes: **claims**, **articles**, **users**, **networks**, and **websites**. We consider these scopes in the context of online services—for instance, articles and websites whose links might be posted to a social media platform.[1] In our literature review, we consider the *stated* versus *actual* target of the classification task that paper authors describe: While the goal of misinformation detection is the identification of *incorrect information*, we find that many methods instead target (e.g.) sentiment and virality as proxies for the presence of misinformation [26, 27]. Targets are specific to information scope: we present them, with scope definitions, in Figure 1. We also describe methodological errors with respect to the step of the experiment design in which they occur: **dataset curation**, **model choice**, **feature selection** and **model evaluation**. These design considerations are generally scope-agnostic: we present them in Figure 2. A summary of takeaways, organized by design steps, is in Figure 3. By-scope takeaways are in the gray boxes at the end of each subsection in Section 4.

Feasibility of detection. Most of the work we review presupposes that automated detection of misinformation is possible at all. We note that this is, in itself, a strong assumption to make: Published work in linguistics and philosophy of language has previously called into question the feasibility of using semantic and syntactic features to determine the veracity of text statements [28, 29]. On the other hand, some researchers maintain that misinformative texts are characterized by indelible "fingerprints" that distinguish them from

[1]Works that make classification decisions about more than one information scope receive multiple labels: works targeting social media posts, for instance, are generally labeled (Ⓒ *and* Ⓤ) or (Ⓒ *and* Ⓝ) in Table 3.

Scopes and detection targets.
Targets are specific to scope. Target discussion sections in the literature review are denoted by a 🎯.

(C) Claim. The smallest semantic unit of fact or misinformation, comprising a subject, predicate, and object, at minimum.

Detection targets:
i. Semantic content signals;
ii. Knowledge graph topology;
iii. "Checkability" or "checkworthiness."

(A) Article. News-oriented writing of length 100 words or more.

Detection targets:
i. Sentiment and style signals, including tone and genre;
ii. Topic-aware detection of stance and relevance to known rumoring topics.

(U) User. All data and metadata associated with a single user's account, as defined by a social media platform (e.g., an account corresponding to a handle on Twitter/X; or a page or profile corresponding to one business, individual, or organization on Facebook).

Detection targets:
i. Account metadata, including bio text, images, and account age;
ii. Single account behaviors, including comments on posts and published content.

(N) Network. A set of users, and interactions between those users, as represented by a social graph.

Detection targets:
i. Propagation patterns across social graphs;
ii. Timestamped records of user-user interactions.

(W) Website. A news site, including its hosting infrastructure and text and image contents.

Detection targets:
i. Text and URL/domain semantics;
ii. Site visitor demographics;
iii. Suspicious UI elements, including ads;
iv. Hosting infrastructure, including DNS certificates and site age.

Figure 1: Descriptions of scopes and unique per-scope targets. When referenced elsewhere, scope targets are assigned alphanumeric labels comprising the scope abbreviation and the Roman numeral corresponding to the target. **C.ii**, for instance, references the second target in the **Claim**-scoped section of this taxonomy ("Knowledge graph topology)."

Design taxonomy.
Design steps are agnostic to scope. Design discussions are in the literature review are denoted by a circled number (e.g., ①).

① **Dataset curation.** The source, size, and contents of datasets used for model training and testing.

Dataset properties:
i. Dataset age;
ii. Evidence of temporal or feature leakage;
iii. Dependencies, including those relating to authorship, source, and topic;
iv. Availability of original data.

② **Model choice.** The choice of ML model used by paper authors for their detection task, as well as their motivation for this choice. These codes reflect common desiderata for social media platforms, as discussed in Section 2.

Models and approaches:
i. Distance calculations on embeddings;
ii. "Traditional" ML, including KNN, naive Bayes, and gradient boosting;
iii. Deep learning;
iv. Graph algorithms;
v. Ensemble classifiers.

③ **Feature selection.** The choice of feature(s) that paper authors use to train their ML models, as well as their motivation for this choice.

Feature types:
i. Textual features, including semantics and syntax;
ii. Network-based features, including user interactions and posting behaviors;
iii. User-based features, including bios, profile images, and account age;
iv. Infrastructural features, including site UI.

④ **Model evaluation.** Paper authors' approach to benchmarking method performance after initial training and reported scores.

Evaluation criteria:
i. Evidence of real-time testing on live content feeds;
ii. Generalizability to topic domains and timestamps not included in training data;
iii. Robustness to evasion, attacks, and distributional shifts in data.

Figure 2: Descriptions of design steps and our codes. When referenced elsewhere, these codes are assigned Arabic-Roman numeric labels comprising the design step and subtaxa code. For instance, **4.i** references the first code in the **Model evaluation** section, or "Evidence of real-time testing on live content feeds."

Detection targets	Some **claim**-scoped methods perform semantic verification on semantic embeddings. Other scopes target *proxy signals* for semantic accuracy. For **article**-scoped methods, these proxies include sentiment and stance on known rumoring topics. **User** detection targets the semantics of account metadata, such as bios and posts made. Authors of **network**- and **website**-scoped methods, recognizing the risks of text-dependent targets, use semantics-agnostic targets, such as website age and user commenting activity.
① **Dataset curation**	Data availability issues pervade the literature across **all five scopes**. **Claim**-scoped methods rely on outdated datasets and inconsistent taxonomies for classification tasks. **Article**-scoped methods assign single labels to newswriting where more granular labels would be more descriptive. **User**- and **network**-scoped methods rely on social media data and human-annotated data that has since become inaccessible to the public. **Website**-scoped data is necessarily multimodal, and is difficult to control for distribution shifts over time.
② **Model choice**	**Claim**-scoped methods that use knowledge graph search have limited efficacy when performing detection on novel information domains. Improvements achieved by unsupervised methods at the **article** scope might be due to dependencies in text data. At the **user** and **network** level, semantics-agnostic models cannot infer the intent (malicious or otherwise) of suspicious behavior. We observe that **website**-scoped methods that use traditional supervised ML models outperform unsupervised models on two-way classification tasks.
③ **Feature selection**	**Claim**-scoped methods require structured inputs that are frequently unavailable in late-breaking scenarios. **Article**-scoped methods rarely control for author- and source-induced dependencies. **User**-scoped methods that use political affiliation and demographic characteristics as features risk reinforcing existing biases. **Network**-scoped methods make strong assumptions about the behaviors of users in a graph, including the likelihood that users will believe misinformative online content. Multimodal feature sets in **website**-scoped methods are not normalized, despite differences in distribution, prevalence, and medium.
④ **Model evaluation**	Across **all five scopes**, testing in real-time is rare, though many authors cite the slow pace of human fact-checking as motivation for their work. In addition, **all scopes** have generalizability issues. **Claim**-scoped methods on knowledge graphs are limited by the contents of known relationships between graph nodes. **Article**-scoped methods that perform stance detection depend on *a priori* knowledge of known rumoring topics. **Network**- and **user**-scoped methods can only detect behaviors that emulate known suspicious behaviors. **Website**-scoped methods are susceptible to shifts in distributions of hosted content.

Figure 3: Summaries of takeaways, organized by target and design step.

non-misinformative texts: for instance, emotional language and reduced lexical diversity [30, 31]. This debate motivates our discussion of non-textual feature sets for detection.

4 Systematization of Literature

Paper selection. To seed our corpus, we manually curated a selection of 23 highly-cited survey papers that provide comprehensive overviews of the state of automated misinformation detection at the time of writing [2, 32–53]. We relied on these papers to identify detection methods that have been well-received by the research community. We searched for survey papers in Google Scholar with queries "survey misinformation detection" and "survey fake news detection" and collected the most-frequently cited papers within the past 10 years.[2] We then inspected each paper's references for related work; we read the abstracts for these works in order to confirm relevance. We supplemented this corpus of papers with publications surfaced by Google Scholar queries. We queried the following terms: "`misinformation detection` [x]" and "`automated fact checking` [x]," where $x \in$ {claims, news articles, accounts, networks, websites, influence operations[3]}. We collected the 50 most highly-cited papers in the set resulting from the union of search results returned by both search queries for each x. To counter potential bias toward older publications, we collected papers with the highest citation rates *per year*. These search terms are deliberately over-inclusive; we manually review all works for relevance after the initial sampling step. After removal of out-of-scope works (see *In- and out-of-scope work*) from this set of 250 papers, 219 eligible papers remained. To ensure that security-oriented approaches to detection were represented in our corpus, we conducted a separate snowball sampling search for work published in four A* [54] security research venues (USENIX Security, IEEE S&P, NDSS, and ACM CCS), using the keywords listed previously. This resulted in the addition of 29 works, mostly addressing the detection of accounts and networks that spread misinformation. Our final corpus comprises

[2] This time frame was naturally enforced by a lack of well-cited older publications, and was not fixed before we began our sampling process.

[3] We include "influence operations" as a separate search term because this term of art is relatively new.

248 papers published between 2009 and 2024, inclusive.

Corpus curation and coding. We used automated keyword matching to confirm that all collected papers address misinformation and automated fact-checking methods; we read the abstracts of papers surfaced by this check to confirm relevance. We then annotated research papers that passed this check in accordance with our taxonomy. One reader made three coding passes over the corpus: First, she identified a taxonomy of information scopes; second, she noted actual versus stated detection targets and approaches to evaluation; finally, she noted method design approaches and errors. Two coders read and independently coded a random subset (30 papers) of the full corpus. Fleiss's kappa for this subset was > 0.80, indicating strong agreement.

"Full" and "focus" paper corpora. We complete our paper coding at two different levels of granularity. We develop our target and design taxonomies from iterative coding of the whole paper corpus. To develop our focus corpus, we identified papers with sufficient responses to each coding field in our design taxonomy: One coder manually reviewed annotations for the full corpus and noted papers that provide at least general descriptions of their choice of dataset, features, and model. From these papers, we then sampled works from each scope in proportion to that scope's representation in the full corpus, oversampling within each scope for diverse methods and data. This corpus comprised 87 papers. For each of these, we perform deep-coding of methods: we additionally note dataset age, availability, contents (including evidence of leakage or dependencies); motivation for model choice; and feature types (resp. steps ①, ②, and ③ in Section 3).[4] A partial summary of this deep-coding is available in Table 3. Our motivation for developing this focus set is practical: Many of the works we review do not include complete discussions of every section of our overall coding taxonomy. For all papers in the full corpus (248 works), we note targets and evaluation.

In- and out-of-scope work. We consider a number of security-oriented approaches to misinformation detection in this work. Sybil and botnet detection methods that target IOs are categorized as account- and network-scoped detection methods in our corpus. *Commercial approaches* to misinformation detection are in-scope for this project. In view of data accessibility issues, we are unable to provide an in-depth analysis of commercial methodologies in our literature review, and include a market survey of commercial fact-checking providers in Appendix A. *LLM-powered detection* is in-scope for this work. Accessibility to code bases for LLMs is similarly limited, and inhibits evaluation by researchers [58, 59]. While we discuss LLM-powered detection approaches in our appendix, we defer more extensive discussion to future work.

Notation. We denote design steps with circled icons (e.g., ①, ④) and subtaxa within those categories with Roman numerals (**iv**). While *targets* are generally unique to information scope (and are referenced by an alphanumeric label (e.g., **(A.i)**), design steps are *not* scope-specific, and are denoted by Arabic-Roman numeral pairs (e.g., **(2.i)**). As an example, the stated target of [60] is "troll" accounts on Reddit. The actual target of this work is **(U.ii)** *account interactions with known troll accounts*. Account identifiers have been scrubbed, impeding manual verification of performance; this is a **(1.iv)** *data accessibility issue*. The method uses a random forest classifier, **(2.ii)** *a classic ML model*, and considers counts of posts and replies—these are **(3.ii)** *network-based features*. The method is not **(4.i)** *tested in real-time*, and **(4.ii)** *cannot generalize* to behavior not captured in a known trolls list.

4.1 Claims

About 70% of paper authors within this scope cite the 2016 U.S. presidential election as motivation for the development of their methods [61–63]; about 30% cite COVID-19 misinformation [64–66]. All works mention the speed and volume of *social media* misinformation in particular [67]. 30% cite the relatively slow pace of manual fact-checking [62, 67, 68]—and the need for faster, automated approaches—as motivation. Claim-scoped papers form 15% of our corpus.

Detection targets. Claim-scoped detection methods generally target the semantic contents of text statements. This is done in the following ways: **(C.i)** *distance calculations on semantic embeddings* to perform textual entailment or stance detection [69, 70]; and **(C.ii)** *search on a knowledge graph topology* [68, 71] to determine if these reference sources corroborate or refute the claim to be checked. A small class of approaches explicitly detect **(C.iii)** "*checkworthiness*," [5] and are intended to surface checkable statements to human fact-checkers for manual verification [62, 67, 71].

Targets such as author credibility are *proxy targets* for the presence of misinformation: non-semantic signals which do not directly indicate text veracity, but suggest (lack of) veracity by association with an external heuristic. About 60% of papers at this scope use proxy targets, including the **(A.i)** *syntactic and/or stylometric qualities of text* [72], **(U.i)** *source reputation* [73], or **(U.ii)** *contextual indicators*, such as commenter responses [74], to classify social media posts.

① **Dataset curation.** Claim-scoped papers in our corpus that propose to perform fact verification **(1.i)** *rely on outdated existing datasets* of labeled statements in order to establish ground truth. LIAR [75] and PolitiFact [76] were the most popular datasets among claim-scoped works, and, taken together, were used by approximately half of all papers within this scope [77, 78]. LIAR is a static political news dataset that was published in 2017; on average, papers that cite LIAR were published two years after LIAR's release. Topic detection and word frequency models trained on LIAR are likely ineffective in contemporary fact-checking contexts,

[4] We take inspiration from [55–57], which perform similar deep-coding.

[5] We include these works in our corpus because they *do* take topic and source credibility into consideration in the process of ranking checkability.

and for non-political subject matter [61, 73]. Additionally, class labels across datasets are inconsistent: PolitiFact uses a six-point labeling scale (pants-on-fire; false; barely-true; half-true; mostly-true; true), FEVER uses a three-point scale (supported; refuted; notenoughinfo), and GossipCop uses an eleven-point scale (ratings from 0 to 10).

② **Model selection.** For methods that pre-construct knowledge graphs or other reference databases, models perform some form of **(2.i)** *shortest-path search* on the KG topology [71], and approximate logical inference via transitive closure on graph edges [68]. For methods that perform information retrieval at query time—e.g., to match corroborating sources to a claim—text statements are first converted to bag-of-words or TF-IDF embeddings [79, 80]. For methods that perform detection of misinformative posts on social media, **(2.v)** *stacked ensemble classifiers* are a common approach to incorporating multiple feature modalities.

③ **Feature selection.** Semantic feature analysis is limited to the **(3.i)** *identification of structured statements* as a precursor to knowledge graph (KG) construction. These claims take the form of subject-predicate-object (SPO) statements (e.g., "I like pie") [68, 71, 97]. Detection versatility is determined by the size of the source dataset and the granularity of the relationships encoded by graph edges [71]. Supervised methods that detect linguistic cues use **(3.i)** *hand-crafted word or topic lists* [67]). Authors using supervised methods claim that their approach permits highly customized targeting of specific rumoring narratives [67], though this also assumes that the method developer has prior knowledge of the contents of test data; authors using unsupervised methods claim that their approaches detect contextual and language features that cannot be easily extracted by common features such as word frequency or sentiment [77]. Methods detecting social media data consider **(3.ii)** *network* and **(3.iii)** *user* interaction features (post likes, shares, comments) [98–100].

④ **Evaluation.** Though a majority of claim-scoped methods cite the speed of social media misinformation as motivation, **(4.i)** *only one method within this scope reported results from testing in real time* [101]. KG-based methods are **(4.ii)** *non-generalizable by design*: the approaches we survey require structured inputs for graph construction, and rely on published datasets for ground truth [68, 71]. Though this approach permits semantic verification of statements, it is difficult to perform iterative updates to source databases in real-time, particularly in the types of online settings where claim-checking might be most usefully deployed (e.g., during breaking news events, where no source of ground truth is immediately available). We observe that a majority of works at this scope do not test on novel or out-of-domain data; we discuss overfitting issues in greater detail in the next section.

Takeaways: 🎯 The efficacy of claim-scoped methods is determined by the coverage conferred by a reference database. ① Datasets are frequently out-of-date, and taxonomies are inconsistent. ② Though some inference is possible via transitive closure on knowledge graph edges, this capability is, in general, limited. ③ Knowledge-graph-based methods require structured inputs, which might not be readily available in a breaking news setting, where ground truth references are not yet available. ④ Few authors test in real-time, despite citing the slow pace of manual fact-checking as motivation.

4.2 Articles

We consider all news-oriented writing of length 100 words or greater to fall within this scope. 25% of papers within this scope cite growing distrust of mainstream news media outlets as motivation for their methods, which promise to deliver fast labeling of news stories that appear on social media [53]. Text-based credibility classifiers have been shown to have limited efficacy, however: while unsupervised approaches can identify *bias* with high accuracy, this performance degrades in misinformation and credibility classification tasks [102, 103]. This scope comprises 24% of our paper corpus.

🎯 **Detection targets.** In contrast to single claims, full-length news articles have sufficient text contents to make semantic verification difficult and some off-the-shelf NLP approaches practicable. These approaches are distinct from direct claim verification and qualify as proxy detection methods: For instance, Bhutani et al. associate strongly negative sentiment with the presence of false information [79], and Horne et al. find that satire and misinformation share stylistic similarities [85]. Article-scoped methods target proxy signals, and adopt at least one of the following approaches to detection: **(A.i)** *NLP analysis of article contents to identify language features particular to writing styles heuristically associated with misinformation* (genre detection, sentiment analysis) [104]; and **(A.ii)** *analysis of article contents and headlines to identify discussion related to known misinformation narratives* (topic detection) [42, 67]. Respectively, these approaches 1) simultaneously assume and detect a heuristic (e.g., strong emotion indicating the presence of misinformation); and 2) assume prior knowledge of rumoring topics.

① **Dataset curation.** While well-annotated, current datasets are in short supply across all scopes, this deficit is particularly glaring at the article scope. This is due in large part to definitional ambiguities that prevent fine-grained labeling of longer texts for classifier training. As a result, 44% of papers within this scope **(1.i)** *use public datasets released years prior to the start of research* [75, 85, 105, 106]. These datasets (LIAR [75] and Buzzfeed-Webis [103]) include news links, speaker credibility scores, and other metadata that **(1.ii)** *constitute serious sources of leakage* for methods that use contextual features to infer true/false labels. The remaining 56% of papers use article corpora curated by asking crowdworkers

Paper: Work	Scope	① Target	② Dataset	③ Model	④ Features: Textual	Network	Author	Infra.	⑤ Performance: Accuracy/AUROC
1. Ajao et al. [26]	C A	Sentiment (A.i)	PHEME [81]	LSTM, DT, RF, SVM	●				0.86 (Acc.)
2. All-Tanvir et al. [82]	C A N	Content (C.i)	Twitter (API)	NB, RNN, LSTM, SVM, Logit	●	●			0.89 (Acc.)
3. Bhutani et al. [79]	C A	Content (C.i); sentiment (A.i)	Twitter (API), PolitiFact [76]	Naive Bayes, RF	●				0.60 (AUC)
1. Afroz et al. [83]	A	Content (C.i); syntax (A.i)	Brennan-Greenstadt	SVM, J48 Decision Trees	●				0.97 (F1)
2. Ahmed et al. [84]	A	Syntax (A.i)	Twitter, Kaggle, Horne and Adali [85]	SVM	●				0.92 (Acc.)
3. Bourgonje et al. [86]	A	Stance (A.ii)	Fake News Challenge Data	Logit	●				0.90 (Acc.)
1. Cao et al. [87]	U N	Acct. cred. (U.i); prop. (N.i)	Tuenti social network	Louvain clustering		●	●		0.90+ (TP)
2. Ezzeddine et al. [88]	U	Acct. behaviors (U.ii)	DATA	LSTM		●	●		0.91 (AUC)
3. Hamdi et al. [89]	U N	Account metadata (U.i); prop. (N.i)	CREDBANK	LDA, Bayes, Logit, SVM		●	●		0.99 (AUC)
1. Alizadeh et al. [90]	N A U	Propagation (N.i); syntax (A.i); acct metadata (U.i)	Twitter (API), Reddit IRA troll list	RF	●	●	●		0.70+ (F1)
2. Antoniadis et al. [91]	U A	Acct metadata (U.i); syntax (A.i)	Hurricane Sandy tweet dataset	J48, RF, KNN, Bayes	●		●		0.79 (Avg. Prec.)
3. Buntain et al. [92]	N U A	Time (N.ii); acct metadata (U.i); sentiment (A.i)	CREDBANK, Buzzfeed	RF	●	●	●		0.65 (Acc.)
1. Baly et al. [93]	W A N	Source rep. (W.i); Site infra. (W.ii); (A.i)	MediaBiasFactCheck[94]	SVM	●			●	0.7152 (Acc.)
2. Castelo et al. [95]	W A	Site infra. (W.ii); syntax (A.i)	Celebrity, US-Election2016	SVM, kNN, RF	●			●	0.86 (Acc.)
3. Chen et al. [96]	W A	Hosting infra. (URL) (W.ii); syntax (A.i)	PoliticalFakeNews	Clustering	●			●	0.97 (AUC)

Table 1: **Focus corpus by scope and target.** Coding of 15 papers in our focus set, sorted by information scope. Codes for the full focus set appear in Appendix C. Values in parentheses in "Target" field correspond to highlighted subcategories presented in Section 4 (e.g., "C.i" denotes target (i) in "Claims," Section 4.1). If authors present evaluation results for multiple models, we underline the most performant model and record its corresponding performance score.

to generate misinformative text [107], selectively editing true news articles (e.g., via verb inversion) [108], or compiling articles from authoritative news sources and known satire sites [109]. These techniques **(1.iii)** *introduce additional dependencies and shortcuts* to datasets for which such variables are already difficult to control. Style [103], for instance, is an emergent quality of writing that cannot be easily marginalized out of a text embedding.

② **Model selection.** Among papers that report testing with multiple models—including **(2.ii)** *classical ML models* [84], **(2.iii)** *unsupervised NN models* [105], and **(2.v)** *stacked ensemble classifiers* [110]—there is no clear correlation between model choice and actual performance. We note that, in instances where authors test on two- *and* multi (i.e., > 2)-way classification tasks, performance declines sharply in the latter case [83]. In those instances, reported performance scores are for two-way tasks. For this reason, as well, classical ML models (logit, SVM) oftentimes *appear* to be most performant.

③ **Feature selection.** Among supervised methods that disclose feature sets, we find that **(3.i)** *word frequency, sentiment, and genre* were among the most commonly used features, and were collectively used by 80% of works within this scope; these features can also be sources of dependency-induced noise. It is difficult to quantify the impact of dependencies related to tone and source on classifier performance, particularly in the case of unsupervised learning methods, which comprise 59% of methods at this scope. We evaluate an unsupervised learning method (and consider possible style-related dependencies) in our replication of Nasir et al. (5.1) [111].

④ **Evaluation.** Article-scoped methods risk overfitting to single topics: 42% of authors select **(4.ii)** *one or more narratives of interest* (e.g., the 2016 presidential election), train a classifier on these topics, then test this classifier on a different set of texts that discuss the *same* topic [111–113]. This approach, while valid for evaluating classifier performance on closed datasets, lacks ecological validity for the use cases that authors claim that their methods will address: high-quality annotations of relevant articles are generally unavailable in breaking news scenarios [114]. 60% of authors at this scope compare the performance of their detection method to published approaches or ML models, but **(4.ii)** *neglect to test on novel datasets*. These methods do well when tested on in-domain texts, and in comparison to a selection of older ML models; many report accuracy well above 80% [115, 116]. Only one paper within the article scope tested in an adversarial setting: Its authors found that, while stylometry-based misinformation detection had an accuracy rate greater than 80% on routine tasks, this score dropped to about 50% in adversarial cases [83].

Takeaways: 🎯 In the absence of semantic definitions of misinformation, *proxy* detection targets are common but easy to evade. ① Well-labeled datasets are rare; those datasets that are available are at least several years old at time of writing. ② Unsupervised methods show marginal improvements over classical ML models in some cases; it is unclear if these improvements are 1) significant or 2) sustainable across different datasets. ③ Detection methods at the article scope are uniquely susceptible to text-based dependencies. ④ Inflated performance scores can often be attributed to testing on same-topic news articles.

4.3 Users

Evidence of foreign interference during the 2016 U.S. presidential election triggered a resurgent interest in malicious account detection [60, 88]. As such, 90% of security- and social-science-oriented works that we include within this scope (a dozen papers) explicitly discuss Russian trolls or other influence operations conducted by nation state actors and train classifiers on published lists of such accounts [18, 60, 117–119]. Account-scoped papers formed approximately 15% of our corpus (40 papers).

🎯 **Detection targets.** Papers within this scope target source reputation, and **(U.i)** *inspect account metadata*, such as bios, account age, and profile images; or distinguish suspicious accounts by **(U.ii)** *a single user's social behaviors*, such as their comments on posts (**n.b.** this target is distinct from **(N.i)**). The security literature we review discusses trolls and bots deployed for astroturfing, misinformation campaigns, and IOs [60, 90, 120–122]. In the absence of rigorous definitions of these account types, however, actual detection targets are tautological: A troll is an account that exhibits troll-like behavior, or that interacts with confirmed troll accounts [60].

① **Dataset curation.** All troll and bot account detection works we reviewed relied on published lists of "known" troll accounts for model training, but **(1.iv)** *neglected to mention the heavily manual investigation required to produce these original lists* [123, 124]. Researchers who compiled some of these account lists, including a set of several hundred Twitter accounts with possible links to a known Russian troll farm (the Internet Research Agency, or IRA) manually examined suspicious accounts and tweet contents in order to produce detailed account and content taxonomies; notably, these classifications required external intelligence about account activity that was not published alongside account lists [124]. Two well-cited troll lists compiled by the U.S. government, comprising thousands of suspicious Twitter and Facebook accounts, were curated using proprietary non-public information [125].

② **Model selection.** We note that, regardless of model choice, if classifier training data *and* feature selection reflect a heuristic about suspicious behavior, the resulting classifier will simply learn this heuristic: The methods we review can be used to detect accounts whose behaviors conform to heuristic assumptions, but cannot be used to surface novel malicious behaviors, and are not resistant to attacks or evasion [126]; we explore this further in our replication analysis of Saeed et al. (Section 5.2) [60]. Some methods at this scope and the network scope formulate detection as a (**2.iv**) *graph cut problem*, and describe approaches to identifying optimal cuts for isolating suspicious accounts [87, 127].

③ **Feature selection.** Methods that define suspicious accounts by intrinsic *properties* of these accounts (e.g., user handles and profile images) target the **(3.iii)** *semantics of this account metadata*, and detect evidence of manipulation in (e.g.) image metadata and bios; or text outputs, such as posts and links [128, 129]. Methods that define suspicious accounts by account *activity* target **(3.ii)** *networked behaviors*, such as liking and resharing statistics [130]. Feature sets for some methods in the first category include demographic data for users, such as inferred political party affiliation or race [131] (these *n*th order assumptions are dangerous to make [132]; see our discussion about proxy signals, in Section 4.2).

④ **Evaluation.** We observe accuracy scores above 80% for **(4.ii)** *confirmatory detection of like accounts* for all methods that reference a seed list of known trolls [60, 88, 133]; de novo detection of behaviors not represented within training data is not possible, by the self-admission of 10% of authors within this scope [17, 60]. As we discuss further in the next subsection (*Networks*), the increasingly hybrid nature of IOs requires more nuanced taxonomies for classification: *extent* of coordination, rather than *existence*, might be a more appropriate measure of possible manipulation. Some authors of bot detection methods acknowledge that their approaches are **(4.iii)** *trivially easy to evade* if account holders 1) avoid interacting with known suspicious accounts or 2) vary their account identity and posting semantics [120, 134].

> **Takeaways**: 🎯 Targets are frequently tautological. ① Hand-annotated training datasets are the result of intensive fact-finding on the part of human researchers, and often require information that is not publicly available. ② Classifiers can only detect accounts resembling those in seed lists. ③ Features that attempt to infer user credibility from demographic information risk reinforcing existing biases. ④ Current methods cannot detect novel malicious behaviors.

4.4 Networks

Within the security literature, a growing awareness of hybrid networks, which use a combination of automated and manual approaches to disseminate content, has encouraged a turn toward *network-based* bot detection methods, and away from detection of individual accounts [120, 135]. We observe a parallel turn toward network-based methods in AI, ML, and NLP venues as a result of growing recognition of overfitting and generalizability issues in purely text-based detection methods (see *Articles*) [95]. The common assumption, across disciplines, is that coordinated networks leave more detectable evidence of manipulation than do individual accounts, and that these footprints should be identifiable regardless of attack type or rumoring topic [15, 95]. Network-scoped methods, including relevant security literature, form 20% of our corpus.

🎯 **Detection targets.** *All* methods at this scope identify patterns of user interaction and content propagation as targets; these methods associate virality with the existence of rumoring narratives [27, 106, 136] and temporally anomalous activity with evidence of coordination [17, 118, 120]. The corresponding targets for these approaches are **(N.i)** *propagation patterns across social graph topologies* and **(N.ii)** *temporal records of user-user interactions*. Within non-security misinformation literature, we note that virality assumptions disallow *early* detection of misinformation [137, 138]. Similarly, in the security literature, anomalous patterns of account registration and user interaction serve as proxies for the presence of Sybils and botnets [134, 139]; early detection requires that authors formulate a priori assumptions about these patterns.

① **Dataset curation.** We conducted an author outreach survey for works within this scope in an attempt to locate hard-to-find social media datasets. We found that **(1.v)** *accessibility issues* were exacerbated by the shutdown of the Twitter API [140]. In total, we attempted to locate datasets and code for 50 papers (see Section 5 for our methods). Thirty-six (72%) of these analyzed tweet corpora, and 42 (84%) of these targeted social media users and posts. We were able to independently source complete methods or data for fourteen (28%) of these. Of the 27 authors we eventually contacted about providing partial or dehydrated datasets, nine responded; six of those authors were able to provide method code or partial datasets.

② **Model selection.** The network-scoped methods we review formulate detection as 1) a structured content classification problem [90, 141], and/or 2) a clustering problem on social graphs [37, 142]. In the former case, authors use an assortment of **(2.ii, 2.iii)** *supervised and unsupervised models* to detect suspicious language across multiple accounts. In the latter case, authors use **(2.vi)** *Louvain or K-means clustering or K-nearest neighbors* to detect neighborhoods of suspicious accounts, as determined by user-user interactions. Though methods in the latter category advertise themselves as content-agnostic, we note that published methods [143, 144] access datasets of social media posts that were already sorted by rumoring topic or event [145, 146].

③ **Feature selection.** 55% of papers within this scope make normative theoretical assumptions about user behaviors: In keeping with an epidemiological model[6] of misinformation spread [149, 150], Nguyen et al. assume a homogeneous population of newsreaders, with **(3.ii)** *identical probabilities of "infection" and reinfection* [151]. Similarly, in the security literature, techniques for detecting bots and Sybils identify behaviors that align with heuristics determined a priori by researchers: These methods assume, for instance, that Sybils will form well-connected neighborhoods [139], or (seemingly contradictorily) that compromised Sybils will refrain from connecting with additional Sybils, to avoid detection [134]. [7]

④ **Evaluation.** Ferrara et al. called attention to the false positive rate problem in botnet detection in 2016, noting that classifiers for bot detection only work when there is a clear-cut distinction between bot and non-bot accounts [153]. This distinction is becoming increasingly blurred, however. Sophisticated network-based attacks try to engage non-bot accounts in organic interactions with bot accounts (e.g., astroturfing attacks) [17, 120], **(4.v)** *rendering even positive detection results insufficient or meaningless*: coordinated activity need not be inauthentic, and inauthentic activity need not be malicious.

[6] Some authors have argued that the disease metaphor promotes an overly simplistic model of information spread [147, 148].

[7] In fact, [152] states that such attacks cannot be prevented unless assumptions are made about account behaviors .

Takeaways: 🎯 Pattern- and virality-based detection approaches disallow early detection of rumors. ① Current social media data is difficult to obtain. ② "Topic-agnostic" classifier design occurs downstream of topic-aware dataset design. ③ Epidemiological models of information spread make strong assumptions about opinion formation and user behaviors; feature sets reflect these notions. ④ Though network-scoped methods are less susceptible to content-based dependencies than are content-aware methods, they cannot infer intent of the behaviors they detect.

4.5 Websites

Methods within this scope apply credibility, factuality, or (political) bias scores to news sites; authors claim that site-wide labels can be used to quickly infer the quality of individual news articles produced by these sites [93, 95, 96, 154]. As with article-scoped methods, we noted intervention fit issues at the whole website scope. Asr et al. found that source labels were insufficient proxies for the factuality of single articles, and elided subject-specific variations in reporting quality [155].

🎯 **Detection targets.** In 50% of works that we review in this scope, authors reduce the task of whole-site credibility labeling to a significantly smaller, unimodal classification task: Chen et al. [96] **(W.i)** *detect suspicious domain semantics* (in essence, a text classification task on URLs); Ribeiro et al. [131] **(W.ii)** *infer site bias from site visitor demographics*; Castillo et al. [138] **(W.iii)** *detect suspicious ad interfaces and markup features*; Baly et al. [5, 93] and Hounsel et al. [154] present methods incorporating **(W.iv)** *infrastructural features*, though the overall performance of the method of Baly et al. is strongly determined by performance on the text classification task alone (thus, in practice, the method closely resembles the article-scoped detection methods we examine, and is susceptible to the same dependencies that we observe in that scope). We demonstrate this via an ablation analysis in our replication study of their method (see Section 5.3).

① **Dataset curation.** For both training and testing, all methods scoped to whole website detection rely on published lists of websites with credibility scores [5, 93, 95, 154]; common reference sites include Media Bias/Fact Check, Snopes, and FactCheck.org [94, 156, 157]. As discussed in Section 4.1, however, these references do not have uniform taxonomies for classifying site credibility. Additionally, pre-labeled lists are 1) **(1.i)** *biased towards older, more visible real news and fake news outlets* [156], 2) **(1.iii)** *are restricted to specific information domains* [5], or 3) **(1.i)** *include inactive websites within their labeled datasets* [94]. Some papers perform infrastructure analysis on contemporaneous snapshots (circa 2019) of websites in their corpora, even though the text-based features for the same analysis were drawn from datasets published in 2016 and 2017 [95]. This constitutes a serious source of **(1.ii)** *temporal leakage*. No works within this scope discuss approaches to accounting for uneven distributions in training data, or how they might account for shifts in baseline distributions during the lifecycle of a website. Hounsel et al. train their classifier on a reference list in which 34% of misinformation training set sites were active and *all* websites in the real news training set were active, possibly resulting in overfitting to features specific to inactive websites [154].

② **Model selection.** Four of the seven methods we reviewed within this scope used (**2.ii**) *SVM classifiers*; in two of those cases, SVM outperformed other, more complex unsupervised models [95, 155]. These results accord with our earlier observation, in Section 4.2, that SVM classifiers are comparatively performant on two-way classification tasks.

③ **Feature selection.** Works within this scope use multimodal feature sets comprising a mix of **(3.i)** *textual*, **(3.ii)** *network-based*, and **(3.iv)** *infrastructural signals*: for instance, Baly et al. [93] consider network traffic, URL semantics, and site contents; and Hounsel et al. [154] consider TLS/SSL certificates, web hosting configurations, and domain registrations. None of the detection methods we reviewed discusses the computational costs of deploying their methods at scale. Though all works discuss their feature selection process (via leave-one-out and use-one-only evaluations), none describes a process for normalizing or weighting features according to dataset distribution or detection setting needs.

④ **Evaluation.** Methods within this scope that propose to perform whole-site labeling from analysis of a selection of news articles or infrastructural features are susceptible to distributional imbalances (e.g., between news verticals represented in an article corpus). Baly et al. train their model on **(4.ii)** *political news websites only*, and their credibility labels are strongly correlated with political bias scores. Hounsel et al. perform **(4.i)** *testing in real time*—one of the few studies we reviewed, and one of two studies at this scope, that did so [96, 154]. Both of these works report significant performance dropoffs between experimental and real-time tests, most likely as a result of distributional differences between real-world and experimental datasets (in practice, most websites do not host news-related content at all) [96, 154].

Takeaways: Methods claiming to classify news sites generally reduce this task to simpler, unimodal tasks, such as URL classification. ① The works we review do not consider shifts in feature distribution over time. ② We note that SVM classifiers perform well on two-way classification tasks, and even outperform more sophisticated unsupervised models. ③ Feature normalization largely undiscussed. ④ Authors who conduct testing in separate real-time settings report significant performance dropoffs with respect to experimental results.

5 REPLICATION STUDIES

We choose three distinct targets and scopes to replicate issues identified in our literature review. Our targets are 1) (**A.i**) syntax-based text features; 2) (**U.ii**) user behaviors; and 3) (**W.iv**) multimodal whole-website features, including hosted content and infrastructure. These works are highly representative of their respective information scopes: Nasir et al. [111] (Section 5.1) train a neural network on corpora of true and misinformative news articles; Saeed et al. [60] (Section 5.2) use published lists of trolls to infer the presence of other troll accounts on Reddit; and Baly et al. [5] (Section 5.3) examine a multimodal feature set comprising infrastructure, content, and network-based features in order to infer whole-site credibility. For each work, we evaluate replicability and generalizability (RQ3, RQ4). Where possible, we 1) replicate reported results; 2) inspect datasets for potential dependencies; 3) perform ablation analyses to understand individual feature performance; and 4) test on current data.

Paper selection criteria and author outreach. We sorted our full text corpus by information scope. Within each scope, we sorted papers in order of decreasing citation count. We then proceeded as follows:

1. We attempted to source the full methods and datasets for the most cited paper within each information scope.
2. If we were unable to find this information during an independent web search, we reached out to the paper's lead author(s) to request access.
3. If this request was unsuccessful—if the author did not respond, or confirmed that the dataset or code was no longer available—we returned to step 1 for the next most highly-cited paper in our corpus within that scope.

Replication analyses. In order to reproduce results published in a selection of papers and perform cross-cutting analyses on out-of-domain and out-of-sample datasets, we conduct a series of replication analyses on a subset of papers. We chose representative methods from disparate information scopes, and which consider a variety of different feature types. We reproduced results reported in each paper on the datasets mentioned therein, contacting authors when necessary to obtain datasets and code. We evaluated reproducibility and generalizability as follows:

- **Reproducibility.** We reproduce published results with code and data reported in the original publication. We evaluate availability of code and data and, where possible, compare our analysis outcomes with those reported in the original paper (RQ3).
- **Explainability.** Toward understanding the contributions of specific feature types to overall classifier performance, and why certain approaches work, we perform feature ablation studies when appropriate.
- **Replicability and generalizability.** Model performance on novel datasets is useful for determining the generalizability of existing detection methods to different contexts (RQ4). For models that were explicitly tested on specific misinformation narratives (e.g., 2020 stolen election narratives), on specific timeframes, or on specific types of misinformation (e.g., parody, satire), we

Dataset (size)	Features	Acc.	FPR	FNR
ISOT (45,000)	`original`	0.995	0.00	0.00
	`scrubbed`	0.987	0.00	0.01
FA-KES (804)	`original`	0.521	0.843	0.151
Reuters	`original`	1	0	–
	`modified`	0	–	1
NYTimes	`original`	0.70	0.28	–
	`modified`	0.28	–	0.72

Table 2: Replication analysis of Nasir et al. (2021). We tested the method of Nasir et al. on both datasets discussed in the original paper, and on novel datasets from Reuters and The New York Times [111].

develop updated datasets to test method performance on diverse information domains.

5.1 Detection of suspicious language

In view of rampant data dependency issues identified in our survey of article-scoped literature, we reproduced results from a representative study published in 2021 by Nasir et al [111]. The authors propose a neural net-based approach to the classification of *news articles*. The method employs a hybrid deep learning model that combines convolutional and recurrent neural networks for the classification of real and fake news. The authors report results for tests on two datasets: the ISOT dataset (45,000 news stories, equally distributed across true and false categories, as labeled by PolitiFact) and the FA-KES dataset (804 news stories about the Syrian war, 426 true and 376 false) [158, 159]. We were able to replicate original paper results and run the method on updated datasets of news articles. Motivated by the prevalence of methods trained on few- or single-source datasets in the article scope, we test possible dependencies related to journalistic house style, as *all true articles in the training corpus were sourced from Reuters.*

House style as a confounder. To investigate house style as a possible confounder for misinformation detection, we excerpted 100 articles from both Reuters and The New York Times, two news outlets with distinctive (and different) reporting styles. We randomly selected these articles from both outlets' RSS feeds in May 2023. We sourced articles for both corpora from the following verticals: U.S. and world politics, economics, science, and entertainment. Excerpt lengths ranged from 100 to 300 words. These corpora, each comprising 100 true news stories, are labeled "Reuters-`original`" and "NYTimes-`original`" in Table 2.

We then selectively edited 50% of the news articles within each corpus. We changed proper nouns, negated verbs, and altered reported statistics so that the factual content of these articles was no longer accurate but house style and tone were preserved. We call these altered text corpora "Reuters-`modified`" and "NYTimes-`modified`" in Table 2. The classifier had 1.0 accuracy on Reuters-`original` and 0.70 accuracy on NYTimes-`original`; this difference is significant ($p < 0.05$). Additionally, the classifier had about 0.50 (random) accuracy on both modified datasets (all 100 articles), and far worse performance on the 50 modified articles alone (accuracy 0 and 0.28, respectively, on Reuters and the NYT). For the most part, misinformative NYTimes articles were classified as factual (FNR = 0.72 for NYTimes-`modified`), while *all* misinformative Reuters articles were classified as factual (FNR = 1 for Reuters-`modified`). These results suggest that the classifier was possibly overfitting to Reuters style signals in factual excerpts: All modified Reuters excerpts, though semantically *false*, were classified as *true*. Additionally, *within* each modified corpus, the classifier did not effectively differentiate between true and untrue news articles. We note that classifier performance in an adversarial setting was no better than random, and that style appears to have a significant impact on classifier sensitivity.

5.2 Detection of suspicious accounts

We reproduce results from a study published in 2022 by Saeed et al. [60]. In summary, the authors propose a method, called TrollMagnifier, for the identification of Reddit accounts that exhibit troll-like behaviors. Like all other account-scoped methods we analyze in the security literature, TrollMagnifier is trained on posting and reply statistics for non-troll and known Russian troll accounts (as identified by Reddit) [60].

Reproducibility of published results. Study authors provided us with pre-processed datasets and classifier code upon email request. The full Reddit Pushshift dataset is freely available online [160]. We were able to replicate original paper results using these materials. Original account handles were anonymized; as such, we were unable to verify if the accounts identified in the original study appeared to be troll-like.

Tautological targets. As described in preceding sections (Section 4.3), suspicious account detection suffers from a lack of clear and consistent definitions. Troll accounts cannot be described by degree of automation (while some trolls are bot-like, many others are operated by humans) [120] or the nature of the information they spread (this might be political misinformation, ads, or anything in between) [161]; as such, the definition implicitly offered is that a troll is an account that exhibits *troll-like behavior*: i.e., interacts with known troll accounts, or appears on the same posts or threads as these accounts. The authors note that this approach cannot be used to detect novel trolling behaviors, and requires a seed list of known troll accounts for every new detection task.

Implied versus actual data dimensionality. While the original TrollMagnifier paper strongly implied that the proposed method would leverage networked behaviors to identify

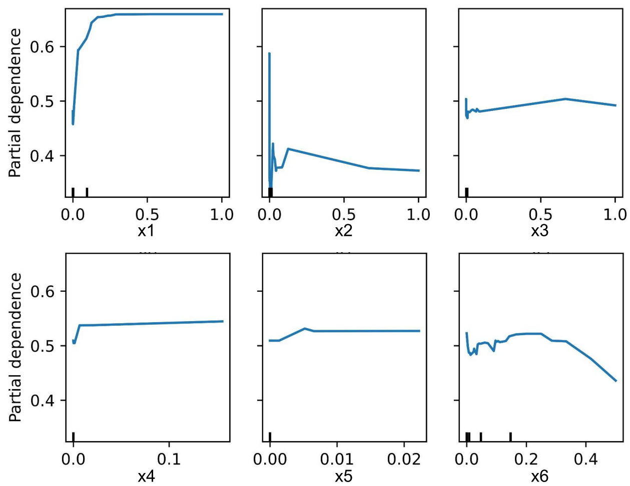

Figure 4: Partial dependence plots for each TrollMagnifier feature. Respectively, x1 = "comments on posts that trolls commented on," x2 = "comments on posts that trolls started," x3 = "direct comment in reply to troll post," x4 = "threaded comment in reply to troll comments," x5 = "threaded comment in reply to troll comments on a troll post" x6 = "same title post as troll."

troll accounts acting in coordination online, we found that, in actuality, the features under analysis lacked any sort of temporal component and were limited in scope. There were six features in all (described in full in Figure 2), each corresponding to an aggregate engagement statistic. Longitudinal data and timestamps were not available; as such, it was not possible to perform time series analysis. Account names were not available, disallowing construction of user graphs.

Feature importance. In our partial dependence analysis, we find that feature x1—commenting statistics—was the sole feature that consistently produced classification accuracy greater than 0.6 (most other features had accuracy no better than random). Per our earlier observation that many account-scoped methods target behaviors that are difficult to distinguish from routine online activity, we recommend that feature engineering for account- and network-scoped methods reflect some intuition about the nature of actually suspicious behaviors. Furthermore, the performance of the current feature set suggests that manual classification might be as effective as (or even more effective than) an automated approach that detects a content-agnostic heuristic.

5.3 Detection of suspicious websites

We reproduce results from a study published in 2018 by Baly et al [5]. In summary: the authors propose a multimodal approach to the classification of *news websites*. This method is particularly representative of works within this information scope: 50% of works within this scope use a similar mixed-modalities approach to detecting misinformation websites, and Baly et al. include site-specific feature types, including domain and traffic-based features, in their analysis. Baly et al. analyzed website contents, associated social media accounts, and Wikipedia pages in order to perform two classification tasks: fact and political bias classification. The authors developed a dataset of 1066 websites manually labeled for their political leaning (extreme-left, left, left-center, center, right-center, right, and extreme-right) and degree of credibility (low, mixed, high). These labels were extracted from the Media Bias/Fact Check (MBFC) database [94]. We were able to reproduce original study results and perform ablation analyses on existing datasets. We were unable to run the method on an updated dataset, as feature extraction code was not available.

Reproducibility of published results. All features extracted for the original analysis were captured in a series of json files. While we were able to readily reproduce results reported in the paper, certain elements of the dataset (follower counts on social media, Wikipedia page contents) were out of date. As we did not find documentation in the method repository for re-extraction of these features, we were restricted to conducting our tests on data that were already available. We binned bias labels into *left*, *center*, and *right* categories, as the seven-way taxonomy initially applied to the dataset by MBFC yielded small label classes. Classifier performance on the resulting three-way bias classification task accords with the results reported by Baly et al. on the same task.

Multimodal features: help or hindrance? We performed an ablation study of the method on the EMNLP18 dataset and analyzed the method's *bias* and *fact-checking* classification functions separately [5]. Specifically, we stratified the original EMNLP18 dataset by political leaning and credibility, as labeled by MBFC, and analyzed the performance of 1) the full feature set, 2) individual features and 3) ablated feature sets (removing one feature type per test). Our results are summarized in Table 4. We find that, on 11 out of 12 test datasets, classifier performance using only text-based features (`articles` and `wikipedia`, derived from articles randomly sampled from the website in question, and the site's corresponding Wikipedia page, respectively) was comparable to performance on the full feature set. On five out six datasets, bias classification accuracy on text-only features actually outperformed bias classification on the whole feature set (see the bottom half of Table 4), suggesting that the full-site classifier of Baly et al. was effectively a text content classifier.

6 Discussion

We present conclusions, discuss limitations of this work, and propose possible directions for future research.

6.1 Conclusions

RQ1: **Fit.** Very few methods that claimed to detect misinformation performed actual fact verification: Instead, they targeted proxy signals that were frequently steps away from promised detection targets. These differences were particularly noticeable in methods that relied heavily on text- and network-based features to perform classification. In those cases, semantic/syntactic signatures and propagation patterns served as proxies for the existence of misinformation. We demonstrated, through our own replication studies, that it is easy to circumvent approaches that rely on style-based cues to perform proxy detection.

RQ2: **Data curation and model explainability.** Lack of access to current, well-annotated datasets remains a serious problem for current and future misinformation research. Across existing datasets, taxonomies for classifying misinformation were inconsistent. Testing on contemporaneous data was uncommon among those papers we annotated, and testing in real-time settings was even rarer. Proof-of-concept experiments oftentimes did not control for data dependencies. Authors describing black-box methods—particularly those using neural nets or other forms of unsupervised learning—did not disclose feature sets retrieved by their methods.

RQ3: **Reproducibility.** We noted widespread code and data availability issues: Fewer than 30% of our attempts to locate code *and* data, or obtain this information from authors, were successful. The code and datasets we *were* able to retrieve were frequently unusable or out of date. We were able to reproduce and replicate results on published *and* current data for only one of our replication studies. In that case, we found that the article-scoped method performed no better than random (0.50 accuracy) on a current, mixed-domain dataset (5.1).

RQ4: **Generalizability.** Methods that were trained on single-source or single-domain datasets appeared to perform well on data from the same source, or within the same domain; these methods, unsurprisingly, performed poorly on multi-source or out-of-domain topics. These discrepancies are closely tied to undisclosed or uncontrolled-for data dependencies, which we discuss in RQ2.

6.2 Limitations

As we completed this work, we were limited by data (in)availability and definitional inconsistencies across papers. As a result of inaccessible or incomplete datasets, we were only able to reproduce works representing four of our five information scopes; moreover, those replications were not generalizable to newer datasets, as authors did not always include feature extraction code. We were unable to perform in-depth analysis of detection algorithms deployed in commercial settings owing to a similar lack of data and code transparency. Finally, the use of different proxy targets across works led to inconsistent task formulations that hindered apples-to-apples comparisons between papers in our full literature corpus.

As well, we were conscious of the limitations of common metrics in AI/ML research. The AI/ML research community has actively sought to introduce novel metrics that are more descriptive of potential sources of bias and class distributions in data. These include the Matthews correlation coefficient [162] for binary classification tasks and balanced accuracy [163]. In-depth discussion of these and other alternatives is beyond the scope of this work; briefly, both metrics attempt to improve upon existing measures by considering the size of positive and negative data classes in their reporting.

6.3 Recommendations for future research

We propose the following as directions for future research:

Designing for intervention fit. *All article-scoped methods detect proxy targets for misinformation. Some of these proxy targets, including house style and text sentiment, are easily circumvented (Section 5.1).* At the point of task formulation, researchers might consider the following: What signals—or proxy signals—will the intervention detect? What is the actual detection target, and how closely does it approximate the classification task to be performed? What are the potential risks of a false negative or false positive result (and are they worth the potential benefits of e.g., greater speed)?

Understanding hybrid detection. *While academic misinformation detection favors fully automated methods, commercial checking services and online platforms use hybrid methods.* A longstanding problem in usability research concerns the role of the automated system in human-computer interactions: should the ML method augment or supplant the abilities of the human moderator? We recommend approaches in the former vein that leave the final moderation decision to the human. The automated intervention could take many forms: detection of misinformative posts resembling content already identified by human annotators, similar to Facebook's SimSearchNet++ [164]; "triaging" content by topic [165]; or scheduling moderation tasks for human review [166].

Investigating dependencies and distributional shifts. *Fewer than 10% of methods within our whole corpus tested on contemporaneous data, and fewer than 5% tested in a real-time setting.* To understand how an ML intervention might work *in practice*, researchers must understand how robust their methods are to changes in the distribution of available data during training versus testing. These shifts are particular to medium (for instance, whole site classifiers are susceptible to fluctuations in news vertical coverage and author). Notably, the few works in our corpus that tested in real-time reported precipitous performance dropoffs. We recommend that researchers rigorously measure potential drift due to temporal and distributional shifts. This might entail training a model on a sliding window of timestamps, t_1 through t_{n-1}, testing on t_n, and recording any change in performance as a result of these time progressions. In supervised models, researchers can understand the contributions of individual feature classes to model performance through a partial dependence analysis (see, for instance, Section 5.2); or clustering datasets based on known dependencies (e.g., authorship), selecting whole clusters for inclusion in training or testing data, and ensuring no overlap between test and train.

Addressing nuanced detection challenges. *Many detection methods elide subtleties that are particular to medium*: for instance, article-level detection methods generally apply broad 'true'/'false' classifications to a text under analysis where only single sentences might be slightly inaccurate. Additionally, the language signals that most methods detect are fairly unsubtle. Suggestion, insinuation, and leading questions are powerful rhetorical tools that might render a newsreader more susceptible to actual misinformation, or that might suggest misinformative ideas using indirect means; no works within our literature corpus targeted these signals, however.

7 Ethics considerations

Our author outreach survey design was approved by the Princeton University IRB. All statistics published from those communications are in aggregate, with no personal identifiers attached to individual author responses. Datasets used for this work were already publicly available or were obtained with permission from study authors.

8 Open science

Code and data used for our analyses, as well as our full paper corpus, are included in our online SI (available at http://doi.org/10.5281/zenodo.15613696).

Acknowledgments

The authors would like to thank Anne Kohlbrenner for assisting with paper coding; Ben Kaiser and Sarah Scheffler for productive conversations, and for reviewing multiple drafts of this manuscript; and Mona Wang for retrieving data used for a replication study.

References

[1] Prateek Dewan and Ponnurangam Kumaraguru. "Towards automatic real time identification of malicious posts on Facebook". In: *2015 13th Annual Conference on Privacy, Security and Trust (PST)*. IEEE. 2015, pp. 85–92.

[2] Firoj Alam et al. "A Survey on Multimodal Disinformation Detection". en. In: *arXiv:2103.12541 [cs]* (Mar. 2021). URL: http://arxiv.org/abs/2103.12541 (visited on 03/29/2021).

[3] Darko Androcec. "Machine learning methods for toxic comment classification: a systematic review". In: *Acta Universitatis Sapientiae, Informatica* 12.2 (2020), pp. 205–216.

[4] Hee-Eun Lee et al. "Detecting child sexual abuse material: A comprehensive survey". In: *Forensic Science International: Digital Investigation* 34 (2020), p. 301022.

[5] Ramy Baly et al. "Predicting Factuality of Reporting and Bias of News Media Sources". In: *EMNLP 2018* (2018), pp. 3528–3539. URL: https://aclanthology.org/D18-1389/.

[6] Numa Dhamani et al. "Using Deep Networks and Transfer Learning to Address Disinformation". en. In: *arXiv:1905.10412 [cs]* (May 2019). URL: http://arxiv.org/abs/1905.10412 (visited on 06/03/2019).

[7] *Generalizing scaled misinformation detection*. 2023.

[8] Ina Fried. "OpenAI touts GPT-4 for content moderation". In: (2023).

[9] Daniel Arp et al. "Dos and Don'ts of Machine Learning in Computer Security". In: *31st USENIX Security Symposium (USENIX Security 22)*. Boston, MA: USENIX Association, Aug. 2022, pp. 3971–3988. ISBN: 978-1-939133-31-1. URL: https://www.usenix.org/conference/usenixsecurity22/presentation/arp.

[10] A. S. Jacobs et al. "AI/ML and Network Security: The Emperor has no Clothes". In: CCS '22. Los Angeles, CA, USA: Association for Computing Machinery, 2022.

[11] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and machine learning: Limitations and opportunities*. MIT press, 2023.

[12] Stephanie Chan et al. "Machine learning in dermatology: current applications, opportunities, and limitations". In: *Dermatology and therapy* 10 (2020), pp. 365–386.

[13] *How Automated Content Moderation Works (Even When It Doesn't)*. 2024. URL: https://themarkup.org/automated-censorship/2024/03/01/how-automated-content-moderation-works-even-when-it-doesnt-work (visited on 03/01/2024).

[14] Rosalie Gillett, Joanne E Gray, and D Bondy Valdovinos Kaye. "'Just a little hack': Investigating cultures of content moderation circumvention by Facebook users". In: *New Media & Society* 26.11 (2024), pp. 6183–6204.

[15] Sadegh M Milajerdi et al. "Holmes: real-time apt detection through correlation of suspicious information flows". In: *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE. 2019, pp. 1137–1152.

[16] Samuel T King and Peter M Chen. "SubVirt: Implementing malware with virtual machines". In: *2006 IEEE Symposium on Security and Privacy (S&P'06)*. IEEE. 2006, 14–pp.

[17] Brett Stone-Gross et al. "Your botnet is my botnet: analysis of a botnet takeover". In: *ACM CCS*. 2009, pp. 635–647.

[18] Shujaat Mirza et al. "Tactics, threats & targets: Modeling disinformation and its mitigation". In: *NDSS 2023*. 2023.

[19] Liang Wu et al. “Misinformation in Social Media: Definition, Manipulation, and Detection”. en. In: *ACM SIGKDD Explorations Newsletter* 21.2 (Nov. 2019), pp. 80–90. ISSN: 1931-0145, 1931-0153. DOI: 10.1145/3373464.3373475. URL: https://dl.acm.org/doi/10.1145/3373464.3373475 (visited on 03/16/2023).

[20] Axel Gelfert. “Fake News: A Definition”. en. In: *Informal Logic* 38.1 (Mar. 2018), pp. 84–117. ISSN: 0824-2577, 0824-2577. DOI: 10.22329/il.v38i1.5068. URL: https://ojs.uwindsor.ca/index.php/informal_logic/article/view/5068 (visited on 02/01/2019).

[21] Leonie Haiden and Jente Althuis. “The Definitional Challenges of Fake News”. In: *SBP-BRiMS 18*. June 2018. URL: http://sbp-brims.org/2018/proceedings/papers/challenge_papers/SBP-BRiMS_2018_paper_116.pdf (visited on 02/01/2019).

[22] David Klein and Joshua Wueller. *Fake News: A Legal Perspective*. en. SSRN Scholarly Paper ID 2958790. Rochester, NY: Social Science Research Network, Mar. 2017. URL: https://papers.ssrn.com/abstract=2958790 (visited on 02/01/2019).

[23] Edson C. Tandoc, Zheng Wei Lim, and Richard Ling. “Defining “Fake News”: A typology of scholarly definitions”. en. In: *Digital Journalism* 6.2 (Feb. 2018), pp. 137–153. ISSN: 2167-0811, 2167-082X. DOI: 10.1080/21670811.2017.1360143. URL: https://www.tandfonline.com/doi/full/10.1080/21670811.2017.1360143 (visited on 11/05/2018).

[24] Joshua Habgood-Coote. ““The term “fake news” is doing great harm””. In: *The Conversation* (July 2018). DOI: https://theconversation.com/the-term-fake-news-is-doing-great-harm-100406.

[25] Jasper Jackson. ““The term “fake news” is doing great harm””. In: *The Bureau of Investigative Journalism* (July 2023). DOI: https://www.thebureauinvestigates.com/stories/2023-07-27/what-are-influence-operations-and-why-are-we-investigating-them/.

[26] Oluwaseun Ajao, Deepayan Bhowmik, and Shahrzad Zargari. “Sentiment Aware Fake News Detection on Online Social Networks”. In: *ICASSP*. ISSN: 2379-190X. May 2019, pp. 2507–2511. DOI: 10.1109/ICASSP.2019.8683170.

[27] Marcella Tambuscio et al. “Fact-checking Effect on Viral Hoaxes: A Model of Misinformation Spread in Social Networks”. In: *WebConf*. WWW ’15 Companion. New York, NY, USA: Association for Computing Machinery, May 2015, pp. 977–982. ISBN: 978-1-4503-3473-0. DOI: 10.1145/2740908.2742572. URL: https://doi.org/10.1145/2740908.2742572 (visited on 03/16/2023).

[28] Sille Obelitz Søe. “Algorithmic detection of misinformation and disinformation: Gricean perspectives”. In: *Journal of Documentation* 74.2 (2017), pp. 309–332.

[29] Aldert Vrij. *Detecting lies and deceit: Pitfalls and opportunities*. John Wiley & Sons, 2008.

[30] Barbara G Amado, Ramón Arce, and Francisca Fariña. “Undeutsch hypothesis and Criteria Based Content Analysis: A meta-analytic review”. In: *The European Journal of Psychology Applied to Legal Context* 7.1 (2015), pp. 3–12.

[31] Carlos Carrasco-Farré. “The fingerprints of misinformation: how deceptive content differs from reliable sources in terms of cognitive effort and appeal to emotions”. In: *Humanities and Social Sciences Communications* 9.1 (2022), pp. 1–18.

[32] Junaed Younus Khan et al. “A benchmark study of machine learning models for online fake news detection”. en. In: *Machine Learning with Applications* 4 (June 2021), p. 100032. ISSN: 2666-8270. DOI: 10.1016/j.mlwa.2021.100032. URL: https://www.sciencedirect.com/science/article/pii/S266682702100013X (visited on 03/16/2023).

[33] Don Fallis. “A Functional Analysis of Disinformation”. In: *iConference 2014 Proceedings* (Mar. 2014). Publisher: iSchools. DOI: 10.9776/14278. URL: https://hdl.handle.net/2142/47258 (visited on 03/16/2023).

[34] Xinyi Zhou and Reza Zafarani. “A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities”. In: *ACM Computing Surveys* 53.5 (Sept. 2020), 109:1–109:40. ISSN: 0360-0300. DOI: 10.1145/3395046. URL: https://doi.org/10.1145/3395046 (visited on 03/16/2023).

[35] Ray Oshikawa, Jing Qian, and William Yang Wang. “A Survey on Natural Language Processing for Fake News Detection”. In: *arXiv:1811.00770 [cs]* (Nov. 2018). URL: http://arxiv.org/abs/1811.00770 (visited on 06/04/2019).

[36] Nadia K. Conroy, Victoria L. Rubin, and Yimin Chen. “Automatic deception detection: Methods for finding fake news”. en. In: *Proceedings of the Association for Information Science and Technology* 52.1 (2015), pp. 1–4. ISSN: 2373-9231. DOI: 10.1002/pra2.2015.145052010082. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/pra2.2015.145052010082 (visited on 03/16/2023).

[37] Karishma Sharma et al. “Combating Fake News: A Survey on Identification and Mitigation Techniques”. In: *ACM Transactions on Intelligent Systems and Technology* 10.3 (Apr. 2019), 21:1–21:42. ISSN: 2157-6904. DOI: 10.1145/3305260. URL: https://doi.org/10.1145/3305260 (visited on 03/16/2023).

[38] Md Rafiqul Islam et al. “Deep learning for misinformation detection on online social networks: a survey and new perspectives”. en. In: *Social Network Analysis and Mining* 10.1 (Sept. 2020), p. 82. ISSN: 1869-5469. DOI: 10.1007/s13278-020-00696-x. URL: https://doi.org/10.1007/s13278-020-00696-x (visited on 03/16/2023).

[39] Alim Al Ayub Ahmed et al. “Detecting Fake News using Machine Learning: A Systematic Literature Review”. In: *Psychology (Savannah, Ga.)* 58 (Jan. 2021), pp. 1932–1939. DOI: 10.17762/pae.v58i1.1046.

[40] Kai Shu et al. “Fake News Detection on Social Media: A Data Mining Perspective”. en. In: (), p. 15.

[41] Ammara Habib et al. “False information detection in online content and its role in decision making: a systematic literature review”. en. In: *Social Network Analysis and Mining* 9.1 (Sept. 2019), p. 50. ISSN: 1869-5469. DOI: 10.1007/s13278-019-0595-5. URL: https://doi.org/10.1007/s13278-019-0595-5 (visited on 03/16/2023).

[42] Jiawei Zhang et al. “Fake News Detection with Deep Diffusive Network Model”. In: *arXiv:1805.08751 [cs, stat]* (May 2018). URL: http://arxiv.org/abs/1805.08751 (visited on 06/04/2019).

[43] Nicollas R. de Oliveira et al. “Identifying Fake News on Social Networks Based on Natural Language Processing: Trends and Challenges”. en. In: *Information* 12.1 (Jan. 2021). Number: 1 Publisher: Multidisciplinary Digital Publishing Institute, p. 38. ISSN: 2078-2489. DOI: 10.3390/info12010038. URL: https://www.mdpi.com/2078-2489/12/1/38 (visited on 03/16/2023).

[44] Kai Shu et al. “Mining Disinformation and Fake News: Concepts, Methods, and Recent Advancements”. en. In: *Disinformation, Misinformation, and Fake News in Social Media: Emerging Research Challenges and Opportunities*. Ed. by Kai Shu et al. Lecture Notes in Social Networks. Cham: Springer International Publishing, 2020, pp. 1–19. ISBN: 978-3-030-42699-6. DOI: 10.1007/978-3-030-42699-6_1. URL: https://doi.org/10.1007/978-3-030-42699-6_1 (visited on 03/16/2023).

[45] Fernando Miró-Llinares and Jesús C. Aguerri. “Misinformation about fake news: A systematic critical review of empirical studies on the phenomenon and its status as a ‘threat’”. en. In: *European Journal of Criminology* (Apr. 2021), p. 1477370821994059. ISSN: 1477-3708. DOI: 10.1177/1477370821994059. URL: https://doi.org/10.1177/1477370821994059 (visited on 05/04/2021).

[46] Yimin Chen, Nadia K. Conroy, and Victoria L. Rubin. “News in an online world: The need for an “automatic crap detector””. en. In: *Proceedings of the Association for Information Science and Technology* 52.1 (2015), pp. 1–4. ISSN: 2373-9231. DOI: 10.1002/pra2.2015.145052010081. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/pra2.2015.145052010081 (visited on 03/16/2023).

[47] Miriam Fernandez and Harith Alani. “Online Misinformation: Challenges and Future Directions”. en. In: *Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW ’18*. Lyon, France: ACM Press, 2018, pp. 595–602. ISBN: 978-1-4503-5640-4. DOI: 10.1145/3184558.3188730. URL: http://dl.acm.org/citation.cfm?doid=3184558.3188730 (visited on 03/16/2023).

[48] Diego A Martin, Jacob N Shapiro, and Michelle Nedashkovskaya. “Recent Trends in Online Foreign Influence Efforts”. en. In: (), p. 34.

[49] Sushila Shelke and Vahida Attar. “Source detection of rumor in social network – A review”. en. In: *Online Social Networks and Media* 9 (Jan. 2019), pp. 30–42. ISSN: 2468-6964. DOI: 10.1016/j.osnem.2018.12.001. URL: https://www.sciencedirect.com/science/article/pii/S2468696418300934 (visited on 03/16/2023).

[50] Bin Guo et al. *The Future of Misinformation Detection: New Perspectives and Trends*. arXiv:1909.03654 [cs]. Sept. 2019. DOI: 10.48550/arXiv.1909.03654. URL: http://arxiv.org/abs/1909.03654 (visited on 03/16/2023).

[51] James Thorne and Andreas Vlachos. “Automated Fact Checking: Task Formulations, Methods and Future Directions”. en. In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, NM, USA, Aug. 2018, pp. 3346–3359.

[52] Arkaitz Zubiaga et al. “Analysing How People Orient to and Spread Rumours in Social Media by Looking at Conversational Threads”. en. In: *PLOS ONE* 11.3 (Mar. 2016). Publisher: Public Library of Science, e0150989. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0150989. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0150989 (visited on 03/16/2023).

[53] Victoria L. Rubin, Yimin Chen, and Nadia K. Conroy. “Deception detection for news: Three types of fakes”. en. In: *Proceedings of the Association for Information Science and Technology* 52.1 (2015). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pra2.2015.145052010083, pp. 1–4. ISSN: 2373-9231. DOI: 10.1002/pra2.2015.145052010083. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/pra2.2015.145052010083 (visited on 03/16/2023).

[54] L Padgham et al. ““CORE Rankings.”” In: (). DOI: https://www.core.edu.au/conference-portal.

[55] Miranda Wei et al. “‘SoK (or SoLK?): On the Quantitative Study of Sociodemographic Factors and Computer Security Behaviors’”. In: *USENIX Security* (2024). DOI: https://www.usenix.org/system/files/usenixsecurity24-wei-miranda-solk.pdf.

[56] Sarah Scheffler and Jonathan Mayer. “SoK: Content Moderation for End-to-End Encryption”. In: *Proceedings on Privacy Enhancing Technologies* (2023). DOI: https://petsymposium.org/popets/2023/popets-2023-0060.pdf.

[57] Noel Warford et al. “SoK: A Framework for Unifying At-Risk User Research”. In: *IEEE S&P* (2021). DOI: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9833643.

[58] Steve Mollman. *OpenAI is getting trolled for its name after refusing to be open about its A.I.* Mar. 2023. URL: https://fortune.com/2023/03/17/sam-altman-rivals-rip-openai-name-not-open-artificial-intelligence-gpt-4/.

[59] Matteo Wong. *There was never such a thing as “open” ai*. Jan. 2024. URL: https://www.theatlantic.com/technology/archive/2024/01/ai-transparency-meta-microsoft/677022/.

[60] Mohammad Hammas Saeed et al. “Trollmagnifier: Detecting state-sponsored troll accounts on reddit”. In: *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE. 2022, pp. 2161–2175.

[61] Amar Debnath et al. “A Hierarchical Learning Model for Claim Validation”. en. In: *Conference on Computational Intelligence*. Ed. by Mohammad Shorif Uddin and Jagdish Chand Bansal. Algorithms for Intelligent Systems. Springer Singapore, 2020, pp. 431–441. ISBN: 9789811375644.

[62] Naeemul Hassan et al. “Data in, fact out: automated monitoring of facts by FactWatcher”. In: *Proceedings of the VLDB Endowment* 7.13 (Aug. 2014), pp. 1557–1560. ISSN: 2150-8097. DOI: 10.14778/2733004.2733029. URL: https://doi.org/10.14778/2733004.2733029 (visited on 11/06/2022).

[63] Dimitrios Katsaros, George Stavropoulos, and Dimitrios Papakostas. “Which machine learning paradigm for fake news detection?” In: *International Conference on Web Intelligence (WI)*. Oct. 2019, pp. 383–387.

[64] Ziyi Kou et al. “HC-COVID: A Hierarchical Crowdsource Knowledge Graph Approach to Explainable COVID-19 Misinformation Detection”. In: *Proceedings of the ACM on Human-Computer Interaction* 6.GROUP (Jan. 2022), 36:1–36:25. DOI: 10.1145/3492855. URL: https://doi.org/10.1145/3492855 (visited on 01/19/2022).

[65] Kevin Roitero et al. “Can the Crowd Judge Truthfulness? A Longitudinal Study on Recent Misinformation about COVID-19”. en. In: *arXiv:2107.11755 [cs]* (July 2021). URL: http://arxiv.org/abs/2107.11755 (visited on 08/03/2021).

[66] Max Glockner, Yufang Hou, and Iryna Gurevych. *Missing Counter-Evidence Renders NLP Fact-Checking Unrealistic for Misinformation*. arXiv:2210.13865 [cs]. Oct. 2022. DOI: 10.48550/arXiv.2210.13865. URL: http://arxiv.org/abs/2210.13865 (visited on 03/16/2023).

[67] Yavuz Selim Kartal, Busra Guvenen, and Mucahid Kutlu. “Too Many Claims to Fact-Check: Prioritizing Political Claims Based on Check-Worthiness”. en. In: *arXiv:2004.08166 [cs]* (Apr. 2020). URL: http://arxiv.org/abs/2004.08166 (visited on 04/24/2020).

[68] Ciampaglia GL et al. “Computational Fact Checking from Knowledge Networks”. In: *PLoS ONE* 10.6 (June 2015). DOI: 10.1371/journal.pone.0128193.

[69] Tanik Saikh et al. “A Novel Approach Towards Fake News Detection: Deep Learning Augmented with Textual Entailment Features”. en. In: *Natural Language Processing and Information Systems*. Ed. by Elisabeth Métais et al. Lecture Notes in Computer Science. Springer International Publishing, 2019, pp. 345–358. ISBN: 978-3-030-23281-8.

[70] Luís Borges, Bruno Martins, and Pável Calado. “Combining Similarity Features and Deep Representation Learning for Stance Detection in the Context of Checking Fake News”. In: *arXiv:1811.00706 [cs, stat]* (Nov. 2018). URL: http://arxiv.org/abs/1811.00706 (visited on 06/04/2019).

[71] Prashant Shiralkar et al. “Finding Streams in Knowledge Graphs to Support Fact Checking”. In: Nov. 2017, pp. 859–864. DOI: 10.1109/ICDM.2017.105.

[72] Shrutika S. Jadhav and Sudeep D. Thepade. “Fake News Identification and Classification Using DSSM and Improved Recurrent Neural Network Classifier”. In: *Applied Artificial Intelligence* 33.12 (Oct. 2019). Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/08839514.2019.1661579, pp. 1058–1068. ISSN: 0883-9514. DOI: 10.1080/08839514.2019.1661579. URL: https://doi.org/10.1080/08839514.2019.1661579 (visited on 03/16/2023).

[73] Tayyaba Rasool et al. “Multi-Label Fake News Detection using Multi-layered Supervised Learning”. In: *Proceedings of the 2019 11th International Conference on Computer and Automation Engineering*. ICCAE 2019. New York, NY, USA: Association for Computing Machinery, Feb. 2019, pp. 73–77. ISBN: 978-1-4503-6287-0. DOI: 10.1145/3313991.3314008. URL: https://doi.org/10.1145/3313991.3314008 (visited on 03/16/2023).

[74] Lin Tian et al. “Early Detection of Rumours on Twitter via Stance Transfer Learning”. en. In: *Advances in Information Retrieval*. Ed. by Joemon M. Jose et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020, pp. 575–588. ISBN: 978-3-030-45439-5. DOI: 10.1007/978-3-030-45439-5_38.

[75] William Yang Wang. “"Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection”. en. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Vancouver, Canada: Association for Computational Linguistics, 2017, pp. 422–426. DOI: 10.18653/v1/P17-2067. URL: http://aclweb.org/anthology/P17-2067 (visited on 05/29/2019).

[76] “PolitiFact”. In: (). URL: https://www.politifact.com/.

[77] Yangqian Wang et al. “Learning Contextual Features with Multi-head Self-attention for Fake News Detection”. en. In: *Cognitive Computing – ICCC 2019*. Ed. by Ruifeng Xu, Jianzong Wang, and Liang-Jie Zhang. Lecture Notes in Computer Science. Springer International Publishing, 2019, pp. 132–142. ISBN: 978-3-030-23407-2.

[78] Michael Soprano et al. “The Many Dimensions of Truthfulness: Crowdsourcing Misinformation Assessments on a Multidimensional Scale”. en. In: *arXiv:2108.01222 [cs]* (Aug. 2021). URL: http://arxiv.org/abs/2108.01222 (visited on 08/11/2021).

[79] Bhavika Bhutani et al. “Fake News Detection Using Sentiment Analysis”. In: *2019 Twelfth International Conference on Contemporary Computing (IC3)*. ISSN: 2572-6129. Aug. 2019, pp. 1–5. DOI: 10.1109/IC3.2019.8844880.

[80] Kai Shu et al. "dEFEND: Explainable Fake News Detection". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. KDD '19. New York, NY, USA: Association for Computing Machinery, July 2019, pp. 395–405. ISBN: 978-1-4503-6201-6. DOI: 10.1145/3292500.3330935. URL: https://dl.acm.org/doi/10.1145/3292500.3330935 (visited on 03/16/2023).

[81] Arkaitz Zubiaga et al. "PHEME dataset of rumours and non-rumours". In: (2016).

[82] Abdullah-All-Tanvir et al. "Detecting Fake News using Machine Learning and Deep Learning Algorithms". In: *2019 7th International Conference on Smart Computing & Communications (ICSCC)*. June 2019, pp. 1–5. DOI: 10.1109/ICSCC.2019.8843612.

[83] Sadia Afroz, Michael Brennan, and Rachel Greenstadt. "Detecting Hoaxes, Frauds, and Deception in Writing Style Online". In: *2012 IEEE Symposium on Security and Privacy*. ISSN: 2375-1207. May 2012, pp. 461–475. DOI: 10.1109/SP.2012.34.

[84] Hadeer Ahmed, Issa Traore, and Sherif Saad. "Detecting opinion spams and fake news using text classification". en. In: *SECURITY AND PRIVACY* 1.1 (2018). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/spy2.9, e9. ISSN: 2475-6725. DOI: 10.1002/spy2.9. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/spy2.9 (visited on 03/16/2023).

[85] Benjamin D. Horne and Sibel Adali. "This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News". In: *AAAI CWSM'17*. Mar. 2017. URL: https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15772 (visited on 01/07/2019).

[86] Peter Bourgonje, Julian Moreno Schneider, and Georg Rehm. "From Clickbait to Fake News Detection: An Approach based on Detecting the Stance of Headlines to Articles". en. In: *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*. Copenhagen, Denmark: Association for Computational Linguistics, 2017, pp. 84–89. DOI: 10.18653/v1/W17-4215. URL: http://aclweb.org/anthology/W17-4215 (visited on 06/04/2019).

[87] Qiang Cao et al. "'Aiding the detection of fake accounts in large scale social online services'". In: *USENIX NSDI* (2012). DOI: https://www.usenix.org/system/files/conference/nsdi12/nsdi12-final42_2.pdf.

[88] Fatima Ezzeddine et al. "Exposing influence campaigns in the age of LLMs: a behavioral-based AI approach to detecting state-sponsored trolls". In: *EPJ Data Science* 12.1 (2023), p. 46.

[89] Tarek Hamdi et al. "A Hybrid Approach for Fake News Detection in Twitter Based on User Features and Graph Embedding". en. In: *Distributed Computing and Internet Technology*. Ed. by Dang Van Hung and Meenakshi D´Souza. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020, pp. 266–280. ISBN: 978-3-030-36987-3. DOI: 10.1007/978-3-030-36987-3_17.

[90] M. Alizadeh et al. "Content-based features predict social media influence operations". In: *Science Advances* (July 2020). DOI: 10.1126/sciadv.abb5824.

[91] Sotirios Antoniadis, Iouliana Litou, and Vana Kalogeraki. "A Model for Identifying Misinformation in Online Social Networks". en. In: *On the Move to Meaningful Internet Systems: OTM 2015 Conferences*. Ed. by Christophe Debruyne et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2015, pp. 473–482. ISBN: 978-3-319-26148-5. DOI: 10.1007/978-3-319-26148-5_32.

[92] Cody Buntain and Jennifer Golbeck. "Automatically Identifying Fake News in Popular Twitter Threads". In: *2017 IEEE International Conference on Smart Cloud (SmartCloud)* (Nov. 2017), pp. 208–215. DOI: 10.1109/SmartCloud.2017.40. URL: http://arxiv.org/abs/1705.01613 (visited on 06/04/2019).

[93] Ramy Baly et al. "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*. Online: Association for Computational Linguistics, July 2020, pp. 3364–3374. DOI: 10.18653/v1/2020.acl-main.308. URL: https://aclanthology.org/2020.acl-main.308.

[94] *Media Bias/Fact Check News*. July 2021. URL: https://mediabiasfactcheck.com/.

[95] Sonia Castelo et al. *A Topic-Agnostic Approach for Identifying Fake News Pages*. San Francisco, USA, 2019. DOI: 10.1145/3308560.3316739. URL: http://doi.acm.org/10.1145/3308560.3316739.

[96] Zhouhan Chen and Juliana Freire. "Proactive Discovery of Fake News Domains from Real-Time Social Media Feeds". In: *Companion Proceedings of the Web Conference 2020*. WWW '20. Taipei, Taiwan: Association for Computing Machinery, Apr. 2020, pp. 584–592. ISBN: 978-1-4503-7024-0. DOI: 10.1145/3366424.3385772. URL: https://doi.org/10.1145/3366424.3385772 (visited on 07/31/2020).

[97] Amr Magdy and Nayer Wanas. "Web-Based Statistical Fact Checking of Textual Documents". In: *Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents*. SMUC '10. Toronto, ON, Canada: Association for Computing Machinery, 2010, pp. 103–110. ISBN: 9781450303866. DOI: 10.1145/1871985.1872002. URL: https://doi.org/10.1145/1871985.1872002.

[98] Pujan Paudel et al. "'LAMBRETTA: Learning to Rank for Twitter Soft Moderation'". In: *IEEE Security & Privacy* (2023).

[99] Aditi Gupta et al. *TweetCred: Real-Time Credibility Assessment of Content on Twitter*. arXiv:1405.5490 [physics]. Jan. 2015. DOI: 10.48550/arXiv.1405.5490. URL: http://arxiv.org/abs/1405.5490 (visited on 03/16/2023).

[100] Limeng Cui, Suhang Wang, and Dongwon Lee. "SAME: Sentiment-Aware Multi-Modal Embedding for Detecting Fake News". In: *2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. ISSN: 2473-991X. Aug. 2019, pp. 41–48. DOI: 10.1145/3341161.3342894.

[101] Naeemul Hassan et al. "Toward Automated Fact-Checking: Detecting Check-worthy Factual Claims by ClaimBuster". en. In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '17*. Halifax, NS, Canada: ACM Press, 2017, pp. 1803–1812. ISBN: 978-1-4503-4887-4. DOI: 10.1145/3097983.3098131. URL: http://dl.acm.org/citation.cfm?doid=3097983.3098131 (visited on 06/04/2019).

[102] James Fairbanks et al. "Credibility assessment in the news: do we need to read". In: *Proc. of the MIS2 Workshop held in conjuction with 11th Int'l Conf. on Web Search and Data Mining*. ACM. 2018, pp. 799–800.

[103] Martin Potthast et al. "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)*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 231–240. DOI: 10.18653/v1/P18-1022. URL: https://aclanthology.org/P18-1022.

[104] Rada Mihalcea and Carlo Strapparava. "The Lie Detector: Explorations in the Automatic Recognition of Deceptive Language". In: *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*. Suntec, Singapore: Association for Computational Linguistics, Aug. 2009, pp. 309–312. URL: https://aclanthology.org/P09-2078.

[105] Adrian M. P. Braşoveanu and Răzvan Andonie. "Semantic Fake News Detection: A Machine Learning Perspective". en. In: *Advances in Computational Intelligence*. Ed. by Ignacio Rojas, Gonzalo Joya, and Andreu Catala. Lecture Notes in Computer Science. Springer International Publishing, 2019, pp. 656–667. ISBN: 978-3-030-20521-8.

[106] Kai Shu, Suhang Wang, and Huan Liu. "Beyond News Contents: The Role of Social Context for Fake News Detection". en. In: *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining - WSDM '19*. Melbourne VIC, Australia: ACM Press, 2019, pp. 312–320. ISBN: 978-1-4503-5940-5. DOI: 10.1145/3289600.3290994. URL: http://dl.acm.org/citation.cfm?doid=3289600.3290994 (visited on 05/29/2019).

[107] Verónica Pérez-Rosas et al. "Automatic Detection of Fake News". In: *arXiv:1708.07104 [cs]* (Aug. 2017). URL: http://arxiv.org/abs/1708.07104 (visited on 06/04/2019).

[108] Rada Mihalcea and Carlo Strapparava. "The lie detector: explorations in the automatic recognition of deceptive language". en. In: *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers on - ACL-IJCNLP '09*. Suntec, Singapore: Association for Computational Linguistics, 2009, p. 309. DOI: 10.3115/1667583.1667679. URL: http://portal.acm.org/citation.cfm?doid=1667583.1667679 (visited on 02/13/2019).

[109] Victoria Rubin et al. "Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News". In: *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*. San Diego, California: Association for Computational Linguistics, June 2016, pp. 7–17. DOI: 10.18653/v1/W16-0802. URL: https://aclanthology.org/W16-0802 (visited on 03/16/2023).

[110] Natali Ruchansky, Sungyong Seo, and Yan Liu. "CSI: A Hybrid Deep Model for Fake News Detection". In: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. CIKM '17. New York, NY, USA: ACM, 2017, pp. 797–806. ISBN: 978-1-4503-4918-5. DOI: 10.1145/3132847.3132877. URL: http://doi.acm.org/10.1145/3132847.3132877 (visited on 05/29/2019).

[111] Jamal Abdul Nasir, Osama Subhani Khan, and Iraklis Varlamis. "Fake news detection: A hybrid CNN-RNN based deep learning approach". en. In: *International Journal of Information Management Data Insights* 1.1 (Apr. 2021), p. 100007. ISSN: 2667-0968. DOI: 10.1016/j.jjimei.2020.100007. URL: https://www.sciencedirect.com/science/article/pii/S2667096820300070 (visited on 03/16/2023).

[112] "Rumors, False Flags, and Digital Vigilantes: Misinformation on Twitter after the 2013 Boston Marathon Bombing". en. In: *iConference 2014 Proceedings*. iSchools, Mar. 2014. ISBN: 978-0-9884900-1-7. DOI: 10.9776/14308. URL: https://www.ideals.illinois.edu/handle/2142/47257 (visited on 03/16/2023).

[113] Zhiwei Jin et al. "Detection and Analysis of 2016 US Presidential Election Related Rumors on Twitter". en. In: *Social, Cultural, and Behavioral Modeling*. Ed. by Dongwon Lee et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2017, pp. 14–24. ISBN: 978-3-319-60240-0. DOI: 10.1007/978-3-319-60240-0_2.

[114] Kai Shu et al. "Detecting fake news with weak social supervision". In: *IEEE Intelligent Systems* 36.4 (2020), pp. 96–103.

[115] Chaowei Zhang et al. "Detecting Fake News for Reducing Misinformation Risks Using Analytics Approaches". In: *European Journal of Operational Research* (June 2019). ISSN: 0377-2217. DOI: 10.1016/j.ejor.2019.06.022. URL: http://www.sciencedirect.com/science/article/pii/S0377221719304977 (visited on 06/18/2019).

[116] Amila Silva et al. "Embracing Domain Differences in Fake News: Cross-domain Fake News Detection using Multimodal Data". en. In: *arXiv:2102.06314 [cs]* (Feb. 2021). URL: http://arxiv.org/abs/2102.06314 (visited on 02/22/2021).

[117] Savvas Zannettou et al. "Disinformation warfare: Understanding state-sponsored trolls on Twitter and their influence on the web". In: *Companion proceedings of the 2019 world wide web conference*. 2019, pp. 218–226.

[118] Josephine Lukito. "Coordinating a multi-platform disinformation campaign: Internet Research Agency Activity on three US Social Media Platforms, 2015 to 2017". In: *Political Communication* 37.2 (2020), pp. 238–255.

[119] Franziska B Keller et al. "Political astroturfing on twitter: How to coordinate a disinformation campaign". In: *Political communication* 37.2 (2020), pp. 256–280.

[120] Stefano Cresci. "A Decade of Social Bot Detection". In: *Commun. ACM* 63.10 (Sept. 2020), pp. 72–83. ISSN: 0001-0782. DOI: 10.1145/3409116. URL: https://doi.org/10.1145/3409116.

[121] Zi Chu et al. "Detecting automation of twitter accounts: Are you a human, bot, or cyborg?" In: *IEEE Transactions on dependable and secure computing* 9.6 (2012), pp. 811–824.

[122] Jinxue Zhang et al. "The rise of social botnets: Attacks and countermeasures". In: *IEEE Transactions on Dependable and Secure Computing* 15.6 (2016), pp. 1068–1082.

[123] Craig Silverman et al. *Facebook groups topped 10,000 daily attacks on election before Jan. 6, analysis shows*. Jan. 2022. URL: https://www.washingtonpost.com/technology/2022/01/04/facebook-election-misinformation-capitol-riot/.

[124] Darren L Linvill and Patrick L Warren. "Troll factories: Manufacturing specialized disinformation on Twitter". In: *Political Communication* 37.4 (2020), pp. 447–467.

[125] URL: https://democrats-intelligence.house.gov/uploadedfiles/exhibit_b.pdf.

[126] Gang Wang et al. "Man vs. machine: Practical adversarial detection of malicious crowdsourcing workers". In: *23rd USENIX Security Symposium (USENIX Security 14)*. 2014, pp. 239–254.

[127] Huiling Zhang et al. "Misinformation in Online Social Networks: Detect Them All with a Limited Budget". In: *ACM Transactions on Information Systems* 34.3 (Apr. 2016), 18:1–18:24. ISSN: 1046-8188. DOI: 10.1145/2885494. URL: https://doi.org/10.1145/2885494 (visited on 03/16/2023).

[128] Stefan Helmstetter and Heiko Paulheim. "Weakly Supervised Learning for Fake News Detection on Twitter". In: *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (Aug. 2018). DOI: https://ieeexplore.ieee.org/document/8508520.

[129] Giuseppe Sansonetti et al. "'Unreliable Users Detection in Social Media: Deep Learning Techniques for Automatic Detection'". In: *IEEE Access* (2020). DOI: 2020.

[130] Diogo Pacheco et al. "Uncovering Coordinated Networks on Social Media". en. In: *arXiv:2001.05658 [physics]* (Jan. 2020). URL: http://arxiv.org/abs/2001.05658 (visited on 01/22/2020).

[131] Filipe Ribeiro et al. "Media Bias Monitor: Quantifying Biases of Social Media News Outlets at Large-Scale". en. In: *Proceedings of the International AAAI Conference on Web and Social Media* 12.1 (June 2018). Number: 1. ISSN: 2334-0770. DOI: 10.1609/icwsm.v12i1.15025. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/15025 (visited on 03/16/2023).

[132] Deen Freelon et al. "Black Trolls Matter: Racial and Ideological Asymmetries in Social Media Disinformation". en. In: *Social Science Computer Review* (Apr. 2020), p. 0894439320914853. ISSN: 0894-4393. DOI: 10.1177/0894439320914853. URL: https://doi.org/10.1177/0894439320914853 (visited on 04/12/2020).

[133] Aseel Addawood et al. "Linguistic Cues to Deception: Identifying Political Trolls on Social Media". en. In: *Proceedings of the Thirteenth International AAAI Conference on Web and Social Media*. 2019, p. 11.

[134] George Danezis and Prateek Mittal. "Sybilinfer: Detecting sybil nodes using social networks." In: *Ndss*. San Diego, CA. 2009, pp. 1–15.

[135] Christian Grimme, Dennis Assenmacher, and Lena Adam. "Changing perspectives: Is it sufficient to detect social bots?" In: *Social Computing and Social Media. User Experience and Behavior: 10th International Conference, SCSM 2018, Held as Part of HCI International 2018, Las Vegas, NV, USA, July 15-20, 2018, Proceedings, Part I 10*. Springer. 2018, pp. 445–461.

[136] Soroush Vosoughi, Deb Roy, and Sinan Aral. "The spread of true and false news online". In: *Science* 359.6380 (Mar. 2018), pp. 1146–1151. DOI: 10.1126/science.aap9559.

[137] Federico Monti et al. *Fake News Detection on Social Media using Geometric Deep Learning*. 2019. arXiv: 1902.06673 [cs.SI].

[138] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. "Information credibility on twitter". In: *Proceedings of the 20th international conference on World wide web*. WWW '11. New York, NY, USA: Association for Computing Machinery, Mar. 2011, pp. 675–684. ISBN: 978-1-4503-0632-4. DOI: 10.1145/1963405.1963500. URL: https://doi.org/10.1145/1963405.1963500 (visited on 03/16/2023).

[139] Dong Yuan et al. "Detecting fake accounts in online social networks at the time of registrations". In: *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security*. 2019, pp. 1423–1438.

[140] *Developer policy – twitter developers | twitter developer platform*. URL: https://developer.twitter.com/en/developer-terms/policy#4-e.

[141] Jacob Ratkiewicz et al. "Detecting and tracking political abuse in social media". In: *Proceedings of the International AAAI Conference on Web and social media*. Vol. 5. 1. 2011, pp. 297–304.

[142] Dennis Assenmacher et al. “A two-phase framework for detecting manipulation campaigns in social media”. In: *Social Computing and Social Media. Design, Ethics, User Behavior, and Social Network Analysis: 12th International Conference, SCSM 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part I 22*. Springer. 2020, pp. 201–214.

[143] Yang Liu and Yi-Fang Brook Wu. “Early Detection of Fake News on Social Media Through Propagation Path Classification with Recurrent and Convolutional Networks”. en. In: (), p. 8.

[144] Jing Ma, Wei Gao, and Kam-Fai Wong. “Detect Rumors in Microblog Posts Using Propagation Structure via Kernel Learning”. en. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, 2017, pp. 708–717. DOI: 10.18653/v1/P17-1066. URL: http://aclweb.org/anthology/P17-1066 (visited on 02/13/2019).

[145] Jing Ma et al. “Detecting rumors from microblogs with recurrent neural networks”. In: *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence* (2016). DOI: https://www.ijcai.org/Proceedings/16/Papers/537.pdf.

[146] Xiamo Liu et al. “Real-time Rumor Debunking on Twitter”. In: *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (CIKM ’15)* (2015). DOI: https://doi.org/10.1145/2806416.2806651.

[147] Daniel Williams. *Misinformation is the symptom, not the disease: Daniel Williams*. Dec. 2023. URL: https://iai.tv/articles/misinformation-is-the-symptom-not-the-disease-daniel-walliams-auid-2690.

[148] Li Qian Tay et al. “Thinking clearly about misinformation”. In: *Communications Psychology* 2.1 (2024), p. 4.

[149] Fang Jin et al. “Epidemiological modeling of news and rumors on Twitter”. en. In: *Proceedings of the 7th Workshop on Social Network Mining and Analysis*. Chicago Illinois: ACM, Aug. 2013, pp. 1–9. ISBN: 978-1-4503-2330-7. DOI: 10.1145/2501025.2501027. URL: https://dl.acm.org/doi/10.1145/2501025.2501027 (visited on 01/05/2023).

[150] Sander van der Linden. *Misinformation: Susceptibility, spread, and interventions to immunize the public*. Mar. 2022. URL: https://www.nature.com/articles/s41591-022-01713-6.

[151] Nam P. Nguyen et al. “Containment of misinformation spread in online social networks”. In: *Proceedings of the 4th Annual ACM Web Science Conference*. WebSci ’12. New York, NY, USA: Association for Computing Machinery, June 2012, pp. 213–222. ISBN: 978-1-4503-1228-8. DOI: 10.1145/2380718.2380746. URL: https://doi.org/10.1145/2380718.2380746 (visited on 03/16/2023).

[152] John R Douceur. “The sybil attack”. In: *International workshop on peer-to-peer systems*. Springer. 2002, pp. 251–260.

[153] Emilio Ferrara et al. “The rise of social bots”. In: *Communications of the ACM* 59.7 (2016), pp. 96–104.

[154] Austin Hounsel et al. “Identifying Disinformation Websites Using Infrastructure Features”. In: *FOCI* (2020). URL: https://www.usenix.org/system/files/foci20-paper-hounsel.pdf.

[155] Fatemeh Torabi Asr and Maite Taboada. “The Data Challenge in Misinformation Detection: Source Reputation vs. Content Veracity”. In: *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 10–15. DOI: 10.18653/v1/W18-5502. URL: https://aclanthology.org/W18-5502 (visited on 03/16/2023).

[156] *Snopes*. URL: https://www.snopes.com/.

[157] *FactCheck.org*. URL: https://www.factcheck.org/.

[158] *Fake news detection datasets*. URL: https://onlineacademiccommunity.uvic.ca/isot/2022/11/27/fake-news-detection-datasets/.

[159] Fatima K. Abu Salem et al. *FA-Kes: A fake news dataset around the Syrian War*. Jan. 2019. URL: https://zenodo.org/record/2607278.

[160] Jason Baumgartner et al. “The pushshift reddit dataset”. In: *Proceedings of the international AAAI conference on web and social media*. Vol. 14. 2020, pp. 830–839.

[161] Kai-Cheng Yang, Emilio Ferrara, and Filippo Menczer. “Botometer 101: Social bot practicum for computational social scientists”. In: *Journal of Computational Social Science* 5.2 (2022), pp. 1511–1528.

[162] Davide Chicco and Giuseppe Jurman. “The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation”. In: *BMC genomics* 21 (2020), pp. 1–13.

[163] Kay Henning Brodersen et al. “The balanced accuracy and its posterior distribution”. In: *2010 20th international conference on pattern recognition*. IEEE. 2010, pp. 3121–3124.

[164] Tekla S. Perry. “How Facebook’s AI Tools Tackle Misinformation”. In: *IEEE Spectrum* (2020).

[165] Connie Moon Sehat et al. “Misinformation as a harm: structured approaches for fact-checking prioritization”. In: *Proceedings of the ACM on Human-Computer Interaction* 8.CSCW1 (2024), pp. 1–36.

[166] Thodoris Lykouris and Wentao Weng. “Learning to defer in content moderation: The human-ai interplay”. In: *arXiv preprint arXiv:2402.12237* (2024).

[167] Simone Leonardi, Giuseppe Rizzo, and Maurizio Morisio. “‘Automated Classification of Fake News Spreaders to Break the Misinformation Chain’”. In: *Information* (2021). DOI: https://doi.org/10.3390/info12060248.

[168] Giovanni Santia, Munif Mujib, and Jake Williams. “‘Detecting Social Bots on Facebook in an Information Veracity Context’”. In: *ICWSM* (2019). DOI: https://doi.org/10.1609/icwsm.v13i01.3244.

[169] Weiling Chen et al. "'Unsupervised rumor detection based on users' behaviors using neural networks'". In: *Pattern Recognition Letters* (2018). DOI: https://doi.org/10.1016/j.patrec.2017.10.014.

[170] Kelvin Chan, Barbara Ortutay, and Nicholas Riccardi. "Meta eliminates fact-checking in latest bow to Trump". In: *Associated Press* (2025).

[171] Julio CS Reis et al. "Supervised learning for fake news detection". In: *IEEE Intelligent Systems* 34.2 (2019), pp. 76–81.

[172] Roney Santos et al. "Measuring the impact of readability features in fake news detection". In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. 2020, pp. 1404–1413.

[173] Gang Wang et al. "You are how you click: Clickstream analysis for sybil detection". In: *22nd USENIX security symposium (USENIX Security 13)*. 2013, pp. 241–256.

[174] Haifeng Yu et al. "Sybillimit: A near-optimal social network defense against sybil attacks". In: *2008 IEEE Symposium on Security and Privacy (sp 2008)*. IEEE. 2008, pp. 3–17.

[175] Thomas Magelinski, Lynnette Hui Xian Ng, and Kathleen M Carley. "A synchronized action framework for responsible detection of coordination on social media". In: *arXiv preprint arXiv:2105.07454* (2021).

[176] Karishma Sharma et al. "Identifying coordinated accounts on social media through hidden influence and group behaviours". In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 2021, pp. 1441–1451.

[177] Margaret Mitchell et al. "Model Cards for Model Reporting". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency* (2019). DOI: https://doi.org/10.1145/3287560.3287596.

[178] Jim Maddock et al. "Characterizing Online Rumoring Behavior Using Multi-Dimensional Signatures". In: *CSCW* (Mar. 2015). DOI: 10.1145/2675133.2675280.

[179] Adrien Friggeri et al. "Rumor Cascades". In: *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014* 8 (May 2014), pp. 101–110. DOI: 10.1609/icwsm.v8i1.14559.

[180] James Thorne et al. "FEVER: a Large-scale Dataset for Fact Extraction and VERification". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 809–819. DOI: 10.18653/v1/N18-1074. URL: https://aclanthology.org/N18-1074.

[181] Srijan Kumar, Robert West, and Jure Leskovec. "Disinformation on the Web: Impact, Characteristics, and Detection of Wikipedia Hoaxes". In: *Proceedings of the 25th International Conference on World Wide Web*. WWW '16. Montréal, Québec, Canada: International World Wide Web Conferences Steering Committee, 2016, pp. 591–602. ISBN: 9781450341431. DOI: 10.1145/2872427.2883085. URL: https://doi.org/10.1145/2872427.2883085.

[182] Vahed Qazvinian et al. "Rumor has it: Identifying Misinformation in Microblogs". In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Edinburgh, Scotland, UK.: Association for Computational Linguistics, July 2011, pp. 1589–1599. URL: https://aclanthology.org/D11-1147.

[183] Pew Research Center. "Social Media Fact Sheet". In: *Pew Research Center* (Sept. 2022). URL: https://www.pewresearch.org/journalism/fact-sheet/social-media-and-news-fact-sheet/.

[184] Press Association. "Blue ticks for all: Twitter allows users to apply to be verified". In: *The Guardian* (July 2016). URL: https://www.theguardian.com/technology/2016/jul/19/blue-ticks-for-all-twitter-allows-all-users-to-be-verified.

[185] Hannah Rashkin et al. "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*. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 2931–2937. DOI: 10.18653/v1/D17-1317. URL: https://aclanthology.org/D17-1317.

[186] Justin Cheng et al. "Can Cascades Be Predicted?" In: *Proceedings of the 23rd International Conference on World Wide Web*. New York, NY, USA: Association for Computing Machinery, 2014, pp. 925–936. ISBN: 9781450327442. DOI: 10.1145/2566486.2567997. URL: https://doi.org/10.1145/2566486.2567997.

[187] Tal Schuster et al. "Limitations of stylometry for detecting machine-generated fake news". In: *Computational Linguistics* 46 (June 2020), pp. 499–510. URL: https://direct.mit.edu/coli/article/46/2/499/93369/The-Limitations-of-Stylometry-for-Detecting.

[188] Craig Silverman and Jeff Kao. *Infamous Russian troll farm appears to be source of Anti-Ukraine propaganda*. Mar. 2022. URL: https://www.propublica.org/article/infamous-russian-troll-farm-appears-to-be-source-of-anti-ukraine-propaganda.

[189] Lia Bozarth, Aparajita Saraf, and Ceren Budak. "Higher Ground? How Groundtruth Labeling Impacts Our Understanding of Fake News about the 2016 U.S. Presidential Nominees". In: *Proceedings of the International AAAI Conference on Web and Social Media* 46.2 (May 2020), pp. 48–59. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/7278.

[190] Andreas Pitsillidis et al. "Botnet Judo: Fighting Spam with Itself." In: *NDSS*. 2010.

[191] Fan Yang et al. "Automatic detection of rumor on Sina Weibo". en. In: *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics - MDS '12*. Beijing, China: ACM Press, 2012, pp. 1–7. ISBN: 978-1-4503-1546-3. DOI: 10.1145/2350190.2350203. URL: http://dl.acm.org/citation.cfm?doid=2350190.2350203 (visited on 02/13/2019).

[192] Srijan Kumar, Robert West, and Jure Leskovec. "Disinformation on the Web: Impact, Characteristics, and Detection of Wikipedia Hoaxes". en. In: *Proceedings of the 25th International Conference on World Wide Web - WWW '16*. Montréal, Québec, Canada: ACM Press, 2016, pp. 591–602. ISBN: 978-1-4503-4143-1. DOI: 10.1145/2872427.2883085. URL: http://dl.acm.org/citation.cfm?doid=2872427.2883085 (visited on 02/13/2019).

[193] Shan Jiang and Christo Wilson. "Linguistic Signals under Misinformation and Fact-Checking: Evidence from User Comments on Social Media". en. In: *Proceedings of the ACM on Human-Computer Interaction* 2.CSCW (Nov. 2018), pp. 1–23. ISSN: 25730142. DOI: 10.1145/3274351. URL: http://dl.acm.org/citation.cfm?doid=3290265.3274351 (visited on 03/28/2019).

[194] Jeppe Norregaard, Benjamin D. Horne, and Sibel Adali. "NELA-GT-2018: A Large Multi-Labelled News Dataset for The Study of Misinformation in News Articles". en. In: *arXiv:1904.01546 [cs]* (Apr. 2019). URL: http://arxiv.org/abs/1904.01546 (visited on 04/04/2019).

[195] Jooyeon Kim et al. "Leveraging the Crowd to Detect and Reduce the Spread of Fake News and Misinformation". en. In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining - WSDM '18*. Marina Del Rey, CA, USA: ACM Press, 2018, pp. 324–332. ISBN: 978-1-4503-5581-0. DOI: 10.1145/3159652.3159734. URL: http://dl.acm.org/citation.cfm?doid=3159652.3159734 (visited on 05/29/2019).

[196] Kai Shu, Suhang Wang, and Huan Liu. "Understanding User Profiles on Social Media for Fake News Detection". en. In: *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. Miami, FL: IEEE, Apr. 2018, pp. 430–435. ISBN: 978-1-5386-1857-8. DOI: 10.1109/MIPR.2018.00092. URL: https://ieeexplore.ieee.org/document/8397048/ (visited on 05/29/2019).

[197] Feng Qian et al. "Neural User Response Generator: Fake News Detection with Collective User Intelligence". en. In: *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*. Stockholm, Sweden: International Joint Conferences on Artificial Intelligence Organization, July 2018, pp. 3834–3840. ISBN: 978-0-9992411-2-7. DOI: 10.24963/ijcai.2018/533. URL: https://www.ijcai.org/proceedings/2018/533 (visited on 05/29/2019).

[198] Yimin Chen, Niall J. Conroy, and Victoria L. Rubin. "Misleading Online Content: Recognizing Clickbait As "False News"". In: *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*. WMDD '15. New York, NY, USA: ACM, 2015, pp. 15–19. ISBN: 978-1-4503-3987-2. DOI: 10.1145/2823465.2823467. URL: http://doi.acm.org/10.1145/2823465.2823467 (visited on 05/29/2019).

[199] Xinyi Zhou and Reza Zafarani. "Fake News: A Survey of Research, Detection Methods, and Opportunities". In: *arXiv:1812.00315 [cs]* (Dec. 2018). URL: http://arxiv.org/abs/1812.00315 (visited on 06/04/2019).

[200] Julio CS Reis et al. "Explainable machine learning for fake news detection". In: *Proceedings of the 10th ACM conference on web science*. 2019, pp. 17–26.

[201] Diego Sáez-Trumper. "Fake tweet buster: a webtool to identify users promoting fake news ontwitter". In: *HT*. 2014. DOI: 10.1145/2631775.2631786.

[202] Nguyen Vo and Kyumin Lee. "The Rise of Guardians: Fact-checking URL Recommendation to Combat Fake News". en. In: *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval - SIGIR '18*. Ann Arbor, MI, USA: ACM Press, 2018, pp. 275–284. ISBN: 978-1-4503-5657-2. DOI: 10.1145/3209978.3210037. URL: http://dl.acm.org/citation.cfm?doid=3209978.3210037 (visited on 06/04/2019).

[203] Julio Amador, Axel Oehmichen, and Miguel Molina-Solana. "Characterizing Political Fake News in Twitter by its Meta-Data". In: *arXiv:1712.05999 [cs, stat]* (Dec. 2017). URL: http://arxiv.org/abs/1712.05999 (visited on 06/05/2019).

[204] Yang Yang et al. "TI-CNN: Convolutional Neural Networks for Fake News Detection". In: *arXiv:1806.00749 [cs]* (June 2018). URL: http://arxiv.org/abs/1806.00749 (visited on 06/04/2019).

[205] Melanie Tosik, Antonio Mallia, and Kedar Gangopadhyay. "Debunking Fake News One Feature at a Time". In: *arXiv:1808.02831 [cs]* (Aug. 2018). URL: http://arxiv.org/abs/1808.02831 (visited on 06/04/2019).

[206] Xinyi Zhou et al. "Fake News Early Detection: A Theory-driven Model". In: *arXiv:1904.11679 [cs]* (Apr. 2019). URL: http://arxiv.org/abs/1904.11679 (visited on 06/04/2019).

[207] Shuo Yang et al. "Unsupervised Fake News Detection on Social Media: A Generative Approach". In: Feb. 2019.

[208] Marco Della Vedova et al. "Automatic Online Fake News Detection Combining Content and Social Signals". In: May 2018. DOI: 10.23919/FRUCT.2018.8468301.

[209] Kashyap Popat et al. "DeClarE: Debunking Fake News and False Claims using Evidence-Aware Deep Learning". en. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: ACL, Nov. 2018, pp. 22–32.

[210] Eugenio Tacchini et al. "Some Like it Hoax: Automated Fake News Detection in Social Networks". In: *arXiv:1704.07506 [cs]* (Apr. 2017). URL: http://arxiv.org/abs/1704.07506 (visited on 06/04/2019).

[211] Eugenio Tacchini et al. "Automated Fake News Detection in Social Networks". en. In: (2017), p. 15.

[212] S. Krishnan and M. Chen. "Cloud-Based System for Fake Tweet Identification". In: *2019 IFIP/IEEE Symposium on Integrated Network and Service Management (IM)*. Apr. 2019, pp. 720–721.

[213] Svitlana Volkova et al. "Explaining Multimodal Deceptive News Prediction Models". en. In: 2019, p. 4.

[214] Shamoz Shah and Madhu Goyal. "Anomaly Detection in Social Media Using Recurrent Neural Network". en. In: *Computational Science – ICCS 2019*. Ed. by João M. F. Rodrigues et al. Lecture Notes in Computer Science. Springer International Publishing, 2019, pp. 74–83. ISBN: 978-3-030-22747-0.

[215] Sung Soo Park and Kun Chang Lee. "A Comparative Study of Text analysis and Network embedding Methods for Effective Fake News Detection". ko. In: *Journal of Digital Convergence* 17.5 (May 2019), pp. 137–143. DOI: 10.14400/JDC.2019.17.5.137. URL: https://doi.org/10.14400/JDC.2019.17.5.137 (visited on 06/12/2019).

[216] Duc Minh Nguyen et al. "Fake News Detection using Deep Markov Random Fields". 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)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 1391–1400. URL: https://www.aclweb.org/anthology/N19-1141 (visited on 06/12/2019).

[217] Laura Burbach et al. "Who Shares Fake News in Online Social Networks?" en. In: *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization - UMAP '19*. Larnaca, Cyprus: ACM Press, 2019, pp. 234–242. ISBN: 978-1-4503-6021-0. DOI: 10.1145/3320435.3320456. URL: http://dl.acm.org/citation.cfm?doid=3320435.3320456 (visited on 06/17/2019).

[218] Maryam Ramezani et al. "News Labeling as Early as Possible: Real or Fake?" en. In: *arXiv:1906.03423 [cs]* (June 2019). URL: http://arxiv.org/abs/1906.03423 (visited on 06/17/2019).

[219] Xinyi Zhou and Reza Zafarani. "Network-based Fake News Detection: A Pattern-driven Approach". en. In: *arXiv:1906.04210 [cs]* (June 2019). URL: http://arxiv.org/abs/1906.04210 (visited on 06/17/2019).

[220] Adrian Rauchfleisch and Jonas Kaiser. "The false positive problem of automatic bot detection in social science research". In: *PloS one* 15.10 (2020), e0241045.

[221] *GARM Brand Safety Floor and Suitability Framework*. URL: https://wfanet.org/l/library/download/urn:uuid:7d484745-41cd-4cce-a1b9-a1b4e30928ea/garm+brand+safety+floor+suitability+framework+23+sept.pdf.

[222] Zefr. July 2022. URL: https://www.prnewswire.com/news-releases/zefr-acquires-israeli-ai-firm-adverifai-bolstering-technology-led-approach-to-identifying-and-defunding-misinformation-301590009.html.

[223] Saleem Alhabash et al. "Pathways to virality: Psychophysiological responses preceding likes, shares, comments, and status updates on Facebook". In: *Media Psychology* 22.2 (2019), pp. 196–216.

[224] Daniel Kapellmann Zafra et al. *How to understand and action Mandiant's intelligence on information operations*. Oct. 2022. URL: https://www.mandiant.com/resources/blog/understand-action-intelligence-information-operations.

[225] Santhosh Srinivasan. *The global disinformation index*. URL: https://www.disinformationindex.org/.

[226] Kevin Aslett et al. "News credibility labels have limited average effects on news diet quality and fail to reduce misperceptions". In: *Science advances* 8.18 (2022), eabl3844.

[227] Mingkun Gao et al. "To label or not to label: The effect of stance and credibility labels on readers' selection and perception of news articles". In: *Proceedings of the ACM on Human-Computer Interaction* 2.CSCW (2018), pp. 1–16.

[228] Mark Stencel, Erica Ryan, and Joel Luther. *Misinformation spreads, but fact-checking has leveled off*. June 2023. URL: https://reporterslab.org/misinformation-spreads-but-fact-checking-has-leveled-off/.

[229] Xishuang Dong et al. "Deep Two-path Semi-supervised Learning for Fake News Detection". en. In: *arXiv:1906.05659 [cs]* (June 2019). URL: http://arxiv.org/abs/1906.05659 (visited on 06/18/2019).

[230] Javier Sánchez-Junquera et al. "Unmasking Bias in News". en. In: *arXiv:1906.04836 [cs]* (June 2019). URL: http://arxiv.org/abs/1906.04836 (visited on 06/18/2019).

[231] Anoop K, Deepak P, and Lajish V. L. "Emotion Cognizance Improves Fake News Identification". en. In: *arXiv:1906.10365 [cs]* (June 2019). URL: http://arxiv.org/abs/1906.10365 (visited on 07/02/2019).

[232] Munyeong Lim and Sungbum Park. "A Study on the Preemptive Measure for Fake News Eradication Using Data Mining Algorithms : Focused on the M Online Community Postings". kor. In: *Journal of Information Technology Services* 18.1 (2019), pp. 219–234. ISSN: 1975-4256. DOI: 10.9716/KITS.2019.18.1.219. URL: http://www.koreascience.or.kr/article/JAKO201915561990199.page (visited on 07/02/2019).

[233] Atif Ahmad et al. "Strategically-Motivated Advanced Persistent Threat: Definition, Process, Tactics and a Disinformation Model of Counterattack". In: *Computers & Security* (July 2019). ISSN: 0167-4048. DOI: 10.1016/j.cose.2019.07.001. URL: http://www.sciencedirect.com/science/article/pii/S0167404818310988 (visited on 07/11/2019).

[234] Michelle Seref and Onur Seref. "Rhetoric Mining for Fake News: Identifying Moves of Persuasion and Disinformation". In: *AMCIS 2019 Proceedings* (July 2019). URL: https://aisel.aisnet.org/amcis2019/rhetoric_social_media_disinformation/rhetoric_social_media_disinformation/1.

[235] Nicolas Papernot et al. "SoK: Security and Privacy in Machine Learning". In: *IEEE Explore* (2018). DOI: 10.1109/EuroSP.2018.00035.

[236] Amanda Storey. *Our ongoing work to fight misinformation online*. Oct. 2023. URL: https://blog.google/around-the-globe/google-europe/our-ongoing-work-to-fight-misinformation-online/.

[237] URL: https://graphika.com/how-it-works.

[238] OpenAI. *Using machine learning to reduce toxicity online*. URL: https://perspectiveapi.com/.

[239] Anya Schiffrin. *Using AI to combat MIS/disinformation – an evolving story*. Oct. 2023. URL: https://www.techpolicy.press/using-ai-to-combat-mis-disinformation-an-evolving-story/.

[240] Kai Shu and Huan Liu. "Detecting Fake News on Social Media". In: *Synthesis Lectures on Data Mining and Knowledge Discovery* 11.3 (July 2019), pp. 1–129. ISSN: 2151-0067. DOI: 10.2200/S00926ED1V01Y201906DMK018. URL: https://www.morganclaypool.com/doi/abs/10.2200/S00926ED1V01Y201906DMK018 (visited on 07/11/2019).

[241] Ben Nimmo. *Meta's adversarial threat report, fourth quarter 2022*. Feb. 2023. URL: https://about.fb.com/news/2023/02/metas-adversarial-threat-report-q4-2022/.

[242] Nov. 2020. URL: https://ai.meta.com/blog/heres-how-were-using-ai-to-help-detect-misinformation/.

[243] Hibaq Farah. *Diary of a TikTok moderator: "we are the people who sweep up the mess"*. Dec. 2023. URL: https://www.theguardian.com/technology/2023/dec/21/diary-of-a-tiktok-moderator-we-are-the-people-who-sweep-up-the-mess#:~:text=TikTok%20says%20it%20has%20more,in%20more%20than%2070%20languages..

[244] Issie Lapowsky. *Inside the research lab teaching Facebook about its trolls*. Aug. 2018. URL: https://www.wired.com/story/facebook-enlists-dfrlab-track-trolls/.

[245] URL: https://www.reuters.com/technology/musk-owned-xs-content-moderation-shift-complicated-effort-win-back-brands-former-2023-09-07/.

[246] Microsoft Threat Intelligence Microsoft Incident Response. *Dev-0537 criminal actor targeting organizations for data exfiltration and destruction*. Sept. 2023. URL: https://www.microsoft.com/en-us/security/blog/2022/03/22/dev-0537-criminal-actor-targeting-organizations-for-data-exfiltration-and-destruction/.

[247] Inc. Integral Ad Science. *IAS expands AI-driven brand safety and suitability measurement to Meta*. Feb. 2024. URL: https://www.prnewswire.com/news-releases/ias-expands-ai-driven-brand-safety-and-suitability-measurement-to-meta-302052737.html.

[248] URL: https://zefr.com/misinformation.

[249] Issie Lapowsky. *After sending content moderators home, YouTube doubled its video removals*. Aug. 2020. URL: https://www.protocol.com/youtube-content-moderation-covid-19.

[250] Frank Pasquine. *What advertisers need to know about surges in online hate speech*. Feb. 2022. URL: https://doubleverify.com/online-hate-speech-surges-amid-protests/.

[251] The YouTube Team. *Responsible policy enforcement during COVID-19*. Aug. 2020. URL: https://blog.youtube/inside-youtube/responsible-policy-enforcement-during-covid-19/.

[252] Niamh McIntyre, Rosie Bradbury, and Billy Perrigo. *Behind TikTok's boom: A legion of traumatised, $10-a-day content moderators*. Dec. 2023. URL: https://www.thebureauinvestigates.com/stories/2022-10-20/behind-tiktoks-boom-a-legion-of-traumatised-10-a-day-content-moderators.

[253] Stefan Wojcik et al. "Birdwatch: Crowd wisdom and bridging algorithms can inform understanding and reduce the spread of misinformation". In: *arXiv preprint arXiv:2210.15723* (2022).

[254] URL: https://www.facebook.com/business/help/866941431052802?id=2520940424820218.

[255] Common Thread. *Four truths about bots*. Sept. 2021. URL: https://blog.twitter.com/common-thread/en/topics/stories/2021/four-truths-about-bots.

[256] Jesus Reyes and Leon Palafox. "Detection of Fake News based on readability". en. In: (), p. 6.

[257] Michael A. Stefanone, Matthew Vollmer, and Jessica M. Covert. "In News We Trust?: Examining Credibility and Sharing Behaviors of Fake News". In: *Proceedings of the 10th International Conference on Social Media and Society*. SMSociety '19. New York, NY, USA: ACM, 2019, pp. 136–147. ISBN: 978-1-4503-6651-9. DOI: 10.1145/3328529.3328554. URL: http://doi.acm.org/10.1145/3328529.3328554 (visited on 07/11/2019).

[258] Aarash Heydari et al. "YouTube Chatter: Understanding Online Comments Discourse on Misinformative and Political YouTube Videos". en. In: (), p. 32.

[259] Jozef Kapusta, Ľubomír Benko, and Michal Munk. "Fake News Identification Based on Sentiment and Frequency Analysis". en. In: *Innovation in Information Systems and Technologies to Support Learning Research*. Ed. by Mohammed Serrhini, Carla Silva, and Sultan Aljahdali. Learning and Analytics in Intelligent Systems. Cham: Springer International Publishing, 2020, pp. 400–409. ISBN: 978-3-030-36778-7. DOI: 10.1007/978-3-030-36778-7_44.

[260] Seyedmehdi Hosseinimotlagh and Evangelos E Papalexakis. “Unsupervised Content-Based Identification of Fake News Articles with Tensor Decomposition Ensembles”. en. In: (), p. 8.

[261] Kashyap Popat. “"Credibility Analysis of Textual Claims with Explainable Evidence"”. en. In: (), p. 134.

[262] Leonardo Nizzoli et al. “Coordinated behavior on social media in 2019 UK general election”. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 15. 2021, pp. 443–454.

[263] Onur Varol et al. “Early detection of promoted campaigns on social media”. In: *EPJ data science* 6 (2017), pp. 1–19.

[264] Rohit Kumar Kaliyar et al. “FNDNet- A Deep Convolutional Neural Network for Fake News Detection”. en. In: *Cognitive Systems Research* (Jan. 2020). ISSN: 1389-0417. DOI: 10.1016/j.cogsys.2019.12.005. URL: http://www.sciencedirect.com/science/article/pii/S1389041720300085 (visited on 01/25/2020).

[265] Ning Xin Nyow and Hui Na Chua. “Detecting Fake News with Tweets’ Properties”. In: *2019 IEEE Conference on Application, Information and Network Security (AINS)*. Nov. 2019, pp. 24–29. DOI: 10.1109/AINS47559.2019.8968706.

[266] Francesco Pierri, Carlo Piccardi, and Stefano Ceri. “Topology comparison of Twitter diffusion networks effectively reveals misleading information”. en. In: *Scientific Reports* 10.1 (Jan. 2020), pp. 1–9. ISSN: 2045-2322. DOI: 10.1038/s41598-020-58166-5. URL: https://www.nature.com/articles/s41598-020-58166-5 (visited on 02/04/2020).

[267] Harita Reddy et al. “Text-mining-based Fake News Detection Using Ensemble Methods”. en. In: *International Journal of Automation and Computing* (Feb. 2020). ISSN: 1751-8520. DOI: 10.1007/s11633-019-1216-5. URL: https://doi.org/10.1007/s11633-019-1216-5 (visited on 02/22/2020).

[268] Juan Cao et al. “Exploring the Role of Visual Content in Fake News Detection”. en. In: *arXiv:2003.05096 [cs]* (Mar. 2020). DOI: 10.1007/978-3-030-42699-6. URL: http://arxiv.org/abs/2003.05096 (visited on 03/17/2020).

[269] Kai Shu et al. “Leveraging Multi-Source Weak Social Supervision for Early Detection of Fake News”. en. In: *arXiv:2004.01732 [cs, stat]* (Apr. 2020). URL: http://arxiv.org/abs/2004.01732 (visited on 04/12/2020).

[270] Harika Kudarvalli and Jinan Fiaidhi. “Detecting Fake News using Machine Learning Algorithms”. en. In: (Apr. 2020). ISSN: doi:10.36227/techrxiv.12089133.v1. DOI: 10.36227/techrxiv.12089133.v1. URL: https://www.techrxiv.org/articles/Detecting_Fake_News_using_Machine_Learning_Algorithms/12089133 (visited on 04/12/2020).

[271] Anmol Uppal, Vipul Sachdeva, and Seema Sharma. “Fake news detection using discourse segment structure analysis”. In: *2020 10th International Conference on Cloud Computing, Data Science Engineering (Confluence)*. Jan. 2020, pp. 751–756. DOI: 10.1109/Confluence47617.2020.9058106.

[272] Marialaura Previti et al. “Fake News Detection Using Time Series and User Features Classification”. en. In: *Applications of Evolutionary Computation*. Ed. by Pedro A. Castillo, Juan Luis Jiménez Laredo, and Francisco Fernández de Vega. Vol. 12104. Cham: Springer International Publishing, 2020, pp. 339–353. ISBN: 978-3-030-43721-3 978-3-030-43722-0. DOI: 10.1007/978-3-030-43722-0_22. URL: http://link.springer.com/10.1007/978-3-030-43722-0_22 (visited on 04/15/2020).

[273] Yue Zhou, Yan Zhang, and JingTao Yao. “Satirical News Detection with Semantic Feature Extraction and Game-theoretic Rough Sets”. en. In: *arXiv:2004.03788 [cs]* (Apr. 2020). URL: http://arxiv.org/abs/2004.03788 (visited on 04/15/2020).

[274] Pepa Atanasova et al. “Generating Fact Checking Explanations”. en. In: *arXiv:2004.05773 [cs]* (Apr. 2020). URL: http://arxiv.org/abs/2004.05773 (visited on 04/17/2020).

[275] Joma George, Shintu Mariam Skariah, and T Aleena Xavier. “Role of Contextual Features in Fake News Detection: A Review”. In: *2020 International Conference on Innovative Trends in Information Technology (ICITIIT)*. Feb. 2020, pp. 1–6. DOI: 10.1109/ICITIIT49094.2020.9071524.

[276] Mitchell L Gordon et al. “The Disagreement Deconvolution: Bringing Machine Learning Performance Metrics In Line With Reality”. en. In: (2021), p. 14.

[277] L. Kurasinski and R.-C. Mihailescu. “Towards Machine Learning Explainability in Text Classification for Fake News Detection”. In: *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*. Dec. 2020, pp. 775–781. DOI: 10.1109/ICMLA51294.2020.00127. URL: https://ieeexplore.ieee.org/abstract/document/9356374.

[278] Paula Carvalho et al. *Assessing News Credibility: Misinformation Content Indicators*. en. preprint. In Review, Mar. 2021. DOI: 10.21203/rs.3.rs-173067/v1. URL: https://www.researchsquare.com/article/rs-173067/v1 (visited on 03/08/2021).

[279] Manolis Chalkiadakis et al. “The Rise and Fall of Fake News sites: A Traffic Analysis”. en. In: *arXiv:2103.09258 [cs]* (Mar. 2021). URL: http://arxiv.org/abs/2103.09258 (visited on 03/23/2021).

[280] Benjamin Krämer. “Stop studying “fake news” (we can still fight against disinformation in the media)”. In: *Studies in Communication and Media* 10.1 (2021), pp. 6–30. ISSN: 2192-4007. DOI: 10.5771/2192-4007-2021-1-6. URL: https://www.nomos-elibrary.de/index.php?doi=10.5771/2192-4007-2021-1-6 (visited on 04/06/2021).

[281] Nayeon Lee et al. "On Unifying Misinformation Detection". en. In: *arXiv:2104.05243 [cs]* (Apr. 2021). URL: http://arxiv.org/abs/2104.05243 (visited on 04/16/2021).

[282] Kellin Pelrine, Jacob Danovitch, and Reihaneh Rabbany. "The Surprising Performance of Simple Baselines for Misinformation Detection". en. In: *arXiv:2104.06952 [cs]* (Apr. 2021). URL: http://arxiv.org/abs/2104.06952 (visited on 04/21/2021).

[283] Bahruz Jabiyev et al. "FADE: Detecting Fake News Articles on the Web". en. In: (2021), p. 10.

[284] Jakub Simko et al. "Towards Continuous Automatic Audits of Social Media Adaptive Behavior and its Role in Misinformation Spreading". en. In: *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*. Utrecht Netherlands: ACM, June 2021, pp. 411–414. ISBN: 978-1-4503-8367-7. DOI: 10.1145/3450614.3463353. URL: https://dl.acm.org/doi/10.1145/3450614.3463353 (visited on 07/15/2021).

[285] Yunkang Yang, Trevor Davis, and Matthew Hindman. "Visual Misinformation on Facebook". en. In: (), p. 8.

[286] Kevin Roitero et al. "The COVID-19 Infodemic: Can the Crowd Judge Recent Misinformation Objectively?" en. In: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (Oct. 2020), pp. 1305–1314. DOI: 10.1145/3340531.3412048. URL: http://arxiv.org/abs/2008.05701 (visited on 08/03/2021).

[287] Antino Kim, Patricia Moravec, and Alan Dennis. "Behind the Stars: The Effects of News Source Ratings on Fake News in Social Media". In: *SSRN Electronic Journal* (Jan. 2017). DOI: 10.2139/ssrn.3090355.

[288] Paul Resnick et al. "Informed Crowds Can Effectively Identify Misinformation". en. In: *arXiv:2108.07898 [cs]* (Aug. 2021). URL: http://arxiv.org/abs/2108.07898 (visited on 08/25/2021).

[289] Prerna Juneja and Tanushree Mitra. "Auditing E-Commerce Platforms for Algorithmically Curated Vaccine Misinformation". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 186. New York, NY, USA: Association for Computing Machinery, May 2021, pp. 1–27. ISBN: 978-1-4503-8096-6. URL: https://doi.org/10.1145/3411764.3445250 (visited on 09/17/2021).

[290] Maria Janicka, Maria Pszona, and Aleksander Wawer. "Cross-Domain Failures of Fake News Detection". en. In: *Computación y Sistemas* 23.3 (Oct. 2019). ISSN: 2007-9737, 1405-5546. DOI: 10.13053/cys-23-3-3281. URL: https://www.cys.cic.ipn.mx/ojs/index.php/CyS/article/view/3281 (visited on 10/29/2021).

[291] Ran Wang et al. "RumorLens: Interactive Analysis and Validation of Suspected Rumors on Social Media". In: *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI EA '22. New York, NY, USA: Association for Computing Machinery, Apr. 2022, pp. 1–7. ISBN: 978-1-4503-9156-6. DOI: 10.1145/3491101.3519712. URL: https://doi.org/10.1145/3491101.3519712 (visited on 11/06/2022).

[292] James Thorne and Andreas Vlachos. "Automated Fact Checking: Task Formulations, Methods and Future Directions". en. In: (), p. 14.

[293] Li Zeng, Kate Starbird, and Emma Spiro. "#Unconfirmed: Classifying Rumor Stance in Crisis-Related Social Media Messages". en. In: *Proceedings of the International AAAI Conference on Web and Social Media* 10.1 (2016). Number: 1, pp. 747–750. ISSN: 2334-0770. DOI: 10.1609/icwsm.v10i1.14788. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14788 (visited on 01/05/2023).

[294] Gregory Eady et al. "Exposure to the Russian Internet Research Agency foreign influence campaign on Twitter in the 2016 US election and its relationship to attitudes and voting behavior". en. In: *Nature Communications* 14.1 (Jan. 2023). Number: 1 Publisher: Nature Publishing Group, p. 62. ISSN: 2041-1723. DOI: 10.1038/s41467-022-35576-9. URL: https://www.nature.com/articles/s41467-022-35576-9 (visited on 01/09/2023).

[295] Abdullah Talha Kabakuş and Mehmet Şimşek. "An Analysis of the Characteristics of Verified Twitter Users". en. In: *Sakarya University Journal of Computer and Information Sciences* 2.3 (Dec. 2019), pp. 180–186. ISSN: 2636-8129. DOI: 10.35377/saucis.02.03.649708. URL: https://dergipark.org.tr/en/doi/10.35377/saucis.02.03.649708 (visited on 01/12/2023).

[296] Alison Hearn. "Verified: Self-presentation, identity management, and selfhood in the age of big data". In: *Popular Communication* 15.2 (Apr. 2017). Publisher: Routledge _eprint: https://doi.org/10.1080/15405702.2016.1269909, pp. 62–77. ISSN: 1540-5702. DOI: 10.1080/15405702.2016.1269909. URL: https://doi.org/10.1080/15405702.2016.1269909 (visited on 01/12/2023).

[297] *Twitter Verification requirements - how to get the blue check*. en. URL: https://help.twitter.com/en/managing-your-account/about-twitter-verified-accounts (visited on 01/12/2023).

[298] *Prolific · Quickly find research participants you can trust*. en. URL: https://www.prolific.co/ (visited on 01/12/2023).

[299] Ahmer Arif et al. "How Information Snowballs: Exploring the Role of Exposure in Online Rumor Propagation". In: *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. CSCW '16. New York, NY, USA: Association for Computing Machinery, Feb. 2016, pp. 466–477. ISBN: 978-1-4503-3592-8. DOI: 10.1145/2818048.2819964. URL: https://doi.org/10.1145/2818048.2819964 (visited on 01/12/2023).

[300] Sander van der Linden. “Misinformation: susceptibility, spread, and interventions to immunize the public”. en. In: *Nature Medicine* 28.3 (Mar. 2022). Number: 3 Publisher: Nature Publishing Group, pp. 460–467. ISSN: 1546-170X. DOI: 10.1038/s41591-022-01713-6. URL: https://www.nature.com/articles/s41591-022-01713-6 (visited on 01/12/2023).

[301] *RumorLens: Interactive Analysis and Validation of Suspected Rumors on Social Media | Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*. URL: https://dl.acm.org/doi/10.1145/3491101.3519712 (visited on 03/16/2023).

[302] Ala Mughaid et al. “An intelligent cybersecurity system for detecting fake news in social media websites”. en. In: *Soft Computing* 26.12 (June 2022), pp. 5577–5591. ISSN: 1433-7479. DOI: 10.1007/s00500-022-07080-1. URL: https://doi.org/10.1007/s00500-022-07080-1 (visited on 03/16/2023).

[303] Firoj Alam et al. *A Survey on Multimodal Disinformation Detection*. arXiv:2103.12541 [cs]. Sept. 2022. DOI: 10.48550/arXiv.2103.12541. URL: http://arxiv.org/abs/2103.12541 (visited on 03/16/2023).

[304] Fernando Miró-Llinares and Jesús C. Aguerri. “Misinformation about fake news: A systematic critical review of empirical studies on the phenomenon and its status as a ‘threat’”. en. In: *European Journal of Criminology* 20.1 (Jan. 2023). Publisher: SAGE Publications, pp. 356–374. ISSN: 1477-3708. DOI: 10.1177/1477370821994059. URL: https://doi.org/10.1177/1477370821994059 (visited on 03/16/2023).

[305] Paul Resnick et al. *Informed Crowds Can Effectively Identify Misinformation*. arXiv:2108.07898 [cs]. Feb. 2022. DOI: 10.48550/arXiv.2108.07898. URL: http://arxiv.org/abs/2108.07898 (visited on 03/16/2023).

[306] *Assessing News Credibility: Misinformation Content Indicators*. en. Mar. 2021. DOI: 10.21203/rs.3.rs-173067/v1. URL: https://www.researchsquare.com (visited on 03/16/2023).

[307] Kevin Roitero et al. “Can the Crowd Judge Truthfulness? A Longitudinal Study on Recent Misinformation about COVID-19”. In: *Personal and Ubiquitous Computing* 27.1 (Feb. 2023). arXiv:2107.11755 [cs], pp. 59–89. ISSN: 1617-4909, 1617-4917. DOI: 10.1007/s00779-021-01604-6. URL: http://arxiv.org/abs/2107.11755 (visited on 03/16/2023).

[308] Feng Yu et al. “A Convolutional Approach for Misinformation Identification”. In: (2017), pp. 3901–3907. URL: https://www.ijcai.org/proceedings/2017/545 (visited on 03/16/2023).

[309] Arjun Roy et al. “A Deep Ensemble Framework for Fake News Detection and Multi-Class Classification of Short Political Statements”. In: *Proceedings of the 16th International Conference on Natural Language Processing*. International Institute of Information Technology, Hyderabad, India: NLP Association of India, Dec. 2019, pp. 9–17. URL: https://aclanthology.org/2019.icon-1.2 (visited on 03/16/2023).

[310] Bashar Al Asaad and Madalina Erascu. “A Tool for Fake News Detection”. In: *2018 20th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*. Sept. 2018, pp. 379–386. DOI: 10.1109/SYNASC.2018.00064.

[311] A. Conrad Nied et al. “Alternative Narratives of Crisis Events: Communities and Social Botnets Engaged on Social Media”. en. In: *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. Portland Oregon USA: ACM, Feb. 2017, pp. 263–266. ISBN: 978-1-4503-4688-7. DOI: 10.1145/3022198.3026307. URL: https://dl.acm.org/doi/10.1145/3022198.3026307 (visited on 03/16/2023).

[312] Xichen Zhang and Ali A. Ghorbani. “An overview of online fake news: Characterization, detection, and discussion”. en. In: *Information Processing & Management* 57.2 (Mar. 2020), p. 102025. ISSN: 0306-4573. DOI: 10.1016/j.ipm.2019.03.004. URL: https://www.sciencedirect.com/science/article/pii/S0306457318306794 (visited on 03/16/2023).

[313] Jake Lever. “Classification evaluation: It is important to understand both what a classification metric expresses and what it hides”. In: *Nature methods* 13.8 (2016), pp. 603–605.

[314] Bruce R Lindsay. “Social media for emergencies and disasters: Overview and policy considerations”. In: *Congressional Research Service* (2016).

[315] Limeng Cui and Dongwon Lee. “Coaid: Covid-19 healthcare misinformation dataset”. In: *arXiv preprint arXiv:2006.00885* (2020).

[316] Jayawant N Mandrekar. “Receiver operating characteristic curve in diagnostic test assessment”. In: *Journal of thoracic oncology* 5.9 (2010), pp. 1315–1316.

[317] Vivek Singh et al. “Automated Fake News Detection Using Linguistic Analy- sis and Machine Learning”. en. In: ().

[318] Sawinder Kaur, Parteek Kumar, and Ponnurangam Kumaraguru. “Automating fake news detection system using multi-level voting model”. en. In: *Soft Computing* 24.12 (June 2020), pp. 9049–9069. ISSN: 1433-7479. DOI: 10.1007/s00500-019-04436-y. URL: https://doi.org/10.1007/s00500-019-04436-y (visited on 03/16/2023).

[319] Luis Vargas, Patrick Emami, and Patrick Traynor. “On the detection of disinformation campaign activity with network analysis”. In: *Proceedings of the 2020 ACM SIGSAC Conference on Cloud Computing Security Workshop*. 2020, pp. 133–146.

[320] Karishma Sharma et al. “Identifying coordinated accounts on social media through hidden influence and group behaviours”. In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 2021, pp. 1441–1451.

[321] Tanushree Mitra and Eric Gilbert. “CREDBANK: A Large-Scale Social Media Corpus With Associated Credibility Annotations”. en. In: *Proceedings of the International AAAI Conference on Web and Social Media* 9.1 (2015). Number: 1, pp. 258–267. ISSN: 2334-0770. DOI: 10.1609/icwsm.v9i1.14625. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14625 (visited on 03/16/2023).

[322] Christopher G. Harris. “Detecting Deceptive Opinion Spam Using Human Computation”. In: July 2012. URL: https://www.semanticscholar.org/paper/Detecting-Deceptive-Opinion-Spam-Using-Human-Harris/471dd1fd942e3fc83e0201f7e51415c4310cdfa1 (visited on 03/16/2023).

[323] Monther Aldwairi and Ali Alwahedi. “Detecting Fake News in Social Media Networks”. en. In: *Procedia Computer Science*. The 9th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN-2018) / The 8th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH-2018) / Affiliated Workshops 141 (Jan. 2018), pp. 215–222. ISSN: 1877-0509. DOI: 10.1016/j.procs.2018.10.171. URL: https://www.sciencedirect.com/science/article/pii/S1877050918318210 (visited on 03/16/2023).

[324] Supanya Aphiwongsophon and Prabhas Chongstitvatana. “Detecting Fake News with Machine Learning Method”. In: *2018 15th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*. July 2018, pp. 528–531. DOI: 10.1109/ECTICon.2018.8620051.

[325] Boris A. Galitsky. “Detecting Rumor and Disinformation by Web Mining”. In: Mar. 2015. URL: https://www.semanticscholar.org/paper/Detecting-Rumor-and-Disinformation-by-Web-Mining-Galitsky/449790743c7de01916182e96369ebaedb5ebe1a6 (visited on 03/16/2023).

[326] Arkaitz Zubiaga et al. “Detection and Resolution of Rumours in Social Media: A Survey”. In: *ACM Computing Surveys* 51.2 (Feb. 2018), 32:1–32:36. ISSN: 0360-0300. DOI: 10.1145/3161603. URL: https://dl.acm.org/doi/10.1145/3161603 (visited on 03/16/2023).

[327] Hadeer Ahmed, Issa Traore, and Sherif Saad. “Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques”. en. In: *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments*. Ed. by Issa Traore, Isaac Woungang, and Ahmed Awad. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2017, pp. 127–138. ISBN: 978-3-319-69155-8. DOI: 10.1007/978-3-319-69155-8_9.

[328] Suman Kalyan Maity et al. “Detection of Sockpuppets in Social Media”. In: *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. CSCW ’17 Companion. New York, NY, USA: Association for Computing Machinery, Feb. 2017, pp. 243–246. ISBN: 978-1-4503-4688-7. DOI: 10.1145/3022198.3026360. URL: https://doi.org/10.1145/3022198.3026360 (visited on 03/16/2023).

[329] Peter Hernon. “Disinformation and misinformation through the internet: Findings of an exploratory study”. en. In: *Government Information Quarterly* 12.2 (Jan. 1995), pp. 133–139. ISSN: 0740-624X. DOI: 10.1016/0740-624X(95)90052-7. URL: https://www.sciencedirect.com/science/article/pii/0740624X95900527 (visited on 03/16/2023).

[330] Lennart van de Guchte et al. “Near Real-Time Detection of Misinformation on Online Social Networks”. en. In: *Disinformation in Open Online Media*. Ed. by Max van Duijn et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020, pp. 246–260. ISBN: 978-3-030-61841-4. DOI: 10.1007/978-3-030-61841-4_17.

[331] Manqing Dong et al. “DUAL: A Deep Unified Attention Model with Latent Relation Representations for Fake News Detection: 19th International Conference, Dubai, United Arab Emirates, November 12-15, 2018, Proceedings, Part I”. In: Nov. 2018, pp. 199–209. ISBN: 978-3-030-02921-0. DOI: 10.1007/978-3-030-02922-7_14.

[332] Maria Konte, Nick Feamster, and Jaeyeon Jung. “Dynamics of Online Scam Hosting Infrastructure”. en. In: *Passive and Active Network Measurement*. Ed. by Sue B. Moon, Renata Teixeira, and Steve Uhlig. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2009, pp. 219–228. ISBN: 978-3-642-00975-4. DOI: 10.1007/978-3-642-00975-4_22.

[333] Kate Starbird. “Examining the Alternative Media Ecosystem Through the Production of Alternative Narratives of Mass Shooting Events on Twitter”. en. In: *Proceedings of the International AAAI Conference on Web and Social Media* 11.1 (May 2017). Number: 1, pp. 230–239. ISSN: 2334-0770. DOI: 10.1609/icwsm.v11i1.14878. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14878 (visited on 03/16/2023).

[334] Kai Shu, Suhang Wang, and Huan Liu. “Exploiting Tri-Relationship for Fake News Detection”. In: (Dec. 2017).

[335] Todor Mihaylov et al. *Exposing Paid Opinion Manipulation Trolls*. arXiv:2109.13726 [cs]. Sept. 2021. DOI: 10.48550/arXiv.2109.13726. URL: http://arxiv.org/abs/2109.13726 (visited on 03/16/2023).

[336] Chandra Mouli Madhav Kotteti et al. “Fake news detection enhancement with data imputation”. In: *Computer Information Systems Faculty Publications* (Oct. 2018). URL: https://digitalcommons.pvamu.edu/computer-information-facpubs/46.

[337] Sebastian Tschiatschek et al. “Fake News Detection in Social Networks via Crowd Signals”. en. In: *Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW ’18*. Lyon, France: ACM Press, 2018, pp. 517–524. ISBN: 978-1-4503-5640-4. DOI: 10.1145/3184558.3188722. URL: http://dl.acm.org/citation.cfm?doid=3184558.3188722 (visited on 03/16/2023).

[338] Ankit Kesarwani, Sudakar Singh Chauhan, and Anil Ramachandran Nair. "Fake News Detection on Social Media using K-Nearest Neighbor Classifier". In: *2020 International Conference on Advances in Computing and Communication Engineering (ICACCE)*. June 2020, pp. 1–4. DOI: 10.1109/ICACCE49060.2020.9154997.

[339] Yunfei Long et al. "Fake News Detection Through Multi-Perspective Speaker Profiles". In: *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Taipei, Taiwan: Asian Federation of Natural Language Processing, Nov. 2017, pp. 252–256. URL: https://aclanthology.org/I17-2043 (visited on 03/16/2023).

[340] Sachin Kumar et al. "Fake news detection using deep learning models: A novel approach". en. In: *Transactions on Emerging Telecommunications Technologies* 31.2 (2020), e3767. ISSN: 2161-3915. DOI: 10.1002/ett.3767. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/ett.3767 (visited on 03/16/2023).

[341] Iftikhar Ahmad et al. "Fake News Detection Using Machine Learning Ensemble Methods". en. In: *Complexity* 2020 (Oct. 2020). Publisher: Hindawi, e8885861. ISSN: 1076-2787. DOI: 10.1155/2020/8885861. URL: https://www.hindawi.com/journals/complexity/2020/8885861/ (visited on 03/16/2023).

[342] Mykhailo Granik and Volodymyr Mesyura. "Fake news detection using naive Bayes classifier". In: *2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)*. May 2017, pp. 900–903. DOI: 10.1109/UKRCON.2017.8100379.

[343] Aswini Thota et al. "Fake News Detection: A Deep Learning Approach". en. In: 1.3 (2018).

[344] Mehrdad Farajtabar et al. "Fake News Mitigation via Point Process Based Intervention". en. In: *Proceedings of the 34th International Conference on Machine Learning*. ISSN: 2640-3498. PMLR, July 2017, pp. 1097–1106. URL: https://proceedings.mlr.press/v70/farajtabar17a.html (visited on 03/16/2023).

[345] Amitabha Dey et al. "Fake News Pattern Recognition using Linguistic Analysis". In: *2018 Joint 7th International Conference on Informatics, Electronics & Vision (ICIEV) and 2018 2nd International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*. June 2018, pp. 305–309. DOI: 10.1109/ICIEV.2018.8641018.

[346] Lianwei Wu et al. "False Information Detection on Social Media via a Hybrid Deep Model". en. In: *Social Informatics*. Ed. by Steffen Staab, Olessia Koltsova, and Dmitry I. Ignatov. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018, pp. 323–333. ISBN: 978-3-030-01159-8. DOI: 10.1007/978-3-030-01159-8_31.

[347] Srijan Kumar and Neil Shah. "False Information on Web and Social Media: A Survey". In: (Apr. 2018).

[348] Ke Wu, Song Yang, and Kenny Q. Zhu. "False rumors detection on Sina Weibo by propagation structures". In: *2015 IEEE 31st International Conference on Data Engineering*. ISSN: 2375-026X. Apr. 2015, pp. 651–662. DOI: 10.1109/ICDE.2015.7113322.

[349] *FANG | Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. URL: https://dl.acm.org/doi/abs/10.1145/3340531.3412046 (visited on 03/16/2023).

[350] Niraj Sitaula et al. "Credibility-Based Fake News Detection". en. In: *Disinformation, Misinformation, and Fake News in Social Media: Emerging Research Challenges and Opportunities*. Ed. by Kai Shu et al. Lecture Notes in Social Networks. Cham: Springer International Publishing, 2020, pp. 163–182. ISBN: 978-3-030-42699-6. DOI: 10.1007/978-3-030-42699-6_9. URL: https://doi.org/10.1007/978-3-030-42699-6_9 (visited on 03/16/2023).

[351] Huiling Zhang et al. "Fight Under Uncertainty: Restraining Misinformation and Pushing out the Truth". In: *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. ISSN: 2473-991X. Aug. 2018, pp. 266–273. DOI: 10.1109/ASONAM.2018.8508402.

[352] Todor Mihaylov, Georgi Georgiev, and Preslav Nakov. "Finding Opinion Manipulation Trolls in News Community Forums". In: *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*. Beijing, China: Association for Computational Linguistics, July 2015, pp. 310–314. DOI: 10.18653/v1/K15-1032. URL: https://aclanthology.org/K15-1032 (visited on 03/16/2023).

[353] Yang Liu and Yi-Fang Brook Wu. "FNED: A Deep Network for Fake News Early Detection on Social Media". In: *ACM Transactions on Information Systems* 38.3 (May 2020), 25:1–25:33. ISSN: 1046-8188. DOI: 10.1145/3386253. URL: https://doi.org/10.1145/3386253 (visited on 03/16/2023).

[354] Nitesh Chawla and Wei Wang, eds. *Proceedings of the 2017 SIAM International Conference on Data Mining*. en. Philadelphia, PA: Society for Industrial and Applied Mathematics, June 2017. ISBN: 978-1-61197-497-3. DOI: 10.1137/1.9781611974973. URL: https://epubs.siam.org/doi/book/10.1137/1.9781611974973 (visited on 03/16/2023).

[355] Liang Wu et al. "Gleaning Wisdom from the Past: Early Detection of Emerging Rumors in Social Media". In: June 2017, pp. 99–107. ISBN: 978-1-61197-497-3. DOI: 10.1137/1.9781611974973.12.

[356] Bo Ni et al. *Improving Generalizability of Fake News Detection Methods using Propensity Score Matching*. arXiv:2002.00838 [cs, stat]. Jan. 2020. DOI: 10.48550/arXiv.2002.00838. URL: http://arxiv.org/abs/2002.00838 (visited on 03/16/2023).

[357] Qiang Liu et al. “Mining Significant Microblogs for Misinformation Identification: An Attention-Based Approach”. In: *ACM Transactions on Intelligent Systems and Technology* 9.5 (Apr. 2018), 50:1–50:20. ISSN: 2157-6904. DOI: 10.1145/3173458. URL: https://doi.org/10.1145/3173458 (visited on 03/16/2023).

[358] Steven Lloyd Wilson and Charles Wiysonge. “Social media and vaccine hesitancy”. In: *BMJ global health* 5.10 (2020).

[359] Hamid Karimi et al. “Multi-Source Multi-Class Fake News Detection”. In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 1546–1557. URL: https://aclanthology.org/C18-1131 (visited on 03/16/2023).

[360] Zhiwei Jin et al. “News Credibility Evaluation on Microblog with a Hierarchical Propagation Model”. In: *2014 IEEE International Conference on Data Mining*. ISSN: 2374-8486. Dec. 2014, pp. 230–239. DOI: 10.1109/ICDM.2014.91.

[361] Dariusz Jemielniak and Yaroslav Krempovych. “An analysis of AstraZeneca COVID-19 vaccine misinformation and fear mongering on Twitter”. In: *Public Health* 200 (2021), pp. 4–6.

[362] Zhiwei Jin et al. “News Verification by Exploiting Conflicting Social Viewpoints in Microblogs”. en. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 30.1 (Mar. 2016). Number: 1. ISSN: 2374-3468. DOI: 10.1609/aaai.v30i1.10382. URL: https://ojs.aaai.org/index.php/AAAI/article/view/10382 (visited on 03/16/2023).

[363] Thomas Magelinski, Lynnette Ng, and Kathleen Carley. “A synchronized action framework for detection of coordination on social media”. In: *Journal of Online Trust and Safety* 1.2 (2022).

[364] Sarthak Jindal et al. “NewsBag: A Multimodal Benchmark Dataset for Fake News Detection”. en. In: ().

[365] Zhiwei Jin et al. “Novel Visual and Statistical Image Features for Microblogs News Verification”. In: *IEEE Transactions on Multimedia* 19.3 (Mar. 2017). Conference Name: IEEE Transactions on Multimedia, pp. 598–608. ISSN: 1941-0077. DOI: 10.1109/TMM.2016.2617078.

[366] Michela Del Vicario et al. “Polarization and Fake News: Early Warning of Potential Misinformation Targets”. In: *ACM Transactions on the Web* 13.2 (Mar. 2019), 10:1–10:22. ISSN: 1559-1131. DOI: 10.1145/3316809. URL: https://doi.org/10.1145/3316809 (visited on 03/16/2023).

[367] Ceren Budak, Divyakant Agrawal, and Amr El Abbadi. “Limiting the spread of misinformation in social networks”. In: *Proceedings of the 20th international conference on World wide web*. WWW ’11. New York, NY, USA: Association for Computing Machinery, Mar. 2011, pp. 665–674. ISBN: 978-1-4503-0632-4. DOI: 10.1145/1963405.1963499. URL: https://doi.org/10.1145/1963405.1963499 (visited on 03/16/2023).

[368] *Web-based statistical fact checking of textual documents | Proceedings of the 2nd international workshop on Search and mining user-generated contents*. URL: https://dl.acm.org/doi/10.1145/1871985.1872002 (visited on 03/16/2023).

[369] Sejeong Kwon et al. “Prominent Features of Rumor Propagation in Online Social Media”. In: *2013 IEEE 13th International Conference on Data Mining*. ISSN: 2374-8486. Dec. 2013, pp. 1103–1108. DOI: 10.1109/ICDM.2013.61.

[370] Chris Grier et al. “@ spam: the underground on 140 characters or less”. In: *Proceedings of the 17th ACM conference on Computer and communications security*. 2010, pp. 27–37.

[371] Shirin Nilizadeh et al. “Poised: Spotting twitter spam off the beaten paths”. In: *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. 2017, pp. 1159–1174.

[372] Taeri Kim et al. “Phishing URL Detection: A Network-based Approach Robust to Evasion”. In: *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*. 2022, pp. 1769–1782.

[373] Hongyu Gao et al. “Detecting and characterizing social spam campaigns”. In: *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*. 2010, pp. 35–47.

[374] Kurt Thomas et al. “Design and evaluation of a real-time url spam filtering service”. In: *2011 IEEE symposium on security and privacy*. IEEE. 2011, pp. 447–462.

[375] Colin Whittaker, Brian Ryner, and Marria Nazif. “Large-scale automatic classification of phishing pages”. In: (2010).

[376] Mihai Christodorescu et al. “Semantics-aware malware detection”. In: *2005 IEEE symposium on security and privacy (S&P’05)*. IEEE. 2005, pp. 32–46.

[377] Fabio Giglietto et al. “It takes a village to manipulate the media: coordinated link sharing behavior during 2018 and 2019 Italian elections”. In: *Information, Communication & Society* 23.6 (2020), pp. 867–891.

[378] Qiang Zhang et al. “Reply-Aided Detection of Misinformation via Bayesian Deep Learning”. In: *The World Wide Web Conference*. WWW ’19. New York, NY, USA: Association for Computing Machinery, May 2019, pp. 2333–2343. ISBN: 978-1-4503-6674-8. DOI: 10.1145/3308558.3313718. URL: https://doi.org/10.1145/3308558.3313718 (visited on 03/16/2023).

[379] Sejeong Kwon, Meeyoung Cha, and Kyomin Jung. “Rumor Detection over Varying Time Windows”. en. In: *PLOS ONE* 12.1 (Jan. 2017). Publisher: Public Library of Science, e0168344. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0168344. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0168344 (visited on 03/16/2023).

[380] Han Guo et al. “Rumor Detection with Hierarchical Social Attention Network”. In: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. CIKM ’18. New York, NY, USA: Association for Computing Machinery, Oct. 2018, pp. 943–951. ISBN: 978-1-4503-6014-2. DOI: 10.1145/3269206.3271709. URL: https://doi.org/10.1145/3269206.3271709 (visited on 03/16/2023).

[381] Xiang Lin et al. “Rumor Detection with Hierarchical Recurrent Convolutional Neural Network”. en. In: *Natural Language Processing and Chinese Computing*. Ed. by Jie Tang et al. Vol. 11839. Series Title: Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019, pp. 338–348. ISBN: 978-3-030-32235-9 978-3-030-32236-6. DOI: 10.1007/978-3-030-32236-6_30. URL: http://link.springer.com/10.1007/978-3-030-32236-6_30 (visited on 03/16/2023).

[382] Li Zeng, Kate Starbird, and Emma S. Spiro. “Rumors at the Speed of Light? Modeling the Rate of Rumor Transmission During Crisis”. In: *2016 49th Hawaii International Conference on System Sciences (HICSS)*. ISSN: 1530-1605. Jan. 2016, pp. 1969–1978. DOI: 10.1109/HICSS.2016.248.

[383] Kai-Cheng Yang et al. “Scalable and generalizable social bot detection through data selection”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 34. 01. 2020, pp. 1096–1103.

[384] Laijun Zhao et al. “SIR rumor spreading model in the new media age”. en. In: *Physica A: Statistical Mechanics and its Applications* 392.4 (Feb. 2013), pp. 995–1003. ISSN: 0378-4371. DOI: 10.1016/j.physa.2012.09.030. URL: https://www.sciencedirect.com/science/article/pii/S037843711200934X (visited on 03/16/2023).

[385] Shivangi Singhal et al. “SpotFake: A Multi-modal Framework for Fake News Detection”. In: *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*. Sept. 2019, pp. 39–47. DOI: 10.1109/BigMM.2019.00-44.

[386] Arjun Mukherjee, Bing Liu, and Natalie Glance. “Spotting fake reviewer groups in consumer reviews”. en. In: *Proceedings of the 21st international conference on World Wide Web*. Lyon France: ACM, Apr. 2012, pp. 191–200. ISBN: 978-1-4503-1229-5. DOI: 10.1145/2187836.2187863. URL: https://dl.acm.org/doi/10.1145/2187836.2187863 (visited on 03/16/2023).

[387] Daron Acemoglu, Asuman Ozdaglar, and Ali ParandehGheibi. “Spread of (mis)information in social networks”. en. In: *Games and Economic Behavior* 70.2 (Nov. 2010), pp. 194–227. ISSN: 0899-8256. DOI: 10.1016/j.geb.2010.01.005. URL: https://www.sciencedirect.com/science/article/pii/S0899825610000217 (visited on 03/16/2023).

[388] Inggrid Yanuar Risca Pratiwi, Rosa Andrie Asmara, and Faisal Rahutomo. “Study of hoax news detection using naïve bayes classifier in Indonesian language”. In: *2017 11th International Conference on Information & Communication Technology and System (ICTS)*. ISSN: 2338-185X. Oct. 2017, pp. 73–78. DOI: 10.1109/ICTS.2017.8265649.

[389] Kai Shu, H. Russell Bernard, and Huan Liu. *Studying Fake News via Network Analysis: Detection and Mitigation*. arXiv:1804.10233 [cs]. Apr. 2018. DOI: 10.48550/arXiv.1804.10233. URL: http://arxiv.org/abs/1804.10233 (visited on 03/16/2023).

[390] Julio C. S. Reis et al. “Supervised Learning for Fake News Detection”. en. In: *IEEE Intelligent Systems* 34.2 (Mar. 2019), pp. 76–81. ISSN: 1541-1672, 1941-1294. DOI: 10.1109/MIS.2019.2899143. URL: https://ieeexplore.ieee.org/document/8709925/ (visited on 03/16/2023).

[391] Kai Shu et al. “The role of user profiles for fake news detection”. In: *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ASONAM ’19. New York, NY, USA: Association for Computing Machinery, Jan. 2020, pp. 436–439. ISBN: 978-1-4503-6868-1. DOI: 10.1145/3341161.3342927. URL: https://dl.acm.org/doi/10.1145/3341161.3342927 (visited on 03/16/2023).

[392] Chengcheng Shao et al. “The spread of low-credibility content by social bots”. en. In: *Nature Communications* 9.1 (Nov. 2018). Number: 1 Publisher: Nature Publishing Group, p. 4787. ISSN: 2041-1723. DOI: 10.1038/s41467-018-06930-7. URL: https://www.nature.com/articles/s41467-018-06930-7 (visited on 03/16/2023).

[393] Jiwei Li et al. “Towards a General Rule for Identifying Deceptive Opinion Spam”. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Baltimore, Maryland: Association for Computational Linguistics, June 2014, pp. 1566–1576. DOI: 10.3115/v1/P14-1147. URL: https://aclanthology.org/P14-1147 (visited on 03/16/2023).

[394] Suchita Jain, Vanya Sharma, and Rishabh Kaushal. “Towards automated real-time detection of misinformation on Twitter”. In: *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. Sept. 2016, pp. 2015–2020. DOI: 10.1109/ICACCI.2016.7732347.

[395] Victoria L Rubin, Niall J Conroy, and Yimin Chen. “Towards News Verification: Deception Detection Methods for News Discourse”. en. In: *HICSS2015* (2015).

[396] Victoria L. Rubin and Tatiana Lukoianova. “Truth and deception at the rhetorical structure level: Truth and Deception at the Rhetorical Structure Level”. en. In: *Journal of the Association for Information Science and Technology* 66.5 (May 2015), pp. 905–917. ISSN: 23301635. DOI: 10.1002/asi.23216. URL: https://onlinelibrary.wiley.com/doi/10.1002/asi.23216 (visited on 03/16/2023).

[397] Qiang Cao et al. “Aiding the detection of fake accounts in large scale social online services”. In: *9th USENIX Symposium on Networked Systems Design and Implementation (NSDI 12)*. 2012, pp. 197–210.

[398] Savvas Zannettou et al. “Who let the trolls out? towards understanding state-sponsored trolls”. In: *Proceedings of the 10th acm conference on web science*. 2019, pp. 353–362.

[399] Gang Wang et al. “You are how you click: Clickstream analysis for sybil detection”. In: *22nd USENIX Security Symposium (USENIX Security 13)*. 2013, pp. 241–256.

[400] Haifeng Yu et al. “Sybillimit: A near-optimal social network defense against sybil attacks”. In: *2008 IEEE Symposium on Security and Privacy (sp 2008)*. IEEE. 2008, pp. 3–17.

[401] Zhi Yang, Christo Wilson, and Xiao Wang. “Uncovering Social Network Sybils in the Wild”. en. In: ().

[402] Kareem Darwish et al. *Unsupervised User Stance Detection on Twitter*. arXiv:1904.02000 [cs]. May 2020. DOI: 10.48550/arXiv.1904.02000. URL: http://arxiv.org/abs/1904.02000 (visited on 03/16/2023).

[403] Brendan Nyhan and Jason Reifler. “When Corrections Fail: The Persistence of Political Misperceptions”. en. In: *Political Behavior* 32.2 (June 2010), pp. 303–330. ISSN: 1573-6687. DOI: 10.1007/s11109-010-9112-2. URL: https://doi.org/10.1007/s11109-010-9112-2 (visited on 03/16/2023).

[404] Matthew Hindman and Vlad Barash. *Disinformation, ‘Fake News’ and Influence Campaigns on Twitter*. Oct. 2018.

[405] Pujan Paudel et al. “Lambretta: learning to rank for Twitter soft moderation”. In: *2023 IEEE Symposium on Security and Privacy (SP)*. IEEE. 2023, pp. 311–326.

[406] Mary Ellen Zurko. “Disinformation and reflections from usable security”. In: *IEEE Security & Privacy* 20.3 (2022), pp. 4–7.

[407] Hunt Allcott and Matthew Gentzkow. “Social Media and Fake News in the 2016 Election”. In: *Journal of Economic Perspectives* 31.2 (May 2017), pp. 211–36. DOI: 10.1257/jep.31.2.211. URL: https://www.aeaweb.org/articles?id=10.1257/jep.31.2.211.

[408] Tavish Vaidya et al. “Does Being Verified Make You More Credible? Account Verification’s Effect on Tweet Credibility”. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI ’19. Glasgow, Scotland Uk: Association for Computing Machinery, 2019, pp. 1–13. ISBN: 9781450359702. DOI: 10.1145/3290605.3300755. URL: https://doi.org/10.1145/3290605.3300755.

[409] Government News. “The Weibo rumor-refuting and co-governance platform is launched, giving government accounts the right to directly refute rumors”. In: (Nov. 2018). URL: https://weibo.com/ttarticle/p/show?id=2309404301992787852969.

[410] Ivan Mehta and Manish Singh. *Twitter to end free access to its API in Elon Musk’s latest monetization push*. Feb. 2023. URL: https://techcrunch.com/2023/02/01/twitter-to-end-free-access-to-its-api..

[411] Jon Porter. *Twitter announces new API pricing, posing a challenge for small developers*. Mar. 2023. URL: https://www.theverge.com/2023/3/30/23662832/twitter-api-tiers-free-bot-novelty-accounts-basic-enterprice-monthly-price.

[412] Simone McCarthy. *China’s promotion of Russian disinformation indicates where its loyalties lie*. Mar. 2022. URL: https://www.cnn.com/2022/03/10/china/china-russia-disinformation-campaign-ukraine-intl-dst-hnk/index.html.

[413] Archive Team. *The Twitter Stream Grab*. URL: https://archive.org/details/twitterstream.

[414] Kai Shu et al. *FakeNewsNet: A data repository with news content, social context and spatialtemporal information for studying fake news on social media*. Mar. 2019. URL: https://arxiv.org/abs/1809.01286.

[415] Rafael Evangelista and Fernanda Bruno. *WhatsApp and political instability in Brazil: Targeted messages and political radicalisation*. Dec. 2019. URL: https://policyreview.info/articles/analysis/whatsapp-and-political-instability-brazil-targeted-messages-and-political.

[416] Kiran Garimella and Dean Eckles. *Images and misinformation in political groups: Evidence from WhatsApp in India: HKS Misinformation Review*. July 2022. URL: https://misinforeview.hks.harvard.edu/article/images-and-misinformation-in-political-groups-evidence-from-whatsapp-in-india/.

[417] Slick. *Commit to transparency - sign up for the international fact-checking network’s code of Principles*. URL: https://ifcncodeofprinciples.poynter.org/.

[418] Sayash Kapoor and Arvind Narayanan. *Leakage and the Reproducibility Crisis in ML-based Science*. 2022. DOI: 10.48550/ARXIV.2207.07048. URL: https://arxiv.org/abs/2207.07048.

[419] Shachar Kaufman et al. “Leakage in Data Mining: Formulation, Detection, and Avoidance”. In: *ACM Trans. Knowl. Discov. Data* 6.4 (Dec. 2012). ISSN: 1556-4681. DOI: 10.1145/2382577.2382579. URL: https://doi.org/10.1145/2382577.2382579.

[420] R. Nisbet, J. Elder, and G. Miner. *Handbook of Statistical Analysis and Data Mining Applications*. Association for Computing Machinery, 2009. ISBN: 978-0-12-374765-5.

[421] Song Feng, Ritwik Banerjee, and Yejin Choi. “Syntactic Stylometry for Deception Detection”. In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Jeju Island, Korea: Association for Computational Linguistics, July 2012, pp. 171–175. URL: https://aclanthology.org/P12-2034.

[422] Felix Soldner, Verónica Pérez-Rosas, and Rada Mihalcea. "Box of Lies: Multimodal Deception Detection in Dialogues". 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)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 1768–1777. DOI: 10.18653/v1/N19-1175. URL: https://aclanthology.org/N19-1175.

[423] Joseph B. Bak-Coleman et al. *Combining interventions to reduce the spread of viral misinformation*. June 2022. URL: https://www.nature.com/articles/s41562-022-01388-6.

[424] Cristiano Lima. *A whistleblower's power: Key Takeaways from the Facebook papers*. Mar. 2022. URL: https://www.washingtonpost.com/technology/2021/10/25/what-are-the-facebook-papers/.

[425] *Content fact-checkers prioritize*. URL: https://transparency.fb.com/en-gb/features/content-fact-checkers-prioritize/.

[426] Fabiana Zollo and Walter Quattrociocchi. "Misinformation Spreading on Facebook". In: *Complex Spreading Phenomena in Social Systems: Influence and Contagion in Real-World Social Networks*. Ed. by Sune Lehmann and Yong-Yeol Ahn. Cham: Springer International Publishing, 2018, pp. 177–196. ISBN: 978-3-319-77332-2. DOI: 10.1007/978-3-319-77332-2_10. URL: https://doi.org/10.1007/978-3-319-77332-2_10.

[427] Alexandre Bovet and Hernán A. Makse. *Influence of fake news in Twitter during the 2016 US presidential election*. Jan. 2019. URL: https://www.nature.com/articles/s41467-018-07761-2#citeas.

[428] Sunday Oluwafemi Oyeyemi, Elia Gabarron, and Rolf Wynn. "Ebola, Twitter, and misinformation: a dangerous combination?" In: *Bmj* 349 (2014).

[429] Wasim Ahmed et al. "COVID-19 and the 5G conspiracy theory: social network analysis of Twitter data". In: *Journal of medical internet research* 22.5 (2020), e19458.

[430] Darsh J Shah, Tal Schuster, and Regina Barzilay. "Automatic Fact-guided Sentence Modification". In: *AAAI*. 2020. URL: https://arxiv.org/abs/1909.13838.

[431] Robert Geirhos et al. "Shortcut learning in deep neural networks". In: *Nature Machine Intelligence* 2.11 (Nov. 2020), pp. 665–673. DOI: 10.1038/s42256-020-00257-z. URL: https://doi.org/10.1038%2Fs42256-020-00257-z.

[432] Meta. *How Meta's third-party fact-checking program works*. URL: https://www.facebook.com/formedia/blog/third-party-fact-checking-how-it-works.

[433] Amber Jamieson and Olivia Solon. *Facebook to begin flagging fake news in response to mounting criticism*. Dec. 2016. URL: https://www.theguardian.com/technology/2016/dec/15/facebook-flag-fake-news-fact-check.

[434] Colin Crowell. *Our approach to bots and misinformation*. URL: https://blog.twitter.com/official/en_us/topics/company/2017/Our-Approach-Bots-Misinformation.html.

[435] Nur Ibrahim. *Why facebook won't be fact-checking Trump now that he's announced candidacy*. Nov. 2022. URL: https://www.snopes.com/news/2022/11/17/facebook-not-fact-checking-trump/.

[436] Ben Kaiser and Jonathan Mayer. *'It's the Algorithm: A Large-Scale Comparative Field Study of Misinformation Interventions'*. 2023.

[437] *GossipCop*. URL: https://web.archive.org/web/20221102102621/https://gossipcop.com/.

[438] Oliver Darcy. *BuzzFeed News will shut down | CNN business*. Apr. 2023. URL: https://www.cnn.com/2023/04/20/media/buzzfeed-news-shuts-down/index.html.

[439] Tiffany Hsu. *As covid-19 continues to spread, so does misinformation about it*. Dec. 2022. URL: https://www.nytimes.com/2022/12/28/technology/covid-misinformation-online.html.

[440] Emily Bell. *The fact-check industry*. Dec. 2019. URL: https://www.cjr.org/special_report/fact-check-industry-twitter.php.

[441] Ryan Woo Liangping Gao. *China's factory activity falls faster than expected as recovery stumbles*. May 2023. URL: https://www.reuters.com/markets/asia/chinas-factory-activity-falls-faster-than-expected-weak-demand-pmi-2023-05-31/.

[442] David Snelling. *Xperia XZ4 release this month - five things every Sony fan should K...* Feb. 2019. URL: https://www.express.co.uk/life-style/science-technology/1084439/Sony-Xperia-XZ4-release-date-price-specs-mobile-world-congress.

[443] Charles F Bond and Bella M DePaulo. *Accuracy of deception judgments*. 2006. URL: https://pubmed.ncbi.nlm.nih.gov/16859438/.

[444] OpenAI. *ChatGPT*. URL: https://openai.com/product/chatgpt.

[445] Matthew R DeVerna et al. "Artificial intelligence is ineffective and potentially harmful for fact checking". In: *arXiv preprint arXiv:2308.10800* (2023).

[446] James Vincent. *OpenAI isn't doing enough to make CHATGPT's limitations clear*. May 2023. URL: https://www.theverge.com/2023/5/30/23741996/openai-chatgpt-false-information-misinformation-responsibility.

[447] Shirin Ali. *Facebook's formula prioritized anger and ended up spreading misinformation*. Oct. 2021. URL: https://thehill.com/changing-america/enrichment/arts-culture/578724-5-points-for-anger-1-for-a-like-how-facebooks/.

[448] Wall Street Journal staff. *The Facebook Files*. Oct. 2021. URL: https://www.wsj.com/articles/the-facebook-files-11631713039.

[449] Devin Coldewey. *Deconstructing "the Twitter files"*. Jan. 2023. URL: https://techcrunch.com/2023/01/13/deconstructing-the-twitter-files/.

Appendix

A Commercial fact-checking services

We include here a brief market survey of commercial and LLM-powered fact-checking and IO detection services. In general, these services fall into five categories: 1) media fact-checking organizations; 2) brand safety and suitability services; 3) trust & safety operations at large social media platforms, 4) threat detection operations, and 5) analytics organizations unaffiliated with a media outlet that offer research capacity to governments and businesses. We define each service category and (with the exception of the first category, which comprises human media workers and fact-checkers) discuss automated content moderation operations deployed by three prominent exemplars within each service category. In general, in instances where such information is made available, we observe that at-scale content moderation businesses *at least* employ human-labeled datasets to train classifiers, and some retain subject-area experts to adjudicate complex moderation decisions. On social media platforms, in particular, human moderators and automated systems appear to work hand-in-hand: automated systems surface potentially misinformative content that receives final verification from a human moderator. For IO detection, specialized knowledge (pertaining to specific geographies, languages, or political climates) is often invoked.

Media fact-checking. Human fact-checkers and content moderators affiliated with news outlets, or who work as freelance fact-checkers. *The International Fact-Checking Network (IFCN)* is a professional network of media workers and fact-checkers; IFCN is also the de facto standards setting body for media fact-checking, and maintains a fact-checking code of ethics [417]. In general, media fact-checking organizations with IFCN affiliations are established news organizations, non-profits, and watchdog organizations that employ human journalists and fact-checkers. Furthermore, (human) fact-checkers can receive IFCN compliance certificates after passing a qualifying exam.

Brand safety and suitability companies. B2B companies that detect categories of potentially harmful speech on websites where ads might appear. Advertisers wishing to protect "brand safety" contract with these services to ensure that their ads do not appear alongside problematic content. The Global Alliance for Responsible Media (GARM) is the standards-setting body for brand safety and suitability companies [221].

- *Zefr*, a GARM member company, deploys AI to detect material that falls within predefined subcategories of problematic content (e.g., explicit content, misinformation, spam). In a press release for Zefr's acquisition of an AI-driven content moderation company (AdVerif.ai) from 2022, the company disclosed that AdVerif.ai is "powered by fact-checking data from more than 50 IFCN-certified organizations around the globe" [222]—that is, AdVerif.ai trains its models on labeled datasets produced by (human) IFCN affiliates.

- *DoubleVerify*, a GARM member company, "uses sophisticated approaches that rely on a combination of AI and comprehensive human review" [250]. According to the company's documentation, human assessors (a "semantic science team") evaluate site infrastructure and contents; AI is used to scale their assessments.

- *Integral Ad Science* (IAS), a GARM member company, deploys AI to detect low-quality sites via infrastructure features. The company's data sources, and deployment methodology were not immediately evident upon web search; IAS recently announced a new partnership with Meta for ad placement management on Facebook [247].

Trust & safety operations. In-house content moderation teams at large social media platforms.

- *Twitter* has deployed a crowd-sourced annotations platform called Community Notes (formerly Birdwatch) since 2021 [253].

- Until recently, *Facebook* partnered with IFCN affiliates to perform third-party manual checking of possibly misinformative content; first-line automated methods detect potentially harmful speech and surface near-duplicates of known problematic image (SimSearchNet++) and text content [242, 432, 448]. Meta announced in January 2025 that it was sunsetting its third-party checking program in favor of a Community Notes-like system [170].

- *TikTok* employs thousands of content moderators across the globe who "work alongside automated moderation systems" [243, 252].

Threat intelligence services. At-scale detection of advanced persistent threats, foreign influence operations, and other cyberattacks oftentimes perpetrated by nation state actors.

- *Mandiant* strongly implies the use of hybrid detection methods, and disclaims that "defenders must constantly explore different techniques and leverage both subject matter expertise and technical capabilities to filter and uncover malicious activity") [224].

- *Microsoft Threat Intelligence* strongly implies the use of hybrid detection methods; in a report from September 2023, MTI cites the work of in-house "Microsoft

Security teams" which are tracking an advanced social engineering attack [246]. Other details—including possible use of automated methods—are undisclosed.

- *Facebook Coordinated Inauthentic Behavior* reports share quarterly updates about Meta's takedown of coordinated activities across its platforms *and* others, including local news outlets. In a report from February 2023, Meta describes a CIB network in Serbia that used local news media to create the impression of grassroots support for the Serbian Progressive Party; while the nature of the detection methodology is unspecified, the complexity and geographic specificity of the CIB described suggest that specialists with country-level expertise were likely consulted [241].

Analytics firms. For- and non-profit organizations that offer checking services and research capacity to governments and businesses.

- *The Global Disinformation Index (GDI)* "reviews news domains based on various metadata and computational signals." Content, however, is manually reviewed by a "country expert," who analyzes a random sample of 10 articles from a news site to determine veracity [225].
- *DFRLabs (Digital Forensic Research Lab)* has disclosed that it employs human subject-area experts, and primarily addresses technology and policy issues pertaining to global and international affairs. In 2018, Facebook contracted its services to detect online trolls [244].
- *Graphika Labs* leverages network analysis to identify influence operations online. On its own website and in the popular press, Graphika has disclosed that it uses AI to map online networks and trace information flows [237, 239].

LLM-driven detection. A few LLM-powered detection methods have been discussed in the popular press, including those advertised by Google [236] and OpenAI [8], but these deployments appear to be mostly experimental, or have required additional adjudication from human moderators. OpenAI in particular has advertised content moderation tools that address misinformation-adjacent tasks, such as toxic speech detection [238]. Misinformation and toxic speech detection are not equivalent tasks, however, and the latter is narrowly defined in the Perspective training data documentation as a four-way classification task (the four class labels are "profanity/obscenity," "identity-based negativity," "insults," and "threatening" language).

B Common datasets

Dataset	Size	Source	Type	Topic(s)	Label classes	Balance
BuzzFeed-Webis [103]	1.6K	3 each of centrist and left- and ring-wing sources (e.g., ABC, Occupy Democrats, Eagle Rising)	Articles	Political news	`True`, `False`, `Mix`, `n/a`	0.78 / 0.05 / 0.13 / 0.04
CoAID [100]	3.8K articles, 400 claims	Twitter, news sites	Tweets, news articles	COVID, healthcare	`True`, `Fake`	0.06 / 0.94 (articles); 0.06 / 0.94 (claims)
CREDBANK [321]	60M	Twitter.com (2015)	Tweets	Mixed	{`Accurate`, `Inaccurate`} × {`Uncertain`, `Probably`, `Certainly`}	0.75 / 0.25
FakeNewsNet* [414]	22K, 1K	Gossipcop.com, Politifact.com	URLs and tweets	Celebrity gossip, Political news	`Real`, `Fake`	0.76 / 0.24, 0.6 / 0.4
FEVER [180]	185K	Wikipedia	Claims, short statements	Mixed	`Supports`, `Refutes`, `Not Enough Info`	0.33 / 0.33 / 0.33
LIAR [75]	12.8K	Politifact.com	Short statements	Political news, politicians, elections	`pants-fire`, `false`, `barelytrue`, `half-true`, `mostly-true`, `true`	0.08 / 0.18 / 0.18 / 0.18 / 0.18 / 0.18
PHEME [81]	6K	Twitter	Tweets, organized by rumoring events	Mass casualty events, political figures	`True`, `False`, `Unverified`	0.38 / 0.62
Site						
GossipCop [437]	n/a (defunct fact-checking site)	Celebrity tabloids	Articles	Celebrity gossip	`Real`, `Fake`	n/a
MediaBiasFactCheck [94]	n/a (fact-checking site)	Other news sites	Whole site credibility	Political news	`Very High`, `High`, `Mostly Factual`, `Mixed`, `Low`, `Very Low`	n/a
PolitiFact [76]	n/a (fact-checking site)	Other news sites	Claims, short statements	Political news	`pants-fire`, `false`, `barelytrue`, `half-true`, `mostly-true`, `true`	n/a

Figure 5: Commonly used datasets and sources for misinformation detection tasks. Class distributions in "Balance" correspond to the labels appearing in "Label classes."

C Full focus set

Paper		① Target	② Dataset	③ Model	④ Features				⑤ Performance
Work	Scope				Textual	Network-based	Author-based	Infrastructural	Accuracy/AUROC
1. Ajao et al. [26]	Ⓒ Ⓐ	Sentiment **(A.i)**	PHEME [81]	LSTM, DT, RF, SVM	●				0.86 (Acc.)
2. Abulldah-All-Tanvir et al. [82]	Ⓒ Ⓐ Ⓝ	Content **(C.i)**	Twitter (API)	NB, RNN, LSTM, SVM, Logit	●	●			0.89 (Acc.)
3. Bhutani et al. [79]	Ⓒ Ⓐ	Content **(C.i)**; sentiment **(A.i)**	Twitter (API), PolitiFact [76]	Naive Bayes, RF	●				0.60 (AUC)
4. Bozarth et al. [189]	Ⓒ	Contents **(C.i)**	PolitiFact [76], Daily Dot, Zimdars, MBFC	LDA	●				n/a
5. Ciampaglia et al. [68]	Ⓒ Ⓝ	Shortest path search **(C.ii)**	DBpedia	kNN, RF	●				0.97 (AUC)
6. Cui et al. [100]	Ⓒ Ⓐ Ⓤ	Content **(C.i)**; sentiment **(A.i)**; user response **(U.ii)**	PolitiFact[76], GossipCop[437]	KNN, SVM, CSI [110], RMSprop	●				0.82 (F1)
7. Debnath et al. [61]	Ⓒ	Content **(C.i)**	LIAR [75]	CNN	●				0.27 (Acc.)
8. Dey et al. [345]	Ⓒ Ⓐ	Content **(C.i)**; sentiment **(A.i)**	Twitter (API)	Clustering (kNN)	●				0.67 (Acc.)
9. Galitsky et al. [325]	Ⓒ	Content **(C.i)**	Amazon reviews	Parse thicket	●				0.81 (Prec.)
10. Glockner et al. [66]	Ⓒ	Content **(C.i)**	PolitiFact[76], Snopes[156], MultiFC	CNN, DNN	●				0.58 (Acc.)
11. Gordon et al. [276]	Ⓒ Ⓐ	Content **(C.i)**; source rep **(U.i)**	Credibility-Factors2020	SVD	●	●			0.63 (Acc.)
12. Gupta et al. [99]	Ⓒ Ⓐ Ⓝ	Content **(C.i)**; stance **(A.iii)**	Twitter (API)	SVM	●	●			0.60 (Agreement)
13. Hassan et al. [62]	Ⓒ	Content **(C.i)**; checkability **(C.iii)**	NBA, weather datasets	Frequency	●				n/a
14. Jain et al. [394]	Ⓒ	Content **(C.i)**; sentiment **(A.i)**	Twitter (API)	Gensim/TextBlob	●				0.77 (Acc.)
15. Jiang et al. [193]	Ⓒ	Content **(C.i)**; ling./syntax **(C.ii)**	PolitiFact[76], Snopes[156]	SVM	●				0.81 (Acc.)
16. Karimi et al. [359]	Ⓒ	Content **(C.i)**	LIAR[75]	LSTM, CNN	●				0.39 (Acc.)
17. Kartal et al. [67]	Ⓒ	Content **(C.i)**; checkability **(C.iii)**	Check That! dataset	Logit, SVM, RF	●				0.26 (MAP)
18. Kou et al. [64]	Ⓒ	Content **(C.i)**	CoAID [315], CONSTRAINT	Knowledge graph	●				0.90 (Acc.)
19. Paudel et al. [98]	Ⓒ Ⓝ Ⓐ	Keyword detection **(A.ii)**	Abilov et al. dataset, Twitter (API)	AdaRank, ListNet, RF	●				0.79 (MAP)
20. Popat et al. [261]	Ⓒ	Content **(C.i)**	PolitiFact[76], Snopes[156], NewsTrust	biLSTM, CNN	●				0.88 (AUC)
21. Shiralkar et al. [71]	Ⓒ Ⓝ	KG search **(C.ii)**	DBpedia	Knowledge graph	●				1.00 (AUC)
22. Shu et al. [80]	Ⓒ Ⓤ	Word encoding **(C.i)**; user response **(U.ii)**	GossipCop [437], PolitiFact [76]	RNN/RMSprop, CSI [110], LSTM, CNN	●		●		0.93 (F1)
23. Tian et al. [74]	Ⓒ Ⓤ	Content **(C.i)**; user response **(U.ii)**	Twitter15, Twitter16	CNN-biLSTM	●		●		0.82 (F1)
24. Zhang et al. [378]	Ⓒ Ⓤ	Content **(C.i)**; user response **(U.ii)**	RumourEval, PHEME[81]	biLSTM, Multitask, SVM, CNN,	●	●			0.89 (Acc.)
1. Afroz et al. [83]	Ⓐ	Content **(C.i)**; syntax **(A.i)**	Brennan-Greenstadt	SVM, J48 Decision Trees	●				0.97 (F1)
2. Ahmed et al. [84]	Ⓐ	Syntax **(A.i)**	Twitter, Kaggle, Horne and Adali [85]	SVM	●				0.92 (Acc.)
3. Bourgonje et al. [86]	Ⓐ	Stance **(A.ii)**	Fake News Challenge Data	Logit	●				0.90 (Acc.)
4. Brasoveanu et al. [105]	Ⓐ Ⓒ	Sentiment **(A.i)**; keywords **(A.ii)**	LIAR[75]	CNN, LSTM, CN	●	●			0.64 (Acc.)
5. Della Vedova et al. [208]	Ⓐ Ⓝ	Content **(C.i)**; virality **(N.iv)**	FakeNewsNet, Buzzfeed	Logit	●	●			0.82 (Acc.)
6. Horne et al. [85]	Ⓐ	Syntax **(A.i)**, headline **(A.ii)**	Buzzfeed, Burfoot & Baldwin	SVM	●				0.78 (Acc.)
7. Jabiyev et al. [283]	Ⓐ Ⓦ	Topic detection **(A.ii)**; site cred. **(W.iv)**	Snopes[156], FactCheck, PolitiFact[76]	SVM, DT, RF	●		●		0.87 (Acc.)
8. Jadhav et al. [72]	Ⓐ	Content **(C.i)**; syntax **(A.i)**	LIAR [75]	DSSM/RNN	●				0.99 (Acc.)
9. Jin et al. [113]	Ⓐ Ⓝ Ⓤ	**(A.ii)**; **(N.i)**; suspicious accounts **(U.i)**	Tweets; articles	n/a	●	●	●		0.87 (Prec.)
10. Kapusta et al. [259]	Ⓐ	Sentiment & word freq. **(A.i)**	MBFC and custom	n/a	●				n/a
11. Kumar et al. [192]	Ⓐ Ⓦ Ⓤ	**(A.i)**; UI **(W.iii)**; Author cred. **(U.i)**	20K Wiki Hoaxes	Random forest	●	●	●	●	0.87 (AUC)
12. Magdy et al. [97]	Ⓐ	Content **(C.i)**	NYT Corpus [**nyt_corpus**], 100 Wikis	Pattern recog.	●				0.99 (Recall)
13. Monti et al. [137]	Ⓐ Ⓝ Ⓤ	**(A.ii)**; **(N.i)**; suspicious accounts **(U.i)**	Tweets; articles	RNN/CNN	●	●	●		0.927 (AUC)
14. Nasir et al. [111]	Ⓐ	Syntax **(A.i)**	ISOT [158]; FAKES [159]	RNN/CNN	●				0.99 (Acc.)
15. Perez-Rosas et al. [107]	Ⓐ	Syntax **(A.i)**	FakeNewsAMT; Celebrity	SVM	●				0.74 (Acc.)
16. Potthast et al. [103]	Ⓐ	Syntax, sentiment, readability **(A.i)**	Buzzfeed-Webis	Bag-of-words	●				0.46 (F1)
17. Reis et al. [171]	Ⓐ	Syntax **(A.i)**; contents **(C.i)**; source rep. **(U.i)**; timing **(N.ii)**; URL **(W.i)**	BuzzFeed	GBM	●	●	●	●	0.85 (AUC)
18. Rubin et al. [395]	Ⓐ	Syntax **(A.i)**	AMT	Clustering	●				0.67 (Agreement)
19. Ruchansky et al. [110]	Ⓐ Ⓤ	**(A.i)**; Responses **(A.ii)**; acct metadata **(U.i)**	Twitter/Weibo posts	RNN/LSTM	●		●		0.95 (Acc.)
20. Santos et al. [172]	Ⓐ	Readability **(A.i)**	Fake.Br corpus	SVM	●				0.92 (Acc.)
21. Silva et al. [116]	Ⓐ Ⓝ	Topic detection **(A.ii)**; propagation **(N.i)**	PolitiFact[76], GossipCop[437], CoAID [315]	Clustering	●	●			0.88 (Acc.)
21. Singh et al. [317]	Ⓐ	Syntax **(A.i)**	Kaggle Fake News	SVM	●				0.87 (Acc.)
22. Uppal et al. [271]	Ⓐ	Discourse structure **(A.i)**	Buzzfeed, PolitiFact[76]	GRU, Dependency tree	●				0.74 (Acc.)
1. Cao et al. [87]	Ⓤ Ⓝ	Acct. cred. **(U.i)**; prop. **(N.i)**	Tuenti social network	Louvain clustering		●	●		0.90+ (TP)
2. Danezis et al. [134]	Ⓤ Ⓝ	Acct. cred. **(U.i)**; prop. **(N.i)**	LiveJournal data	Bayesian inf.		●	●		n/a*
3. Ezzeddine et al. [88]	Ⓤ	Acct. behaviors **(U.ii)**	DATA	LSTM		●	●		0.91 (AUC)
4. Hamdi et al. [89]	Ⓤ Ⓝ	Account metadata **(U.i)**; prop. **(N.i)**	CREDBANK [321]	LDA, Bayes, Logit, SVM		●	●		0.99 (AUC)
5. Helmstetter et al. [128]	Ⓤ Ⓝ Ⓐ	Acct metadata **(U.i)**; post sharing data **(U.ii)**	Public site cred. lists	SVM, NB, DT, RF	●	●	●		0.936 (F1)
6. Jain et al. [394]	Ⓤ Ⓐ	Acct metadata **(U.i)**; topic det. **(A.ii)**	Twitter (API)	Gensim, TextBlob	●		●		0.77 (Acc.)
7. Leonardi et al. [167]	Ⓤ Ⓝ Ⓐ	Acct metadata **(U.i)**; prop. **(N.i)**	CoAID [315]	RF	●	●	●		0.81 (F1)
8. Saeed et al. [60]	Ⓤ	User behavior **(U.ii)**	Reddit Pushshift; Reddit IRA trolls list	RF			●		0.98 (Acc.)
9. Sansonetti et al. [129]	Ⓤ Ⓒ	Acct metadata **(U.i)**; acct activity **(U.ii)**	PolitiFact [76], Twitter (API)	LSTM-CNN, SVM, KNN	●		●		0.92 (Acc.)
10. Santia et al. [168]	Ⓤ Ⓐ	Source rep. **(U.i)**; user response **(U.ii)**; syntax **(A.i)**	BuzzFeed	SVM, RF, DT, NB	●	●	●		0.77 (Prec.)
11. Shu et al. [334]	Ⓤ Ⓝ Ⓐ	User behavior **(U.ii)**; prop. **(N.i)** syntax **(A.i)**	Buzzfeed, PolitiFact	Gibbs sampling	●	●	●		0.85+ (Acc.)
12. Vargas et al. [319]	Ⓤ Ⓝ Ⓐ	Prop. **(N.i)**; topic det. **(A.ii)**	Twitter (API)	RF	●	●			0.98 (F1)
13. Wang et al. [173]	Ⓤ Ⓝ	User behavior **(U.ii)**	Renren data	SVM		●	●		0.99 (Acc.)
14. Yu et al. [174]	Ⓤ Ⓝ	User behavior **(U.i)**; prop. **(N.i)**	LiveJournal, Friendster, DBLP accounts	Random route		●	●		n/a*
15. Yuan et al. [139]	Ⓤ Ⓝ	Acct metadata **(U.i)**; timing **(N.ii)**	WeChat data	Clustering		●	●		0.90+ (Prec.)
16. Zhang et al. [127]	Ⓤ Ⓝ	User behavior **(U.ii)**; prop. **(N.i)**	Twitter, Slashdot, Epinion	Graph cut		●	●		n/a
17. Zhou et al. [219]	Ⓤ Ⓝ	User suscept. **(U.i)**; prop **(N.i)**	PolitiFact [76], BuzzFeed	SVM, KNN, NB, DT, RF		●	●		0.93 (Acc.)
1. Alizadeh et al. [90]	Ⓝ Ⓐ Ⓤ	Propagation **(N.i)**; syntax **(A.i)**; acct metadata **(U.i)**	Twitter (API), Reddit IRA troll list	RF	●	●	●		0.70+ (F1)
2. Antoniadis et al. [91]	Ⓤ Ⓐ	Acct metadata **(U.i)**; syntax **(A.i)**	Hurricane Sandy tweet dataset	J48, RF, KNN, Bayes	●		●		0.79 (Avg. Prec.)
3. Assenmacher et al. [142]	Ⓝ Ⓐ	Propagation **(N.i)**; topic det. **(A.ii)**	Twitter (API)	Clustering	●	●			not reported
4. Buntain et al. [92]	Ⓝ Ⓤ Ⓐ	Time **(N.ii)**; acct metadata **(U.i)**; sentiment **(A.i)**	CREDBANK [321], Buzzfeed	RF	●	●	●		0.65 (Acc.)
5. Castillo et al. [138]	Ⓤ Ⓐ	Syntax **(A.i)**; user behavior **(U.ii)**	Twitter Monitor events	SVM, DT	●	●	●		0.874 (P)
6. Chen et al. [169]	Ⓝ Ⓤ Ⓐ	Social graph **(N.iii)**; syntax **(A.i)**; user behavior **(U.ii)**	Weibo	RNN	●	●	●		0.92 (Acc.)
7. Guo et al. [380]	Ⓝ Ⓒ	Prop. **(N.i)**; semantics **(C.i)**	Twitter, Weibo	LSTM	●	●			0.9 (Acc.)
8. Jin et al. [362]	Ⓝ Ⓐ	Propagation **(N.i)**; stance **(A.iii)**	Sina Weibo posts	Clustering	●	●			0.84 (Acc.)
9. Liu et al. [143]	Ⓝ	Propagation **(N.i)**	Weibo, Twitter15, Twitter16	RNN, CNN		●			0.897 (Acc.)
10. Ma et al. [144]	Ⓝ	Propagation **(N.i)**	Kochina, Ma, Shu Twitter datasets	RNN/BiLSTM		●			0.75 (Acc.)
11. Magelinski et al. [175]	Ⓝ Ⓤ	Prop. **(N.i)**; timing **(N.ii)**; user behavior **(U.ii)**	Twitter (API)	–		●	●		– n/a
12. Nguyen et al. [216]	Ⓝ Ⓐ	Prop. **(N.i)**; semantics **(A.i)**	Twitter, Weibo, PHEME [81]	SVM, RNN, DT	●	●			0.970 (Acc.)
13. Pacheco et al. [130]	Ⓤ Ⓝ	Account metadata **(U.i)**; timing **(N.i)**	Twitter (API)	Clustering		●	●		0.8+ (Prec.)
14. Ratkiewicz et al. [141]	Ⓝ Ⓐ Ⓤ	Prop. **(N.i)**; keywords **(A.ii)**; user behavior **(U.ii)**	Twitter (API)	AdaBoost, SVM	●	●	●		0.96 (Acc.)
15. Sharma et al. [176]	Ⓝ Ⓐ Ⓤ	Prop. **(N.i)**; keywords **(A.ii)**; user behavior **(U.ii)**	Twitter (API); Twitter IRA trolls list	GMM, Kmeans, NN	●	●	●		0.94 (AUC)
16. Tschiatschek et al. [337]	Ⓝ Ⓤ	Prop. **(N.i)**; user rep. **(U.i)**	Facebook dataset	Bayes inf.		●	●		n/a
17. Zeng et al. [293]	Ⓝ Ⓐ	Prop. **(N.i)**; stance **(A.iii)**	4.3K tweets (from API)	Logit, NB, RF	●	●			0.88 (Acc.)
1. Asr et al. [155]	Ⓦ Ⓐ	Source rep. **(W.i)**; Syntax **(A.i)**	BuzzfeedUSE, Snopes, Rashkin, Rubin	CNN, SVM, NB	●	●		●	not reported
2. Baly et al. [5]	Ⓦ Ⓐ Ⓝ	Source rep. **(W.i)**; Site infra. **(W.ii)**; **(A.i)**	MediaBiasFactCheck[94]	SVM	●			●	0.66 (Acc.)
3. Baly et al. [93]	Ⓦ Ⓐ Ⓝ	Source rep. **(W.i)**; Site infra. **(W.ii)**; **(A.i)**	MediaBiasFactCheck[94]	SVM	●			●	0.7152 (Acc.)
4. Castelo et al. [95]	Ⓦ Ⓐ	Site infra. **(W.ii)**; syntax **(A.i)**	Celebrity, US-Election2016	SVM, kNN, RF	●			●	0.86 (Acc.)
5. Chen et al. [96]	Ⓦ Ⓐ	Hosting infra. (URL) **(W.ii)**; syntax **(A.i)**	PoliticalFakeNews	Clustering	●			●	0.97 (AUC)
6. Hounsel et al. [154]	Ⓦ	Site infra. **(W.i)**	FactCheck, Snopes, PolitiFact, Buzzfeed	RF				●	0.98 (AUC)
7. Ribeiro et al. [131]	Ⓦ Ⓤ	Site infra. **(W.i)**; User demog. **(U.i)**	Facebook API	Graph search		●	●	●	0.97 (PCC)

Table 3: **Focus corpus by scope and target.** Coding of a focus set of 87 papers, sorted by information scope. Values in parentheses in "Target" field correspond to highlighted subcategories presented in Section 4 (e.g., "C.i" denotes target (i) in "Claims," Section 4.1). If authors present evaluation results for multiple models, we underline the most performant model and record its corresponding performance score.

D Baly ablation analysis

Dataset (size)	All features	articles		traffic		twitter		wikipedia		url	
		–	+	–	+	–	+	–	+	–	+
Full corpus (1066)	0.654	0.631	0.644	0.654	0.508	0.648	0.550	0.627	0.606	0.638	0.533
Med. corpus (400)	0.623	0.608	0.630	0.620	0.488	0.635	0.500	0.590	0.588	0.623	0.495
Small corpus (250)	0.636	0.632	0.596	0.632	0.524	0.608	0.512	0.588	0.536	0.624	0.516
Left bias (398)	0.691	0.683	0.671	0.688	0.668	0.686	0.628	0.678	0.683	0.678	0.636
Center (263)	0.913	0.810	0.890	0.913	0.700	0.924	0.741	0.920	0.776	0.890	0.635
Right bias (405)	0.279	0.267	0.252	0.279	0.173	0.286	0.230	0.274	0.205	0.272	0.121
Full corpus (1066)	0.569	0.523	0.595	0.569	0.399	0.580	0.440	0.552	0.538	0.577	0.373
Med. corpus (400)	0.563	0.517	0.580	0.560	0.420	0.578	0.478	0.585	0.545	0.568	0.360
Small corpus (250)	0.456	0.424	0.560	0.452	0.364	0.500	0.408	0.400	0.496	0.444	0.436
Low cred. (256)	0.590	0.516	0.633	0.590	0.641	0.629	0.445	0.609	0.633	0.590	0.473
Mixed cred. (268)	0.407	0.340	0.474	0.407	0.0522	0.414	0.258	0.362	0.276	0.414	0.198
High cred. (542)	0.349	0.336	0.408	0.349	0.255	0.369	0.271	0.341	0.316	0.351	0.218

Table 4: Replication analysis of Baly et al.: Dropout(–) and feature importance(+) analyses of subsets of Baly et al.'s EMNLP18 dataset, stratified by political leaning and credibility. Most (secondmost) performant feature, as determined by its contribution to overall classifier accuracy on the full feature set, is highlighted in darker (lighter) hues. Fact and bias classification task performances are reported in the top and bottom halves of the table, respectively.