[SoK] The Challenges of Machine Learning for Trust and Safety: A Case Study on Misinformation Detection

***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 [275]. 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 [154], toxic speech classification [357], and other core trust and safety challenges (e.g., [358]). 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 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 performance [43], [44].

In the past year, news items from major tech companies have tempered these expectations. In June of 2023, OpenAI announced the deployment of a content moderation system based on GPT-4 to detect problematic content online; in the same press release, OpenAI's head of safety systems admitted that human advisors would still need to "adjudicate borderline cases" [355]. In October of 2023, in a Bluesky post commenting on Twitter's user-driven Community Notes program, 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 [380]. These disclosures accord with our observation that, in practice, trust and safety functions at online services remain heavily manual: driven by user reports and carried out by 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 [331], [336]. 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. Are there identifiable missteps related to target selection, dataset curation, feature selection, and evaluations of method performance in ML-driven misinformation detection studies?
- RQ3. How reproducible are published ML-driven misinformation detection methods?
- RQ4. How generalizable are published ML-driven misinformation detection methods to out-of-domain data (i.e., to data types 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.

Through this work, we hope to underscore 1) the prevalence and severity of reproducibility and ML-driven method design issues in the existing misinformation detection literature, 2) the need for careful and preemptive evaluation of ML-driven methods at the point of problem formulation, and 3) the importance of method explainability and data accessibility. Based on these observations, we provide recommendations for future work that proposes ML models to address trust and safety concerns. **We contribute information taxonomies, our annotated corpus of 248 research papers, recent datasets, and frameworks for evaluating ML-driven content moderation tasks.**

2. MOTIVATION

The detection of misinformation and "influence operations" (IOs) is a topic of convergent interest for security, social science, and AI/ML researchers. Methods developed in recent years for the detection of influence operations (e.g., astroturfing, coordinated misinformation campaigns) resemble techniques previously employed by security researchers for the detection of botnets, advanced persistent threats (APTs), and malware [359], [378], [381]. As such, we believe that the security community is uniquely well-positioned to shepherd the responsible development of misinformation detection methods. In addition, misinformation research stands to benefit a good deal from the formal structure of security research methodology: In a survey study from 2022 conducted by Mirza et al., human fact-checkers and journalists expressed interest in adopting rigorous threat modeling practices similar to those found in academic security literature [360].

3. PRELIMINARIES

Definitions. In the absence of agreed-upon definitions of *misinformation, disinformation, malinformation,* and *influence operations*, we refrain from advancing singular definitions of the same[1]. Instead, 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 [23], [25]–[27], [258]. We limit our analysis to text-based English-language misinformation.[2] We 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 the sake of completeness and concision, we refer to intentionally *and* unintentionally false information as *misinformation* for this work unless otherwise qualified. We occasionally use *false rumors, false news,* and *false information* interchangeably, but avoid *fake news*, a politically charged term [424]. We use "influence operations" (IOs) or "coordinated campaigns" in our discussion of collaborative attempts to disseminate misinformation across networks [425].

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 psychology, 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 and the usefulness of binary true/false classifications of information, particularly in the absence of unified definitions of misinformation, disinformation, and information [201], [354]. On the other hand, some researchers maintain that misinformative texts are characterized by indelible "fingerprints" that distinguish them from non-misinformative texts: for instance, emotional language and reduced lexical diversity [177], [327]. We use this debate to motivate our discussion of non-textual feature sets for misinformation detection.

Taxonomies. Through an iterative process of reading and inductive coding of our papers, we develop a taxonomy of five "information scopes." As we read each paper in our corpus, we noted the *operative unit* of misinformation detection for the classifier or method described—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. Our final taxonomy comprises five information scopes: claims, news articles, social media accounts, networks, and websites. We consider these scopes *in the context of* online services—for instance, news articles and websites whose links might be posted to

1. While certain taxonomies [153]–[155] provide relative definitions of these terms—distinguishing *misinformation* from *disinformation*, for instance, via the presence or absence of intent—we emphasize that, to the best of our knowledge, there is no robust invariant for identifying a statement as (mis)informative on the basis of *semantics alone*.

2. We acknowledge that the nature of misinformation narratives and misinformation spread varies with language and geography.

a social media platform.[3]

Ⓒ **Claims**. The smallest semantic unit of fact or misinformation, comprising a subject, predicate, and object, at minimum.

Ⓐ **Articles**. News-oriented writing of length 100 words or more.

Ⓤ **Users**. 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).

Ⓝ **Networks**. The set of users and interactions between these users as represented by a networked graph of behaviors.

Ⓦ **Websites**. A news site, including its hosting infrastructure and text and image contents.

We describe errors in method design with respect to the step of the development pipeline in which they occur. We break this process down as follows: target selection, dataset curation, model choice, feature set selection and model evaluation. Development step definitions follow:

① **Target selection.** The *stated* versus *actual* objective(s) of the classification task that paper authors describe: for instance, detection of emotional valence of a text (angry, sad, happy); verification of semantic accuracy (true, false); or characterizing degree of virality (e.g., as measured by number of reshares on a post).

② **Dataset curation.** The source, size, and contents of datasets used for model training and testing. Provenance information includes temporal labels for 1) the date of the dataset's production and 2) the date of dataset access by paper authors.

③ **Model choice.** The choice of ML model used by paper authors for their detection task, as well as their motivation for this choice.

④ **Feature selection.** The choice of feature(s) that paper authors use to train their ML models, as well as their motivation for this choice.

⑤ **Model evaluation.** Paper authors' approach to benchmarking method performance after initial training. This includes 1) choice of test dataset; 2) performance statistics (e.g., ROC values, true and false positive rates); and 3) ecological validity of test cases in relation to proposed deployment setting.

Paper organization. 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 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) in order of operations of method development. 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 1) a discussion of findings from our literature review and replication studies and 2) generalizable recommendations for evaluating ML-driven interventions for trust and safety.

3. We label certain works as members of *multiple* scopes when they make classification decisions about more than one type of detection unit: works targeting social media posts, for instance, are generally categorized as falling within (Ⓒ *and* Ⓤ or (Ⓒ *and* Ⓝ) in Table 1.

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 [151]–[157], [159]–[174]. We relied on these papers to ground our own understanding of existing approaches to automated misinformation detection, and 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.[4] We then inspected each paper's reference section for related papers; we read the abstracts for these references in order to confirm relevance and fit. We supplemented this core corpus of highly-cited papers with publications surfaced by Google Scholar queries. We queried the following terms on Google Scholar: "`misinformation detection` [x]" and "`automated fact checking` [x]," where $x \in$ {claims, news articles, accounts, networks, websites, influence operations[5]}. 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*. We note that these search terms are deliberately broad and over-inclusive; we manually review all potential works for relevance after the initial sampling step. After removal of out-of-scope works (see *In- and out-of-scope work.*) from this initial 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 security venues: Using the same keywords listed in the previous subsection, we oversampled publications from four prominent[6] security research venues (USENIX Security, IEEE S&P, NDSS, and ACM CCS). This sampling process resulted

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

5. We include "influence operations" as a separate search term because this term of art is a relatively new one—we group most of these works with our network-scoped methods.

6. A*-rated venues, as determined by CORE [255]

in the addition of 29 additional works, most of which address the detection of accounts and networks that spread misinformative content (e.g., botnets and trolls); as such, most of these works are categorized as network- or account-scoped detection methods. Our final corpus comprises 248 papers published between 2009 and 2023, inclusive.

"Full" and "focus" paper corpora. We conduct our literature survey at two different levels of granularity. For all papers in the full corpus (248 works), we note detection targets and model evaluation (steps ① and ⑤ in Section 3). For a subset of these (87), we perform deep-coding of methods: we note dataset design choices, model selection, and feature sets (resp. steps ②, ③, and ④ in Section 3).[7] A summary of this deep-coding is available in Table 1. Our motivation for developing this focus set is practical: A good many of the works we review do not include in-depth discussions of model choice, weights, and datasets. As such, each per-scope discussion is book-ended by comprehensive overviews of the full paper corpus, with more focused analyses of method design details in between.

Corpus curation. As a first-pass relevancy check, we used automated keyword matching to confirm that all collected papers did, in fact, address misinformation and automated fact-checking methods; we read the abstracts of papers surfaced by this check to confirm relevance and fit. We then annotated research papers that passed this check in accordance with our codebook (see SI). One reader made three separate coding passes over the corpus, varying the objectives of her annotation on each pass: In the first pass, she specifically sought to identify taxonomies for classification and detection; on the second, she noted actual versus stated detection targets and approaches to evaluation; on the final pass, she noted method design approaches and errors.

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 specifically target influence operations online 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, however, we are unable to provide an in-depth analysis of commercial methodologies in our main literature review, and instead include a market survey of commercial fact-checking providers in our online supplement. (In general, commercial vendors do not make code and training data publicly available.) *LLM-powered detection* is in-scope for this work. Accessibility to code bases and training sets for LLMs is similarly limited (this has been discussed at length in the popular press, as well [393], [394]), and disallows extensive testing and evaluation by researchers. As such, while we do briefly discuss LLM-powered detection approaches in our supplement, we defer a more extensive discussion to future work.

Notation. In these sections, we denote information scope and pipeline steps with circled icons (Ⓐ, ①) and sub-taxa within those categories with lowercase bolded Roman numerals (**iv**). We note that, while *targets* are generally unique to each information scope (and are referenced by an alphanumeric label (e.g., **(A.i)**), critiques of steps ②-⑤ are *not* scope-specific, and are always denoted by Arabic-Roman numeral pairs (e.g., **(2.i)**). We highlight each sub-taxon in-text, with a different color corresponding to each development step. (The full taxonomy is in Section 7.) A summary of findings for each scope is available in the **Takeaways** box at the end of each subsection.

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 [84], [130], [297]; about 30% cite COVID-19 misinformation [121], [128], [142]. All works mention the speed and volume of *social media* misinformation in particular [108]. 30% cite the relatively slow pace of manual fact-checking [22], [108], [130]—and the need for faster, automated approaches—as motivation. Claim-scoped papers form 15% of our corpus.

① **Detection targets.** Across all information scopes, we find that claim-scoped detection methods are most consistent in their attempts to verify 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 [48], [77]; and **(C.ii)** *search on a knowledge graph topology* [22], [56] to determine if these reference sources corroborate or refute the claim to be checked. A small class of approaches explicitly detect **(C.iii)** "*checkworthiness*," [8] and are intended to surface checkable statements to human fact-checkers for manual verification [56], [108], [130]. Targets such as author credibility and language cues are *proxy targets* for the presence of misinformation: signals which may not be sufficient for determining text veracity in isolation, but which are indicative of (lack of) veracity by association with an external heuristic. About 60% of papers within the full corpus at this scope employ non-semantic targets, including the **(A.i)** *syntactic and/or stylometric qualities of text* [238], **(U.i)** *source reputation* [383], or **(U.ii)** *contextual indicators*, such as commenter responses [106], to classify social media posts.

② **Datasets.** Claim-scoped papers in our corpus that propose to perform fact verification **(2.i)** *rely on outdated existing datasets* of labeled statements in order to establish ground truth. LIAR [35][9] and PolitiFact [299] were the most popular datasets among claim-scoped works, and, taken together, were used by approximately half of all papers within this scope [76], [123]. 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

7. We take inspiration from [2]–[4], who perform a similar deep-coding of a subset of their whole-paper corpus.

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

9. We note that the LIAR dataset is actually a collection of 27.8K labeled PolitiFact statements—so, in a sense, PolitiFact is the dominant data source.

TABLE 1. **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.1" 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.

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. [278]	(C) (A)	Sentiment **(A.i)**	PHEME [146]	LSTM, DT, RF, SVM	●				0.86 (Acc.)
2. Abulldah-All-Tanvir et al. [210]	(C) (A) (N)	Content **(C.i)**	Twitter (API)	NB, RNN, LSTM, SVM, Logit	●	●			0.89 (Acc.)
3. Bhutani et al. [236]	(C) (A)	Content **(C.i)**; sentiment **(A.i)**	Twitter (API), PolitiFact [299]	Naive Bayes, RF	●				0.60 (AUC)
4. Bozarth et al. [20]	(C)	Contents **(C.i)**	PolitiFact [299], Daily Dot, Zimdars, MBFC	LDA	●				n/a
5. Ciampaglia et al. [22]	(C) (N)	Shortest path search **(C.ii)**	DBpedia	kNN, RF	●				0.97 (AUC)
6. Cui et al. [277]	(C) (A) (U)	Content **(C.i)**; sentiment **(A.i)**; user response **(U.ii)**	PolitiFact [299], GossipCop [343]	KNN, SVM, CSI [37], RMSprop	●				0.82 (F1)
7. Debnath et al. [84]	(C)	Content **(C.i)**	LIAR [35]	CNN	●				0.27 (Acc.)
8. Dey et al. [241]	(C) (A)	Content **(C.i)**; sentiment **(A.i)**	Twitter (API)	Clustering (kNN)	●				0.67 (Acc.)
9. Galitsky et al. [215]	(C)	Content **(C.i)**	Amazon reviews	Parse thicket	●				0.81 (Prec.)
10. Glockner et al. [142]	(C)	Content **(C.i)**	PolitiFact [299], Snopes [298], MultiFC	CNN, DNN	●				0.58 (Acc.)
11. Gordon et al. [110]	(C) (A)	Content **(C.i)**; source rep **(U.i)**	Credibility-Factors2020	SVD	●	●			0.63 (Acc.)
12. Gupta et al. [294]	(C) (A) (N)	Content **(C.i)**; stance **(A.iii)**	Twitter (API)	SVM	●	●			0.60 (Agreement)
13. Hassan et al. [130]	(C)	Content **(C.i)**; checkability **(C.iii)**	NBA, weather datasets	Frequency	●				n/a
14. Jain et al. [291]	(C)	Content **(C.i)**; sentiment **(A.i)**	Twitter (API)	Gensim/TextBlob	●				0.77 (Acc.)
15. Jiang et al. [32]	(C)	Content **(C.i)**; ling./syntax **(C.ii)**	PolitiFact [299], Snopes [298]	SVM	●				0.81 (Acc.)
16. Karimi et al. [260]	(C)	Content **(C.i)**	LIAR [35]	LSTM, CNN	●				0.39 (Acc.)
17. Kartal et al. [108]	(C)	Content **(C.i)**; checkability **(C.iii)**	Check That! dataset	Logit, SVM, RF	●				0.26 (MAP)
18. Kou et al. [128]	(C)	Content **(C.i)**	CoAID, CONSTRAINT	Knowledge graph	●				0.90 (Acc.)
19. Paudel et al. [185]	(N) (A)	Keyword detection **(A.ii)**	Abilov et al. dataset, Twitter (API)	AdaRank, ListNet, RF	●				0.79 (MAP)
20. Popat et al. [90]	(C)	Content **(C.i)**	PolitiFact [299], Snopes [298], NewsTrust	biLSTM, CNN	●				0.88 (AUC)
21. Shiralkar et al. [56]	(C) (N)	KG search **(C.ii)**	DBpedia	Knowledge graph	●				1.00 (AUC)
22. Shu et al. [207]	(C) (U)	Word encoding **(C.i)**; user response **(U.ii)**	GossipCop [343], PolitiFact [299]	RNN/RMSprop, CSI [37], LSTM, CNN	●		●		0.93 (F1)
23. Tian et al. [106]	(C) (U)	Content **(C.i)**; user response **(U.ii)**	Twitter15, Twitter16	CNN-biLSTM	●		●		0.82 (F1)
24. Zhang et al. [382]	(C) (U)	Content **(C.i)**; user response **(U.ii)**	RumourEval, PHEME [146]	biLSTM, Multitask, SVM, CNN,	●	●			0.89 (Acc.)
1. Afroz et al. [212]	(A)	Content **(C.i)**; syntax **(A.i)**	Brennan-Greenstadt	SVM, J48 Decision Trees	●				0.97 (F1)
2. Ahmed et al. [214]	(A)	Syntax **(A.i)**	Twitter, Kaggle, Horne and Adali [24]	SVM	●				0.92 (Acc.)
3. Bourgonje et al. [55]	(A)	Stance **(A.ii)**	Fake News Challenge Data	Logit	●				0.90 (Acc.)
4. Brasoveanu et al. [68]	(A) (C)	Sentiment (**A.i**); keywords **(A.ii)**	LIAR [35]	CNN, LSTM, CN	●	●			0.64 (Acc.)
5. Della Vedova et al. [58]	(A) (N)	Content **(C.i)**; virality **(N.iv)**	FakeNewsNet, Buzzfeed	Logit	●	●			0.82 (Acc.)
6. Horne et al. [24]	(A)	Syntax (**A.i**), headline **(A.ii)**	Buzzfeed, Burfoot & Baldwin	SVM	●				0.78 (Acc.)
7. Jabiyev et al. [118]	(A) (W)	Topic detection (**A.ii**); site cred. (**W.iv**)	Snopes [298], FactCheck, PolitiFact [299]	SVM, DT, RF	●		●		0.87 (Acc.)
8. Jadhav et al. [238]	(A)	Content **(C.i)**; syntax **(A.i)**	LIAR [35]	DSSM/RNN	●				0.99 (Acc.)
9. Jin et al. [216]	(A) (N) (U)	(**A.ii**); (**N.i**); suspicious accounts **(U.i)**	Tweets; articles	n/a	●	●	●		0.87 (Prec.)
10. Kapusta et al. [88]	(A)	Sentiment & word freq. (**A.i**)	MBFC and custom	n/a	●				n/a
11. Kumar et al. [29]	(A) (W) (U)	(**A.i**); UI (**W.iii**); Author cred. (**U.i**)	20K Wiki Hoaxes	Random forest	●	●	●	●	0.87 (AUC)
12. Magdy et al. [18]	(A)	Content **(C.i)**	NYT Corpus [224], 100 Wikis	Pattern recog.	●				0.99 (Recall)
13. Monti et al. [315]	(A) (N) (U)	(**A.ii**); (**N.i**); suspicious accounts **(U.i)**	Tweets; articles	RNN/CNN	●	●	●		0.927 (AUC)
14. Nasir et al. [141]	(A)	Syntax **(A.i)**	ISOT [312]; FAKES [313]	RNN/CNN	●				0.99 (Acc.)
15. Perez-Rosas et al. [63]	(A)	Syntax **(A.i)**	FakeNewsAMT; Celebrity	SVM	●				0.74 (Acc.)
16. Potthast et al. [17]	(A)	Syntax, sentiment, readability **(A.i)**	Buzzfeed-Webis	Bag-of-words	●				0.46 (F1)
17. Reis et al. [286]	(A)	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. [292]	(A)	Syntax **(A.i)**	AMT	Clustering	●				0.67 (Agreement)
19. Ruchansky et al. [37]	(A) (U)	(**A.i**); Responses (**A.ii**); acct metadata **(U.i)**	Twitter/Weibo posts	RNN/LSTM	●		●		0.95 (Acc.)
20. Santos et al. [19]	(A)	Readability (**A.i**)	Fake.Br corpus	SVM	●				0.92 (Acc.)
21. Silva et al. [111]	(A) (N)	Topic detection (**A.ii**); propagation (**N.i**)	PolitiFact [299], GossipCop [343], CoAID	Clustering	●	●			0.88 (Acc.)
21. Singh et al. [203]	(A)	Syntax **(A.i)**	Kaggle Fake News	SVM	●				0.87 (Acc.)
22. Uppal et al. [102]	(A)	Discourse structure **(A.i)**	Buzzfeed, PolitiFact [299]	GRU, Dependency tree	●				0.74 (Acc.)
1. Cao et al. [101]	(U) (N)	Acct. cred. (**U.i**); prop. (**N.i**)	Tuenti social network	Louvain clustering		●	●		0.90+ (TP)
2. Danezis et al. [373]	(U) (N)	Acct. cred. (**U.i**); prop. (**N.i**)	LiveJournal data	Bayesian inf.		●	●		n/a*
3. Ezzeddine et al. [366]	(U)	Acct. behaviors (**U.ii**)	DATA	LSTM		●	●		0.91 (AUC)
4. Hamdi et al. [91]	(U) (N)	Account metadata (**U.i**); prop. (**N.i**)	CREDBANK	LDA, Bayes, Logit, SVM		●	●		0.99 (AUC)
5. Helmstetter et al. [5]	(U) (N) (A)	Acct metadata (**U.i**); post sharing data (**U.ii**)	Public site cred. lists	SVM, NB, DT, RF	●	●	●		0.936 (F1)
6. Jain et al. [291]	(U) (A)	Acct metadata (**U.i**); topic det. (**A.ii**)	Twitter (API)	Gensim, TextBlob	●		●		0.77 (Acc.)
7. Leonardi et al. [38]	(U) (N) (A)	Acct metadata (**U.i**); prop. (**N.i**)	CoAID	RF	●	●	●		0.81 (F1)
8. Saeed et al. [361]	(U)	User behavior (**U.ii**)	Reddit Pushshift; Reddit IRA trolls list	RF			●		0.98 (Acc.)
9. Sansonetti et al. [384]	(U) (C)	Acct metadata (**U.i**); acct activity (**U.ii**)	PolitiFact [299], Twitter (API)	LSTM-CNN, SVM, KNN	●		●		0.92 (Acc.)
10. Santia et al. [21]	(U) (A)	Source rep. (**U.i**); user response (**U.ii**); syntax (**A.i**)	BuzzFeed	SVM, RF, DT, NB	●	●	●		0.77 (Prec.)
11. Shu et al. [226]	(U) (N) (A)	User behavior (**U.ii**); prop. (**N.i**) syntax (**A.i**)	Buzzfeed, PolitiFact	Gibbs sampling	●	●	●		0.85+ (Acc.)
12. Vargas et al. [376]	(N) (A)	Prop. (**N.i**); topic det. (**A.ii**)	Twitter (API)	RF	●	●			0.98 (F1)
13. Wang et al. [191]	(U) (N)	User behavior (**U.ii**)	Renren data	SVM		●	●		0.99 (Acc.)
14. Yu et al. [200]	(U) (N)	User behavior (**U.i**); prop. (**N.i**)	LiveJournal, Friendster, DBLP accounts	Random route		●	●		n/a*
15. Yuan et al. [375]	(U) (N)	Acct metadata (**U.i**); timing (**N.ii**)	WeChat data	Clustering		●	●		0.90+ (Prec.)
16. Zhang et al. [257]	(U) (N)	User behavior (**U.ii**); prop. (**N.i**)	Twitter, Slashdot, Epinion	Graph cut		●	●		n/a
17. Zhou et al. [72]	(U) (N)	User suscept. (**U.i**); prop (**N.i**)	PolitiFact [299], BuzzFeed	SVM, KNN, NB, DT, RF		●	●		0.93 (Acc.)
1. Alizadeh et al. [370]	(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. [178]	(U) (A)	Acct metadata (**U.i**); syntax (**A.i**)	Hurricane Sandy tweet dataset	J48, RF, KNN, Bayes	●		●		0.79 (Avg. Prec.)
3. Assenmacher et al. [391]	(N) (A)	Propagation (**N.i**); topic det. (**A.ii**)	Twitter (API)	Clustering	●	●			not reported
4. Buntain et al. [50]	(N) (U) (A)	Time (**N.ii**); acct metadata (**U.i**); sentiment (**A.i**)	CREDBANK, Buzzfeed	RF	●	●	●		0.65 (Acc.)
5. Castillo et al. [253]	(U) (A)	Syntax (**A.i**); user behavior (**U.ii**)	Twitter Monitor events	SVM, DT	●	●	●		0.874 (P)
6. Chen et al. [183]	(N) (U) (A)	Social graph (**N.iii**); syntax (**A.i**); user behavior (**U.ii**)	Weibo	RNN	●	●	●		0.92 (Acc.)
7. Guo et al. [272]	(N) (C)	Prop. (**N.i**); semantics (**C.i**)	Twitter, Weibo	LSTM	●	●			0.9 (Acc.)
8. Jin et al. [263]	(N) (A)	Propagation (**N.i**); stance (**A.iii**)	Sina Weibo posts	Clustering	●	●			0.84 (Acc.)
9. Liu et al. [40]	(N)	Propagation (**N.i**)	Weibo, Twitter15, Twitter16	RNN, CNN		●			0.897 (Acc.)
10. Ma et al. [30]	(N)	Propagation (**N.i**)	Kochina, Ma, Shu Twitter datasets	RNN/BiLSTM		●			0.75 (Acc.)
11. Magelinski et al. [262]	(N) (U)	Prop. (**N.i**); timing (**N.ii**); user behavior (**U.ii**)	Twitter (API)	–		●	●		– n/a
12. Nguyen et al. [69]	(N) (A)	Prop. (**N.i**); semantics (**A.i**)	Twitter, Weibo, PHEME [146]	SVM, RNN, DT	●	●			0.970 (Acc.)
13. Pacheco et al. [92]	(U) (N)	Account metadata (**U.i**); timing (**N.i**)	Twitter (API)	Clustering		●	●		0.8+ (Prec.)
14. Ratkiewicz et al. [388]	(N) (A) (U)	Prop. (**N.i**); keywords (**A.ii**); user behavior (**U.ii**)	Twitter (API)	AdaBoost, SVM	●	●	●		0.96 (Acc.)
15. Sharma et al. [158]	(N) (A) (U)	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. [230]	(N) (U)	Prop. (**N.i**); user rep. (**U.i**)	Facebook dataset	Bayes inf.		●	●		n/a
17. Zeng et al. [133]	(N) (A)	Prop. (**N.i**); stance (**A.iii**)	4.3K tweets (from API)	Logit, NB, RF	●	●			0.88 (Acc.)
1. Asr et al. [287]	(W) (A)	Source rep. (**W.i**); Syntax (**A.i**)	BuzzfeedUSE, Snopes, Rashkin, Rubin	CNN, SVM, NB	●	●		●	not reported
2. Baly et al. [43]	(W) (A) (N)	Source rep. (**W.i**); Site infra. (**W.ii**); (**A.i**)	MediaBiasFactCheck [311]	SVM	●			●	0.66 (Acc.)
3. Baly et al. [316]	(W) (A) (N)	Source rep. (**W.i**); Site infra. (**W.ii**); (**A.i**)	MediaBiasFactCheck [311]	SVM	●			●	0.7152 (Acc.)
4. Castelo et al. [302]	(W) (A)	Site infra. (**W.ii**); syntax (**A.i**)	Celebrity, US-Election2016	SVM, kNN, RF	●			●	0.86 (Acc.)
5. Chen et al. [109]	(W) (A)	Hosting infra. (URL) (**W.ii**); syntax (**A.i**)	PoliticalFakeNews	Clustering	●			●	0.97 (AUC)
6. Hounsel et al. [1]	(W)	Site infra. (**W.i**)	FactCheck, Snopes, PolitiFact, Buzzfeed	RF				●	0.98 (AUC)
7. Ribeiro et al. [254]	(W) (U)	Site infra. (**W.i**); User demog. (**U.i**)	Facebook API	Graph search		●	●	●	0.97 (PCC)

LIAR are likely ineffective in contemporary fact-checking contexts, and for non-political subject matter [84], [383].[10]

10. Choice of ground truth site or labeled dataset can significantly influence the outcome of analysis: Bozarth et al. found that perceived prevalence of misinformation in a corpus of 2016 election news varied from 2% to 40%, depending on choice of ground-truth reference website [20].

Additionally, misinformation taxonomies across reference sites are inconsistent: PolitiFact employs a six-point labeling scale (pants-on-fire; false; barely-true; half-true; mostly-true; true), FEVER employs a three-point scale (supported; refuted; notenoughinfo), and GossipCop employs an eleven-point scale (ratings from 0 to 10). This divergence is likely a symptom of definitional issues (Section 3).

③ **Model selection.** Model choice follows target choice for claim-scoped methods: for methods that pre-construct knowledge graphs or other reference databases, models perform some form of **(3.i)** *shortest-path search* on the KG topology [22], [56], and approximate logical inference via transitive closure on graph edges. For methods that perform information retrieval at query time—e.g., to match corroborating sources to a claim to be checked—model training generally follows conversion of the text statement to a bag-of-words or TF-IDF embedding; choice of model is highly variable [207], [236], and does not appear to be predictive of performance. For methods that perform detection of misinformative posts on social media, **(3.v)** *stacked ensemble classifiers* are a common approach to incorporating multiple feature modalities.

④ **Feature selection.** At the level of single claims, semantic feature analysis is limited to the **(4.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") [18], [22], [56]. Detection versatility is determined by the size of the source dataset and the granularity of the relationships encoded by graph edges [56]. Supervised methods that detect linguistic cues employ **(4.i)** *hand-crafted word or topic lists* [108]). Authors employing supervised methods claim that their approach permits highly customized targeting of specific rumoring narratives [108], though this also assumes that the method developer has prior knowledge of the contents of test data; authors employing 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 [76]. Methods detecting social media data consider **(4.ii)** *network* and **(4.iii)** *user* interaction features (post likes, shares, comments) [185], [277], [294].

⑤ **Evaluation.** Though a majority of claim-scoped methods cite the speed of social media misinformation as motivation, **(5.i)** *only one method within this scope reported results from testing in real time* [54]. KG-based methods are **(5.ii)** *non-generalizable by design*: the approaches we survey require structured inputs for graph construction, and rely on published datasets for ground truth [22], [56]. 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). Additionally, 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: ① The efficacy of claim-scoped methods is completely determined by the depth and breadth of coverage conferred by a reference database. ② Datasets are frequently out-of-date, and taxonomies are inconsistent. ③ Though some inference is possible via transitive closure on knowledge graph edges, this capability is, in general, limited. ④ Knowledge-graph-based methods require structured inputs, which might not be readily available in a breaking news setting, or in scenarios where ground truth references are not yet available. ⑤ Few authors test in real-time, despite citing the slow pace of manual fact-checking as motivation for their work.

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 [174]. Text-based credibility classifiers have been shown to have limited efficacy, however: while unsupervised approaches can identify *bias* with high accuracy, this performance degrades significantly in misinformation and credibility classification tasks [17], [395]. This scope comprises 24% of our full paper corpus.

① **Detection targets.** In contrast to single claims, full-length news articles have sufficient text contents to make semantic verification difficult—and certain 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 [236], and Horne et al. find that satire and misinformation share stylistic similarities [24]. All article-scoped methods target proxy signals, and adopt at least one of the following three 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) [324]; and **(A.ii)** *analysis of article contents and headlines to identify potentially clickbait-y titles or discussion related to known misinformation narratives* (topic detection) [108], [163]. 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.

② **Datasets.** While well-annotated, current datasets are in short supply across all information 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 paper authors within this scope **(2.i)** *use public datasets released years prior to the start of their research* [24], [35], [41], [68]. These datasets (LIAR [35], Buzzfeed-Webis [17], and PolitiFact [299]) include news links, speaker

credibility scores, and other metadata that **(2.ii)** *constitute serious sources of leakage* for methods that use contextual features to infer true/false labels. The remaining 56% of authors curate their own article corpora by asking crowdworkers to generate misinformative text [63], selectively editing true news articles (e.g., via verb inversion or noun replacement) [31], or compiling articles from authoritative news sources and known satire sites [240]. These data curation techniques **(2.iii)** *introduce additional dependencies and shortcuts* to textual datasets for which such variables are already difficult to control. Style [17] and genre [89], for instance, are emergent qualities of writing that cannot be easily marginalized out of a text embedding.

③ **Model selection.** Among papers that report testing with multiple models—including **(3.ii)** *classical ML models* [214], **(3.iii)** *unsupervised NN models* [68], and **(3.v)** *stacked ensemble classifiers* [37]—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 [212]. 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.

④ **Feature selection.** Among supervised methods that disclose their feature sets, we find that **(4.i)** *word frequency, sentiment, and genre* were among the most commonly used features, and were collectively employed 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 voice, house style, 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 its performance in light of possible style-related dependencies) in our replication analysis of Nasir et al. (Section 5.1) [141].

⑤ **Evaluation.** Misinformation detection methods scoped to full texts risk overfitting to single topics: 42% of authors select **(5.ii)** o*ne or more narratives of interest* (e.g., the 2016 presidential election, the Boston Marathon bombing), train a classifier on these topics, then test this classifier on a different set of texts that discuss the *same* topic [141], [216], [276]. 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: Rapid topic identification and high-quality annotation of relevant articles are generally unavailable in breaking news scenarios on social media platforms [396]. 60% of authors at this scope compare the performance of their detection method to other published approaches or ML models, but **(5.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% [73], [111]. 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 [212]. We demonstrate such a performance dropoff in our replication analysis (Section 5.1).

> **Takeaways**: ① In the absence of semantic definitions of misinformation, *proxy* detection targets are common but easy to evade. ② Well-labeled datasets are rare; those datasets that are available are at least several years old at time of writing. ③ 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. ④ Detection methods at the article scope are uniquely susceptible to text-based dependencies that are difficult to control for. ⑤ 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 [361], [366]. 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 [360]–[362], [364], [365]. Account-scoped papers formed approximately 15% of our corpus (40 papers).

① **Detection targets.** Papers within this scope target source reputation, and **(U.i)** *inspect account metadata*, such 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 [361], [367]–[370]. In the absence of rigorous definitions of these account types, however, actual detection targets are tautological: A troll or bot is an account that exhibits troll- or bot-like behavior, or that interacts with confirmed troll or bot accounts [361], [367].

② **Datasets.** All troll and bot account detection works we reviewed relied on published lists of "known" troll accounts for model training, but **(2.iv)** *neglected to mention the heavily manual investigation required to produce these original lists* [300], [385]. 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 [385]. Two other highly-cited troll lists compiled by the U.S. government, comprising several thousand suspicious Twitter and Facebook accounts, were curated through the use of proprietary non-public information [423].

③ **Model selection.** Detection proceeds via classifier training on a list of "known" suspicious accounts and application of this classifier to a dataset of novel accounts [361], [364]. 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 [371]; we explore this further in our replication analysis of Saeed et al. (Section 5.2) [361]. A subset of methods at this scope and the network scope formulate detection as a **(3.iv)** *graph cut or influence maximization problem*, and describe approaches to identifying optimal cuts for isolating suspicious accounts [101], [257].

④ **Feature selection.** Methods that define suspicious accounts by intrinsic *properties* of these accounts (e.g., user handles and profile images) target the **(4.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 [5], [384]. Methods that define suspicious accounts by account *activity* target **(4.ii)** *networked behaviors*, such as liking and resharing statistics [92]. Feature sets for some methods in the first category include demographic data for users, such as inferred political party affiliation or race [254] (these *n*th order assumptions are dangerous to make [100]; see our discussion about proxy signals for detection, in Section 4.2.

⑤ **Evaluation.** We observe accuracy scores above 80% for **(5.ii)** *confirmatory detection of like accounts* for all methods that reference a seed list of known trolls [66], [361], [366]; de novo detection of behaviors not represented within training data is not possible, by the self-admission of 10% of authors within this scope [359], [361]. 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. Formerly, Botometer did this by assigning a "bot-ness" score to individual Twitter accounts that gauged extent of automation [422]. Though currently defunct post-Twitter API shutdown, Botometer performance had already degraded significantly on Twitter data from 2020, per Rauchfleisch et al. [372].[11] Some authors of bot detection methods acknowledge that their approaches are **(5.iii)** *trivially easy to evade* if account holders 1) avoid interacting with known suspicious accounts or 2) vary their account identity and posting semantics [367], [373].

11. In a blog post on bot detection from 2021, Twitter "debunked" four common heuristics commonly used to identify bot accounts, including several detected by methods in our corpus, and described a "forensic team of investigators" who manually verify bot-ness of suspect accounts [398].

Takeaways: ① Targets are frequently tautological. ② 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. ③ Classifiers can only detect accounts resembling those in seed lists. ④ Features that attempt to infer user credibility from demographic information risk reinforcing existing biases. ⑤ Current methods cannot detect novel malicious behaviors.

4.4. Networks

Within the security literature, a growing awareness of hybrid networks, which employ 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 [367], [374]. 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*) [302]. 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 [302], [378]. Network-scoped methods, including relevant work in the security literature, form 20% of our corpus.

① **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 [6], [41], [228] and temporally anomalous activity with evidence of coordination [359], [364], [367]. 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 [253], [315]. Similarly, within the security literature, anomalous patterns of account registration and user interaction serve as proxies for the presence of Sybils and botnets [373], [375]; early detection requires that authors formulate a priori assumptions about the nature of these patterns.

② **Datasets.** 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 **(2.v)** *accessibility issues* were exacerbated by the shutdown of the Twitter API [308]. In total, we attempted to locate datasets and code for 50 different papers (see Section 5 for our methodology). Thirty-six (72%) of these analyzed tweet corpora, and 42 (84%) of these targeted social media users and posting contents. 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 (as we were unable to locate these datasets ourselves), nine responded; six of those authors were able to provide method code or partial datasets.

③ **Model selection.** The network-scoped methods we review formulate detection as 1) a structured content classification problem [370], [388], and/or 2) a clustering problem on social graphs [157], [391]. In the former case, authors employ an assortment of **(3.ii, 3.iii)** *supervised and unsupervised models* to detect suspicious language across multiple accounts. In the latter case, authors use **(3.vi)** *Louvain or Kmeans 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 [30], [40] access datasets of social media posts that were already sorted by rumoring topic or event [189], [202].

④ **Feature selection.** 55% of papers within this scope make normative theoretical assumptions about user behaviors: In keeping with an epidemiological model[12] of misinformation spread [132], [328], Nguyen et al. assume a homogeneous population of newsreaders, with **(4.ii)** identical probabilities of "infection" and reinfection [205]. 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 [375], or (seemingly contradictorily) that compromised Sybils will refrain from connecting with additional Sybils, to avoid detection [373].[13]

⑤ **Evaluation.** Ferrara et al. [379] called attention to the false positive rate problem in botnet detection in 2016, noting that classifiers for bot detection only work well in instances where there is a clear-cut distinction between bot and non-bot accounts. 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) [359], [367], **(5.v)** *rendering even positive detection results insufficient or meaningless*: coordinated activity need not be inauthentic, and inauthentic activity need not be malicious.

Takeaways: ① Pattern- and virality-based detection approaches disallow early detection of rumors. ② Current social media data is difficult to obtain. ③ "Topic-agnostic" classifier design occurs downstream of topic-aware dataset design. ④ Epidemiological models of information spread make strong assumptions about opinion formation and user behaviors; feature sets reflect these *a priori* notions. ⑤ Though network-scoped methods are less susceptible to content-based dependencies than are content-aware methods, they cannot infer intent or authenticity of the behaviors they detect.

12. Some academics have argued that the disease metaphor for misinformation promotes an overly simplistic model of information spread and opinion formation [399], [400].

13. In fact, the landmark paper by Douceur that characterized Sybil attacks stated that such attacks cannot be prevented unless special assumptions are made about account behaviors [401].

4.5. Websites

Methods within this scope apply credibility, factuality, or (political) bias scores to whole news sites; authors claim that site-wide labels can be used to quickly infer the quality of individual news articles produced by these sites [1], [109], [302], [316]. As with article-scoped methods, we noted intervention fit issues at the whole website scope. Asr et al. found that whole-source labels were insufficient proxies for the truthfulness of single news articles, and elided subject-specific variations in reporting quality [287]. (We discuss dataset distribution in greater depth in *Datasets*.)

① **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. [109] **(W.i)** *detect suspicious domain semantics* (in essence, a text classification task on URLs); Ribeiro et al. [254] **(W.ii)** *infer site bias from site visitor demographics*; Castillo et al. [253] **(W.iii)** *detect suspicious ad interfaces and markup features*; Baly et al. [43], [316] and Hounsel et al. [1] 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).

② **Datasets.** For both training and testing, all methods scoped to whole website detection rely on published lists of websites with accompanying credibility scores [1], [43], [302], [316]; common reference sites include Media Bias/Fact Check, Snopes, and FactCheck.org [298], [311], [402]. As discussed in Section 4.1, however, these references do not have uniform taxonomies for classifying site credibility. Additionally, pre-labeled lists are 1) **(2.i)** biased towards older, more visible real news and fake news outlets [298], 2) **(2.iii)** are restricted to specific information domains [43], or 3) **(2.i)** include inactive websites within their labeled datasets [311].[14] We note papers that perform website 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 [302]. This constitutes a serious source of **(2.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 an active website. Hounsel et al., for instance, 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 those inactive websites [1].

14. The median lifespan of a set of 283 misinformation news sites is 4 years, per a survey conducted by Chalkiadakis et al. in 2021 [114].

③ **Model selection.** Four of the seven methods we reviewed within this scope employed **(3.ii)** *SVM classifiers*; in two of those cases, SVM outperformed other, more complex unsupervised models [287], [302]. These results accord with our earlier observation, in Section 4.2, that SVM classifiers are comparatively fairly performant on two-way classification tasks.

④ **Feature selection.** Works within this scope employ multi-modal feature sets comprising a mix of **(4.i)** *textual*, **(4.ii)** *network-based*, and **(4.iv)** *infrastructural signals*: for instance, Baly et al. [316] consider network traffic, URL semantics, and site contents; and Hounsel et al. [1] 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.

⑤ **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 **(5.ii)** *political news websites only*, and their credibility labels are strongly correlated with political bias scores. Hounsel et al. perform **(5.i)** *testing in real time*—one of the few studies we reviewed, and one of two studies at this scope, that did so [1], [109]. 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) [1], [109].

Takeaways: ① Methods claiming to classify news sites generally reduce this task to simpler, unimodal ones, such as URL classification. ② The works we review do not consider shifts in feature distribution over time. ③ We note that SVM classifiers perform well on two-way classification tasks, and even outperform more sophisticated unsupervised models. ④ Feature normalization largely undiscussed. ⑤ 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. [141] (Section 5.1) train a neural network on corpora of true and misinformative news articles; Saeed et al. [361] (Section 5.2) use published lists of trolls to infer the presence of other troll accounts on Reddit; and Baly et al. [43] (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 develop updated datasets to test method performance on diverse information domains.

Ethical Considerations. Datasets used for this work were already publicly available or were obtained with permission from study authors. Our outreach study was approved by our institutional review board.

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 [141].

Dataset (size)	Features	Acc.	FPR	FNR
ISOT (45,000)	`original`	0.995	0.00	0.00
	`scrubbed`	0.987	0.01	0.00
FA-KES (804)	`original`	0.521	0.151	0.843
Reuters (100)	`original`	0.727	0.71	–
	`modified`	0.507	0.02	1
NYTimes (100)	`original`	0.673	0.705	–
	`modified`	0.492	0.735	0.286
ChatGPT (250)	`original`	0.939	0.02	–

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 [141]. 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) [312], [313]. 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 Nasir et al. 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 0.727 accuracy on Reuters-`original` and 0.673 accuracy on NYTimes-`original`; this difference is not significant ($p > 0.05$). Additionally, the classifier had about 0.50 (random) accuracy on both modified datasets. The classifier's false positive and false negative rates on modified and original Reuters and NYTimes corpora tell a more interesting story, however: Overwhelmingly, `false` NYTimes articles were classified as `true` (FPR = 0.02 and 0.735 for Reuters-`modified` and NYTimes-`modified`, respectively), while all `true` Reuters articles were classified as `false` (FNR = 1 and 0.286 for Reuters-`modified` and NYTimes-`modified`, respectively). These results indicate that the classifier was significantly more conservative in its assignment of `true` labels for the Reuters dataset than it was for the NYTimes dataset; additionally, *within* each modified corpus, the classifier did not effectively differentiate between true and untrue news articles. We did not investigate the stylistic features of each corpus that might have produced (anti-)conservative label distributions; for this work, however, it is sufficient to note that classifier performance in an adversarial setting was no better than random, and that house style appears to have a significant impact on classifier sensitivity to misinformative texts.

5.2. Detection of suspicious accounts

We reproduce results from a study published in 2022 by Saeed et al. [361]. 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) [361].

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 [421]. We were able to replicate original paper results using these materials. Original account handles were anonymized; as such, we were unable to manually 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) [367] or the nature of the information they spread (this might be political misinformation, ads, or anything in between) [422]; as such, the clearest definition that Saeed et al. implicitly offer 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 message threads as these accounts. The authors note in their own work 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 (we note in our discussion of *Account*-scoped data curation practices that the process of identifying these seed accounts is actually a fairly manual one).

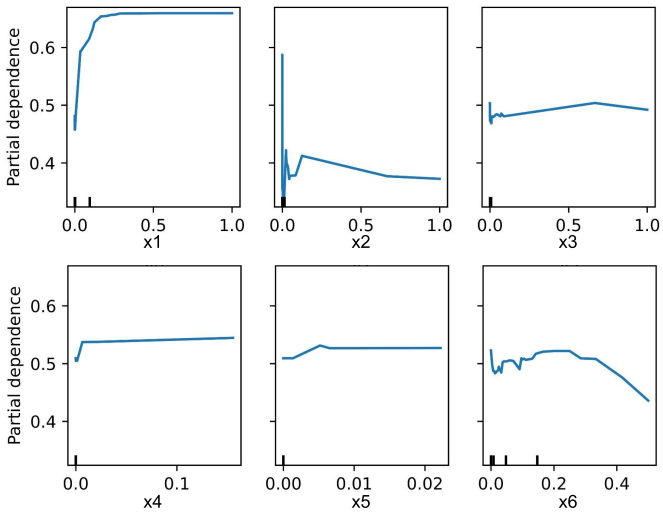

Figure 1. Partial dependence plots for each TrollMagnifier feature. Respectively, x1 = "comments on posts that trolls commented on," x2 = "comments on posts that trolls started," x3 = "direct comment in reply to troll post," x4 = "threaded comment in reply to troll comments on a troll post," x5 = "same title post as troll," x6 = "same title post as troll."

Implied versus actual data dimensionality. While the original TrollMagnifier paper strongly implied that the proposed method would leverage networked behaviors to identify 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. Additionally, account names were not available, which disallowed 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; two features, x2 and x6, had performance below 0.5). 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—no better than random—suggests that manual classification might be as effective as (or even more effective than) an automated approach that only detects a content-agnostic heuristic.

5.3. Detection of suspicious websites

We reproduce results from a study published in 2018 by Baly et al [43]. 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 [311]. 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 [43]. 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 3. 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 3), suggesting that the full-site classifier of Baly et al. was effectively a text content classifier.

6. DISCUSSION

We focus our discussion of results on those issues identified by our literature review and investigated in greater depth in our replication analyses. Additionally, we provide recommendations for evaluating ML-driven trust and safety interventions and link these recommendations to major takeaways of the present study.

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. Additionally, in the absence of current or complete social media post data, we argued that most network-based methods are poorly-equipped to perform early detection of false rumors—this is by design, as the propagation patterns these methods purport to detect are frequently only noticeable after virality.

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 employing neural nets or other forms of unsupervised learning—did not disclose feature sets retrieved by their methods. Model explainability for these approaches—if available at all—was very limited. For methods employing multimodal feature sets, temporal consistency across disparate feature sets was poorly controlled, and feature normalization was often questionable.

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. In fact, we were able to reproduce and replicate results on published *and* current data for only one of our replication studies. In that single case, we found that the article-scoped method performed no better than random (0.50 accuracy) on a current, mixed-domain dataset (Section 5.1). Additionally, for the significant number of methods that purported to detect false rumors on Twitter, the shutdown of the Twitter Academic API poses an existential risk to reproducibility of results and usefulness of dehydrated datasets.

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. Though we did not attempt to test detection methods on out-of-scope data—e.g., we did not test claim-scoped methods on datasets of whole news articles—it seems unlikely that these methods would fare well in out-of-scope contexts, particularly in light of their already-poor performance on in-scope datasets.

6.1. Future Directions for Research

From our analysis of academic and commercial approaches to detection tasks, we offer the following suggestions for possible future directions for research.

What works? Though the majority of this project has been devoted to critique of an existing body of work, we choose to highlight here a number of approaches that we found especially promising. In general, we found methods that 1) clearly described their detection targets and heuristics and 2) provided context for how their detection outcomes might be employed in service of fact-checking work the most methodologically sound [17], [108], [130]. As we note in our survey of commercial fact-checking services, the preponderance of high-volume checking work is *hybrid*, with (e.g.) potentially problematic content surfaced by automated means and ultimately vetted by human moderators. We recommend that future research efforts be devoted to understanding how best to facilitate collaboration between human and machine moderation techniques—in particular, how automated methods can support human moderators.

***Definitive*, rather than *descriptive*, targets.** We discuss the knock-on effects of a basic lack of agreement about what does/n't constitute misinformation throughout the main body of this work: In the absence of a well-defined detection target, detection methods detect proxy signals instead. Though the task of defining misinformation should likely *not* be left solely to computer scientists, we can nonetheless set our sights on more manageable targets, such as the three below.

Nuanced detection tasks. We note that, in general, the detection methods we consider in this work elide subtleties that are particular to the medium under consideration: for instance, article-level detection methods generally apply broad 'true'/'false' classifications to a text under analysis where only single sentences or turns of phrase might be slightly inaccurate. Additionally, the language signals that most methods expressly 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 via indirect means; no works within our literature corpus expressly targeted these forms of language, however.

Multimedia detection. The detection methods we consider for this work are solely text-based; we note, however, that a good deal of social media misinformation is transmitted via audio, image, and video. This problem has become especially salient in the current year, as election misinformation transmitted via audiovisual means is rampant on messaging apps worldwide.

Designing for evasion resistance. The AI/NLP literature we review aligns with the current dogma in those subfields, which emphasizes generalizability of single models to all detection settings. We argue that a security framework for method development—wherein attempts at evasion and attacks are assumed, and iterative development is required to stay a step ahead of adversaries—is a more practical approach to misinformation detection research, and will yield more durable outcomes in the long run.

Acknowledgments

The authors would like to thank [redacted] for assisting with paper coding; [redacted] and [redacted] for reviewing early drafts of this manuscript; and [redacted] for retrieving Twitter data used for a replication study.

Availability

Our online SI is available in the following anonymous repo: https://anonymous.4open.science/r/sok_misinformation-B297/).

7. App: Full taxonomy of targets and critiques

1) Targets
   - Ⓒ Claims
     i. Content-based detection via distance calculations on semantic embeddings
     ii. Content-based detection via search on knowledge graph topology
     iii. "Checkability" or "checkworthiness"
   - Ⓐ Articles
     i. Syntactic and stylometric signals, including genre and sentiment
     ii. Topic-aware detection of stance and relevance to known rumoring topics
   - Ⓤ Users
     i. Account metadata (bios, images, account age)
     ii. Single account behaviors (comments on posts, published posts)
   - Ⓝ Networks
     i. Propagation patterns across social graphs
     ii. Timestamped records of user-user interactions
   - Ⓦ Websites
     i. Text and URL/domain semantics
     ii. Site visitor demographics
     iii. Suspicious UI elements
     iv. Hosting infrastructure (DNS certificate, site age)
2) Datasets
   i. Dataset age (is it current or out of date?)
   ii. Leakage (evidence of temporal leakage, feature leakage)
   iii. Data dependencies (are they accounted/controlled for?)
   iv. Availability of information required to reproduce or reconstruct similar datasets (was non-public information required to produce ground-truth training sets?)
   v. Availability of original data
3) Models
   i Distance calculations on semantic embeddings
   ii. "Traditional" ML (SVM, RF, DT)
   iii. Deep learning (CNN, LSTM, GRU)
   iv. Graph cut algorithms
   v Stacked ensemble classifiers
   vi. Graph clustering algorithms
4) Features
   i. Textual
   ii. Network-based
   iii. Author-, user- or source-based
   iv. Infrastructural
5) Evaluation
   i. Testing in real time
   ii. Generalizability of approach
   iii. Evasion-resistance
   iv. Distributional shifts in training or test data
   v. Implications of high false-positive/false-negative rates

References

[1] A. Hounsel, J. Holland, B. Kaiser, K. Borgolte, N. Feamster, and J. Mayer, 'Identifying Disinformation Websites Using Infrastructure Features', FOCI, 2020.

[2] M. Wei, J. Mink, Y. Eiger, Y. Kohno, E. Redmiles, and F. Roesner, 'SoK (or SoLK?): On the Quantitative Study of Sociodemographic Factors and Computer Security Behaviors', USENIX Security. 2024.

[3] N. Warford, T. Matthews, K. Yang, O. Akgul, S. Consolvo, P. Kelley, N. Malkin, M. Mazurek, M. Sleeper, and K. Thomas, 'SoK: A Framework for Unifying At-Risk User Research,' IEEE S&P. 2022.

[4] S. Scheffler and J. Mayer, 'SoK: Content Moderation for End-to-End Encryption', Proceedings on Privacy Enhancing Technologies. 2023.

[5] S. Helmstetter and H. Paulheim, 'Weakly Supervised Learning for Fake News Detection on Twitter', IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Aug. 2018.

[6] S. Vosoughi, D. Roy, and S. Aral, 'The spread of true and false news online', Science, vol. 359, no. 6380, pp. 1146–1151, Mar. 2018.

[7] J. Maddock, K. Starbird, H. Al-Hassani, D. E. Sandoval, M. Orand, and R. M. Mason, 'Characterizing Online Rumoring Behavior Using Multi-Dimensional Signatures', CSCW, Mar. 2015.

[8] A. Friggeri, L. A. Adamic, D. Eckles, and J. Cheng, 'Rumor Cascades', Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014, vol. 8, pp. 101–110, 05 2014.

[9] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, '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), 2018, pp. 809–819.

[10] S. Kumar, R. West, and J. Leskovec, 'Disinformation on the Web: Impact, Characteristics, and Detection of Wikipedia Hoaxes', in Proceedings of the 25th International Conference on World Wide Web, Montréal, Québec, Canada, 2016, pp. 591–602.

[11] V. Qazvinian, E. Rosengren, D. R. Radev, and Q. Mei, 'Rumor has it: Identifying Misinformation in Microblogs', in Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, 2011, pp. 1589–1599.

[12] P. R. Center, 'Social Media Fact Sheet', Pew Research Center, Sep. 2022.

[13] P. Association, 'Blue ticks for all: Twitter allows users to apply to be verified', The Guardian, Jul. 2016.

[14] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. Choi, '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, 2017, pp. 2931–2937.

[15] J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec, 'Can Cascades Be Predicted?', in Proceedings of the 23rd International Conference on World Wide Web, 2014, pp. 925–936.

[16] T. Schuster, R. Schuster, D. J. Shah, and R. Barzilay, 'Limitations of stylometry for detecting machine-generated fake news', Computational Linguistics, vol. 46, pp. 499–510, Jun. 2020.

[17] M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff, and B. Stein, '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), 2018, pp. 231–240.

[18] A. Magdy and N. Wanas, 'Web-Based Statistical Fact Checking of Textual Documents', in Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents, Toronto, ON, Canada, 2010, pp. 103–110.

[19] R. Santos, G. Pedro, S. Leal, O. Vale, T. Pardo, K. Bontcheva, and C. Scarton, 'Measuring the Impact of Readability Features in Fake News Detection', Proceedings of the Twelfth Language Resources and Evaluation Conference, 2020.

[20] L. Bozarth, A. Saraf, and C. Budak, 'Higher Ground? How Groundtruth Labeling Impacts Our Understanding of Fake News about the 2016 U.S. Presidential Nominees', Proceedings of the International AAAI Conference on Web and Social Media, vol. 46, no. 2, pp. 48–59, May 2020.

[21] G. Santia, M. Mujib, and J. Williams, 'Detecting Social Bots on Facebook in an Information Veracity Context', ICWSM. 2019.

[22] G.L. Ciampaglia, P. Shiralkar, LM Rocha, J. Bollen, F. Menczer, and A. Flammini, 'Computational Fact Checking from Knowledge Networks', PLoS ONE, vol. 10, no. 6, Jun. 2015.

[23] E. C. Tandoc, Z. W. Lim, and R. Ling, 'Defining "Fake News": A typology of scholarly definitions', Digital Journalism, vol. 6, no. 2, pp. 137–153, Feb. 2018.

[24] B. D. Horne and S. 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, 2017.

[25] L. Haiden and J. Althuis, 'The Definitional Challenges of Fake News', in SBP-BRiMS 18, 2018.

[26] A. Gelfert, 'Fake News: A Definition', Informal Logic, vol. 38, no. 1, pp. 84–117, Mar. 2018.

[27] D. Klein and J. Wueller, 'Fake News: A Legal Perspective', Social Science Research Network, Rochester, NY, Mar. 2017.

[28] F. Yang, Y. Liu, X. Yu, and M. Yang, 'Automatic detection of rumor on Sina Weibo', in Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics - MDS '12, 2012, pp. 1–7.

[29] S. Kumar, R. West, and J. 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, 2016, pp. 591–602.

[30] J. Ma, W. Gao, and K.-F. Wong, 'Detect Rumors in Microblog Posts Using Propagation Structure via Kernel Learning', in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017, pp. 708–717.

[31] R. Mihalcea and C. Strapparava, 'The lie detector: explorations in the automatic recognition of deceptive language', in Proceedings of the ACL-IJCNLP 2009 Conference Short Papers on - ACL-IJCNLP '09, 2009, p. 309.

[32] S. Jiang and C. Wilson, 'Linguistic Signals under Misinformation and Fact-Checking: Evidence from User Comments on Social Media', Proceedings of the ACM on Human-Computer Interaction, vol. 2, no. CSCW, pp. 1–23, Nov. 2018.

[33] J. Norregaard, B. D. Horne, and S. Adali, 'NELA-GT-2018: A Large Multi-Labelled News Dataset for The Study of Misinformation in News Articles', arXiv:1904. 01546 [cs], Apr. 2019.

[34] J. Kim, B. Tabibian, A. Oh, B. Schölkopf, and M. Gomez-Rodriguez, 'Leveraging the Crowd to Detect and Reduce the Spread of Fake News and Misinformation', in Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining - WSDM '18, 2018, pp. 324–332.

[35] W. Y. Wang, '"Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection', in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2017, pp. 422–426.

[36] K. Shu, S. Wang, and H. Liu, 'Understanding User Profiles on Social Media for Fake News Detection', in 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), 2018, pp. 430–435.

[37] N. Ruchansky, S. Seo, and Y. Liu, 'CSI: A Hybrid Deep Model for Fake News Detection', in Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, 2017, pp. 797–806.

[38] S. Leonardi, G. Rizzo, and M. Morisio, 'Automated Classification of Fake News Spreaders to Break the Misinformation Chain', Information. June 2021.

[39] F. Qian, C. Gong, K. Sharma, and Y. Liu, 'Neural User Response Generator: Fake News Detection with Collective User Intelligence', in Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, 2018, pp. 3834–3840.

[40] Y. Liu and Y.-F. B. Wu, 'Early Detection of Fake News on Social Media Through Propagation Path Classification with Recurrent and Convolutional Networks', p. 8.

[41] K. Shu, S. Wang, and H. Liu, 'Beyond News Contents: The Role of Social Context for Fake News Detection', in Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining - WSDM '19, 2019, pp. 312–320.

[42] Y. Chen, N. J. Conroy, and V. L. Rubin, 'Misleading Online Content: Recognizing Clickbait As "False News"', in Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection, 2015, pp. 15–19.

[43] R. Baly, G. Karadzhov, D. Alexandrov, J. Glass, and P. Nakov, 'Predicting Factuality of Reporting and Bias of News Media Sources', Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 3528–3539, 2018.

[44] N. Dhamani et al., 'Using Deep Networks and Transfer Learning to Address Disinformation', arXiv:1905. 10412 [cs], May 2019.

[45] X. Zhou and R. Zafarani, 'Fake News: A Survey of Research, Detection Methods, and Opportunities', arXiv:1812. 00315 [cs], Dec. 2018.

[46] D. Sáez-Trumper, 'Fake tweet buster: a webtool to identify users promoting fake news ontwitter', in HT, 2014.

[47] N. Vo and K. Lee, 'The Rise of Guardians: Fact-checking URL Recommendation to Combat Fake News', in The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval - SIGIR '18, 2018, pp. 275–284.

[48] L. Borges, B. Martins, and P. Calado, 'Combining Similarity Features and Deep Representation Learning for Stance Detection in the Context of Checking Fake News', arXiv:1811. 00706 [cs, stat], Nov. 2018.

[49] J. Amador, A. Oehmichen, and M. Molina-Solana, 'Characterizing Political Fake News in Twitter by its Meta-Data', arXiv:1712. 05999 [cs, stat], Dec. 2017.

[50] C. Buntain and J. Golbeck, 'Automatically Identifying Fake News in Popular Twitter Threads', 2017 IEEE International Conference on Smart Cloud (SmartCloud), pp. 208–215, Nov. 2017.

[51] Y. Yang, L. Zheng, J. Zhang, Q. Cui, Z. Li, and P. S. Yu, 'TI-CNN: Convolutional Neural Networks for Fake News Detection', arXiv:1806. 00749 [cs], Jun. 2018.

[52] M. Tosik, A. Mallia, and K. Gangopadhyay, 'Debunking Fake News One Feature at a Time', arXiv:1808. 02831 [cs], Aug. 2018.

[53] X. Zhou, A. Jain, V. V. Phoha, and R. Zafarani, 'Fake News Early Detection: A Theory-driven Model', arXiv:1904. 11679 [cs], Apr. 2019.

[54] N. Hassan, F. Arslan, C. Li, and M. Tremayne, 'Toward Automated Fact-Checking: Detecting Check-worthy Factual Claims by ClaimBuster', in Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '17, 2017, pp. 1803–1812.

[55] P. Bourgonje, J. Moreno Schneider, and G. Rehm, 'From Clickbait to Fake News Detection: An Approach based on Detecting the Stance of Headlines to Articles', in Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism, 2017, pp. 84–89.

[56] P. Shiralkar, A. Flammini, F. Menczer, and G. Luca Ciampaglia, 'Finding Streams in Knowledge Graphs to Support Fact Checking', 2017, pp. 859–864.

[57] S. Yang, K. Shu, S. Wang, R. Gu, F. Wu, and H. Liu, 'Unsupervised Fake News Detection on Social Media: A Generative Approach', 2019.

[58] M. Della Vedova, E. Tacchini, S. Moret, G. Ballarin, M. Di Pierro, and L. de Alfaro, 'Automatic Online Fake News Detection Combining Content and Social Signals', 2018.

[59] K. Popat, S. Mukherjee, A. Yates, and G. Weikum, 'DeClarE: Debunking Fake News and False Claims using Evidence-Aware Deep Learning', in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 22–32.

[60] E. Tacchini, G. Ballarin, M. L. Della Vedova, S. Moret, and L. de Alfaro, 'Some Like it Hoax: Automated Fake News Detection in Social Networks', arXiv:1704. 07506 [cs], Apr. 2017.

[61] E. Tacchini, G. Ballarin, M. L. D. Vedova, S. Moret, and L. de Alfaro, 'Automated Fake News Detection in Social Networks', p. 15, 2017.

[62] S. Krishnan and M. Chen, 'Cloud-Based System for Fake Tweet Identification', in 2019 IFIP/IEEE Symposium on Integrated Network and Service Management (IM), 2019, pp. 720–721.

[63] V. Pérez-Rosas, B. Kleinberg, A. Lefevre, and R. Mihalcea, 'Automatic Detection of Fake News', arXiv:1708. 07104 [cs], Aug. 2017.

[64] S. Volkova, E. Ayton, D. L. Arendt, Z. Huang, and B. Hutchinson, 'Explaining Multimodal Deceptive News Prediction Models', 2019, p. 4.

[65] S. Shah and M. Goyal, 'Anomaly Detection in Social Media Using Recurrent Neural Network', in Computational Science – ICCS 2019, 2019, pp. 74–83.

[66] A. Addawood, A. Badawy, K. Lerman, and E. Ferrara, 'Linguistic Cues to Deception: Identifying Political Trolls on Social Media', in Proceedings of the Thirteenth International AAAI Conference on Web and Social Media, 2019, p. 11.

[67] S. S. Park and K. C. Lee, 'A Comparative Study of Text analysis and Network embedding Methods for Effective Fake News Detection', Journal of Digital Convergence, vol. 17, no. 5, pp. 137–143, May 2019.

[68] A. M. P. Braşoveanu and R. Andonie, 'Semantic Fake News Detection: A Machine Learning Perspective', in Advances in Computational Intelligence, 2019, pp. 656–667.

[69] D. M. Nguyen, T. H. Do, R. Calderbank, and N. Deligiannis, '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), 2019, pp. 1391–1400.

[70] L. Burbach, P. Halbach, M. Ziefle, and A. Calero Valdez, 'Who Shares Fake News in Online Social Networks?', in Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization - UMAP '19, 2019, pp. 234–242.

[71] M. Ramezani, M. Rafiei, S. Omranpour, and H. R. Rabiee, 'News Labeling as Early as Possible: Real or Fake?', arXiv:1906. 03423 [cs], Jun. 2019.

[72] X. Zhou and R. Zafarani, 'Network-based Fake News Detection: A Pattern-driven Approach', arXiv:1906. 04210 [cs], Jun. 2019.

[73] C. Zhang, A. Gupta, C. Kauten, A. V. Deokar, and X. Qin, 'Detecting Fake News for Reducing Misinformation Risks Using Analytics Approaches', European Journal of Operational Research, Jun. 2019.

[74] X. Dong, U. Victor, S. Chowdhury, and L. Qian, 'Deep Two-path Semi-supervised Learning for Fake News Detection', arXiv:1906. 05659 [cs], Jun. 2019.

[75] J. Sánchez-Junquera, P. Rosso, M. Montes-y-Gómez, and S. P. Ponzetto, 'Unmasking Bias in News', arXiv:1906. 04836 [cs], Jun. 2019.

[76] Y. Wang, H. Han, Y. Ding, X. Wang, and Q. Liao, 'Learning Contextual Features with Multi-head Self-attention for Fake News Detection', in Cognitive Computing – ICCC 2019, 2019, pp. 132–142.

[77] T. Saikh, A. Anand, A. Ekbal, and P. Bhattacharyya, 'A Novel Approach Towards Fake News Detection: Deep Learning Augmented with Textual Entailment Features', in Natural Language Processing and Information Systems, 2019, pp. 345–358.

[78] Anoop K, Deepak P, and Lajish VL. 'Emotion Cognizance Improves Fake News Identification', arXiv:1906. 10365 [cs], Jun. 2019.

[79] M. Lim and S. Park, 'A Study on the Preemptive Measure for Fake News Eradication Using Data Mining Algorithms: Focused on the M Online Community Postings', Journal of Information Technology Services, vol. 18, no. 1, pp. 219–234, 2019.

[80] A. Ahmad, J. Webb, K. C. Desouza, and J. Boorman, 'Strategically-Motivated Advanced Persistent Threat: Definition, Process, Tactics and a Disinformation Model of Counterattack', Computers & Security, Jul. 2019.

[81] M. Seref and O. Seref, 'Rhetoric Mining for Fake News: Identifying Moves of Persuasion and Disinformation', AMCIS 2019 Proceedings, Jul. 2019.

[82] N. Papernot, P. McDaniel, A. Sinha, and M. Wellman, 'SoK: Security and Privacy in Machine Learning', IEEE Explore, 2018.

[83] K. Shu and H. Liu, 'Detecting Fake News on Social Media', Synthesis Lectures on Data Mining and Knowledge Discovery, vol. 11, no. 3, pp. 1–129, Jul. 2019.

[84] A. Debnath, R. Rahman, M. Mofijul Islam, and M. Abdur Razzaque, 'A Hierarchical Learning Model for Claim Validation', in Proceedings of International Joint Conference on Computational Intelligence, 2020, pp. 431–441.

[85] J. Reyes and L. Palafox, 'Detection of Fake News based on readability', p. 6.

[86] M. A. Stefanone, M. Vollmer, and J. 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, 2019, pp. 136–147.

[87] A. Heydari, J. Zhang, S. Appel, X. Wu, and P. G. Ranade, 'YouTube Chatter: Understanding Online Comments Discourse on Misinformative and Political YouTube Videos', p. 32.

[88] J. Kapusta, Ľ. Benko, and M. Munk, 'Fake News Identification Based on Sentiment and Frequency Analysis', in Innovation in Information Systems and Technologies to Support Learning Research, 2020, pp. 400–409.

[89] S. Hosseinimotlagh and E. E. Papalexakis, 'Unsupervised Content-Based Identification of Fake News Articles with Tensor Decomposition Ensembles', p. 8.

[90] K. Popat, '"Credibility Analysis of Textual Claims with Explainable Evidence"', p. 134.

[91] T. Hamdi, H. Slimi, I. Bounhas, and Y. Slimani, 'A Hybrid Approach for Fake News Detection in Twitter Based on User Features and Graph Embedding', in Distributed Computing and Internet Technology, 2020, pp. 266–280.

[92] D. Pacheco, P.-M. Hui, C. Torres-Lugo, B. T. Truong, A. Flammini, and F. Menczer, 'Uncovering Coordinated Networks on Social Media', arXiv:2001. 05658 [physics], Jan. 2020.

[93] R. K. Kaliyar, A. Goswami, P. Narang, and S. Sinha, 'FNDNet-A Deep Convolutional Neural Network for Fake News Detection', Cognitive Systems Research, Jan. 2020.

[94] N. X. Nyow and H. N. Chua, 'Detecting Fake News with Tweets' Properties', in 2019 IEEE Conference on Application, Information and Network Security (AINS), 2019, pp. 24–29.

[95] F. Pierri, C. Piccardi, and S. Ceri, 'Topology comparison of Twitter diffusion networks effectively reveals misleading information', Scientific Reports, vol. 10, no. 1, pp. 1–9, Jan. 2020.

[96] H. Reddy, N. Raj, M. Gala, and A. Basava, 'Text-mining-based Fake News Detection Using Ensemble Methods', International Journal of Automation and Computing, Feb. 2020.

[97] J. Cao, P. Qi, Q. Sheng, T. Yang, J. Guo, and J. Li, 'Exploring the Role of Visual Content in Fake News Detection', arXiv:2003. 05096 [cs], Mar. 2020.

[98] K. Shu et al., 'Leveraging Multi-Source Weak Social Supervision for Early Detection of Fake News', arXiv:2004. 01732 [cs, stat], Apr. 2020.

[99] H. Kudarvalli and J. Fiaidhi, 'Detecting Fake News using Machine Learning Algorithms', Apr. 2020.

[100] D. Freelon, M. Bossetta, C. Wells, J. Lukito, Y. Xia, and K. Adams, 'Black Trolls Matter: Racial and Ideological Asymmetries in Social Media Disinformation', Social Science Computer Review, p. 0894439320914853, Apr. 2020.

[101] Q. Cao, M. Sirivianos, X. Yang, and T. Pregueiro, 'Aiding the detection of fake accounts in large scale social online services', USENIX NSDI 2012.

[102] A. Uppal, V. Sachdeva, and S. Sharma, 'Fake news detection using discourse segment structure analysis', in 2020 10th International Conference on Cloud Computing, Data Science Engineering (Confluence), 2020, pp. 751–756.

[103] M. Previti, V. Rodriguez-Fernandez, D. Camacho, V. Carchiolo, and M. Malgeri, 'Fake News Detection Using Time Series and User Features Classification', in Applications of Evolutionary Computation, vol. 12104, P. A. Castillo, J. L. Jiménez Laredo, and F. Fernández de Vega, Eds. Cham: Springer International Publishing, 2020, pp. 339–353.

[104] Y. Zhou, Y. Zhang, and J. Yao, 'Satirical News Detection with Semantic Feature Extraction and Game-theoretic Rough Sets', arXiv:2004. 03788 [cs], Apr. 2020.

[105] P. Atanasova, J. G. Simonsen, C. Lioma, and I. Augenstein, 'Generating Fact Checking Explanations', arXiv:2004. 05773 [cs], Apr. 2020.

[106] L. Tian, X. Zhang, Y. Wang, and H. Liu, 'Early Detection of Rumours on Twitter via Stance Transfer Learning', in Advances in Information Retrieval, 2020, pp. 575–588.

[107] J. George, S. M. 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), 2020, pp. 1–6.

[108] Y. S. Kartal, B. Guvenen, and M. Kutlu, 'Too Many Claims to Fact-Check: Prioritizing Political Claims Based on Check-Worthiness', arXiv:2004. 08166 [cs], Apr. 2020.

[109] Z. Chen and J. Freire, 'Proactive Discovery of Fake News Domains from Real-Time Social Media Feeds', in Companion Proceedings of the Web Conference 2020, 2020, pp. 584–592.

[110] M. L. Gordon, K. Zhou, K. Patel, T. Hashimoto, and M. S. Bernstein, 'The Disagreement Deconvolution: Bringing Machine Learning Performance Metrics In Line With Reality', p. 14, 2021.

[111] A. Silva, L. Luo, S. Karunasekera, and C. Leckie, 'Embracing Domain Differences in Fake News: Cross-domain Fake News Detection using Multimodal Data', arXiv:2102. 06314 [cs], Feb. 2021.

[112] 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), 2020, pp. 775–781.

[113] P. Carvalho et al., 'Assessing News Credibility: Misinformation Content Indicators', In Review, Mar. 2021.

[114] M. Chalkiadakis, A. Kornilakis, P. Papadopoulos, E. P. Markatos, and N. Kourtellis, 'The Rise and Fall of Fake News sites: A Traffic Analysis', arXiv:2103. 09258 [cs], Mar. 2021.

[115] B. Krämer, 'Stop studying "fake news" (we can still fight against disinformation in the media)', Studies in Communication and Media, vol. 10, no. 1, pp. 6–30, 2021.

[116] N. Lee et al., 'On Unifying Misinformation Detection', arXiv:2104. 05243 [cs], Apr. 2021.

[117] K. Pelrine, J. Danovitch, and R. Rabbany, 'The Surprising Performance of Simple Baselines for Misinformation Detection', arXiv:2104. 06952 [cs], Apr. 2021.

[118] B. Jabiyev, K. Onarlioglu, S. Pehlivanoglu, and E. Kirda, 'FADE: Detecting Fake News Articles on the Web', p. 10, 2021.

[119] J. Simko et al., 'Towards Continuous Automatic Audits of Social Media Adaptive Behavior and its Role in Misinformation Spreading', in Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization, 2021, pp. 411–414.

[120] Y. Yang, T. Davis, and M. Hindman, 'Visual Misinformation on Facebook', p. 8.

[121] K. Roitero et al., 'Can the Crowd Judge Truthfulness? A Longitudinal Study on Recent Misinformation about COVID-19', arXiv:2107. 11755 [cs], Jul. 2021.

[122] K. Roitero et al., 'The COVID-19 Infodemic: Can the Crowd Judge Recent Misinformation Objectively?', Proceedings of the 29th ACM International Conference on Information & Knowledge Management, pp. 1305–1314, Oct. 2020.

[123] M. Soprano et al., 'The Many Dimensions of Truthfulness: Crowdsourcing Misinformation Assessments on a Multidimensional Scale', arXiv:2108. 01222 [cs], Aug. 2021.

[124] A. Kim, P. Moravec, and A. Dennis, 'Behind the Stars: The Effects of News Source Ratings on Fake News in Social Media', SSRN Electronic Journal, Jan. 2017.

[125] P. Resnick, A. Alfayez, J. Im, and E. Gilbert, 'Informed Crowds Can Effectively Identify Misinformation', arXiv:2108. 07898 [cs], Aug. 2021.

[126] P. Juneja and T. Mitra, 'Auditing E-Commerce Platforms for Algorithmically Curated Vaccine Misinformation', in Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, New York, NY, USA: Association for Computing Machinery, 2021, pp. 1–27.

[127] M. Janicka, M. Pszona, and A. Wawer, 'Cross-Domain Failures of Fake News Detection', Computación y Sistemas, vol. 23, no. 3, Oct. 2019.

[128] Z. Kou, L. Shang, Y. Zhang, and D. Wang, 'HC-COVID: A Hierarchical Crowdsource Knowledge Graph Approach to Explainable COVID-19 Misinformation Detection', Proceedings of the ACM on Human-Computer Interaction, vol. 6, no. GROUP, p. 36:1-36:25, Jan. 2022.

[129] R. 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, 2022, pp. 1–7.

[130] N. Hassan et al., 'Data in, fact out: automated monitoring of facts by FactWatcher', Proceedings of the VLDB Endowment, vol. 7, no. 13, pp. 1557–1560, Aug. 2014.

[131] J. Thorne and A. Vlachos, 'Automated Fact Checking: Task Formulations, Methods and Future Directions', p. 14.

[132] F. Jin, E. Dougherty, P. Saraf, Y. Cao, and N. Ramakrishnan, 'Epidemiological modeling of news and rumors on Twitter', in Proceedings of the 7th Workshop on Social Network Mining and Analysis, 2013, pp. 1–9.

[133] L. Zeng, K. Starbird, and E. Spiro, '#Unconfirmed: Classifying Rumor Stance in Crisis-Related Social Media Messages', Proceedings of the International AAAI Conference on Web and Social Media, vol. 10, no. 1, pp. 747–750, 2016.

[134] G. Eady, T. Paskhalis, J. Zilinsky, R. Bonneau, J. Nagler, and J. A. Tucker, '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', Nature Communications, vol. 14, no. 1, p. 62, Jan. 2023.

[135] A. T. Kabakuş and M. Şimşek, 'An Analysis of the Characteristics of Verified Twitter Users', Sakarya University Journal of Computer and Information Sciences, vol. 2, no. 3, pp. 180–186, Dec. 2019.

[136] A. Hearn, 'Verified: Self-presentation, identity management, and selfhood in the age of big data', Popular Communication, vol. 15, no. 2, pp. 62–77, Apr. 2017.

[137] 'Twitter Verification requirements - how to get the blue check'. Twitter Help Center.

[138] 'Prolific · Quickly find research participants you can trust'. Prolific.co.

[139] A. Arif, K. Shanahan, F.-J. Chou, Y. Dosouto, K. Starbird, and E. S. Spiro, '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, 2016, pp. 466–477.

[140] S. van der Linden, 'Misinformation: susceptibility, spread, and interventions to immunize the public', Nature Medicine, vol. 28, no. 3, pp. 460–467, Mar. 2022.

[141] J. A. Nasir, O. S. Khan, and I. Varlamis, 'Fake news detection: A hybrid CNN-RNN based deep learning approach', International Journal of Information Management Data Insights, vol. 1, no. 1, p. 100007, Apr. 2021.

[142] M. Glockner, Y. Hou, and I. Gurevych, 'Missing Counter-Evidence Renders NLP Fact-Checking Unrealistic for Misinformation'. arXiv, Oct-2022.

[143] 'RumorLens: Interactive Analysis and Validation of Suspected Rumors on Social Media |Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems'.

[144] A. Mughaid et al., 'An intelligent cybersecurity system for detecting fake news in social media websites', Soft Computing, vol. 26, no. 12, pp. 5577–5591, Jun. 2022.

[145] F. Alam et al., 'A Survey on Multimodal Disinformation Detection'. arXiv, Sep-2022.

[146] E. Kochkina, M. Liakata, A. Zubiaga, 'PHEME dataset for Rumour Detection and Veracity Classification'. figshare. Dataset. https://doi.org/10.6084/m9.figshare.6392078.v1

[147] F. Miró-Llinares and J. C. Aguerri, 'Misinformation about fake news: A systematic critical review of empirical studies on the phenomenon and its status as a "threat"', European Journal of Criminology, vol. 20, no. 1, pp. 356–374, Jan. 2023.

[148] P. Resnick, A. Alfayez, J. Im, and E. Gilbert, 'Informed Crowds Can Effectively Identify Misinformation'. arXiv, Feb-2022.

[149] 'Assessing News Credibility: Misinformation Content Indicators'. Mar-2021.

[150] K. Roitero et al., 'Can the Crowd Judge Truthfulness? A Longitudinal Study on Recent Misinformation about COVID-19', Personal and Ubiquitous Computing, vol. 27, no. 1, pp. 59–89, Feb. 2023.

[151] J. Y. Khan, M. T. I. Khondaker, S. Afroz, G. Uddin, and A. Iqbal, 'A benchmark study of machine learning models for online fake news detection', Machine Learning with Applications, vol. 4, p. 100032, Jun. 2021.

[152] D. Fallis, 'A Functional Analysis of Disinformation', iConference 2014 Proceedings, Mar. 2014.

[153] X. Zhou and R. Zafarani, 'A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities', ACM Computing Surveys, vol. 53, no. 5, p. 109:1-109:40, Sep. 2020.

[154] F. Alam et al., 'A Survey on Multimodal Disinformation Detection', arXiv:2103. 12541 [cs], Mar. 2021.

[155] R. Oshikawa, J. Qian, and W. Y. Wang, 'A Survey on Natural Language Processing for Fake News Detection', arXiv:1811. 00770 [cs], Nov. 2018.

[156] N. K. Conroy, V. L. Rubin, and Y. Chen, 'Automatic deception detection: Methods for finding fake news', Proceedings of the Association for Information Science and Technology, vol. 52, no. 1, pp. 1–4, 2015.

[157] K. Sharma, F. Qian, H. Jiang, N. Ruchansky, M. Zhang, and Y. Liu, 'Combating Fake News: A Survey on Identification and Mitigation Techniques', ACM Transactions on Intelligent Systems and Technology, vol. 10, no. 3, p. 21:1-21:42, Apr. 2019.

[158] K. Sharma, Y. Zhang, E. Ferrara, and Y. Liu, 'Identifying Coordinated Accounts on Social Media through Hidden Influence and Group Behaviours', KDD 2021.

[159] M. R. Islam, S. Liu, X. Wang, and G. Xu, 'Deep learning for misinformation detection on online social networks: a survey and new perspectives', Social Network Analysis and Mining, vol. 10, no. 1, p. 82, Sep. 2020.

[160] A. A. A. Ahmed, A. Aljarbouh, P. Donepudi, and M. Choi, 'Detecting Fake News using Machine Learning: A Systematic Literature Review', Psychology (Savannah, Ga.), vol. 58, pp. 1932–1939, Jan. 2021.

[161] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, 'Fake News Detection on Social Media: A Data Mining Perspective', p. 15.

[162] A. Habib, M. Z. Asghar, A. Khan, A. Habib, and A. Khan, 'False information detection in online content and its role in decision making: a systematic literature review', Social Network Analysis and Mining, vol. 9, no. 1, p. 50, Sep. 2019.

[163] J. Zhang, L. Cui, Y. Fu, and F. B. Gouza, 'Fake News Detection with Deep Diffusive Network Model', arXiv:1805. 08751 [cs, stat], May 2018.

[164] N. R. de Oliveira, P. S. Pisa, M. A. Lopez, D. S. V. de Medeiros, and D. M. F. Mattos, 'Identifying Fake News on Social Networks Based on Natural Language Processing: Trends and Challenges', Information, vol. 12, no. 1, p. 38, Jan. 2021.

[165] K. Shu, S. Wang, D. Lee, and H. Liu, 'Mining Disinformation and Fake News: Concepts, Methods, and Recent Advancements', in Disinformation, Misinformation, and Fake News in Social Media: Emerging Research Challenges and Opportunities, K. Shu, S. Wang, D. Lee, and H. Liu, Eds. Cham: Springer International Publishing, 2020, pp. 1–19.

[166] F. Miró-Llinares and J. C. Aguerri, 'Misinformation about fake news: A systematic critical review of empirical studies on the phenomenon and its status as a "threat"', European Journal of Criminology, p. 1477370821994059, Apr. 2021.

[167] Y. Chen, N. K. Conroy, and V. L. Rubin, 'News in an online world: The need for an "automatic crap detector"', Proceedings of the Association for Information Science and Technology, vol. 52, no. 1, pp. 1–4, 2015.

[168] M. Fernandez and H. Alani, 'Online Misinformation: Challenges and Future Directions', in Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW '18, 2018, pp. 595–602.

[169] D. A. Martin, J. N. Shapiro, and M. Nedashkovskaya, 'Recent Trends in Online Foreign Influence Efforts', p. 34.

[170] S. Shelke and V. Attar, 'Source detection of rumor in social network – A review', Online Social Networks and Media, vol. 9, pp. 30–42, Jan. 2019.

[171] B. Guo, Y. Ding, L. Yao, Y. Liang, and Z. Yu, 'The Future of Misinformation Detection: New Perspectives and Trends'. arXiv, Sep-2019.

[172] J. Thorne and A. Vlachos, 'Automated Fact Checking: Task Formulations, Methods and Future Directions', in Proceedings of the 27th International Conference on Computational Linguistics, 2018, pp. 3346–3359.

[173] A. Zubiaga, M. Liakata, R. Procter, G. W. S. Hoi, and P. Tolmie, 'Analysing How People Orient to and Spread Rumours in Social Media by Looking at Conversational Threads', PLOS ONE, vol. 11, no. 3, p. e0150989, Mar. 2016.

[174] V. L. Rubin, Y. Chen, and N. K. Conroy, 'Deception detection for news: Three types of fakes', Proceedings of the Association for Information Science and Technology, vol. 52, no. 1, pp. 1–4, 2015.

[175] F. Yu, Q. Liu, S. Wu, L. Wang, and T. Tan, 'A Convolutional Approach for Misinformation Identification', pp. 3901–3907, 2017.

[176] A. Roy, K. Basak, A. Ekbal, and P. Bhattacharyya, '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, 2019, pp. 9–17.

[177] B. Amado, F. Ramón Fariña, 'Undeutsch hypothesis and Criteria Based Content Analysis: A meta-analytic review', in European Journal of Psychology Applied to Legal Context, 2015., pp. 3-12.

[178] S. Antoniadis, I. Litou, and V. Kalogeraki, 'A Model for Identifying Misinformation in Online Social Networks', in On the Move to Meaningful Internet Systems: OTM 2015 Conferences, 2015, pp. 473–482.

[179] B. Al Asaad and M. Erascu, 'A Tool for Fake News Detection', in 2018 20th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), 2018, pp. 379–386.

[180] A. C. Nied, L. Stewart, E. Spiro, and K. Starbird, 'Alternative Narratives of Crisis Events: Communities and Social Botnets Engaged on Social Media', in Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, 2017, pp. 263–266.

[181] X. Zhang and A. A. Ghorbani, 'An overview of online fake news: Characterization, detection, and discussion', Information Processing & Management, vol. 57, no. 2, p. 102025, Mar. 2020.

[182] M. Christodorescu, S. Jha, S. A. Seshia, D. Song and R. E. Bryant, 'Semantics-aware malware detection,' 2005 IEEE Symposium on Security and Privacy (S&P'05), Oakland, CA, USA, 2005, pp. 32-46, doi: 10.1109/SP.2005.20.

[183] W. Chen, Y. Zhang, C. Yeo, C. Lau, and B. Lee, 'Unsupervised rumor detection based on users' behaviors using neural networks', Pattern Recognition Letters, pp 226-233. 2018.

[184] S. T. King and P. M. Chen, 'SubVirt: implementing malware with virtual machines,' 2006 IEEE Symposium on Security and Privacy (S&P'06), Berkeley/Oakland, CA, 2006, pp. 14 pp.-327, doi: 10.1109/SP.2006.38.

[185] P. Paudel, J. Blackburn, E. De Cristofaro, S. Zannettou and G. Stringhini, 'LAMBRETTA: Learning to Rank for Twitter Soft Moderation,' in 2023 IEEE Security & Privacy, San Francisco, CA, USA, 2023 pp. 311-326.

[186] M.H. Saeed, S. Ali, J. Blackburn, E. De Cristofaro, S. Zannettou, and G. Stringhini, 'TROLLMAGNIFIER: Detecting State-Sponsored Troll Accounts on Reddit,' in 2022 IEEE Security & Privacy (SP), San Francisco, CA, USA, 2022 pp 2161-2175.

[187] M. Zurko, 'Disinformation and Reflections From Usable Security,' in IEEE Security & Privacy, vol. 20, no. 03, pp. 4-7, 2022.

[188] K. Thomas, C. Grier, J. Ma, V. Paxson and D. Song, 'Design and Evaluation of a Real-Time URL Spam Filtering Service,' in 2011 IEEE Security & Privacy, Oakland, CA, USA, 2011, pp. 447-462, doi: 10.1109/SP.2011.25.

[189] J. Ma, W. Gao, P. Mitra, S. Kwon, B. Jansen, K. Wong, and Meeyoung Cha, 'Detecting rumors from microblogs with recurrent neural networks', Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence. 2016.

[190] G. Danezis and P. Mittal, 'SybilInfer: Detecting Sybil Nodes using Social Networks' in NDSS Symposium 2009, San Diego, CA, USA, 2009.

[191] G. Wang, T. Konolige, C. Wilson, X. Wang, H. Zheng, and B. Y. Zhao, 'You are how you click: clickstream analysis for Sybil detection,' in 22nd USENIX conference on Security (SEC'13). USENIX Association, USA, 241–256.

[192] G. Wang, T. Wang, H. Zhang, and B. Y. Zhao, 'Man vs. machine: practical adversarial detection of malicious crowdsourcing workers,' in 23rd USENIX conference on Security Symposium (SEC'14). USENIX Association, USA, 239–254.

[193] T. Kim, N. Park, J. Hong, and S. Kim, 'Phishing URL Detection: A Network-based Approach Robust to Evasion,' Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security (CCS '22). Association for Computing Machinery, New York, NY, USA, 1769–1782. https://doi.org/10.1145/3548606.3560615

[194] D. Yuan, Y. Miao, N. Z. Gong, Z. Yang, Q. Li, D. Song, Q. Wang, and X. Liang. 2019. 'Detecting Fake Accounts in Online Social Networks at the Time of Registrations,' in ACM SIGSAC Conference on Computer and Communications Security (CCS '19). Association for Computing Machinery, New York, NY, USA, 1423–1438. https://doi.org/10.1145/3319535.3363198

[195] S. Nilizadeh, F. Labrèche, A. Sedighian, A. Zand, J. Fernandez, C. Kruegel, G. Stringhini, and G. Vigna, 'POISED: Spotting Twitter Spam Off the Beaten Paths,' in ACM SIGSAC Conference on Computer and Communications Security (CCS '17). Association for Computing Machinery, New York, NY, USA, 1159–1174. https://doi.org/10.1145/3133956.3134055

[196] C. Grier, K. Thomas, V. Paxson, and M. Zhang, '@spam: the underground on 140 characters or less,' in ACM conference on Computer and communications security (CCS '10). Association for Computing Machinery, New York, NY, USA, 27–37. https://doi.org/10.1145/1866307.1866311

[197] C. Whittaker, B. Ryner, M. Nazif, 'Large-Scale Automatic Classification of Phishing Pages,' in NDSS Symposium 2010, San Diego, CA, USA, 2010.

[198] A. Pitsillidis , K. Levchenko, C. Kreibich, C. Kanich, G. M. Voelker, V. Paxson, N. Weaver, S. Savage, 'Botnet Judo: Fighting Spam with Itself,' in NDSS Symposium 2010, San Diego, CA, USA, 2010.

[199] B. Stone-Gross, M. Cova, L. Cavallaro, B. Gilbert, M. Szydlowski, R. Kemmerer, C. Kruegel, and G. Vigna, 'Your botnet is my botnet: analysis of a botnet takeover,' in 6th ACM conference on Computer and communications security (CCS '09). Association for Computing Machinery, New York, NY, USA, 635–647. https://doi.org/10.1145/1653662.1653738

[200] H. Yu, P. B. Gibbons, M. Kaminsky and F. Xiao, 'SybilLimit: A Near-Optimal Social Network Defense against Sybil Attacks,' 2008 IEEE Symposium on Security and Privacy (sp 2008), Oakland, CA, USA, 2008, pp. 3-17, doi: 10.1109/SP.2008.13.

[201] A. Vrij, 'Detecting lies and deceit: pitfalls and opportunities', Wiley Publishers, Chichester, England, 2008.

[202] X. Liu, A. Nourbakhsh, Q. Li, R. Fang, and S. Shah. 2015. Real-time Rumor Debunking on Twitter. In Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (CIKM '15). Association for Computing Machinery, New York, NY, USA, 1867–1870. https://doi.org/10.1145/2806416.2806651

[203] V. Singh, R. Dasgupta, D. Sonagra, K. Raman, and I. Ghosh, 'Automated Fake News Detection Using Linguistic Analy- sis and Machine Learning'.

[204] S. Kaur, P. Kumar, and P. Kumaraguru, 'Automating fake news detection system using multi-level voting model', Soft Computing, vol. 24, no. 12, pp. 9049–9069, Jun. 2020.

[205] N. P. Nguyen, G. Yan, M. T. Thai, and S. Eidenbenz, 'Containment of misinformation spread in online social networks', in Proceedings of the 4th Annual ACM Web Science Conference, 2012, pp. 213–222.

[206] T. Mitra and E. Gilbert, 'CREDBANK: A Large-Scale Social Media Corpus With Associated Credibility Annotations', Proceedings of the International AAAI Conference on Web and Social Media, vol. 9, no. 1, pp. 258–267, 2015.

[207] K. Shu, L. Cui, S. Wang, D. Lee, and H. Liu, 'dEFEND: Explainable Fake News Detection', in Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019, pp. 395–405.

[208] C. G. Harris, 'Detecting Deceptive Opinion Spam Using Human Computation', 2012.

[209] M. Aldwairi and A. Alwahedi, 'Detecting Fake News in Social Media Networks', Procedia Computer Science, vol. 141, pp. 215–222, Jan. 2018.

[210] Abdullah-All-Tanvir, E. M. Mahir, S. Akhter, and M. R. Huq, 'Detecting Fake News using Machine Learning and Deep Learning Algorithms', in 2019 7th International Conference on Smart Computing & Communications (ICSCC), 2019, pp. 1–5.

[211] S. Aphiwongsophon and P. Chongstitvatana, 'Detecting Fake News with Machine Learning Method', in 2018 15th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), 2018, pp. 528–531.

[212] S. Afroz, M. Brennan, and R. Greenstadt, 'Detecting Hoaxes, Frauds, and Deception in Writing Style Online', in 2012 IEEE Symposium on Security and Privacy, 2012, pp. 461–475.

[213] K. Starbird, A. Arif, and T. Wilson, 'Disinformation as Collaborative Work: Surfacing the Participatory Nature of Strategic Information Operations', in 2019 Proceedings of the ACM on Human-Computer Interaction, 2019, pp. 1-26.

[214] H. Ahmed, I. Traore, and S. Saad, 'Detecting opinion spams and fake news using text classification', SECURITY AND PRIVACY, vol. 1, no. 1, p. e9, 2018.

[215] B. A. Galitsky, 'Detecting Rumor and Disinformation by Web Mining', 2015.

[216] Z. Jin, J. Cao, H. Guo, Y. Zhang, Y. Wang, and J. Luo, 'Detection and Analysis of 2016 US Presidential Election Related Rumors on Twitter', in Social, Cultural, and Behavioral Modeling, 2017, pp. 14–24.

[217] A. Zubiaga, A. Aker, K. Bontcheva, M. Liakata, and R. Procter, 'Detection and Resolution of Rumours in Social Media: A Survey', ACM Computing Surveys, vol. 51, no. 2, p. 32:1-32:36, Feb. 2018.

[218] H. Ahmed, I. Traore, and S. Saad, 'Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques', in Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments, 2017, pp. 127–138.

[219] S. K. Maity, A. Chakraborty, P. Goyal, and A. Mukherjee, 'Detection of Sockpuppets in Social Media', in Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, 2017, pp. 243–246.

[220] P. Hernon, 'Disinformation and misinformation through the internet: Findings of an exploratory study', Government Information Quarterly, vol. 12, no. 2, pp. 133–139, Jan. 1995.

[221] L. van de Guchte, S. Raaijmakers, E. Meeuwissen, and J. Spenader, 'Near Real-Time Detection of Misinformation on Online Social Networks', in Disinformation in Open Online Media, 2020, pp. 246–260.

[222] M. Dong, L. Yao, X. Wang, B. Benatallah, Q. Sheng, and H. Huang, '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', 2018, pp. 199–209.

[223] M. Konte, N. Feamster, and J. Jung, 'Dynamics of Online Scam Hosting Infrastructure', in Passive and Active Network Measurement, 2009, pp. 219–228.

[224] E. Sandhaus, 'The New York Times Annotated Corpus'. 2008. Linguistic Data Consortium, Philadelphia.

[225] K. Starbird, 'Examining the Alternative Media Ecosystem Through the Production of Alternative Narratives of Mass Shooting Events on Twitter', Proceedings of the International AAAI Conference on Web and Social Media, vol. 11, no. 1, pp. 230–239, May 2017.

[226] K. Shu, S. Wang, and H. Liu, 'Exploiting Tri-Relationship for Fake News Detection', Dec. 2017.

[227] T. Mihaylov, I. Koychev, G. Georgiev, and P. Nakov, 'Exposing Paid Opinion Manipulation Trolls'. arXiv, Sep-2021.

[228] M. Tambuscio, G. Ruffo, A. Flammini, and F. Menczer, 'Fact-checking Effect on Viral Hoaxes: A Model of Misinformation Spread in Social Networks', in Proceedings of the 24th International Conference on World Wide Web, 2015, pp. 977–982.

[229] C. M. M. Kotteti, X. Dong, N. Li, and L. Qian, 'Fake news detection enhancement with data imputation', Computer Information Systems Faculty Publications, Oct. 2018.

[230] S. Tschiatschek, A. Singla, M. Gomez Rodriguez, A. Merchant, and A. Krause, 'Fake News Detection in Social Networks via Crowd Signals', in Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW '18, 2018, pp. 517–524.

[231] A. Kesarwani, S. S. Chauhan, and A. R. Nair, 'Fake News Detection on Social Media using K-Nearest Neighbor Classifier', in 2020 International Conference on Advances in Computing and Communication Engineering (ICACCE), 2020, pp. 1–4.

[232] Y. Long, Q. Lu, R. Xiang, M. Li, and C.-R. Huang, 'Fake News Detection Through Multi-Perspective Speaker Profiles', in Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers), 2017, pp. 252–256.

[233] S. Kumar, R. Asthana, S. Upadhyay, N. Upreti, and M. Akbar, 'Fake news detection using deep learning models: A novel approach', Transactions on Emerging Telecommunications Technologies, vol. 31, no. 2, p. e3767, 2020.

[234] I. Ahmad, M. Yousaf, S. Yousaf, and M. O. Ahmad, 'Fake News Detection Using Machine Learning Ensemble Methods', Complexity, vol. 2020, p. e8885861, Oct. 2020.

[235] M. Granik and V. Mesyura, 'Fake news detection using naive Bayes classifier', in 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), 2017, pp. 900–903.

[236] B. Bhutani, N. Rastogi, P. Sehgal, and A. Purwar, 'Fake News Detection Using Sentiment Analysis', in 2019 Twelfth International Conference on Contemporary Computing (IC3), 2019, pp. 1–5.

[237] A. Thota, P. Tilak, S. Ahluwalia, and N. Lohia, 'Fake News Detection: A Deep Learning Approach', vol. 1, no. 3, 2018.

[238] S. S. Jadhav and S. D. Thepade, 'Fake News Identification and Classification Using DSSM and Improved Recurrent Neural Network Classifier', Applied Artificial Intelligence, vol. 33, no. 12, pp. 1058–1068, Oct. 2019.

[239] M. Farajtabar et al., 'Fake News Mitigation via Point Process Based Intervention', in Proceedings of the 34th International Conference on Machine Learning, 2017, pp. 1097–1106.

[240] V. Rubin, N. Conroy, Y. Chen, and S. Cornwell, 'Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News', in Proceedings of the Second Workshop on Computational Approaches to Deception Detection, 2016, pp. 7–17.

[241] A. Dey, R. Z. Rafi, S. Hasan Parash, S. K. Arko, and A. Chakrabarty, '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), 2018, pp. 305–309.

[242] L. Wu, Y. Rao, H. Yu, Y. Wang, and A. Nazir, 'False Information Detection on Social Media via a Hybrid Deep Model', in Social Informatics, 2018, pp. 323–333.

[243] S. Kumar and N. Shah, 'False Information on Web and Social Media: A Survey', Apr. 2018.

[244] K. Wu, S. Yang, and K. Q. Zhu, 'False rumors detection on Sina Weibo by propagation structures', in 2015 IEEE 31st International Conference on Data Engineering, 2015, pp. 651–662.

[245] Nguyen, V.-H. A. Sugiyama, K. A. Nakov, P. A. Kan, and Min-Yen, 'FANG |Proceedings of the 29th ACM International Conference on Information & Knowledge Management'. .

[246] N. Sitaula, C. K. Mohan, J. Grygiel, X. Zhou, and R. Zafarani, 'Credibility-Based Fake News Detection', in Disinformation, Misinformation, and Fake News in Social Media: Emerging Research Challenges and Opportunities, K. Shu, S. Wang, D. Lee, and H. Liu, Eds. Cham: Springer International Publishing, 2020, pp. 163–182.

[247] H. Zhang, A. Kuhnle, J. D. Smith, and M. T. Thai, '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), 2018, pp. 266–273.

[248] T. Mihaylov, G. Georgiev, and P. Nakov, 'Finding Opinion Manipulation Trolls in News Community Forums', in Proceedings of the Nineteenth Conference on Computational Natural Language Learning, 2015, pp. 310–314.

[249] Y. Liu and Y.-F. B. Wu, 'FNED: A Deep Network for Fake News Early Detection on Social Media', ACM Transactions on Information Systems, vol. 38, no. 3, p. 25:1-25:33, May 2020.

[250] N. Chawla and W. Wang, Eds., Proceedings of the 2017 SIAM International Conference on Data Mining. Philadelphia, PA: Society for Industrial and Applied Mathematics, 2017.

[251] L. Wu, J. Li, X. Hu, and H. Liu, 'Gleaning Wisdom from the Past: Early Detection of Emerging Rumors in Social Media', 2017, pp. 99–107.

[252] B. Ni, Z. Guo, J. Li, and M. Jiang, 'Improving Generalizability of Fake News Detection Methods using Propensity Score Matching'. arXiv, Jan-2020.

[253] C. Castillo, M. Mendoza, and B. Poblete, 'Information credibility on twitter', in Proceedings of the 20th international conference on World wide web, 2011, pp. 675–684.

[254] F. Ribeiro et al., 'Media Bias Monitor: Quantifying Biases of Social Media News Outlets at Large-Scale', Proceedings of the International AAAI Conference on Web and Social Media, vol. 12, no. 1, Jun. 2018.

[255] L. Padgham, Y. Lee, S. Sadiq, M. Winikoff, A. Fekete, S. MacDonell, D. Kaafar, and S. Zollmann, "CORE Rankings." [Online]. Available: https://www.core.edu.au/conference-portal

[256] Q. Liu, F. Yu, S. Wu, and L. Wang, 'Mining Significant Microblogs for Misinformation Identification: An Attention-Based Approach', ACM Transactions on Intelligent Systems and Technology, vol. 9, no. 5, p. 50:1-50:20, Apr. 2018.

[257] H. Zhang, M. A. Alim, X. Li, M. T. Thai, and H. T. Nguyen, 'Misinformation in Online Social Networks: Detect Them All with a Limited Budget', ACM Transactions on Information Systems, vol. 34, no. 3, p. 18:1-18:24, Apr. 2016.

[258] L. Wu, F. Morstatter, K. M. Carley, and H. Liu, 'Misinformation in Social Media: Definition, Manipulation, and Detection', ACM SIGKDD Explorations Newsletter, vol. 21, no. 2, pp. 80–90, Nov. 2019.

[259] T. Rasool, W. H. Butt, A. Shaukat, and M. U. Akram, 'Multi-Label Fake News Detection using Multi-layered Supervised Learning', in Proceedings of the 2019 11th International Conference on Computer and Automation Engineering, 2019, pp. 73–77.

[260] H. Karimi, P. Roy, S. Saba-Sadiya, and J. Tang, 'Multi-Source Multi-Class Fake News Detection', in Proceedings of the 27th International Conference on Computational Linguistics, 2018, pp. 1546–1557.

[261] Z. Jin, J. Cao, Y.-G. Jiang, and Y. Zhang, 'News Credibility Evaluation on Microblog with a Hierarchical Propagation Model', in 2014 IEEE International Conference on Data Mining, 2014, pp. 230–239.

[262] T. Magelinski, L. Ng, and K. Carley, 'A Synchronized Action Framework for Detection of Coordination on Social Media', Journal of Online Trust and Safety, 2022.

[263] Z. Jin, J. Cao, Y. Zhang, and J. Luo, 'News Verification by Exploiting Conflicting Social Viewpoints in Microblogs', Proceedings of the AAAI Conference on Artificial Intelligence, vol. 30, no. 1, Mar. 2016.

[264] S. Jindal, R. Sood, R. Singh, M. Vatsa, and T. Chakraborty, 'NewsBag: A Multimodal Benchmark Dataset for Fake News Detection'.

[265] Z. Jin, J. Cao, Y. Zhang, J. Zhou, and Q. Tian, 'Novel Visual and Statistical Image Features for Microblogs News Verification', IEEE Transactions on Multimedia, vol. 19, no. 3, pp. 598–608, Mar. 2017.

[266] M. D. Vicario, W. Quattrociocchi, A. Scala, and F. Zollo, 'Polarization and Fake News: Early Warning of Potential Misinformation Targets', ACM Transactions on the Web, vol. 13, no. 2, p. 10:1-10:22, Mar. 2019.

[267] C. Budak, D. Agrawal, and A. El Abbadi, 'Limiting the spread of misinformation in social networks', in Proceedings of the 20th international conference on World wide web, 2011, pp. 665–674.

[268] 'Web-based statistical fact checking of textual documents |Proceedings of the 2nd international workshop on Search and mining user-generated contents.'

[269] S. Kwon, M. Cha, K. Jung, W. Chen, and Y. Wang, 'Prominent Features of Rumor Propagation in Online Social Media', in 2013 IEEE 13th International Conference on Data Mining, 2013, pp. 1103–1108.

[270] Q. Zhang, A. Lipani, S. Liang, and E. Yilmaz, 'Reply-Aided Detection of Misinformation via Bayesian Deep Learning', in The World Wide Web Conference, 2019, pp. 2333–2343.

[271] S. Kwon, M. Cha, and K. Jung, 'Rumor Detection over Varying Time Windows', PLOS ONE, vol. 12, no. 1, p. e0168344, Jan. 2017.

[272] H. Guo, J. Cao, Y. Zhang, J. Guo, and J. Li, 'Rumor Detection with Hierarchical Social Attention Network', in Proceedings of the 27th ACM International Conference on Information and Knowledge Management, 2018, pp. 943–951.

[273] X. Lin, X. Liao, T. Xu, W. Pian, and K.-F. Wong, 'Rumor Detection with Hierarchical Recurrent Convolutional Neural Network', in Natural Language Processing and Chinese Computing, vol. 11839, J. Tang, M.-Y. Kan, D. Zhao, S. Li, and H. Zan, Eds. Cham: Springer International Publishing, 2019, pp. 338–348.

[274] L. Zeng, K. Starbird, and E. 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), 2016, pp. 1969–1978.

[275] P. Dewan and P. Kumaraguru, "Towards automatic real time identification of malicious posts on Facebook," 2015 13th Annual Conference on Privacy, Security and Trust (PST), Izmir, Turkey, 2015, pp. 85-92.

[276] 'Rumors, False Flags, and Digital Vigilantes: Misinformation on Twitter after the 2013 Boston Marathon Bombing', in iConference 2014 Proceedings, 2014.

[277] L. Cui, S. Wang, and D. 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), 2019, pp. 41–48.

[278] O. Ajao, D. Bhowmik, and S. Zargari, 'Sentiment Aware Fake News Detection on Online Social Networks', in ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 2507–2511.

[279] L. Zhao, H. Cui, X. Qiu, X. Wang, and J. Wang, 'SIR rumor spreading model in the new media age', Physica A: Statistical Mechanics and its Applications, vol. 392, no. 4, pp. 995–1003, Feb. 2013.

[280] S. Singhal, R. R. Shah, T. Chakraborty, P. Kumaraguru, and S. Satoh, 'SpotFake: A Multi-modal Framework for Fake News Detection', in 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM), 2019, pp. 39–47.

[281] A. Mukherjee, B. Liu, and N. Glance, 'Spotting fake reviewer groups in consumer reviews', in Proceedings of the 21st international conference on World Wide Web, 2012, pp. 191–200.

[282] D. Acemoglu, A. Ozdaglar, and A. ParandehGheibi, 'Spread of (mis)information in social networks', Games and Economic Behavior, vol. 70, no. 2, pp. 194–227, Nov. 2010.

[283] I. Y. R. Pratiwi, R. A. Asmara, and F. 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), 2017, pp. 73–78.

[284] K. Shu, H. R. Bernard, and H. Liu, 'Studying Fake News via Network Analysis: Detection and Mitigation'. arXiv, Apr-2018.

[285] J. C. S. Reis, A. Correia, F. Murai, A. Veloso, and F. Benevenuto, 'Supervised Learning for Fake News Detection', IEEE Intelligent Systems, vol. 34, no. 2, pp. 76–81, Mar. 2019.

[286] J. C. S. Reis, A. Correia, F. Murai, A. Veloso, and F. Benevenuto, 'Explainable Machine Learning for Fake News Detection,' Web Science '19, Boston, MA, USA, 2019, pp. 17-26.

[287] F. Torabi Asr and M. Taboada, 'The Data Challenge in Misinformation Detection: Source Reputation vs. Content Veracity', in Proceedings of the First Workshop on Fact Extraction and VERification (FEVER), 2018, pp. 10–15.

[288] K. Shu, X. Zhou, S. Wang, R. Zafarani, and H. Liu, '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, 2020, pp. 436–439.

[289] C. Shao, G. L. Ciampaglia, O. Varol, K.-C. Yang, A. Flammini, and F. Menczer, 'The spread of low-credibility content by social bots', Nature Communications, vol. 9, no. 1, p. 4787, Nov. 2018.

[290] J. Li, M. Ott, C. Cardie, and E. Hovy, '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), 2014, pp. 1566–1576.

[291] S. Jain, V. Sharma, and R. Kaushal, 'Towards automated real-time detection of misinformation on Twitter', in 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2016, pp. 2015–2020.

[292] V. L. Rubin, N. J. Conroy, and Y. Chen, 'Towards News Verification: Deception Detection Methods for News Discourse'.

[293] V. L. Rubin and T. Lukoianova, 'Truth and deception at the rhetorical structure level: Truth and Deception at the Rhetorical Structure Level', Journal of the Association for Information Science and Technology, vol. 66, no. 5, pp. 905–917, May 2015.

[294] A. Gupta, P. Kumaraguru, C. Castillo, and P. Meier, 'TweetCred: Real-Time Credibility Assessment of Content on Twitter'. arXiv, Jan-2015. [258] Z. Yang, C. Wilson, and X. Wang, 'Uncovering Social Network Sybils in the Wild'.

[295] K. Darwish, P. Stefanov, M. Aupetit, and P. Nakov, 'Unsupervised User Stance Detection on Twitter'. arXiv, May-2020.

[296] B. Nyhan and J. Reifler, 'When Corrections Fail: The Persistence of Political Misperceptions', Political Behavior, vol. 32, no. 2, pp. 303–330, Jun. 2010.

[297] D. Katsaros, G. Stavropoulos, and D. Papakostas, 'Which machine learning paradigm for fake news detection?', in 2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI), 2019, pp. 383–387.

[298] 'Snopes'. www.snopes.com.

[299] 'PolitiFact'. www.politifact.com.

[300] C. Silverman, C. Timberg, J. Kao, and J. B. Merrill, 'Facebook groups topped 10,000 daily attacks on election before Jan. 6, analysis shows', The Washington Post. WP Company, Jan-2022.

[301] M. Hindman and V. Barash, 'Disinformation, 'Fake News' and Influence Campaigns on Twitter', Knight Foundation. Knight Foundation, Oct-2018.

[302] S. Castelo et al., 'A Topic-Agnostic Approach for Identifying Fake News Pages', Companion Proceedings of The 2019 World Wide Web Conference. ACM, New York, NY, USA, pp. 975–980, 2019.

[303] H. Allcott and M. Gentzkow, 'Social Media and Fake News in the 2016 Election', Journal of Economic Perspectives, vol. 31, no. 2, pp. 211–236, May 2017.

[304] T. Vaidya, D. Votipka, M. L. Mazurek, and M. Sherr, '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, Glasgow, Scotland Uk, 2019, pp. 1–13.

[305] Government News, 'The Weibo rumor-refuting and co-governance platform is launched, giving government accounts the right to directly refute rumors', Nov. 2018.

[306] I. Mehta and M. Singh, 'Twitter to end free access to its API in Elon Musk's latest monetization push', TechCrunch. Feb-2023.

[307] J. Porter, 'Twitter announces new API pricing, posing a challenge for small developers', The Verge. The Verge, Mar-2023.

[308] 'Developer policy – twitter developers | twitter developer platform', Twitter. Twitter.

[309] S. McCarthy, 'China's promotion of Russian disinformation indicates where its loyalties lie', CNN. Cable News Network, Mar-2022.

[310] Archive Team, 'The Twitter Stream Grab', Internet Archive. Internet Archive.

[311] 'Media Bias/Fact Check News', Media Bias/Fact Check. Jul-2021.

[312] 'Fake news detection datasets', ISOT research lab.

[313] F. K. Abu Salem, R. Al Feel, S. Elbassuoni, M. Jaber, and M. Farah, 'FA-Kes: A fake news dataset around the Syrian War', Zenodo. Jan-2019.

[314] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, 'FakeNewsNet: A data repository with news content, social context and spatialtemporal information for studying fake news on social media', arXiv. org. Mar-2019.

[315] F. Monti, F. Frasca, D. Eynard, D. Mannion, and M. M. Bronstein, 'Fake News Detection on Social Media using Geometric Deep Learning', arXiv [cs.SI]. 2019.

[316] R. 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, 2020, pp. 3364–3374.

[317] R. Evangelista and F. Bruno, 'WhatsApp and political instability in Brazil: Targeted messages and political radicalisation', Internet Policy Review. Dec-2019.

[318] K. Garimella and D. Eckles, 'Images and misinformation in political groups: Evidence from WhatsApp in India: HKS Misinformation Review', Misinformation Review. Jul-2022.

[319] Slick, 'Commit to transparency - sign up for the international fact-checking network's code of Principles', IFCN Code of Principles.

[320] S. Kapoor and A. Narayanan, 'Leakage and the Reproducibility Crisis in ML-based Science'. Patterns, 2022.

[321] S. Kaufman, S. Rosset, C. Perlich, and O. Stitelman, 'Leakage in Data Mining: Formulation, Detection, and Avoidance', ACM Trans. Knowl. Discov. Data, vol. 6, no. 4, Dec. 2012.

[322] G. Miner, Handbook of Statistical Analysis and Data Mining Applications. Association for Computing Machinery, 2009.

[323] S. Feng, R. Banerjee, and Y. Choi, 'Syntactic Stylometry for Deception Detection', in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2012, pp. 171–175.

[324] R. Mihalcea and C. Strapparava, 'The Lie Detector: Explorations in the Automatic Recognition of Deceptive Language', in Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, 2009, pp. 309–312.

[325] F. Soldner, V. Pérez-Rosas, and R. 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), 2019, pp. 1768–1777.

[326] J. B. Bak-Coleman et al., 'Combining interventions to reduce the spread of viral misinformation', Nature News. Nature Publishing Group, Jun-2022.

[327] C. Carrasco-Farré, 'The fingerprints of misinformation: how deceptive content differs from reliable sources in terms of cognitive effort and appeal to emotions,' Humanit Soc Sci Commun 9, 162, May-2022.

[328] S. van der Linden, 'Misinformation: Susceptibility, spread, and interventions to immunize the public', Nature News. Nature Publishing Group, Mar-2022.

[329] C. Lima, 'A whistleblower's power: Key Takeaways from the Facebook papers', The Washington Post. WP Company, Mar-2022.

[330] 'Content fact-checkers prioritize', Transparency Center.

[331] D. Arp et al., 'Dos and Don'ts of Machine Learning in Computer Security', in 31st USENIX Security Symposium (USENIX Security 22), 2022, pp. 3971–3988.

[332] F. Zollo and W. Quattrociocchi, 'Misinformation Spreading on Facebook', in Complex Spreading Phenomena in Social Systems: Influence and Contagion in Real-World Social Networks, S. Lehmann and Y.-Y. Ahn, Eds. Cham: Springer International Publishing, 2018, pp. 177–196.

[333] A. Bovet and H. A. Makse, 'Influence of fake news in Twitter during the 2016 US presidential election', Nature News. Nature Publishing Group, Jan-2019.

[334] S. O. Oyeyemi, E. Gabarron, and R. Wynn, 'Ebola, Twitter, and misinformation: a dangerous combination?', Bmj, vol. 349, 2014.

[335] W. Ahmed, J. Vidal-Alaball, J. Downing, F. L. Seguí, and Others, 'COVID-19 and the 5G conspiracy theory: social network analysis of Twitter data', Journal of medical internet research, vol. 22, no. 5, p. e19458, 2020.

[336] A. S. Jacobs, R. Beltiukov, W. Willinger, R. A. Ferreira, A. Gupta, and L. Z. Granville, 'AI/ML and Network Security: The Emperor has no Clothes', in Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security, Los Angeles, CA, USA, 2022.

[337] D. J. Shah, T. Schuster, and R. Barzilay, 'Automatic Fact-guided Sentence Modification', in Proceedings of the AAAI Conference on Artificial Intelligence, 2020.

[338] R. Geirhos et al., 'Shortcut learning in deep neural networks', Nature Machine Intelligence, vol. 2, no. 11, pp. 665–673, Nov. 2020.

[339] Meta, 'How Meta's third-party fact-checking program works', Meta Blog. Meta.

[340] A. Jamieson and O. Solon, 'Facebook to begin flagging fake news in response to mounting criticism', The Guardian. Guardian News and Media, Dec-2016.

[341] C. Crowell, 'Our approach to bots and misinformation', Twitter Blog. Twitter.

[342] N. Ibrahim, 'Why facebook won't be fact-checking Trump now that he's announced candidacy', Snopes. Snopes.com, Nov-2022.

[343] 'GossipCop'. GossipCop.com.

[344] O. Darcy, 'BuzzFeed News will shut down', CNN Business. Cable News Network, Apr-2023.

[345] T. Hsu, 'As COVID-19 continues to spread, so does misinformation about it', The New York Times. The New York Times, Dec-2022.

[346] R. Woo and L. Gao, 'China's factory activity falls faster than expected as recovery stumbles', Reuters. Thomson Reuters, May-2023.

[347] D. Snelling, 'Xperia XZ4 release this month - five things every Sony fan should know', Express. co. uk. Express.co.uk, Feb-2019.

[348] C. F. Bond and B. M. DePaulo, 'Accuracy of deception judgments', Sage Journals. Personality and Social Psychology Review, 2006.

[349] OpenAI, 'ChatGPT', chat.openai.com. OpenAI.

[350] J. Vincent, 'OpenAI isn't doing enough to make ChatGPT's limitations clear', The Verge. The Verge, May-2023.

[351] S. Ali, 'Facebook's formula prioritized anger and ended up spreading misinformation', The Hill. The Hill, Oct-2021.

[352] W. S. J. Staff, 'The Facebook Files', The Wall Street Journal. Dow Jones & Company, Oct-2021.

[353] D. Coldewey, 'Deconstructing "the Twitter files"', TechCrunch. Jan-2023.

[354] S. Søe, 'Algorithmic detection of misinformation and disinformation: Gricean perspectives', Journal of Documentation. Dec-2017.

[355] I. Fried, 'OpenAI touts GPT-4 for content moderation', Axios. Aug-2023.

[356] M. R. DeVerna, H. Y. Yan, K. Yang, and F. Menczer, 'Artificial intelligence is ineffective and potentially harmful for fact checking', arXiv. Aug-2023.

[357] D. Andročec, 'Machine learning methods for toxic comment classification: a systematic review', Informatica. Jan-2021.

[358] H. Lee, T. Ermakov, V. Ververis, and B. Fabian, 'Detecting child sexual abuse material: A comprehensive survey', Forensic Science International: Digital Investigation. Sept-2020.

[359] B. Stone-Gross, M. Cova, L. Cavallaro, B. Gilbert, M. Szydlowski, R. Kemmerer, C. Kruegel, and G. Vigna, 'Your botnet is my botnet: analysis of a botnet takeover', Proceedings of the 16th ACM conference on Computer and communications security. 2009.

[360] S. Mirza, L. Begum, L. Niu, S. Pardo, A. Abouzied, P. Papotti, and C. Pöpper, 'Tactics, threats & targets: Modeling disinformation and its mitigation', ISOC Network and Distributed Systems Security Symposium (NDSS). 2023.

[361] M. Saeed, S. Ali, J. Blackburn, E. De Cristofaro, S. Zannettou, and G. Stringhini. 'Trollmagnifier: Detecting state-sponsored troll accounts on reddit', IEEE Symposium on Security and Privacy (SP). 2022.

[362] S. Zannettou, T. Caulfield, E. De Cristofaro, M. Sirivianos, G. Stringhini, and J. Blackburn. 'Disinformation warfare: Understanding state-sponsored trolls on Twitter and their influence on the web', Companion proceedings of the 2019 world wide web conference. 2019.

[363] S. Zannettou, T. Caulfield, W. Setzer, M. Sirivianos, G. Stringhini, J. Blackburn, 'Who Let The Trolls Out? Towards Understanding State-Sponsored Trolls', 2019 Proceedings of the 10th ACM Conference on Web Science. 2019.

[364] J. Lukito. 'Coordinating a multi-platform disinformation campaign: Internet Research Agency Activity on three US Social Media Platforms, 2015 to 2017', Political Communication. 2020.

[365] F. Keller, D. Schoch, S. Stier, and J. Yang. 'Political astroturfing on twitter: How to coordinate a disinformation campaign', Political communication. 2020.

[366] F. Ezzeddine, O. Ayoub, S. Giordano, G. Nogara, I. Sbeity, E. Ferrara, and L. Luceri. 'Exposing influence campaigns in the age of LLMs: a behavioral-based AI approach to detecting state-sponsored trolls', EPJ Data Science. 2023.

[367] , S. Cresci, 'A Decade of Social Bot Detection', Commun. ACM. October 2020.

[368] Z. Chu, S. Gianvecchio, H. Wang, and J. Sushil, 'Detecting automation of twitter accounts: Are you a human, bot, or cyborg?', IEEE Transactions on dependable and secure computing. 2012.

[369] J. Zhang, R. Zhang, Y. Zhang, and G. Yan. 'The rise of social botnets: Attacks and countermeasures', IEEE Transactions on Dependable and Secure Computing. 2016.

[370] M. Alizadeh, J. Shapiro, C. Buntain, and J. Tucker, 'Content-based features predict social media influence operations', Science Advances. July 2020.

[371] G. Wang, T. Wang, H. Zheng, and B. Zhao, 'Man vs. machine: Practical adversarial detection of malicious crowdsourcing workers', 23rd USENIX Security Symposium (USENIX Security). 2014.

[372] A. Rauchfleisch and J. Kaiser, 'The false positive problem of automatic bot detection in social science research', PloS one. 2020.

[373] G. Danezis and P. Mittal, 'Sybilinfer: Detecting sybil nodes using social networks', NDSS. 2009.

[374] C. Grimme, D. Assenmacher, and L. Adam, 'Changing perspectives: Is it sufficient to detect social bots?', 10th International Conference, SCSM. 2018.

[375] D. Yuan, Y. Miao, N. Gong, Z. Yang, Q. Li, D. Song, Q. Wang, and X. Liang, 'Detecting fake accounts in online social networks at the time of registrations', Proceedings of the 2019 ACM SIGSAC conference on computer and communications security. 2019.

[376] L. Vargas, P. Emami, and P. Traynor, 'On the detection of disinformation campaign activity with network analysis', Proceedings of the 2020 ACM SIGSAC Conference on Cloud Computing Security Workshop. 2020.

[377] S. Alhabash, N. Almutairi, C. Lou, and W. Kim, 'Pathways to virality: Psychophysiological responses preceding likes, shares, comments, and status updates on Facebook', Media Psychology. 2019.

[378] S. Milajerdi, R. Gjomemo, B. Eshete, R. Sekar, and V. Venkatakrishnan, 'Holmes: real-time apt detection through correlation of suspicious information flows', 2019 IEEE Symposium on Security and Privacy (SP). 2019.

[379] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, 'The rise of social bots.' Communications of the ACM. 2016.

[380] Y. Roth. 'Generalizing scaled misinformation detection', Bluesky. October 2023.

[381] S. King and P. Chen, 'SubVirt: Implementing malware with virtual machines', IEEE Symposium on Security and Privacy. 2006.

[382] Q. Zhang, A. Lipani, S. Liang, and E. Yilmaz, 'Reply-aided detection of misinformation via Bayesian deep learning,' The World Wide Web Conference, 2019.

[383] T. Rasool, W. Butt, A. Shaukat, and M. Akram, 'Multi-label fake news detection using multi-layered supervised learning', Proceedings of the 2019 11th International Conference on Computer and Automation Engineering. 2019.

[384] G. Sansonetti, F. Gasparetti, G. D'Aniello, and A. Micarelli, 'Unreliable Users Detection in Social Media: Deep Learning Techniques for Automatic Detection', IEEE Access. 2020.

[385] D. Linvill and P. Warren, 'Troll factories: Manufacturing specialized disinformation on Twitter', Political Communication. 2020.

[386] A. Pitsillidis, K. Levchenko, C. Kreibich, C. Kanich, G. Voelker, V. Paxson, N. Weaver, and S. Savage, 'Botnet Judo: Fighting Spam with Itself', NDSS. 2010.

[387] C. Grier, K. Thomas, V. Paxson, and M. Zhang, '@ spam: the underground on 140 characters or less', Proceedings of the 17th ACM conference on Computer and communications security. 2010.

[388] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, A. Flammini, and F. Menczer, 'Detecting and tracking political abuse in social media', Proceedings of the International AAAI Conference on Web and social media. 2011.

[389] K. Yang, O. Varol, P. Hui, and F. Menczer, 'Scalable and generalizable social bot detection through data selection', Proceedings of the AAAI conference on artificial intelligence. 2020.

[390] K. Yang, E. Ferrara, and F. Menczer, 'Botometer 101: Social bot practicum for computational social scientists', Journal of Computational Social Science. 2022.

[391] D. Assenmacher, L. Clever, J. Pohl, H. Trautmann, and C. Grimme, 'A two-phase framework for detecting manipulation campaigns in social media', 12th International Conference, SCSM. 2020.

[392] F. Giglietto, N. Righetti, L. Rossi, and G. Marino, 'It takes a village to manipulate the media: coordinated link sharing behavior during 2018 and 2019 Italian elections', Information, Communication & Society. 2020.

[393] S. Mollman, 'OpenAI is getting trolled for its name after refusing to be open about its A.I.', Fortune. March 2023.

[394] M. Wong, 'There was never such a thing as "open" ai', The Atlantic. January 2024.

[395] J. Fairbanks, N. Fitch, N. Knauf, and E. Briscoe, 'Credibility assessment in the news: do we need to read', Proc. of the MIS2 Workshop held in conjunction with 11th Int'l Conf. on Web Search and Data Mining. 2018.

[396] K. Shu, S. Dumais, A. Awadallah, and H. Liu, 'Detecting fake news with weak social supervision', IEEE Intelligent Systems. 2020.

[397] C. Silverman and J. Kao, 'Infamous Russian troll farm appears to be source of Anti-Ukraine propaganda', ProPublica. Mar 2022.

[398] Common Thread, 'Four truths about bots', Twitter. Sep 2021.

[399] D. Williams, 'Misinformation is the symptom, not the disease: Daniel Williams', IAI TV. Dec 2023.

[400] L. Tay, S. Lewandowsky, M. Hurlstone, T. Kurz, and U. Ecker, 'Thinking clearly about misinformation', Communications Psychology. 2024.

[401] J. Douceur, 'The sybil attack', 'International workshop on peer-to-peer systems.' 2002.

[402] 'FactCheck.org', FactCheck.org. URL: https://www.factcheck.org/

[403] "GARM Brand Safety Floor and Suitability Framework", Garm: Brand Safety Floor + Suitability Framework.

[404] Zefr, 'Zefr acquires Israeli AI firm Adverif.ai, bolstering technology-led approach to identifying and defunding misinformation', PR Newswire. Jul 2022.

[405] F. Pasquine, 'What advertisers need to know about surges in online hate speech', DoubleVerify. Feb 2022.

[406] Integral Ad Science, Inc. 'IAS expands AI-driven brand safety and suitability measurement to Meta', PR Newswire. Feb 2024.

[407] 'Here's how we're using AI to help detect misinformation', AI at Meta. Nov 2020.

[408] S. Wojcik, S. Hilgard, N. Judd, D. Mocanu, S. Ragain, MB Hunzaker, K. Coleman, and J. Baxter, 'Birdwatch: Crowd wisdom and bridging algorithms can inform understanding and reduce the spread of misinformation', arXiv preprint:2210.115723. 2022.

[409] N. McIntyre, R. Bradbury, and B. Perrigo, 'Behind TikTok's boom: A legion of traumatised, $10-a-day content moderators', The Bureau of Investigative Journalism (en-GB). Dec 2023.

[410] F. Hibaq, 'Diary of a TikTok moderator: "we are the people who sweep up the mess"', The Guardian. Dec 2023.

[411] D. Kapellmann Zafra, R. Serabian, S. Riddell, and N. Brubaker, 'How to understand and action Mandiant's intelligence on information operations', Mandiant. Oct 2022.

[412] Microsoft Incident Response and Microsoft Threat Intelligence, 'Dev-0537 criminal actor targeting organizations for data exfiltration and destruction', Microsoft Security Blog. Sep 2023.

[413] B. Nimmo, 'Meta's adversarial threat report, fourth quarter 2022', Meta. Feb 2023.

[414] S. Srinivasan, 'The global disinformation index', GDI.

[415] I. Lapowsky, 'Inside the research lab teaching Facebook about its trolls', Wired. Aug 2018.

[416] A. Schiffrin, 'sing AI to combat MIS/disinformation – an evolving story', Tech Policy Press. Oct 2023.

[417] 'How it works', Graphika.

[418] A. Storey, 'Our ongoing work to fight misinformation online', Google. Oct 2023.

[419] OpenAI, 'Using machine learning to reduce toxicity online', Perspective.

[420] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. Raji, and T. Gebru, 'Model Cards for Model Reporting', Proceedings of the Conference on Fairness, Accountability, and Transparency. 2019.

[421] J. Baumgartner, S. Zannettou, B. Keegan, M. Squire, and J. Blackburn, 'The pushshift reddit dataset', Proceedings of the international AAAI conference on web and social media. 2020.

[422] K. Yang, E. Ferrara, and F. Menczer, 'Botometer 101: Social bot practicum for computational social scientists', Journal of Computational Social Science. 2022.

[423] Democratics House Permanent Select Committee on Intelligence, 'Exhibit B.' May 2018.

[424] J. Habgood-Coote, "The term "fake news" is doing great harm," The Conversation. July 2018.

[425] J. Jackson, "What are influence operations and why are we investigating them?", The Bureau of Investigative Journalism. July 2023.

Appendix

1. Annotation Guide

Our guide for annotating papers follows below:

Target selection:
- What is the stated target of detection?
- What is the actual target of detection?
- Is the detection method *directly* performing fact-checking and misinformation detection, or is it detecting *proxy signals* instead (e.g., emotional language, topics associated with rumoring narratives)?

Dataset curation:
- Note dataset size, source, and temporality.
- Note *information scope*: what is the "operative unit" of checkable information? **Where this might become more complicated:** some multimodal methods, particularly network-based ones, usually accept some combination of user- and post-scoped data, *but only make a final decision about the veracity of one of those things.*
- Note *information domain*: does the data set coalesce around a topic / a set of topics? Are these pre-defined by the study authors?
- Fine to name multiple scopes if one info scope is a vital parameter / input to the model in its determination of the veracity of another; e.g., a model that performs network analysis of social media posts that share a link to a specific news article; the model might make a final determination about the truthfulness of the news article, but propagation patterns + social media posts were a crucial element of that decision.

Feature set selection:
- How were features selected?
- Are features explainable and accessible? Were they handcrafted? Were they generated by unsupervised models?
- Do positive signals indicate evidence of actually suspicious behavior?
- Do authors account for feature noise?

Test-train split:
- Note how/if this split was observed, and the percentage of the total data set assigned to test and training sets.
- Note any overlap between both data sets (ideally, none).
- Note any evidence of temporal leakage (future data used to make a prediction about past events).

Cross-validation performed:
- Note size and number of folds.
- Note presence of any development data sets (what percentage of total data set was this?).
- Note presence of any hyperparameter tuning during cross-validation process – are these parameters published in-text?

Out-of-sample testing:
- Note presence, if any.
- What data sets were used for this?
- Note any evidence of leakage during this process.

Third-party references:
- Particularly for claim-based studies that require some form of ground-truth reference for fact-checking, what is the ground truth reference employed by the study?
- Is this ground truth site reliable and up-to-date?
- What kind of taxonomy does the ground-truth reference employ for labeling true / false statements?
- What baked-in biases or slants might exist in the ground-truth reference's data set (e.g., reference only labels political news sites; ground-truth reference is run by a known conservative organization)?

Model choice:

Evaluation:
- Note dataset(s) used for evaluating classifier performance.
- How does the evaluation dataset(s) compare to testing and training datasets in terms of size, sourcing, contents, and temporality?

- How do authors account for distribution shifts in rumoring topics or activity types? Is there a road map for iterative updates to their method?

2. 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.

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 [403].

- *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" [404]—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" [405]. 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 [406].

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

- *Facebook* partners 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 [339], [352], [407].
- *Twitter* has deployed a crowd-sourced annotations platform called Community Notes (formerly Birdwatch) since 2021 [408].
- *TikTok* employs thousands of content moderators across the globe who "work alongside automated moderation systems" [409], [410].

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") [411].
- *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 [412]. 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 [413].

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 [414].
- *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 [415].

TABLE 3. 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.

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

- *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 [416], [417].

LLM-driven detection. A few LLM-powered detection methods have been discussed in the popular press, including those advertised by Google [418] and OpenAI [355], 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 [419]. 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).