Emotion and Information Disorder in NLP: A Systematic Mapping and Benchmark Blueprint

Renatha Vieira[1], **Alvaro Figueira**[1,2]
[1]Faculty of Sciences, University of Porto, Porto, Portugal
[2]INESC TEC, Porto, Portugal
{vieirasrept@gmail.com, arfiguei@fc.up.pt}

Abstract

Online misinformation research in NLP has expanded rapidly, including approaches that model affective signals such as sentiment, discrete emotions, and emotion dynamics. However, the Information Disorder framework distinguishes misinformation, disinformation, and mal-information along dimensions of intention, harm, and contextual dependence, which are rarely operationalised in current datasets, tasks, and evaluation protocols. We provide a systematic mapping of 82 studies at the intersection of Information Disorder and emotion-aware NLP (51 model papers, 7 dataset papers, 24 survey/theory papers). Across empirical works (58), veracity-centric supervision dominates (72.4% binary labels), while explicit intention and harm variables appear in only 1.7% each. Evaluation relies mostly on random splits (79.3%), limiting robustness to source and temporal shifts. Emotion is represented in 43.1% of model papers, mostly as static features, with emotion dynamics and audience emotion rare. Based on these findings, we propose an operational taxonomy aligned with Information Disorder and a benchmark blueprint specifying tasks, annotation variables, split strategies, and evaluation protocols to support theory-grounded, comparable progress.

Keywords: misinformation, disinformation, mal-information, emotion, systematic mapping, evaluation blueprint.

1. Introduction

Information Disorder is a framework proposed to organise research on false and harmful information by distinguishing *misinformation*, *disinformation*, and *mal-information*, and by emphasising contextual factors such as intention and potential harm (Wardle and Derakhshan, 2017; Wardle, 2020; Turcilo and Obrenovic, 2020). In parallel, NLP research on misinformation and fake news has expanded rapidly, supported by benchmark datasets and supervised tasks that most often operationalise the problem as veracity classification or related variants of credibility assessment (Shu et al., 2017; Zhou and Zafarani, 2018; Wu et al., 2019; Wang, 2017).

Affective signals (sentiment, discrete emotions, and emotion dynamics) have been increasingly incorporated into such systems, motivated by the intuition that manipulation and engagement are partly mediated through emotion and that diffusion dynamics can amplify emotionally charged misinformation (Vosoughi et al., 2018; Ghanem et al., 2019; Giahanou et al., 2019). However, emotion-aware approaches are typically trained and evaluated against labels that do not explicitly encode intention, harm, or contextual dependence. As a result, systems may achieve strong performance on coarse veracity targets while remaining theoretically misaligned with Information Disorder distinctions, a gap that is particularly salient for *mal-information*, where harm can arise even when content is factually correct (Wardle and Derakhshan, 2017).

We address this misalignment with a systematic mapping of emotion-aware NLP work through an Information Disorder lens and provide reusable artefacts to support comparability. Specifically, we (i) compile and code a curated corpus of studies using a shared coding scheme that captures dataset properties, task formulations, evaluation protocols, and emotion operationalisations along Information Disorder dimensions (Sections 2–3); (ii) synthesise recurrent structural mismatches between common operational settings and Information Disorder constructs (Section 4); and (iii) propose an operational taxonomy and a benchmark blueprint to guide dataset creation and evaluation in a theory-grounded, replicable manner (Section 5).

Research questions. The mapping is guided by the following questions:

1. **RQ1:** Which Information Disorder dimensions (intention, harm, contextual dependence) are explicitly annotated or otherwise operationalised in current datasets and tasks?
2. **RQ2:** How is emotion represented (sentiment, discrete emotions, appraisal, dynamics) and how is it used (feature, auxiliary task, audience response, discourse structure) in NLP approaches to false information?
3. **RQ3:** What evaluation protocols (splits, metrics, robustness tests) are most common, and where do they fail to support Information Disorder-relevant claims?
4. **RQ4:** What minimal benchmark design would enable theory-aligned, replicable comparison of

Proceedings of the 1st Workshop on Information Disorder (InDor) @ LREC 2026, pages 77–84
12 May 2026.

systems for Information Disorder?

The remaining sections of this paper are organised as follows: Section 2 presents the selection protocol and coding scheme; Section 3 reports mapping results; Section 4 summarises misalignments; and Section 5 introduces the taxonomy and benchmark blueprint.

2. Methodology of Reference Selection

The bibliographic corpus was constructed through a structured but intentionally selective mapping process. Our aim was not to exhaustively cover all fake news detection research, but to assemble a conceptually coherent set of studies that makes it possible to analyse how emotion is operationalised in relation to the Information Disorder framework.

The search was conducted between 10 February and 2 March 2026. Semantic Scholar was used as the primary retrieval source, complemented by citation chaining from high-relevance papers and targeted Google Scholar follow-up searches used to close thematic gaps, especially around dataset construction and emotion-aware fake news detection. We explored three query families. Q1 used broad Information Disorder terms (*misinformation, disinformation, malinformation, fake news, information disorder*); Q2 combined the same group with affect terms (*emotion, affect, sentiment*) and NLP terms (*natural language processing, NLP, text classification*); and Q3 flattened these groups into a broader disjunctive query. Query 3 retrieved 254 records and was retained as the main starting set because it offered the best balance between breadth and conceptual relevance.

Selection proceeded in stages. A title-based screening reduced the 254 retrieved records to 60 candidate papers by excluding work not directly related to misinformation/fake news, not connected to affect or emotion, or outside the scope of NLP. Abstracts were then inspected when needed to assess thematic fit with the goals of the study, especially relevance to taxonomy, benchmarking, or methodological grounding, yielding a core set of 43 papers. The corpus was subsequently expanded through backward citation tracking and targeted follow-up searches, producing the final set of 82 included studies.

Inclusion prioritised studies on emotion recognition, sentiment analysis, or affective computing applied to misinformation/disinformation, together with adjacent work on datasets, propagation, contextual signals, and conceptual distinctions that was necessary for interpreting evaluation and taxonomy choices. We excluded generic text classification work unrelated to deceptive information, non-NLP approaches, and domain- or language-specific studies with no clear relevance to the research focus. Title/abstract screening, full-text assessment, and final coding were conducted by the first author. No duplicate records were found during consolidation of the candidate set, and no independent double-coding or formal inter-rater agreement was performed.

2.1. Selection transparency and reproducibility

Our protocol is therefore iterative rather than fully exhaustive: 254 records were retrieved from the main query, 60 were retained after title-based screening, 43 formed the initial thematic core, and the final corpus reached 82 studies after snowballing and targeted complementary searches. Each study was then coded using the shared variables in Table 1. The coded spreadsheet and search strings are available upon request.

2.2. Coding scheme for systematic mapping

Table 1 summarises the coding scheme used to extract comparable metadata from each study/resource.

3. Systematic Mapping Results

This section reports the mapping results using the shared coding scheme (Table 1). We first focus on dataset construction and annotation practices, with emphasis on how intention, harm, and context dependence are represented (RQ1). We then summarise task families and evaluation protocols (RQ3), and finally map how emotion is represented and integrated into detection pipelines (RQ2). Throughout, we use compact tables to support cross-study comparison and to make the link between evidence and the proposed benchmark blueprint explicit.

Corpus overview. The mapped corpus contains 82 studies: 51 model papers, 7 dataset papers, and 24 survey/theory papers. The literature is overwhelmingly English (80/82), with only two multilingual studies. We treat dataset and model papers as empirical works (58) when reporting the operational coverage of Information Disorder variables.

Table 2 summarises two key quantitative findings that drive the rest of the paper: (i) intention and harm are rarely operationalised beyond proxies, and (ii) evaluation relies predominantly on random splits, limiting the validity of claims about generalisation.

Field	Values / notes
Resource type	dataset, benchmark, model paper, survey, tool
Languages	ISO codes; monolingual vs multilingual
Unit of analysis	claim, article, post, thread, user, network
Label type	veracity (binary/ordinal/multi-class), stance, entailment, intent, harm, context-dependence
InDor mapping	misinformation / disinformation / mal-information; or "not operationalised"
Intention	explicit label / proxy / not available
Harm	explicit type(s) / proxy / not available
Context dependence	low/medium/high; required metadata
Emotion representation	sentiment polarity; discrete emotions; appraisal; emotion dynamics; audience emotion
Emotion usage	feature; auxiliary task; analysis-only; propagation signal; discourse structure
Evaluation protocol	random split; source split; temporal split; cross-domain; cross-lingual; robustness
Metrics	macro-F1, AUC, calibration, class-wise recall, etc.
Availability	licence; link; documentation quality; ethical notes

Table 1: Fields used in the structured coding scheme for the systematic mapping.

InDor variable coverage (N=58)			
Var.	**Exp.**	**Proxy**	**None**
Intention	1	22	35
Harm	1	7	50
Context dep.	17	7	34

Evaluation split strategy (#)	
Split	**#**
Random	46
Temporal	5
Cross-domain	3
Cross-lingual	1
Other/none	3

Table 2: Mapping summary for empirical studies (datasets+models, N=58): coverage of Information Disorder variables and evaluation split strategies.

3.1. Current landscape: datasets and annotation practices

To address **RQ1**, we coded how datasets and benchmarks define their label space and which Information Disorder variables (intention, harm, context dependence) are explicitly annotated versus approximated via proxies.

In general, four main construction strategies can be observed. The first is manual curation by specialists, as in LIAR, the COVID-19 Fake News Dataset, and the PolitiFact-Oslo Corpus, where statements or full texts are individually verified by editors or fact-checkers (Wang, 2017; Patwa et al., 2020; Põldvere et al., 2023). The second consists of structured extraction from fact-checking platforms such as PolitiFact and GossipCop, as in FakeNewsNet (Shu et al., 2018). The third strategy is large-scale distant supervision, exemplified by Fakeddit, where labels are automatically inferred based on the nature of the originating subreddit (Nakamura et al., 2020). The fourth approach is theory-guided synthetic generation, as in MegaFake, which uses language models under the so-called LLM-Fake Theory to produce intentional disinformation grounded in principles of social psychology (Wang et al., 2024).

Despite this methodological diversity, most datasets are primarily structured around a criterion of factual veracity. The central organising axis is to determine whether content is true or false, with varying degrees of granularity, but rarely does dataset construction begin from an explicit modelling of communicative intention, type of harm, or contextual dependence.

Although some resources include contextual elements, such as author profiles, media type, network structure, or structured claim–evidence pairs as in Factify 2 (Suryavardan et al., 2023), context appears predominantly as auxiliary data rather than as a structured interpretative dimension. In most cases, there is no explicit annotation of communicative intention, potential harm, or degree of situational dependence.

Across the seven dataset papers (Table 3), intention is never explicitly labelled (5/7 use proxies; 2/7 omit it), and harm is at best proxied in 2/7 resources. Contextual dependence is more often present as structured metadata (explicit in 5/7), but it is rarely surfaced as a variable to be predicted or evaluated. In the next subsection, we examine how these resources are translated into supervised tasks and evaluation protocols.

3.2. Task formulations and evaluation protocols

To address **RQ3**, we grouped studies by task family and by evaluation protocol (split strategy, robustness tests, and metrics). This clarifies which kinds

Resource	Lang.	Unit	Veracity labels	Intent	Harm	Notes
LIAR (Wang, 2017)	EN	claim	6-point	none	none	speaker metadata; veracity-centric
FakeNewsNet (Shu et al., 2018)	EN	article+ social	binary	proxy	proxy	propagation metadata; social context
Fakeddit (Nakamura et al., 2020)	EN	post+ image	2/3/6-way	proxy	none	distant supervision via subreddit
COVID-19 Fake News (Patwa et al., 2020)	EN	post	binary	none	none	health domain; veracity labels
Factify 2 (Suryavardan et al., 2023)	EN	multi-modal	multi-class	proxy	none	multimodal evidence; includes satire
PolitiFact-Oslo (Põldvere et al., 2023)	EN	article	multi-class	proxy	none	full-text corpus; excludes satire
MegaFake (Wang et al., 2024)	EN	article	binary (synthetic)	proxy	proxy	theory-guided LLM generation

Table 3: Datasets/resources in the mapped corpus (N=7) and their alignment with Information Disorder variables (intent and harm coded as explicit/proxy/none).

of claims are empirically supported by a given setup and which Information Disorder dimensions remain outside the supervised target.

In our empirical subset (N=58), supervision remains overwhelmingly veracity-centric: 42 studies use binary veracity labels and 15 use multi-class veracity labels, with only one stance-based formulation. This reinforces a common conflation between Information Disorder categories and truth labels.

Evaluation protocols further constrain Information Disorder-relevant claims. As summarised in Table 2, 46/58 empirical studies rely on random splits; only 5 use temporal splits and 3 use cross-domain evaluation, and just one reports cross-lingual evaluation. Reported metrics concentrate on macro-F1 (35/58) and accuracy (20/58), while calibration and harm-aware error analysis are rare.

These patterns matter for interpreting emotion-aware work: improvements on veracity benchmarks do not necessarily imply better modelling of intention or harm.

3.3. Emotion in false information NLP

To address **RQ2**, we coded how affective information is represented (polarity, discrete emotions, transitions/dynamics, audience response) and how it is used (features, auxiliary objectives, analysis-only, or social/propagation signals).

Across the full corpus (N=82), 28 studies include an explicit emotion component (emotion representation other than *none*). In the model subset (N=51), 22 papers (43.1%) incorporate emotion: most rely on discrete emotion categories (15/51) or sentiment polarity (4/51), while emotion dynamics is rare (2/51) and audience emotion appears in one model. Emotion is most commonly used as feature augmentation (21/51), with fewer analysis-only studies (6/51) and only one discourse-structured approach.

Representative examples include feature-based emotion fusion for detection (Guo et al., 2023; Liu et al., 2024; Zhang et al., 2021), and dynamic modelling of emotion transitions as a marker of manipulation (Vieira and Figueira, 2025; Bian and Zhang, 2024). Crucially, among the 22 emotion-aware model papers, none provides explicit intention or harm labels, and harm is proxied in only 5 cases. This supports the broader pattern that emotion is mostly deployed to improve veracity discrimination rather than to operationalise Information Disorder dimensions.

Emotion repr.	#	Emotion usage	#
None	29	None	23
Discrete	15	Feature	21
Sentiment	4	Analysis-only	6
Dynamics	2	Discourse	1
Audience	1		

Table 4: Emotion operationalisation among model papers (N=51): representation and usage categories.

Overall, emotion is treated as a meaningful component of persuasion and engagement, but its evaluation remains coupled to veracity labels. The next section consolidates how this coupling, together with missing intention/harm/context variables, leads to systematic misalignment with the Information Disorder framework.

4. Structural Misalignment with Information Disorder

The Information Disorder framework distinguishes misinformation, disinformation, and mal-information primarily along intention, harm, and context dependence (Wardle and Derakhshan, 2017). However, the mapping in Section 3 indicates that the operational space of most NLP work remains veracity-centred. The problem is not only terminological: when labels collapse truth, communicative intent, and harm into a single target, improvements in model performance become difficult to interpret from an Information Disorder perspective. This matters especially for emotion-aware systems, whose affective cues may track engagement or rhetorical style without resolving whether content is accidental, strategic, or harmful.

The mismatch can be read at three levels. At the label level, veracity targets collapse communicative situations that the Information Disorder framework treats as distinct. At the feature level, context and emotion are often present as metadata or auxiliary signals but are not tied to explicit claims about intention or harm. At the evaluation level, random partitions reward within-source regularities and make it harder to know whether a model will generalise to new actors, events, or time periods.

Table 5 summarises recurrent gaps, why they matter for Information Disorder claims, and what a minimal benchmark should encode to address them.

Mapping evidence. Table 2 shows that intention is absent in 35/58 empirical studies and explicitly labelled in only one; harm is absent in 50/58 and explicitly labelled in only one; and contextual dependence is explicit in 17/58 but still missing in 34/58. Moreover, 23/82 model papers frame their target as *disinformation*, yet none includes explicit intention or harm variables, so the label is often used as a synonym for falsity. Finally, random splits dominate (46/58), increasing the risk of leakage and overestimating robustness under source and temporal shifts.

This misalignment matters because claims about "disinformation" are sometimes made from experimental setups that cannot distinguish deliberate from accidental falsehoods, or harmful reframing of true information. Consequently, progress in model accuracy does not necessarily translate into progress in theory-grounded Information Disorder detection. Put differently, the field already has useful predictive signals, but too few task formulations that bind those signals to the distinctions the framework actually cares about. The next section translates these gaps into concrete annotation variables, task definitions, and evaluation protocols.

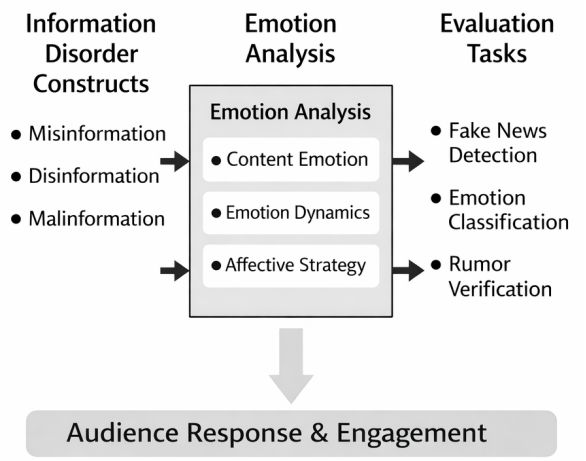

Figure 1: From Information Disorder theory to operational datasets, tasks, and evaluation: proposed bridge for emotion-aware NLP.

5. Proposed Direction

5.1. Operational taxonomy aligned with Information Disorder

We propose a minimal operational taxonomy that decomposes each instance into annotatable variables, rather than treating veracity as a proxy for intention or harm. Variables that are inherently subjective should allow an explicit *unknown* value and, where feasible, an annotator confidence score. This is particularly important for intention and harm, which can vary across cultures and contexts. Rather than forcing annotators to assign a single Information Disorder category from the outset, the taxonomy separates the underlying variables that make those categories interpretable. In practice, this makes it easier to document uncertainty and to distinguish cases that are false but strategically ambiguous from cases that are factually correct yet potentially harmful. The proposed dimensions and their suggested label sets are summarised in Table 6.

Two design principles follow from this decomposition. First, variables that can often be grounded more directly in available evidence, such as veracity, should remain distinct from variables that depend more heavily on pragmatic interpretation, such as intention and harm. Second, context

InDor dimension	Common operationalisation	Failure mode for InDor claims	Benchmark remedy
Intention	absent; inferred via source-level proxies; conflated with falsity	cannot distinguish misinformation (unintentional) from disinformation (strategic)	annotate intention (allow "unknown" + confidence); avoid source-only proxies
Harm	rarely labelled; sometimes implicit via topic (e.g., health)	models optimise for falsity, not for downstream risk or impact	add harm type/severity labels; report harm-aware metrics and error analysis
Context dependence	metadata present but not encoded as a variable; random splits	unclear when context is required; risk of leakage and brittle generalisation	include context-dependence labels; prefer source/temporal splits; report context ablations
Mal-information	typically invisible under veracity labels	harmful true content is misclassified or ignored	decouple truth and harm; include "true but harmful" cases and evaluate separately
Emotion	used as feature amplifier for veracity tasks	affective cues do not map to intent/harm; risk of overfitting emotional stereotypes	define affective strategy labels (incl. dynamics); evaluate robustness across domains and cultures

Table 5: Recurrent misalignments between Information Disorder constructs and common NLP operationalisations, with suggested benchmark remedies.

should be encoded not only as metadata but also as a judgement about how much extra information is required to interpret the case. This makes it possible to compare text-only systems with models that rely on thread, source, or temporal cues, and to report more clearly where uncertainty actually enters the pipeline.

Dimension	Suggested labels / annotation guidance
Veracity	true / false / mixed / unknown (claim- or post-level)
Intention	unintentional / deceptive / unknown (+ confidence)
Harm	harm type(s) (e.g., health, financial, reputational, civic) + severity (low/med/high) + unknown
Context dependence	low/medium/high; record minimal required context (thread, source, time)
Affective strategy	emotion cues (sentiment or discrete emotions), dynamics (transitions), and/or audience emotion
InDor category	derived from veracity+intention+harm when possible; allow "unknown"

Table 6: Operational taxonomy aligned with Information Disorder: annotatable variables and suggested label sets.

Treating the InDor category as a derived label, rather than as the only target to annotate directly, has two practical advantages. It makes annotation decisions more transparent by separating factuality, intention, and harm instead of collapsing them into a single judgement, and it allows uncertainty to remain explicit when one of those dimensions cannot be established reliably.

5.2. Benchmark blueprint

The blueprint in Table 7 translates the remedies in Table 5 into evaluation-ready tasks. It directly targets the main gaps observed in Table 2: scarce intention/harm labels, widespread random splits, and limited modelling of context dependence. The four tasks play complementary roles: T1 turns Information Disorder into a multi-target prediction problem, T2 tests how much contextual information is actually required, T3 isolates affective strategy as an object of analysis rather than a mere auxiliary feature, and T4 checks whether reported gains survive topic, source, and temporal shift.

In practice, T1 should be treated as a structured multi-label setting rather than as a single coarse classification task: some resources may support veracity and context labels but provide only uncertain evidence for intention or harm, so the benchmark should allow partial supervision and report performance per dimension in addition to any derived InDor category. T2 makes context dependence measurable rather than assumed by revealing source, thread, or temporal information incrementally and tracking when predictions stabilise. T3 separates affective strategy from coarse veracity prediction, enabling direct comparison between static emotion features, sequence-aware emotion dynamics, and audience-response signals. T4 acts as a shared stress test across topic, source, language, and time, so that gains are not interpreted only within convenient random partitions.

Task	Input	Output labels	Recommended evaluation
T1: InDor classification	post/thread + minimal context (source, time, conversation)	InDor category + intention + harm + context dependence	source and temporal splits; macro-F1 + calibration; error analysis by harm type
T2: Context dependence	post + progressively revealed context	low/medium/high (or ordinal) context dependence	context ablation curves; report minimal context for stable prediction
T3: Affective strategy	text segments and/or audience reactions	affective strategy type; emotion dynamics (transitions)	cross-domain and cross-lingual robustness; spurious correlation checks
T4: Robustness suite	any of the above	n/a	cross-topic, cross-source, temporal generalisation; leakage checks

Table 7: Benchmark blueprint at a glance: tasks, labels, and evaluation protocols.

Minimal baselines should include (i) text-only models, (ii) text+context metadata models, (iii) text+emotion feature models, and (iv) joint multi-task models predicting the full label set in Table 7. Reporting should go beyond macro-F1 by including calibration and error analysis stratified by harm type and context dependence. For corpora with derived InDor categories, reports should also state explicitly which variables were directly annotated and which were inferred by rule, since this affects both label reliability and comparability across datasets. Benchmark documentation should also record which context fields are available at training and test time, and whether emotion variables come from content, sequential dynamics, or audience response. This makes cross-study comparisons easier to interpret and reduces the risk of attributing gains to affect when they may actually come from metadata or split artefacts.

6. Conclusion

This paper argues that current emotion-aware NLP research on false information is often operationally misaligned with the Information Disorder framework. Through a systematic mapping of 82 studies, we showed that empirical work remains largely veracity-centred: explicit intention and harm labels are almost absent, context dependence is rarely operationalised as a variable, and evaluation relies mostly on random splits that can overestimate robustness. In response, we proposed a minimal operational taxonomy and a benchmark blueprint that make intention, harm, and context dependence first-class targets and that evaluate affective modelling in a way that supports theory-grounded, comparable progress.

Limitations. This mapping is iterative rather than exhaustive, and the resulting corpus may under-represent work outside the main venues indexed by our search sources. The literature we coded is strongly English-dominated (80/82), so cross-lingual generalisations remain limited. Screening and coding were conducted by the first author, without independent double-coding or a formal inter-rater agreement measure; this improves procedural clarity but remains a limitation of the present study. Coding also necessarily simplifies heterogeneous papers into shared categories, so some nuance (e.g., fine-grained task variants) is inevitably lost. Finally, the benchmark blueprint is conceptual and requires new datasets with explicit intention and harm annotation to be fully validated.

Ethics statement. Research on disinformation can be dual-use. Our work is a meta-level mapping and benchmark proposal; it does not release new disinformation content. For future datasets following the proposed blueprint, we recommend minimising the redistribution of harmful material (e.g., sharing identifiers or short excerpts when possible), documenting annotation guidelines and potential biases, and respecting privacy and platform terms when collecting social media data. The intended impact is to support more transparent, theory-grounded evaluation of systems that mitigate information disorder.

Acknowledgements

This work is funded by national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., under the support UID/50014/2025 (https://doi.org/10.54499/UID/50014/2025).

7. Bibliographical References

Haodong Bian and Lisheng Zhang. 2024. Fake news detection incorporating emotion transition in text. *Proceedings of the International Conference on Modeling, Natural Language Processing and Machine Learning*.

Bilal Ghanem, Paolo Rosso, and Francisco Rangel. 2019. An emotional analysis of false information

in social media and news articles. *ACM Transactions on Internet Technology (TOIT)*, 20:1 – 18.

Anastasia Giahanou, Paolo Rosso, and Fabio A. Crestani. 2019. Leveraging emotional signals for credibility detection. *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Quanjiang Guo, Zhao Kang, Ling Tian, and Zhouguo Chen. 2023. Tiefake: Title-text similarity and emotion-aware fake news detection. *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7.

Fei Liu, Xinsheng Zhang, and Qi Liu. 2024. An emotion-aware approach for fake news detection. *IEEE Transactions on Computational Social Systems*, 11:3516–3524.

Kai Nakamura, Sharon Levy, and William Yang Wang. 2020. Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. In *International Conference on Language Resources and Evaluation*.

Parth Patwa, Shivam Sharma, Srinivas Pykl, Vineeth Guptha, Gitanjali Kumari, Md. Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. 2020. Fighting an infodemic: Covid-19 fake news dataset. *ArXiv*, abs/2011.03327.

Nele Põldvere, Md. Zia Uddin, and Aleena Thomas. 2023. The politifact-oslo corpus: A new dataset for fake news analysis and detection. *Inf.*, 14:627.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2018. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big Data*, 8:171 – 188.

Kai Shu, Amy Lynn Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ArXiv*, abs/1708.01967.

S Suryavardan, Shreyash Mishra, Parth Patwa, Megha Chakraborty, Anku Rani, Aishwarya N. Reganti, Aman Chadha, Amitava Das, Amit P. Sheth, Manoj Kumar Chinnakotla, Asif Ekbal, and Srijan Kumar. 2023. Factify 2: A multimodal fake news and satire news dataset. *ArXiv*, abs/2304.03897.

Lejla Turcilo and Mladen Obrenovic. 2020. Misinformation, disinformation, malinformation: Causes, trends, and their influence on democracy. E-paper series “a companion to democracy”, no. 3, Heinrich Böll Foundation. Publicado em agosto de 2020.

Renatha Souza Vieira and Álvaro Figueira. 2025. Emotional sequencing as a marker of manipulation in social media disinformation. *Future Internet*, 17(12).

Soroush Vosoughi, Deb K. Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359:1146 – 1151.

Lionel Z. Wang, Yiming Ma, Renfei Gao, Beichen Guo, Zhuoran Li, Han Zhu, Wenqi Fan, Zexin Lu, and Ka Chung Boris Ng. 2024. Megafake: A theory-driven dataset of fake news generated by large language models. *ArXiv*, abs/2408.11871.

William Yang Wang. 2017. “liar, liar pants on fire”: A new benchmark dataset for fake news detection. In *Annual Meeting of the Association for Computational Linguistics*.

Claire Wardle. 2020. Understanding information disorder. First Draft News (long-form article). Accessed 8 June 2025.

Claire Wardle and Hossein Derakhshan. 2017. Information disorder: Toward an interdisciplinary framework for research and policy making. Technical Report DGI(2017)09, Council of Europe. Council of Europe report, published 27 Sep 2017.

Liang Wu, Fred Morstatter, Kathleen M. Carley, and Huan Liu. 2019. Misinformation in social media: Definition, manipulation, and detection. *SIGKDD Explor.*, 21:80–90.

Xueyao Zhang, Juan Cao, Xirong Li, Qiang Sheng, L. Zhong, and Kai Shu. 2021. Mining dual emotion for fake news detection. *Proceedings of the Web Conference 2021*.

Xinyi Zhou and Reza Zafarani. 2018. Fake news: A survey of research, detection methods, and opportunities. *ArXiv*, abs/1812.00315.