Journal of Sustainable Software Engineering and Information Systems
Volume 2, Issue 2, 167-183.
e_ISSN: 3123-3007
https://e-journal.gomit.id/jsseis
DOI: 10.58723/jsseis.v2i2.261

Balancing Societal Impact and Digital Footprint: A Sustainability-Aware Framework for AI-Based Misinformation Detection

Sofowora Mayowa Abiola
Lead City University, NIGERIA

Niken Parnia Nirda
University of Bengkulu, INDONESIA

Mojeed Taiwo Akolade
Lead City University, NIGERIA

Rajeswari Nakka
Seshadri Rao Gudlavalleru Engineering College, INDIA

Suresh Babu Chandolu
Dhanekula Institute of Engineering and Technology, INDIA

*** Corresponding author:**
Niken Parnia Norda, University of Bengkulu, INDONESIA. ✉nikenpn87@gmail.com

Article Info

Article history:

Received: October 16, 2026
Revised: November 06, 2026
Accepted: November 19, 2026

Keywords:
AI ethics
Digital sustainability
Energy-aware AI
Freedom of information
Misinformation detection
Socio-technical systems

Abstract

Background: AI-based misinformation detection can protect digital information environments, but existing evaluations often separate predictive performance from energy consumption, inference efficiency, operational cost, and societal consequences. This limits assessments of deployment sustainability.
Aims: This study develops a sustainability-aware framework integrating societal benefits and digital footprint to identify deployment configurations balancing misinformation-exposure reduction, information-trust support, predictive performance, and computational efficiency.
Methods: A controlled experiment used 983 unique public news titles from the PolitiFact subset of FakeNewsNet. A word- and character-level TF-IDF classifier was evaluated under four configurations: individual full-precision inference, batched full-precision inference, batched INT8 inference, and compact INT8 edge/offline inference. Metrics included macro-F1, misinformation-exposure reduction, an operational information-trust support proxy, latency, throughput, estimated energy consumption, cloud cost, and serialized model size. Confidence intervals were estimated using 3,000 bootstrap resamples.
Results: Individual full-precision inference achieved a macro-F1 of .873, 88.8% estimated exposure reduction, and an 87.4% social-benefit score, consuming 3.14 Wh per million inferences. Batching preserved predictive and social outcomes while reducing energy consumption to 0.33 Wh. Batched INT8 maintained these outcomes, consumed 0.32 Wh, and reduced model size from 128.2 to 82.2 KB. Compact edge inference consumed 0.05 Wh and reduced model size to 14.5 KB, but macro-F1 declined to .847 and exposure-reduction potential to 83.2%.
Conclusion: Batched INT8 inference offered the strongest overall sustainability balance, preserving societal benefits while reducing energy consumption, latency, cloud cost, and model size. The framework supports integrated social–technical evaluation of sustainable misinformation-detection systems.

To cite this article: Abiola, Sofowora Mayowa., Norda, Niken Parnia., Alokade, Mojeed Taiwo., Nakka, rajeswari., & Chandolu, Suresh Babu. (2026). Balancing Societal Impact and Digital Footprint: A Sustainability-Aware Framework for AI-Based Misinformation Detection. *Journal of Sustainable Software Engineering and Information Systems, 2*(2), 167-183. https://doi.org/10.58723/jsseis.v2i2.261
This article is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License ©2026 by author/s

INTRODUCTION

Misinformation has become a persistent problem within digitally mediated information environments (King et al., 2025; Wardle & Derakhshan, 2017). Social media platforms, online news channels, messaging applications, and algorithmic recommendation systems allow information to circulate at a speed and scale that exceed conventional editorial and institutional verification

mechanisms. Misinformation not only results in misunderstanding among individuals, but has the potential to undermine the public's trust, exacerbate crises, enhance social division, and diminish the quality of collective deliberation. False information spreads easily when it is new, evokes emotion, and aligns with people's existing beliefs (Pennycook & Rand, 2021; Vosoughi et al., 2018).

AI has played a large role in studying misinformation. Present methods extract features from linguistic elements, sources, user interactions, the distribution of information, and even multifarious data types to distinguish genuine information from false information (Lazer et al., 2018; Shu et al., 2017; Zhou & Zafarani, 2020). The FakeNewsNet has advanced this subfield significantly by providing datasets containing information on fact-checked content, context, and social dissemination (Shu et al., 2020). Researchers use these datasets to study classification and evaluate the methods of generating, disseminating, and detecting misinformation.

Most studies of misinformation-detection systems continue to place an emphasis on predictive metrics. Though accuracy, precision, recall, F1-score, area under the receiver operating characteristic curve, and calibration are important, poorly calibrated systems may either fail to detect misinformation or detect misinformation in legitimate content. However, predictive performance of a system is not the sole indicator of its social usefulness (Baqraf et al., 2026; Liu et al., 2025). For example, a classifier with high accuracy may still fail to detect most of the misinformation. Alternatively, an overly aggressive classifier may reduce exposure to false information, at the cost of generating so many false positives that it reduces user confidence in the classifier and increases actual information barriers. The usefulness of detection systems is therefore determined by the error distribution, how those errors determine exposure to different types of trustworthy and untrustworthy information, and how the errors affect the system's intervention (Butler et al., 2024; Hoes et al., 2024).

Like predictive performance, the ethical implications are equally important. Responsible artificial intelligence cannot be evaluated solely through predictive performance; ethical principles, accountability, transparency, and societal consequences should also be considered when evaluating AI systems (Floridi et al., 2018; Ikram et al., 2026; Jobin et al., 2019). This principle is particularly applicable to misinformation detection systems because the classification of information systems may impact the visibility, reputation, and understanding of the information and its integrity. Therefore, an ethically acceptable system is one that creates significant social value while avoiding unnecessary costs and resources, and maintains a high degree of transparency.

Trust is one of the primary social outcomes of information systems. In the case of classification systems, trust is not simply a matter of efficiency. Trust, like other social outcomes, is contextual and is determined by perceptions of consistency, timeliness, sincerity, and credibility (Rachmiatie et al., 2026). A system designed to detect misinformation may decrease trust if it is slow but continually flags credible content. On the other hand, if a system is designed to maintain trusted content while amplifying misinformation, it would improve trust. Trust, therefore, is a socio-technical outcome. It is generated by the interaction of classification systems, the governance of a communication platform, and social interpretations of the communication.

The third limitation of conventional evaluation is the omission of the digital footprint. The development, training, and deployment of AI systems consume energy (Patterson et al., 2021). So do the transfer of data, continual system monitoring, and the numerous inferences the system is tasked with making. Even if a single invocation of the system requires little energy, a high volume of requests can make repeated use burdensome. Because of the environmental impacts of AI, there have been calls to incorporate the reporting of system resources, energy consumption, emissions, and the efficiency of systems into the research on model quality (Dodge et al., 2022; Henderson et al., 2020; Schwartz et al., 2020; Strubell et al., 2019). Digital sustainability is about more than just using digital tools to resolve social and environmental challenges. It is about ensuring that the tools are designed, deployed, and managed for the long term in a way that ensures the digital tools do not compromise the environment, society, and the economy.

Social value and digital footprint are important in continuously running information systems. Creating a misinformation detector may mean thousands or millions of items are processed at a time.

This may make model size, cost, latency, and energy consumption important design criteria. Batching may help address these concerns by using multiple requests to be processed together. Quantisation helps reduce model parameter size and computational overhead by using lower precision values to represent model parameters. Edge or offline inference helps address the burden of constant data transfer and reliance on cloud services, but it may reduce the computational power of the model and the predictive capacity. The Cloud–edge architecture may help address computational burden based on latency, connectivity, privacy, and resource constraints (Fan et al., 2026). Energy-aware edge design also helps increase the accuracy of the model while helping adjust the computational burden within the system (Alawad et al., 2026).

Optimising systems from a purely technical perspective leaves many unanswered questions. For a quantised model, if a considerable drop in accuracy is observed, then the configuration may have been deemed technically successful. A service configured to be deployed in a batched manner is considered successful if throughput is drastically improved. From these framings, the system is successful only if social benefit is reduced. Whereas for a configuration that addresses social benefit, more computational resources may be needed to achieve minimal improvement in system capability.

Sustainable software engineering and information systems research typically consider environmental, social, and economic dimensions; however, empirical integration is still nascent. A recent study published in the target journal showed that the field is dominated by conceptual and review-based studies, and that quantitative research is still scant, and a standard set of sustainability metrics is still not established (Saputra et al., 2026). This gap is particularly evident in socially conscious AI applications. Researchers define social issues, including AI ethics, trust, and computational sustainability, but few researchers integrate them into a unified, repeatable assessment framework.

This paper seeks to fill that gap by introducing a sustainability-oriented AI-based misinformation detection framework. In contrast to other frameworks, this framework integrates societal impact and digital footprint as two interconnected aspects of the evaluation framework. The framework's societal dimension incorporates a proxy for information trust based on the measures of precision, specificity and probability calibration, and a reduction in exposure to misinformation. The framework's technical dimension comprises macro-F1, inference latency, throughput, energy consumption and cloud cost, and size of serialised model in bits. These metrics will be evaluated for full-precision, batched, quantised, and edge-oriented models.

Three research questions are answered in this study. First, how can social outcomes and technical resource indicators be integrated in a single evaluation framework for the detection of misinformation? Second, which optimisation techniques lessen the digital footprint without significantly affecting the estimated social benefit? Third, how does the most preferred choice alter when system selection is done based on energy, latency, model size, and cloud cost, as compared to the case when predictive performance is the only consideration?

There are three main contributions. First, the study builds a socio-technical model integrating information integrity, public trust, responsible AI, and digital sustainability. Second, the study's construction of transparent equations and measurement boundaries provides a means to quantify social and technical indicators. Third, a set of techniques, most notably request batching, INT8 quantisation, and compact edge deployment, yield distinct sustainability profiles when applied to the same detection problem and test data.

METHOD

Research Design

The study used a controlled comparative prototype design. The unit of comparison was a complete inference configuration rather than an isolated algorithm. Each configuration combined a text representation, classifier parameterisation, request-processing strategy, and deployment mode (Memon et al., 2024; Sovrano et al., 2024; Wambsganß et al., 2024). All configurations were evaluated

on the same held-out observations so that differences could be attributed to system design rather than changes in the test corpus.

The research design isolated two types of validity. Predictive validity focused on whether the system distinguished reliable information from false information. Sustainability validity focused on whether the configuration sustained social utility while reducing its operational digital footprint. Because of this distinction, an optimised configuration could not be sustainability valid just because it consumed less energy (Catyanadika et al., 2024; Hina et al., 2024). A configuration was preferable only when it achieved a lower resource consumption, a better performance of detection, and a potentially more positive social outcome (Rawat & Rajavat, 2026).

Sustainability-Aware Conceptual Framework

The conceptual framework connects five elements: information input, AI detector, deployment configuration, societal impact, and digital footprint. Information flows into the system from public news or platform sources. The AI detector captures the information and classifies it. It outputs a label and a probabilistic estimate. The deployment configuration dictates whether inferences are performed one-at-a-time or in batches, whether full-precision or quantised parameters are used, and whether the inference is done on centralised or edge devices.

The elements in the conceptual framework are represented in Figure 1. The branches of societal and technological impacts are separated after the model runs, so they can be assessed separately and then joined in a sustainability-sensitive decision.

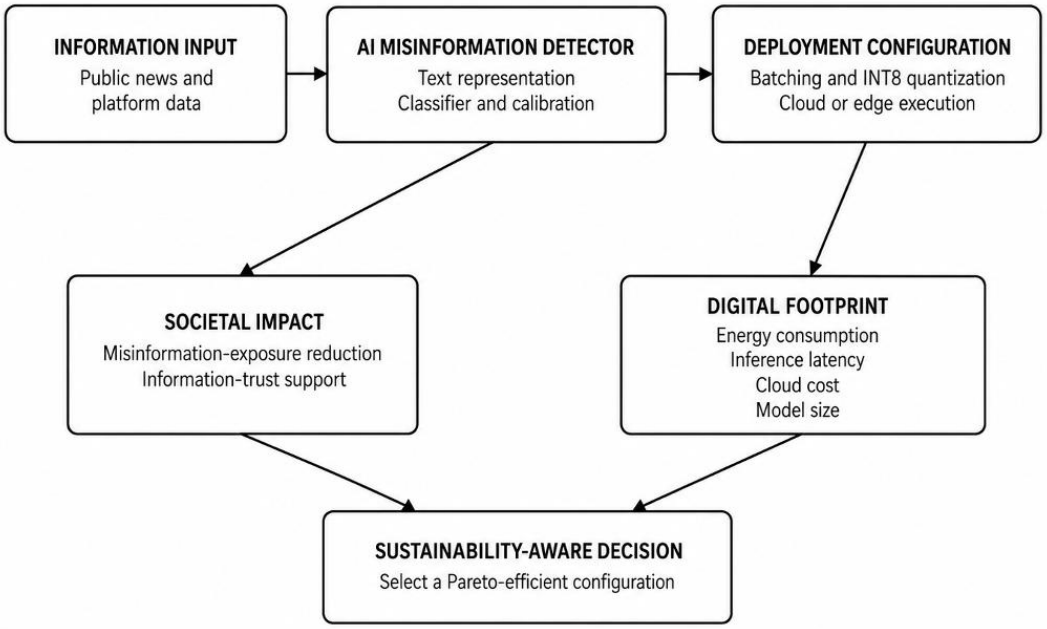


Figure 1. Integrated social–technical sustainability framework for AI-based misinformation detection.

The framework in Figure 1 doesn't assume that the design with the highest predictive score is the optimal design. Rather, the official judgment is made via Pareto reasoning. A configuration is considered attractive when there is no alternative for which society would gain more benefit with less use of technical resources. Deployment contexts with varying priorities for detection coverage, trust support, energy, latency, and cost don't conceal measurement bias with this approach.

Data Source and Sample Construction

The prototype used the public version of FakeNewsNet's minimal data set. The data set contains news articles with PolitiFact and GossipCop ground-truth labels. Wherever possible, the data set contains unique identifiers, source URLs, news article titles, and tweet IDs (Shu et al., 2020). This study used the PolitiFact subset because it provides labels grounded in political fact-checking and represents a socially consequential misinformation context.

The original data set contained 432 records that were labelled false and 624 records that were labelled real. Records that did not contain titles were removed. Duplicate titles of records were retained, but previous observations of records with duplicate titles were discarded. The final data set contained 983 unique titles consisting of 428 false and 555 real records. No tweet content, user identification, private personal messages, or identifying information were used. Therefore, the study was focused on content detection and did not reconstruct social network relationships.

The cleaned data set was used for a stratified 75:25 split with a fixed random seed. The training data set comprised 737 records, and the test data set comprised 246 records. Stratification ensured that the distribution of false and real records was the same for the two sets. The same test data set was used for all configurations.

Three reasons explain why the study utilised news article titles over full articles. First, a minimal dataset may include titles consistently. Second, titles are an example of an early screening input. Titles can be processed before full content is collected. Third, since short text classification is able to operate on the edge with a small model and a small amount of data, it is an ideal candidate for edge analysis. It should be noted that the prototype is not meant to completely replace text-based systems that include articles, sources, content propagation, images, and user involvement.

Text Representation and Baseline Classifier

The principal model utilised word- and character-level TF-IDF features. Word features included unigrams and bigrams, while character features were defined as character sequences of 3 to 5 within word boundaries. Word-level features were used to capture lexical and phrasal repeated patterns, while character-level features were more tolerant of variations in spelling, punctuation, inflexions, and styles.

Features appearing in fewer than two training examples were ignored. With the combined representation, 9,302 features were active. A class-weighted logistic regression classifier with a penalty of 2.0 was trained. Class weighting was employed since real titles outnumbered false titles. Logistic regression was used since it allows for fast CPU inference, has a reasonable basis for lower-precision deployment, and provides interpretable results.

The compact edge configuration used only word-level unigrams and bigrams and had a vocabulary of 1,302 features. The remaining components of the configuration (classifier family, regularisation parameter, class weighting, train–test split, and decision threshold) remained constant. Given the smaller vocabulary, the configuration caused a reduction in both the serialised size of the configuration and the demand for inference while exposing the expected performance trade-offs due to feature compression.

Deployment Configurations

Four configurations were evaluated.

C0: Full-precision individual inference. Each of the transformed and classified titles used the combined word- and character-level representation. The model parameters were maintained in their original full precision. This configuration provided a simple request-by-request service.

C1: Full-precision batched inference. The same representation and model were used; however, requests were processed in batches of 32. Batching only affected the execution model, and, thus, it was expected to be neutral to performance and predictive outputs.

C2: Batched INT8 inference. The combined model was processed in batches of 32; classifier weights and nonzero feature values were represented with signed 8-bit integers. A scale factor was kept to map back integers to the original decision score range. Quantisation was applied with the constraint of reducing the model precision and maintaining the approximation error within acceptable bounds (Jacob et al., 2018).

C3: Compact INT8 edge or offline inference. The compact word-level representation was used in combination with INT8 parameters and a batch size of 64. The configuration was treated as edge or offline, as it required no remote inference call after the model package and vocabulary were deployed.

Cloud processing cost was therefore zero; however, local hardware cost remained represented and was not assumed to disappear in this configuration.

Social Outcome Measures

The framework operationalised misinformation-exposure reduction as the proportion of false items correctly identified:

$$MER = TP / (TP + FN) \times 100$$

where MER is misinformation-exposure reduction potential, TP is the number of false items correctly detected, and FN is the number of false items missed. Under the evaluation assumption that correctly flagged items are intercepted before exposure, MER corresponds to false-class recall. In a live platform, actual exposure reduction would additionally depend on intervention coverage, ranking position, user behaviour, and the time between publication and detection.

Perceived trust was not measured through human participants. Instead, the prototype used an information-trust support proxy. The proxy was designed to represent three properties likely to support trust in a platform-level intervention: correct warning labels, preservation of legitimate content, and calibrated confidence. It was calculated as:

$$ITS = 100[0.50P + 0.30S + 0.20(1 - BS)]$$

where ITS is the information-trust support score, P is precision for false-content predictions, S is specificity for legitimate content, and BS is the Brier score. Precision received the largest weight because unreliable warnings can weaken confidence in the intervention. Specificity represented the system's capacity to avoid incorrectly restricting legitimate information. The calibration term rewarded probabilities that corresponded more closely to observed outcomes.

The proxy is an operational indicator rather than a psychometrically validated measure of perceived trust. It translates classifier behaviour into a testable system property but does not claim to represent the beliefs of actual users. This distinction is important because trust is socially and contextually constructed, as demonstrated by research showing that consistency, timeliness, sincerity, and credibility contribute differently to trust in AI-mediated communication (Rachmiatie et al., 2026).

A composite social-benefit score was calculated as:

$$SBS = 0.60MER + 0.40ITS$$

Reduction in exposure to misinformation received a weight of .60, as the primary function of the system is to reduce exposure. Information-trust support received a weight of .40 to ensure that the evaluation didn't penalise recall while encouraging false positives or poor calibration. Depending on the operational context, flexibility to alternative weights is required.

Technical and Digital-Footprint Measures

For evaluation of predictive performance, macro-F1 was utilised to encompass false and true positives. The other metrics evaluated were accuracy, false-class precision, false-class recall, specificity, the AUC, and the Brier score.

To evaluate inference time, a monotonically increasing high-resolution timer was employed. For single inference, latency was defined as the time required to transform and classify a single item. For the latency of batched inference, the reported service latency was the time required to process all items. Batch-completion latency was solely defined as the time required to complete a batch of requests to differentiate the time required for processing of items from the time required to wait for the requests to complete. Throughput was defined as the total number of test items divided by the total time it took to process all items.

The benchmark ran in a CPU-only virtualised environment with five allocated processor cores identified as AMD EPYC 9V74 virtual processors. No GPU was used. Individual inference was repeated eight times, and batched configurations were repeated 50 times after warm-up execution. Predictive performance metrics were evaluated based on a single held-out run, while latency and energy metrics were evaluated based on the average of multiple runs.

Energy consumption was estimated rather than directly measured at the electrical outlet. The calculation followed the general structure used in computational-footprint estimation methods (Lannelongue et al., 2021):

$$E = (tCPU / 3{,}600) \times P \times PUE$$

where E is energy in watt-hours, tCPU is processor time in seconds, P is an assumed allocation of 10 watts per active virtual processor, and PUE is a power usage effectiveness factor of 1.4. Results were scaled to one million inferences. Because the same power assumptions were applied to all configurations, relative differences primarily reflect measured processor-time differences. Absolute values should be interpreted as estimates rather than hardware-meter readings.

Cloud cost was calculated using a normalised rate of US$0.05 per virtual processor-hour:

$$\text{Cloud cost} = (tCPU / 3{,}600) \times US\$0.05$$

The value was scaled to one million inferences. This was a comparative cost model rather than a quotation from a specific cloud provider. Data-transfer, storage, gateway, monitoring, and managed-service charges were excluded. The compact edge configuration was assigned zero cloud-inference cost because execution occurred locally, but this did not imply zero total ownership cost.

Serialised model size was measured from compressed model packages containing the fitted vocabulary or feature representation and classifier parameters. It therefore reflected the deployable software artefact rather than only the classifier coefficient array.

Evaluation Workflow and Statistical Analysis

Figure 2 summarises the evaluation procedure from data preparation through the final social–technical comparison. The workflow uses the same test observations across all scenarios, preserving a paired design.

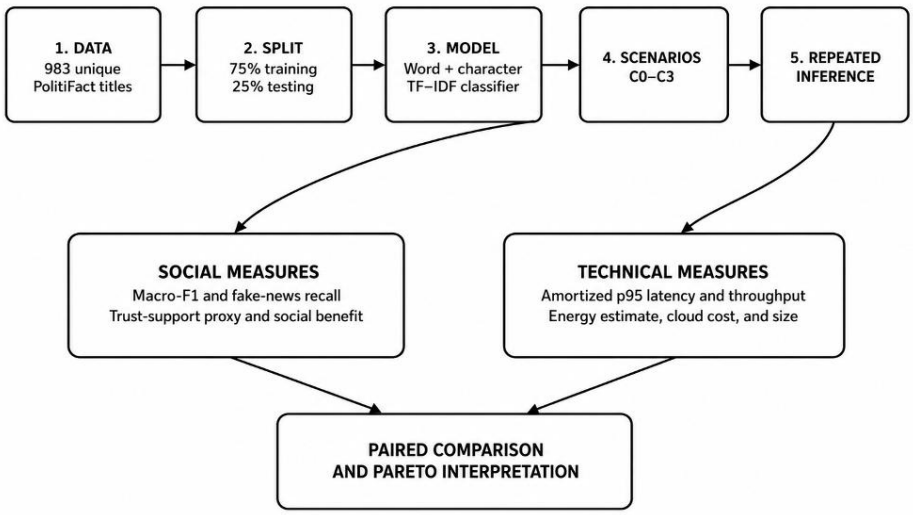


Figure 2. Prototype evaluation workflow and measurement boundary.

The evaluation first established predictive validity, then calculated the societal and technical indicators separately. The two branches were combined only during comparative interpretation. This ordering reduced the risk of hiding a substantial predictive loss behind a composite sustainability score.

Confidence intervals were estimated using 3,000 nonparametric bootstrap samples of the test set. The procedure resampled observations with replacement and recalculated macro-F1, misinformation-exposure reduction, and the social-benefit score. Percentile-based 95% confidence intervals were reported. Paired bootstrap differences were also examined for the contrast between the full and compact configurations.

Ethics, Data Governance, and Reproducibility

This study used public data, did not recruit subjects, and did not evaluate any private accounts; therefore, the BERT-based classifier was not trained on any data that would require user-level tweet identification. Because the research did not make a decision about individual users, the ethical risk of this study is less than that of a live moderation study. However, the labels added to remove misinformation posts may be relevant in a political and social context. For this reason, the prototype must be considered a discussion about the design of an information system and not a case for the automated removal of content.

Governance must include the process and outcome of training data, error analysis, model limits, appeal mechanisms, human review for high-impact interventions, and continuous monitoring for temporal or domain shifts. These requirements are consistent with the need to move AI ethics from broad principles to operational governance mechanisms (Ikram et al., 2026; Technology, 2023).

The public FakeNewsNet files contain everything needed to reproduce this benchmark. The reporters also provided the logic of the dataset split, the features, the formulas, and the configuration. Exact values of latency and energy will vary across different hardware and software. For replication, the reporters should also provide the processor model, number of active threads, all of the software and their versions, the warm-up procedure, how many times the experiment was repeated, the PUE assumption, the method of power estimation, and whether the energy was measured or estimated.

RESULTS AND DISCUSSION

Results:

Predictive Validity of the Baseline System

The combined word- and character-level classifier produced 120 true-negative, 19 false-positive, 12 false-negative, and 95 true-positive predictions on the 246-item test set. The accuracy was .874, the macro-F1 was .873, the false-class precision was .833, the false-class recall was .888, the specificity was .863, and the area under the receiver operating characteristic curve was .936. The Brier score was .116.

These results show that the prototype was able to differentiate false and true labels significantly better than random classification, whilst returning a mostly even distribution between the two classes. The false-negative count was important because it left 12 misleading labels unidentified. The 19 false positives were also important because they represented legitimate labels that were unnecessarily flagged. As a result, the sustainability framework maintained both exposure-reduction and false-positive-sensitive measures, contrary to reporting accuracy.

The baseline macro-F1 for bootstrap was [.829, .913]. The estimated misinformation-exposure reduction was 88.8%, with a 95% confidence interval of [82.6%, 94.4%]. The corresponding information-trust support score was 85.3%, resulting in a composite social benefit score of 87.4% with a bootstrap interval of approximately [82.7%, 91.5%].

Integrated Configuration Results

Table 1 shows the four deployment configurations. C0 and C1 used the same full-precision model, resulting in the same predictive and social results. C2 introduced integer quantisation while still making the same classification decisions at the .50 threshold. C3 reduced complex representation and achieved a below-threshold, although practically useful, social benefit score.

Table 1. Social and technical performance of the misinformation-detection configurations

Configuration	Deployment design	Macro-F1	Exposure reduction (%)	Trust support (%)	Social benefit (%)	Amortized p95 latency (ms)	Throughput (requests/s)	Energy (Wh/1M inferences)	Cloud cost (US$/1M)	Model size (KB)
C0	Full precision,	.873	88.8	85.3	87.4	0.985	1,245	3.14	0.0112	128.2

Configuration	Deployment design	Macro-F1	Exposure reduction (%)	Trust support (%)	Social benefit (%)	Amortized p95 latency (ms)	Throughput (requests/s)	Energy (Wh/1M inferences)	Cloud cost (US$/1M)	Model size (KB)
	individual requests									
C1	Full precision, batch = 32	.873	88.8	85.3	87.4	0.100	11,739	0.33	0.0112	128.2
C2	INT8 quantization, batch = 32	.873	88.8	85.2	87.4	0.101	11,779	0.32	0.0112	82.2
C3	Compact INT8 edge/offline, batch = 64	.847	83.2	84.4	83.7	0.015	76,707	0.05	0.0000*	14.5

Note. Energy values are estimates based on measured CPU time, an assumed allocated processor power of 10 W and PUE = 1.4. Cloud cost excludes network, storage, monitoring, local electricity, and device ownership. C3 has zero cloud-inference cost because inference is local, not because total deployment cost is zero.

The table illustrates that optimisation did not impact each model evenly. Among all models, batching improved efficiency the most while requiring no changes to the model or predictions. Estimated energy for C1 dropped by 89.6%, and throughput increased by a factor of greater than nine. The amortised p95 service latency reduced from 0.985 to 0.100 ms. p95 completion time for a whole, 32-item batch, was approximately 3.0 ms. This demonstrated that the improvement to throughput required a computation time of less than 3.0 ms after batch formation.

C2 preserved the macro-F1, exposure-reduction estimate, and social-benefit score of C0 and C1. The area under the curve of the receiver operating characteristic changed from .9355 to .9353, and the Brier score changed from .1158 to .1159. These changes were very minor and thus were not meaningful enough to influence classification at the chosen threshold. The serialised package was reduced by approximately 35.9%, from 128.2 to 82.2 KB. Thus, model compactness was improved without an important performance penalty in the social dimension in the experiment.

The difference between C1 and C2 was small for the measured inference demand. This is because text transformation was a significant portion of the total work required. This is an important discovery for designing sustainable software: decreasing the size of classifier parameters does not eliminate the inference cost of tokenisation, feature construction, data transfer, orchestration, surrounding services, etc. The entire pipeline should be evaluated for optimisation, as inference demand should not be assumed.

C3 produced the strongest technical efficiency. Its energy estimate was approximately 98.4% lower than C0, serialised size was 88.7% smaller, and throughput was more than 60 times higher. These gains were accompanied by a reduction in macro-F1 from .873 to .847 and misinformation-exposure reduction from 88.8% to 83.2%. The bootstrap 95% interval for C3 macro-F1 was [.798, .890], and its exposure-reduction interval was approximately [75.9%, 90.0%]. The paired bootstrap interval for the C3-minus-C0 macro-F1 difference was approximately [–.059, .007]. This interval includes zero, but its distribution and point estimate indicate a credible risk of reduced detection quality that should not be ignored simply because the compact model is more efficient.

Societal Benefit–Digital Footprint Trade-Off

The joint relationship between estimated social benefit and energy use is shown in Figure 3. The horizontal axis represents the energy estimate for one million inferences, the vertical axis represents the composite social-benefit score, and marker size represents serialised model size.

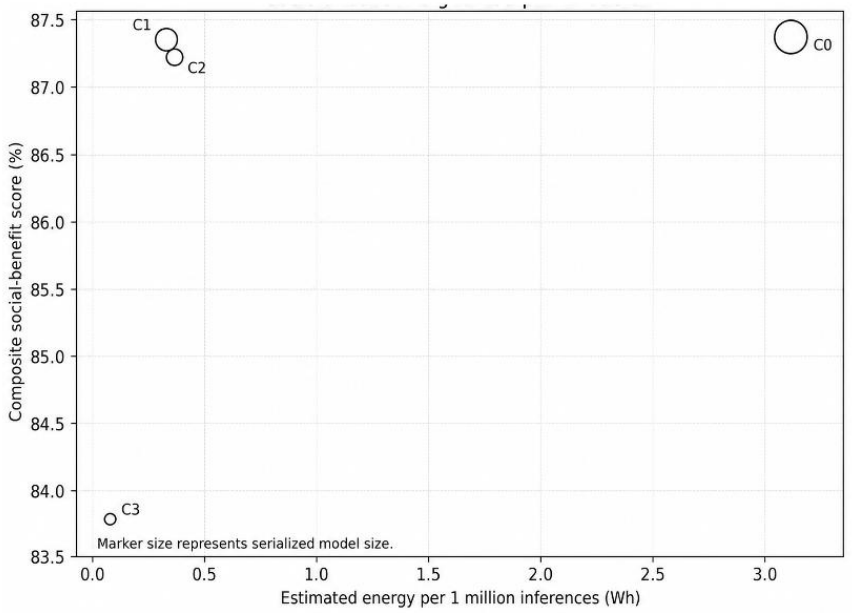


Figure 3. Societal benefit–digital footprint trade-off across four deployment configurations.

Figure 3 shows that C0 was dominated by C1 and C2. Both alternatives preserved the same social-benefit score while consuming far less estimated energy. C2 was generally the best configuration for cloud-related scenarios, as it further reduced the model size relative to C1.

C3 did not fully dominate the other configurations due to the fact that it offered an unprecedented level of energy efficiency and compactness. C3 occupied a distinct area of the trade-off space, offering a low digital footprint and a lower social benefit. As such, its suitability is dependent on deployment context. For instance, if a fact-checking assistant runs on a device with low resources and without continuous connection, C3 may be preferred. However, a national service with a higher degree of intervention, such as news monitoring, may prefer C2, as the high level of misinformation recalled justifies the cost of the resource used.

This case study shows why having a single sustainability ranking is problematic. Sustainability can be multidimensional and context-dependent. This framework does not advocate that each organisation should apply the same weighting. Rather, it describes trade-offs for different weighting schemes. Then, it is up to decision-makers to evaluate if a trade-off of approximately 5.6 percentage points in misinformation-exposure coverage is acceptable, in comparison to an added 84% energy efficiency, on moving from C2 to C3.

Normalised Sustainability Profiles

In Figure 4, configurations are depicted with normalised scores. Social benefit and macro-F1 retain the same unit percentage values as previously described. The rest of the metrics, energy efficiency, latency efficiency, and compactness, are min–max normalised.

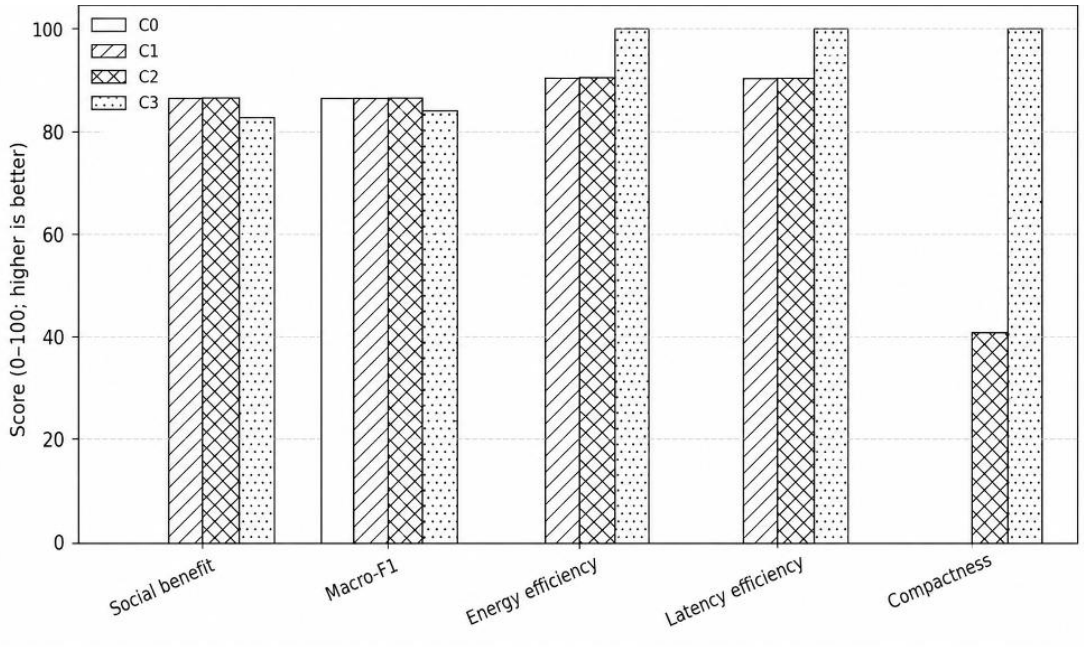


Figure 4. Normalised sustainability profile by configuration.

This analysis suggests that the inclusion of resource indicators in the sustainability framework confirmed C0 as a non-compensatory solution. C1 and C2 showed a trade-off of equal social benefits and predictive performance. However, C2 was more compact. C3 was the most energy-efficient, latency-efficient, and compact of the configurations, but had the poorest social benefit and macro-F1.

The example also shows that the term "energy-aware AI" should not be interpreted as "the lowest-energy AI." A system that minimised energy yet lost its purpose could be environmentally efficient, but socially ineffective. Sustainability-aware design includes a social performance boundary that provides a minimally acceptable level of social outcome. In the case of the data provided, C2 was the preferred balanced design because it achieved improvements in energy, latency, cost, and size while maintaining the baseline level of social outcomes. C3 became the preferred design with an even stronger prioritisation of resource constraints.

Contribution to Sustainable Software and Information Systems

The results provide three contributions to the evaluation of sustainable software. To begin with, they place social sustainability at the level of system outcomes. In contrast to social benefit framed as a broad narrative, the framework links this to behaviour that can be measured. In misinformation recall, we consider the estimate of how much false content could be intercepted. The benefit of the trust-support proxy is to ensure that this is not evaluated in isolation from false warnings and probability calibration. Furthermore, the study evaluates the deployable configuration in its entirety. Sustainable software engineering typically considers isolated properties, such as processor utilisation, energy consumption, model compression, or cloud architecture. The main result of this study is that the main boost in efficiency was from the orchestration of requests and not quantisation alone. This strongly suggests a systems-level approach, in which model selection, representation, batching, deployment, and the associated governance procedures are considered as coupled and interdependent.

Finally, the framework proposes a way to address the lack of standardised and empirically validated sustainability metrics in the JSSEIS research (Saputra et al., 2026). The equations suggested do not attempt to provide absolute standards, but rather offer a grounded basis for assessing and potentially adjusting to other contexts. In this particular case, misinformation exposure could be replaced with another appropriate domain-specific outcome.

Relationship to Responsible AI and Public Trust

The results show that a socially influential AI classifier goes beyond a model that is technically correct. An AI classifier becomes socially significant when its output determines the information a user sees, the content that is labelled, and the claims that are prioritised for review. Governance mechanisms must therefore define acceptable error rates, human-review thresholds, explanation practices, and appeal procedures (Raji et al., 2020).

Ikram et al. (2026) see a continued need for pragmatic AI governance structures. With this in mind, the sustainability-focused framework converts some ethical issues to measurable system components. False-positive-sensitive trust support is linked to proportionality and information access. Exposure reduction is linked to societal benefit. Calibration is related to the transparency and confidence of the system. Energy and cost are related to responsible use of resources and long-term sustainability of the system. The framework shows that trust should not be solely based on metrics. Rachmiatie et al. (2026) show that trust in AI-mediated communication consists of relational and perceptual aspects that cannot be based on classifier outputs. Thus, a real-world implementation of the operational proxy should be coupled with user research. Users can be presented with the same information environments and evaluated using different warning strategies concerning the appearance of the environments, trust in the system, the perception of management, and confidence and willingness to verify the information. The proxy should be considered an indicator of a system's readiness for this type of assessment.

Discussion:

The first implication for system developers is that redesigning a functioning model should be done with care and after consideration of request batching. Model C1 resulted in a significant reduction in processing workload with no predictive change. Dynamic micro-batching might be useful for systems where workloads are (and will be continually arriving, since it should improve system utilisation while preserving system responsiveness. The optimum batching window should be determined based on real-world system traffic. Large batching may increase queueing latency, even though the average service time is expected to be low.

The second implication is that for pipeline-level quantisation, model C2 achieved a trade-off of small energy consumption at the cost of an even smaller size compared to model C1. Smaller quantisation of pipeline parameters might result in greater energy savings in neural architectures with larger sizes. In lightweight text pipelines, the overhead and feature extraction components might dominate the processing workload. Therefore, engineers should report detailed component-level studies, while not assuming that smaller weight format implies smaller operational workload.

The third implication concerns edge deployment. C3 showed that a compact local detector can substantially reduce cloud dependence and remain useful. Edge execution may also support privacy and resilience because content does not need to leave the local device for every inference. However, local models require version control, secure updates, monitoring, and fallback mechanisms. A stale or contextually inappropriate model can create social harm even when its energy use is low.

The fourth implication is that the decision threshold should be treated as a governance parameter. The present study used .50 for comparability, but platforms may select a higher threshold for automatic warning labels and a lower threshold for human-review queues. Threshold selection changes exposure reduction, false positives, trust support, and workload simultaneously. Sustainability-aware evaluation can therefore be extended from configuration selection to policy selection.

Implications:

Organisations deploying misinformation detection should require both model-performance reporting and operational sustainability reporting (Mitchell et al., 2019). A procurement or governance document should specify the evaluation corpus, prevalence assumptions, subgroup or topic performance, calibration, latency distribution, energy-measurement method, cost boundary, model size, update frequency, and expected intervention effect. Policy guidance should avoid equating the presence of an AI detector with a successful response to misinformation. A system can be accurate in a laboratory but ineffective in practice if detection occurs after content has already

circulated widely. Likewise, a system may reduce misleading exposure while undermining trust through unexplained or inconsistent interventions (Pawelczyk et al., 2026; Wang et al., 2026). Evaluation standards should therefore include timing, user-facing communication, review procedures, and post-deployment monitoring. Organisations should also distinguish between estimates and measured values of environmental parameters. Software-based energy estimates may be used for indirect comparisons, but the accuracy of such estimates may vary significantly. A hardware-based approach (e.g., hardware counters or external meters) is recommended for making claims (Kirkeby & Lagermann, 2025; Rojahn & Grum, 2026). Most software energy estimators have demonstrated the ability to approximate overall system energy consumption. Sustainability reporting should disclose these uncertainties.

Research Contribution:

This study contributes to sustainable software engineering and information systems by introducing an empirically evaluated framework that integrates societal impact and digital footprint in assessing AI-based misinformation detection systems. Unlike conventional evaluations that primarily focus on predictive performance, the proposed framework considers misinformation-exposure reduction, information-trust support, computational efficiency, energy consumption, latency, cloud cost, and model size as interconnected evaluation dimensions. This provides a more comprehensive perspective for assessing the sustainability of AI-based information systems. Furthermore, the empirical comparison of four deployment configurations demonstrates that software optimisation strategies can produce different sustainability outcomes even when addressing the same detection task. In particular, request batching and INT8 quantisation achieved substantial improvements in computational efficiency while maintaining predictive performance and estimated societal benefits. These findings highlight the importance of evaluating complete deployment configurations rather than focusing exclusively on model-level optimisation. The study also contributes a measurable and adaptable approach to sustainability assessment through operational indicators and transparent measurement equations. By integrating social and technical dimensions into a unified evaluation process, the framework provides researchers, software developers, and organisations with a practical basis for making deployment decisions that balance information integrity, societal value, and responsible resource utilisation.

Limitations:

The study has a few limitations. First, the corpus contained only the English language versions of the PolitiFact titles. Titles contain less information than full articles, and political claims differ from health, climate, financial, conflict-related and local misinformation. The performance of the fact-checking model should not be generalised beyond the evaluated data without external validation (Maggini et al., 2026). Second, the dataset contains historical fact-checking decisions. The definitions of labels, the distribution of topics, the behaviours of sources, and the patterns of language may all change. Temporal validation is needed because lean detectors of misinformation may learn source- or time-specific cues that weaken when deployed in new environments. Third, the estimated reduction in exposure was based on a user recall assumption; that is, correct detection would result in content being intercepted and not reaching the user. Real reduction would depend on intervention timing, ranking systems, user sharing, repeated uploads, cross-platform diffusion, and user response to system warnings. Fourth, the information-trust support measure was a classifier-derived proxy, which included measures of precision and specificity but not calibration and thus did not measure perceived trust. A future mixed methods or experimental study should validate the weighting scheme and explore if changes in the proxy measure would lead to changes in perceived trust. Fifth, energy values were estimated using processor time and fixed power and PUE assumptions. The values provide a relative comparison that is more transparent compared to the existing certification; however, the values should not be confused for a certified measurement of energy use. Memory energy, embodied carbon, networking, storage, monitoring, model updates, and device manufacturing were all outside the system boundary. Sixth, the prototype used a linear text classifier instead of a transformer, graph, or multimodal architecture. Although the energy and performance trade-offs may be different for computationally expensive systems, the lightweight approach provides a useful proof of concept. Designs should consider whether a more resource-intensive

approach is necessary to provide enhanced social benefits. The Cloud-cost estimate employed a normalised processor-hour rate. This estimate omitted the provider-specific billing, network transfer, orchestration, managed-service premiums, redundancy, and human review. Therefore, the values should be interpreted as comparative compute costs and not as complete estimates of the production budget.

Suggestions:

Future research should extend the proposed framework through several directions. First, the framework should be evaluated using full-text news articles and multilingual datasets to examine whether its effectiveness can be maintained across different languages, topics, and misinformation contexts. Incorporating multimodal information, including images and other content features, may also provide a more comprehensive approach to misinformation detection. Second, future studies should investigate more advanced AI architectures, including transformer-based and multimodal models, to compare their predictive capabilities and computational requirements with lightweight classifiers. Such comparisons could provide further insights into the relationship between model complexity, detection performance, and digital sustainability. Third, energy consumption should be measured directly using hardware-based monitoring tools rather than relying exclusively on processor-time estimates. Future evaluations should also consider additional environmental factors, including carbon emissions, memory consumption, network communication, storage, and the lifecycle impact of hardware and software deployment. Fourth, the proposed information-trust support proxy should be validated through user-based experiments or mixed-methods research. Involving actual users would help determine whether the proposed indicators correspond to perceived trust, confidence in information, and willingness to engage with automated misinformation warnings. Finally, future research should evaluate the framework under real-world deployment conditions using live platform traffic, temporal validation, and cross-domain datasets. This would help assess the adaptability of different deployment configurations and establish more reliable sustainability benchmarks for AI-based information systems in practical environments.

CONCLUSION

This study presents a sustainability-focused and empirically grounded AI framework to detect misinformation. This framework addresses a primary limitation in existing model assessment frameworks by integrating social benefits and digital footprint as co-primary considerations. The study integrates misinformation exposure reduction, trust promotion, macro-F1, calibration, energy, latency, throughput, cloud cost, model size, and model complexity, and aims to provide a more complete basis for system selection.

The impact of the controlled prototype demonstrated that the sustainability outcomes of a deployment design can be improved/affected even if the underlying detection task is similar. Full-precision individual inference resulted in a macro-F1 of .873 and an estimated exposure reduction of 88.8%. However, it was outperformed by several batch variants. Predictive performance and social outcome estimates were improved by order-of-magnitude energy reductions and request batching. With INT8 quantisation, batch variants made an even larger reduction in the model size without affecting classification outcomes. In the evaluated experimental context, INT8 batch quantisation was the best trade-off design.

Compact INT8 edge inference models provided the lowest estimated power consumption, smallest size, highest throughput, and cost-free cloud inference. Although this increase in performance did cause a significant drop in macro-F1 score and an increase in misinformation exposure, edge inference models weren't always inferior. Edge inference was useful in scenarios when users had limited resources, connectivity, or desired more privacy, and with that came a trade-off for the detection coverage.

The first major contribution is not a single aggregate score, but a replicable decision-making process. With respect to transfer learning, models should be optimised for a fixed level of societal value, while cutting down on resources as much as possible. In the future, the framework should be tested for full-

text and multimodal datasets, direct measurements of hardware energy consumption, real-world platform traffic, cross-domain evaluations, temporal testing, and focused trust measurements. Such work can move digital sustainability from a general design aspiration toward an empirically testable property of responsible information systems.

ACKNOWLEDGMENT

The authors acknowledge the researchers who developed and publicly documented FakeNewsNet and the organisations whose fact-checking records support reproducible misinformation research. No external funding was used for the prototype evaluation reported in this manuscript.

AUTHOR CONTRIBUTION STATEMENT

Conceptualisation: development of the sustainability-aware social–technical framework. Methodology: construction of the evaluation configurations and sustainability metrics. Software and formal analysis: implementation of data preparation, classification, quantisation, repeated inference benchmarking, and bootstrap analysis. Visualisation: preparation of the framework, workflow, trade-off, and normalised-profile figures. Writing original draft: preparation of the complete manuscript. Writing review and editing: conceptual, methodological, and linguistic refinement. The final author-specific CRediT allocation should be adjusted to reflect the verified contributions of each listed author before submission.

REFERENCES

Alawad, M. A., Alfilh, R. H. C., Singh, N. S. S., & Kumar, R. (2026). Adaptive energy-aware approximate multiplier with dynamic reconfiguration for IoT edge applications. *Computer Communications*, *250*, 108457. https://doi.org/10.1016/j.comcom.2026.108457

Baqraf, Y., Keikhosrokiani, P., & Cheah, Y. (2026). Enhancing trustworthiness of Arabic online health information quality evaluation using an enhanced BERT architecture with PCA and ICA feature weighting. *Scientific Reports*, *16*(1). https://doi.org/10.1038/s41598-026-43158-8

Butler, L., Prike, T., & Ecker, U. K. H. (2024). Nudge-based misinformation interventions are effective in information environments with low misinformation prevalence. *Scientific Reports*, *14*(1), 11495. https://doi.org/10.1038/s41598-024-62286-7

Catyanadika, P. E., Sabani, A., & Leenders, M. A. A. M. (2024). Exploring Data Veracity Management in a Post-Truth Business Environment. *Journal of Database Management*, *35*(1), 1–46. https://doi.org/10.4018/jdm.361726

Dodge, J., Prewitt, T., Tachet des Combes, R., Odmark, E., Schwartz, R., Strubell, E., Luccioni, A. S., Smith, N. A., DeCario, N., & Buchanan, W. (2022). Measuring the carbon intensity of AI in cloud instances. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1877–1894. https://doi.org/10.1145/3531146.3533234

Fan, B., Luo, K., Wang, P., Moghimi, M. M., & Hafez, M. (2026). Cloud–edge collaborative AI for symmetry and asymmetry in human motion understanding and collaboration. *Journal of Cloud Computing*, *15*, 64. https://doi.org/10.1186/s13677-025-00810-4

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, *28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Henderson, P., Hu, J., Romoff, J., Brunskill, E., Jurafsky, D., & Pineau, J. (2020). Towards the systematic reporting of the energy and carbon footprints of machine learning. *J. Mach. Learn. Res.*, *21*(1). https://dl.acm.org/doi/10.5555/3455716.3455964

Hina, M., Islam, A. K. M. N., & Dhir, A. (2024). Blockchain for sustainable consumption: an affordance

and consumer value-based view. *Internet Research*, *34*(7), 215–250. https://doi.org/10.1108/intr-07-2023-0523

Hoes, E., Aitken, B., Zhang, J., Gackowski, T., & Wojcieszak, M. (2024). Prominent misinformation interventions reduce misperceptions but increase scepticism. *Nature Human Behaviour*, *8*(8), 1545–1553. https://doi.org/10.1038/s41562-024-01884-x

Ikram, M., Mehmood, W., Saleem, S. M. U., & Hanefar, S. B. M. (2026). Exploring the nexus of governance and AI ethics: Using systematic literature review for future direction. *Data and Information Management*, *10*(3), 100122. https://doi.org/10.1016/j.dim.2026.100122

Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., Adam, H., & Kalenichenko, D. (2018). Quantization and training of neural networks for efficient integer-arithmetic-only inference. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2704–2713. https://doi.org/10.1109/CVPR.2018.00286

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

King, C., Phillips, S. C., & Carley, K. M. (2025). A path forward on online misinformation mitigation based on current user behavior. *Scientific Reports*, *15*(1), 9475. https://doi.org/10.1038/s41598-025-93100-7

Kirkeby, M. H., & Lagermann, T. (2025). Power Assumptions Matter: Evaluating End-user Laptop Energy Models for Sustainability Reporting of Browser-Based Web Services. In *arXiv (Cornell University)*. Cornell University. https://doi.org/10.48550/arxiv.2510.12566

Lannelongue, L., Grealey, J., & Inouye, M. (2021). Green algorithms: Quantifying the carbon footprint of computation. *Advanced Science*, *8*(12), 2100707. https://doi.org/10.1002/advs.202100707

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science*, *359*(6380), 1094–1096. https://doi.org/10.1126/science.aao2998

Liu, Y., Shen, X., Zhang, Y., Wang, Z., Tian, Y., Dai, J., & Cao, Y. (2025). A systematic review of machine learning approaches for detecting deceptive activities on social media: methods, challenges, and biases. *arXiv (Cornell University)*, *20*(7), 6157–6182. https://doi.org/10.1007/s41060-025-00850-8

Maggini, M., Merzougui, D. E., Bandyopadhyay, R., Dias, G., Maurel, F., & Gamallo, P. (2026). Are LLMs Enough for Hyperpartisan, Fake, Polarized and Harmful Content Detection? Evaluating In-Context Learning vs. Fine-Tuning. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 20, Nomor 1). https://doi.org/10.1609/icwsm.v20i1.42712

Memon, A. B., Sootahar, D. K., Luhana, K. K., & Meyer, K. (2024). A corpus-based real-time text classification and tagging approach for social data. *Frontiers in Computer Science*, *6*. https://doi.org/10.3389/fcomp.2024.1294985

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229. https://doi.org/10.1145/3287560.3287596

Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M., & Dean, J. (2021). Carbon emissions and large neural network training. In *arXiv*. https://doi.org/10.48550/arXiv.2104.10350

Pawelczyk, F., Dimmery, D., & Yan, P. (2026). Implied Authenticity Effect? The Impact of Explicit Labels on AI-Generated Content. *Proceedings of the International AAAI Conference on Web and Social Media*, *20*(1), 1738–1766. https://doi.org/10.1609/icwsm.v20i1.42721

Pennycook, G., & Rand, D. G. (2021). The psychology of fake news. *Trends in Cognitive Sciences*, *25*(5), 388–402. https://doi.org/10.1016/j.tics.2021.02.007

Rachmiatie, A., Kurniawan, T. B., & Dewi, D. A. (2026). Crisis Communication 4.0: Evaluating the role of artificial intelligence in public trust management. *Studies in Media and Communication*, *14*(3), 159–168. https://doi.org/10.11114/smc.v14i3.8307

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33–44. https://doi.org/10.1145/3351095.3372873

Rawat, H., & Rajavat, A. (2026). Trust-Aware Federated User Representation Learning for Privacy-Preserving and Carbon-Optimized Personalized Recommendation in Large-Scale Social Platforms. *Human-Centric Intelligent Systems*. https://doi.org/10.1007/s44230-026-00145-6

Rojahn, M., & Grum, M. (2026). Green artificial intelligence and its lifecycle, hardware, and measurement dimensions. *Information and Software Technology*, *198*, 108186. https://doi.org/10.1016/j.infsof.2026.108186

Saputra, D., Risdianto, E., Rezvani, M. Q., & Rahayu, P. W. (2026). Software and Information Systems for Sustainability: A Systematic Literature Review of Models, Applications, and Evaluation Metrics. *Journal of Sustainable Software Engineering and Information Systems*, *2*(1), 16–40. https://doi.org/10.58723/jsseis.v2i1.160

Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. *Communications of the ACM*, *63*(12), 54–63. https://doi.org/10.1145/3381831

Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Liu, H. (2020). FakeNewsNet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big Data*, *8*(3), 171–188. https://doi.org/10.1089/big.2020.0062

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, *19*(1), 22–36. https://doi.org/10.1145/3137597.3137600

Sovrano, F., Palmirani, M., Sapienza, S., & Pistone, V. (2024). DiscoLQA: zero-shot discourse-based legal question answering on European Legislation. *Archivio istituzionale della ricerca (Alma Mater Studiorum Università di Bologna)*, *33*(2), 323–359. https://doi.org/10.1007/s10506-023-09387-2

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3645–3650. https://doi.org/10.18653/v1/P19-1355

Technology, N. I. of S. and. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)*. U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

Wambsganß, T., Benke, I., Maedche, A., Koedinger, K. R., & Käser, T. (2024). Evaluating the Impact of Learner Control and Interactivity in Conversational Tutoring Systems for Persuasive Writing. *International Journal of Artificial Intelligence in Education*, *35*(2), 791–822. https://doi.org/10.1007/s40593-024-00409-x

Wang, C., Sturgis, P., & Kadt, D. de. (2026). AI labeling reduces the perceived accuracy of online content but has limited broader effects. *Telematics and Informatics*, *107*, 102416. https://doi.org/10.1016/j.tele.2026.102416

Wardle, C., & Derakhshan, H. (2017). *Information disorder: Toward an interdisciplinary framework for research and policy making*. Council of Europe.

Zhou, X., & Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys*, *53*(5), 109. https://doi.org/10.1145/3395046