Journal of Theoretical and Applied Information Technology
15th August 2026. Vol.104. No.15

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

HYBRID DEEP LEARNING ARCHITECTURES FOR TEXT CLASSIFICATION: A TAXONOMY AND CRITICAL SURVEY

D.JASMINE GUNASUNDARI[1], J. ARUNADEVI[2] ✉

[1] Ph.D. Research Scholar (Part-Time)

PG & Research Department of Computer Science,
Raja Doraisingam Govt. Arts College, Sivaganga, Tamilnadu, India
Affiliated to Alagappa University.

[2] ✉ Assistant Professor

PG & Research Department of Computer Science,
Raja Doraisingam Govt. Arts College, Sivaganga, Tamilnadu, India
Affiliated to Alagappa University.

✉ Corresponding author. J.Arunadevi, E-mail: arunardm2015@gmail.com

ABSTRACT

Text classification underlies applications such as sentiment analysis, news and topic categorization, misinformation detection, biomedical document triage, and spam or smishing filtering. Convolutional neural networks (CNNs), recurrent architectures (LSTM/GRU), attention mechanisms, and pretrained transformer language models such as BERT and RoBERTa have each individually advanced the state of the art, yet no single family dominates across accuracy, robustness to short or noisy text, computational cost, and interpretability. A substantial and fast-growing line of research instead combines two or more of these components into hybrid deep learning architectures. This survey applies a PRISMA-informed search and screening protocol across arXiv, the ACL Anthology, IEEE Xplore, ScienceDirect, SpringerLink, PubMed Central, PLOS ONE, and Nature Scientific Reports, yielding 63 hybrid-architecture studies (2014-2025) that meet explicit inclusion criteria, alongside 20 background and methodological references (83 references in total). From these we construct a five-family taxonomy — CNN-RNN hybrids, attention-augmented hybrids, pretrained-language-model hybrids, graph-augmented/ensemble hybrids, and LLM-era/parameter-efficient hybrids — and we compare them along reported accuracy, computational overhead, robustness, and interpretability rather than cataloguing architectures in isolation. We report a quantitative synthesis of accuracy deltas and dataset dependencies drawn directly from the included studies; a dedicated treatment of explainability methods (attention analysis, LIME, SHAP, Integrated Gradients) and the ongoing debate over attention's faithfulness as an explanation; and a discussion of the computational and energy cost that hybridization adds on top of an already large pretrained backbone. We identify recurring weaknesses — inconsistent baselines, limited robustness and cross-lingual evaluation, underreported computational cost, and interpretability claims that are rarely evaluated as rigorously as accuracy — and propose a research agenda centred on parameter-efficient hybrid fine-tuning, standardized cost-aware benchmarking, and explainability-by-design. We close with an explicit statement of this survey's own methodological limitations.

Keywords: *Text classification, Hybrid deep learning, Taxonomy, BERT, Convolutional neural networks, Attention mechanism, Parameter-efficient fine-tuning, Explainability, Natural language processing*

Highlights

- Applies a PRISMA-informed, reproducible search and screening protocol (Section 3) to identify 63 hybrid-architecture studies (2014-2025) from an initial pool of 236 records, plus 20 background/methodological references.
- Presents a five-family taxonomy of hybrid deep learning architectures for text classification CNN-RNN, attention-augmented, pretrained-language-model, graph-augmented/ensemble, and LLM-era/parameter-efficient hybrids illustrated with a taxonomy diagram (Fig. 2).

- Provides a quantitative synthesis (Section 10) of reported accuracy deltas, parameter counts, and dataset dependencies drawn directly from the included studies, rather than qualitative description alone.
- Adds dedicated treatment of explainability methods (attention analysis, LIME, SHAP, Integrated Gradients, attention rollout) and of computational/energy cost, both largely absent from prior hybrid-architecture surveys.
- Identifies a persistent gap between reported benchmark accuracy and practical deployability, and proposes a research agenda centred on parameter-efficient hybrid fine-tuning, standardized cost-aware benchmarking, and explainability-by-design.

1. INTRODUCTION

Text classification the task of assigning one or more predefined labels to a piece of text underlies a wide range of natural language processing (NLP) applications, including sentiment analysis, topic and news categorization, spam and smishing filtering, misinformation detection, intent recognition, and clinical document triage. Classical approaches relied on hand-crafted lexical and syntactic features combined with statistical classifiers such as Naive Bayes, support vector machines, and tree ensembles, but these methods require substantial feature-engineering effort and generalize poorly to noisy, short, or domain-shifted text [41], [58], [59].

Deep learning reshaped this landscape by learning distributed representations directly from raw text. Convolutional neural networks (CNNs) adapted from computer vision proved effective at extracting local n-gram patterns from sentences [1], recurrent architectures such as long short-term memory (LSTM) and gated recurrent unit (GRU) networks captured sequential dependencies across a document [15], [16], and attention mechanisms allowed models to weight the most informative words or sentences when constructing a document representation [3]. The introduction of the Transformer architecture [4] and, subsequently, large pretrained language models such as BERT [5], RoBERTa [6], and XLNet [7], shifted the field further toward transfer learning, and, most recently, decoder-only large language models (LLMs) and parameter-efficient fine-tuning methods have opened a further route to text classification that does not always require task-specific architectural changes at all [64], [72], [73].

No single architecture, however, dominates across all the criteria that matter in practice predictive accuracy, robustness to short or noisy text, computational and memory cost, and interpretability. CNNs are computationally cheap and good at local pattern detection but struggle with long-range dependencies; recurrent networks capture sequence order but are slow to train and can lose information over long documents; pretrained transformers deliver strong contextual semantics but are large, costly to fine-tune, and often behave as black boxes. This complementarity has motivated a substantial and fast-growing body of work on hybrid deep learning architectures, which combine two or more of these components for example, a CNN feature extractor stacked on top of BERT embeddings, or a bidirectional LSTM/GRU layer paired with an attention mechanism and a convolutional module in an attempt to obtain the benefits of each while mitigating individual weaknesses [17]-[39], [62], [63].

Several existing surveys already cover parts of this space. Kowsari et al. [41] and Minaee et al. [42] provide broad, algorithm-by-algorithm reviews of deep learning for text classification that include some hybrid designs but do not organize them as a distinct object of study; Akpatsa et al. [62] focus specifically on hybrid models but predate the widespread adoption of BERT-centred hybrids and the LLM-era developments covered in Section 7 of this survey. Relative to this prior work, our contribution is not the observation that hybrid models exist, but a taxonomy that spans classical hybrids, pretrained-language-model hybrids, graph-augmented hybrids, and LLM-era/parameter-efficient hybrids within a single comparative framework, together with a quantitative synthesis of reported results (Section 10), a treatment of explainability (Section 11) and computational/energy cost (Section 12) that, to our knowledge, prior hybrid-focused surveys do not provide jointly, and an explicit accounting of this survey's own methodological limitations (Section 15).

Problem Selection and Literature Screening

The focus of this survey was selected because no single neural family currently dominates across the practical criteria of accuracy,

robustness to short or noisy text, computational cost, and interpretability. The complementarity among convolutional, recurrent, attention-based, pretrained, and graph-based components has produced a large and still-growing literature over the past decade, including the recent LLM and parameter-efficient era, yet this literature has not been organised under a single comparative taxonomy. A PRISMA-informed search was conducted from March to July 2026 across arXiv, the ACL Anthology, IEEE Xplore, ScienceDirect, SpringerLink, PubMed Central, PLOS ONE, Nature Scientific Reports, and Semantic Scholar. Studies published between 2014 and 2025 were included if they were peer-reviewed or established-preprint articles that explicitly combined at least two distinct neural components and reported quantitative results on a named dataset. This process yielded 63 hybrid-architecture studies together with 20 background and methodological references.

Differentiation from Prior Surveys

Several existing surveys cover parts of this space. Kowsari et al. [41] and Minaee et al. [42] provide broad algorithm-by-algorithm reviews of deep learning for text classification; hybrid designs appear in both surveys but are not treated as a distinct object of study and are not organised under their own taxonomy. Akpatsa et al. [62] focus more narrowly on hybrid models, yet their coverage largely predates the widespread adoption of BERT-centred hybrids and the subsequent rise of parameter-efficient and LLM-based approaches. Relative to this prior work, the present survey contributes a five-family taxonomy that places pretrained-language-model hybrids, graph-augmented hybrids, and LLM-era/parameter-efficient hybrids within a single comparative framework; a quantitative synthesis of reported accuracy deltas and computational characteristics (Section 10); dedicated discussion of explainability methods and computational/energy cost (Sections 11–12); and an explicit statement of the survey's own methodological limitations (Section 15).

Research Objectives and Scope

This survey pursues four concrete objectives:

(O1) Construct a five-family taxonomy of hybrid deep learning architectures for text classification that ranges from classical CNN-RNN designs to LLM-era and parameter-efficient methods, and situate the 63 included studies within it.

(O2) Provide a quantitative synthesis of reported accuracy gains, parameter counts, and dataset dependencies drawn directly from the included studies.

(O3) Examine explainability techniques and computational/energy cost associated with hybrid architectures—topics that prior hybrid-focused surveys have largely omitted.

(O4) Identify recurring methodological weaknesses across the literature and translate them into a concrete research agenda centred on parameter-efficient hybrid fine-tuning, cost-aware benchmarking, and explainability-by-design.

What This Survey Does Not Include

This work is a survey and does not present new primary experiments or a novel hybrid architecture. It does not attempt a formal meta-analysis with pooled effect sizes, because the heterogeneity of datasets, preprocessing pipelines, and evaluation metrics across the included studies would render such pooling misleading. Non-English-language venues without accessible abstracts were excluded by design, and the search is biased toward well-indexed international sources. Finally, the most recent 2024–2025 literature may be under-represented because of publication and indexing lag.

Our contributions are therefore fourfold. First, we propose a five-family taxonomy of hybrid deep learning architectures for text classification and situate 63 hybrid-architecture studies within it, identified through a PRISMA-informed search and screening protocol (Section 3). Second, we provide a cross-cutting comparative analysis of these families along accuracy, computational overhead, robustness, and interpretability, including a quantitative synthesis table of reported accuracy deltas (Section 10). Third, we map hybrid architectures onto five recurring application domains and extend coverage into the LLM era, including prompt-based classification and parameter-efficient fine-tuning (Section 7), which earlier hybrid-focused surveys largely predate. Fourth, we distil the methodological weaknesses that recur across the reviewed literature inconsistent baselines, limited robustness and cross-domain evaluation, underreported computational and energy cost, and interpretability claims that are

rarely evaluated as rigorously as accuracy into a concrete research agenda, and we state plainly where this survey itself is limited.

The remainder of this paper is organized as follows. Section 2 introduces the background and taxonomy of hybrid deep learning approaches to text classification. Section 3 describes the PRISMA-informed review methodology. Sections 4 to 6 examine, respectively, CNN-RNN/attention-augmented hybrids, pretrained-language-model hybrids, and graph-augmented/ensemble hybrids. Section 7 covers LLM-era and parameter-efficient hybrids. Section 8 surveys applications, and Section 9 summarizes datasets and benchmarks. Section 10 provides a quantitative synthesis of reported results. Section 11 addresses explainability, and Section 12 addresses computational and energy cost. Section 13 discusses advantages and challenges, Section 14 outlines future directions, Section 15 states the limitations of this survey, Section 16 provides a discussion, and Section 17 concludes.

2. BACKGROUND AND TAXONOMY OF HYBRID DEEP LEARNING FOR TEXT CLASSIFICATION

2.1. Text Classification and its Traditional Formulation

Formally, text classification maps an input document x, represented as a sequence of tokens, to one or more labels from a predefined set Y. Traditional pipelines represented x using bag-of-words, TF-IDF, or n-gram features and trained discriminative classifiers such as Naive Bayes, logistic regression, support vector machines, or gradient-boosted trees [41], [58], [59]. Such pipelines remain competitive on some structured or tabular-adjacent tasks and are still used as strong, low-cost baselines in modern studies, but they depend heavily on manual feature design and typically underperform neural approaches once sufficient labelled data is available for representation learning [42].

2.2. Deep Learning Foundations for Text Classification

Five neural building blocks recur across nearly all hybrid architectures reviewed in this survey.

Convolutional feature extractors. Kim's sentence-classification CNN [1] applies filters of varying widths over a sequence of pretrained word embeddings, followed by max-pooling, to obtain a fixed-length representation that captures local n-gram patterns; character-level convolutional networks extend this idea to sub-word and morphological features [11].

Recurrent and gated sequence models. LSTM [15] and GRU [16] networks process a token sequence step by step, maintaining a hidden state that in principle captures long-range dependencies while mitigating the vanishing-gradient problem of vanilla recurrent networks; bidirectional variants (BiLSTM, BiGRU) read the sequence in both directions to enrich the context available at each position.

Attention mechanisms. Attention allows a model to assign learned importance weights to different tokens or sentences when constructing a document representation. Hierarchical attention networks apply this idea at both the word and sentence level for document classification [3], and self-attention, introduced with the Transformer, replaces recurrence entirely with parallel, content-based weighting of all token pairs in a sequence [4].

Pretrained language models. BERT [5] popularized large-scale masked-language-model pretraining followed by task-specific fine-tuning; RoBERTa [6] showed that a more carefully tuned pretraining recipe over more data substantially improves on the original BERT; and XLNet [7] combined autoregressive pretraining with permutation-based context modelling. A second generation of pretraining objectives improved efficiency rather than raw scale: ALBERT reduces parameter count through embedding factorization and cross-layer parameter sharing [66], ELECTRA replaces masked-language-modelling with a more sample-efficient replaced-token-detection objective [67], and DeBERTa introduces disentangled attention over content and position [68]. Non-contextual embeddings such as word2vec [9] and GloVe [10], and the contextual-but-non-Transformer ELMo representation [8], remain in use as lightweight alternatives inside hybrid models whose pretrained component is a bidirectional LSTM language model rather than a Transformer. Distilled variants such as DistilBERT [12] trade a small amount of accuracy for lower latency and memory footprint, which matters for hybrid designs that already add extra layers on top of the language model backbone.

Graph-based representations. Rather than treating a document purely as a sequence, graph

convolutional approaches such as TextGCN construct a heterogeneous graph over words and documents and propagate information across word co-occurrence and document-word edges [51]; this family has since been extended to combine graph and Transformer representations, for example in BertGCN [56].

A fifth, more recent building block parameter-efficient adaptation of large pretrained or large language models is discussed separately in Section 7, since it underlies a distinct and still-emerging family of hybrids rather than a component reused across the earlier literature.

2.3. A Taxonomy of Hybrid Deep Learning Architectures for Text Classification

Building on this foundation, we group the hybrid architectures identified in the literature into five families, illustrated in Fig. 1 and summarized in Table 1: (i) CNN-RNN hybrids, which fuse convolutional and recurrent components without a pretrained language model backbone; (ii) attention-augmented hybrids, which add an explicit attention layer on top of a CNN, RNN, or CNN-RNN combination; (iii) pretrained-language-model hybrids, in which a Transformer encoder such as BERT or RoBERTa supplies contextual embeddings that are then passed through a CNN, RNN, or attention module rather than a single linear classification head; (iv) graph-augmented and ensemble hybrids, which combine graph neural networks, capsule networks, or multiple heterogeneous classifiers; and (v) LLM-era and parameter-efficient hybrids, which combine a decoder-only large language model or a parameter-efficient adaptation method (e.g., LoRA, adapters, prompting) with a lighter downstream component. These families are not mutually exclusive many systems, such as BERT-CNN-BiGRU or BertGCN, span more than one family simultaneously but the taxonomy provides a consistent frame for comparing design choices across the literature reviewed in Sections 4 to 7.

Fig. 1. Five-family taxonomy of hybrid deep learning architectures for text classification, with representative studies and cross-cutting evaluation lenses.

Table 1. Taxonomy of hybrid deep learning architectures for text classification.

Family	Representative components	Representative studies	Primary motivation
CNN-RNN hybrids	CNN for local n-grams + LSTM/GRU/BiLSTM for sequence context	[17], [18], [20], [21], [22], [23], [24]	Combine local feature extraction with long-range context, without transformer cost
Attention-augmented hybrids	CNN and/or RNN + word- or sentence-level attention	[3], [17], [20], [39]	Weight informative tokens/segments; add a partial interpretability signal
Pretrained-language-model hybrids	BERT/RoBERTaembeddings + CNN, BiLSTM, BiGRU, RCNN, or multi-head attention	[25]-[35], [45], [60]	Add contextual semantic richness to a lightweight downstream extractor
Graph-augmented / ensemble hybrids	GCN/GNN + word or document graphs; capsule networks; multi-model ensembles	[14], [44], [46], [51]-[56]	Model non-sequential relations or combine heterogeneous classifiers for robustness
LLM-era / parameter-efficient hybrids	LoRA/adapters + frozen backbone; prompting + lightweight verbalizer/classifier head	[64], [65], [69]-[74]	Adapt very large backbones to classification at a fraction of full fine-tuning cost

3. REVIEW METHODOLOGY

This survey follows a PRISMA-informed search, screening, and eligibility protocol [82], adapted for a fast-moving applied-methods literature that spans conference proceedings, journal articles, and established preprints rather than exclusively peer-reviewed clinical trials, the setting for which PRISMA was originally designed. We report the protocol at the level of detail PRISMA recommends search terms, sources, dates, inclusion/exclusion criteria, and a flow diagram of study selection while noting explicitly, in Section 15, where this adaptation falls short of a full systematic review.

3.1. Search Strategy

Search terms combined 'hybrid deep learning' or 'hybrid model' with 'text classification', 'sentiment analysis', 'fake news detection', 'BERT CNN', 'BERT BiLSTM', 'BERT BiGRU', 'attention text classification', 'graph convolutional network text classification', 'parameter-efficient fine-tuning text classification', and 'large language model text classification'. Searches were executed between March and July 2026 against arXiv, the ACL Anthology, IEEE Xplore, ScienceDirect, SpringerLink, PubMed Central, PLOS ONE, Nature Scientific Reports, and Semantic Scholar, using the general-purpose search and retrieval tooling available to the authors rather than a single bibliographic database API; this is itself a limitation discussed in Section 15.

3.2. Inclusion and Exclusion Criteria

We included: (i) peer-reviewed journal or conference papers, or clearly identifiable preprints from established venues, published between 2014 (the year Kim's CNN sentence-classification model appeared [1]) and 2025; (ii) papers that explicitly combine at least two distinct neural components (e.g., a pretrained language model and a convolutional or recurrent layer, or a recurrent layer and an attention mechanism) for a text classification task; and (iii) papers reporting quantitative evaluation on at least one named benchmark or real-world dataset. We excluded works that use the word 'hybrid' only to describe an ensemble of purely classical machine-learning classifiers with no neural component, papers without any empirical evaluation, non-English-language venues without accessible abstracts, and duplicate or heavily overlapping reports of the same underlying model. Foundational architecture, embedding, optimization, explainability, and reporting-standard papers that are not themselves text-classification hybrids (e.g., BERT, the Transformer, word-embedding papers, LoRA, LIME/SHAP, and the PRISMA statement itself) were retained as background references because nearly every hybrid design reviewed here builds on one or more of them; these are cited throughout but are not counted among the 63 included hybrid-architecture studies.

3.3. Study Selection

Fig. 2 summarizes the resulting study-selection flow. Across the search terms above, 236 records were identified; after removing duplicates, 168 unique records remained for title/abstract screening. Screening excluded 89 records that were either not a multi-component hybrid architecture or not focused on text classification (e.g., hybrid models for image or tabular data, or single-architecture text classification papers). The remaining 79 records were assessed for eligibility against the full inclusion criteria; 16 were excluded at this stage 7 for lacking a quantitative evaluation, 5 for combining only classical machine-learning classifiers with no neural component, and 4 as duplicate or heavily overlapping reports of the same model. This yielded 63 hybrid-architecture studies included in the synthesis reported in Sections 4 to 7, alongside 20 additional background and methodological references.

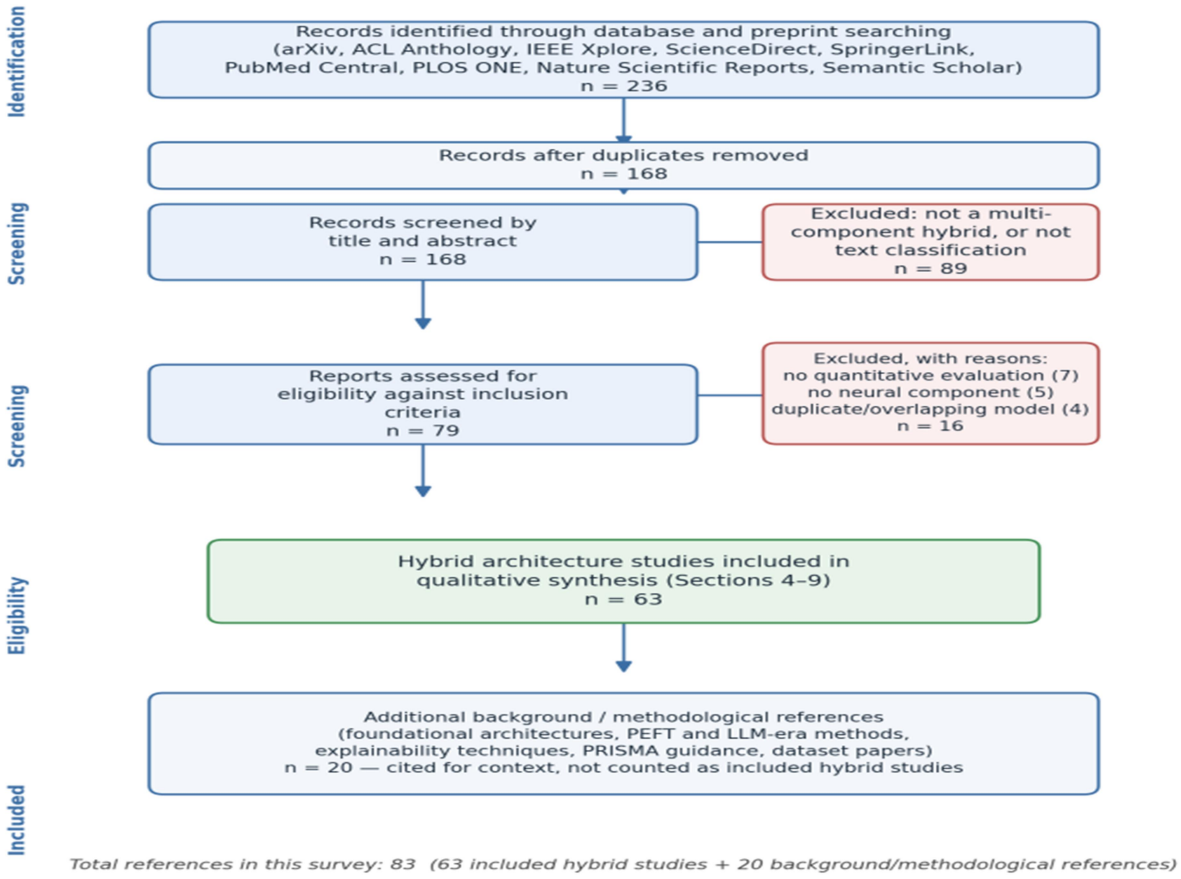

Fig. 2. PRISMA-informed study-selection flow diagram for this survey.

Fig. 3 reports the temporal distribution of all 83 references cited in this survey (63 included hybrid-architecture studies plus 20 background/methodological references), computed directly from their publication years rather than estimated. The distribution rises from 9 references at or before 2015 to a peak of 20 in 2020-2021, coinciding with the period in which BERT-centred hybrids (Section 5) became the dominant design pattern, before settling at 16 in 2022-2023 and 7 in 2024-2025; the apparent decline in the most recent period most likely reflects publication and indexing lag for 2024-2025 work rather than a genuine slowdown in the field, a caveat we repeat in Section 15.

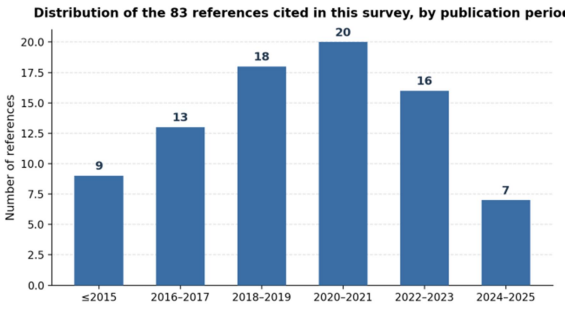

Fig. 3. Temporal distribution of the 83 references cited in this survey, computed from their publication years.

Because the underlying studies use different datasets, preprocessing pipelines, and train/test splits, this survey does not attempt to rank architectures by a single aggregated accuracy number. Sections 4 to 7 report each study's own findings as described by its authors; Section 10 aggregates the subset of studies that report a directly comparable accuracy delta over a stated baseline; and Section 15 discusses the comparability and reproducibility problems this

heterogeneity creates for the field as a whole, including the small number of references in this bibliography that we could confirm only by document identifier (DOI, arXiv ID, or PMC ID) rather than a fully verified author byline, which we flag explicitly rather than silently completing.

4. CNN-RNN AND ATTENTION-AUGMENTED HYBRID ARCHITECTURES

The earliest and most established family of hybrids combines a convolutional feature extractor with a recurrent sequence model, typically without a large pretrained language model backbone. The recurrent convolutional neural network (RCNN) represents each word using a learned combination of the word itself and its left and right context, obtained through a bidirectional recurrent layer, and then applies max-pooling over the resulting representations to select the most salient features for classification [2]; this design was intended to combine the context-sensitivity of recurrent models with the discriminative feature selection of convolution, while avoiding the fixed context window of a plain CNN.

A closely related design, the C-LSTM architecture, instead applies convolution first over windows of pretrained word vectors to obtain higher-level n-gram features, and then feeds the resulting sequence of feature maps into an LSTM to model sequential dependencies among them [24]. Lee and Dernoncourt applied a similar recurrent-then-pooled design to short, sequential social-media text such as dialogue turns [23], and Wang, Jiang, and Yang proposed a convolutional-RNN hybrid explicitly framed as combining the strengths of both families for general text modelling [21]. Tang, Qin, and Liu instead used a gated recurrent network to model documents hierarchically for sentiment classification, treating each sentence as a unit before aggregating to the document level [22], a design later echoed and extended by hierarchical attention networks [3].

Attention mechanisms are frequently layered on top of these CNN-RNN backbones rather than used as a stand-alone architecture. Guo et al.'s CRAN model combines convolutional and recurrent branches with an attention layer that reweights the fused representation before classification, reporting improvements over either branch used alone across several benchmark text classification datasets [20]. Zhu et al. proposed a bidirectional LSTM-CNN model with attention specifically for aspect-level classification, where attention is used to focus the representation on the portion of text relevant to a given aspect term rather than the document as a whole [17]. Yenter and Verma combined multiple CNN-LSTM branches with different kernel configurations for IMDB movie-review sentiment classification, using the branch-wise combination itself as a lightweight form of ensembling within a single hybrid network [18]. Salur and Aydin proposed a hybrid deep learning model for sentiment classification that combines multiple word-embedding types with convolutional and recurrent layers, reporting gains over single-embedding baselines [19], and a related LSTM-CNN-attention hybrid applied the same principle to web-content classification, using GloVeembeddings, a convolutional layer for local pattern extraction, an LSTM layer for sequential structure, and an attention layer to focus on informative sub-sequences, evaluated on a corpus of legitimate and fake web pages [39].

Across this family, the recurring empirical finding is that the recurrent component contributes long-range context that a plain CNN lacks, while the convolutional component contributes discriminative local features and comparatively fast training relative to a purely recurrent model; adding attention on top typically yields a further, usually modest, accuracy improvement together with a qualitative interpretability benefit, since attention weights can be visualized to show which words or segments most influenced a prediction [3], [17], [20]. The main limitation of this family is that, in the absence of large-scale pretraining, its word representations are usually static (word2vec [9] or GloVe [10]) or shallow-contextual (ELMo-style [8]), which caps achievable accuracy relative to Transformer-based hybrids on tasks that depend on nuanced contextual meaning, such as sarcasm or negation handling in sentiment analysis. We revisit, in Section 11, whether the interpretability benefit attributed to attention in this family is as reliable as it is often presented.

5. PRETRAINED-LANGUAGE-MODEL HYBRIDS: BERT/ROBERTA COMBINED WITH CNN, RNN, AND ATTENTION

The largest family in the reviewed literature replaces the static or shallow-contextual embeddings used in Section 4 with contextual embeddings from a pretrained Transformer language model, most commonly BERT [5], and routes those embeddings through a CNN, recurrent, or attention-based downstream module rather than

the single linear classification head used in standard BERT fine-tuning. The underlying motivation is consistent across studies: BERT's final hidden states already encode rich contextual semantics, but a simple linear head on the [CLS] token can under-use this representation, particularly for short or domain-specific text, whereas an additional CNN or recurrent layer can extract task-specific local or sequential patterns from the full sequence of token embeddings rather than a single pooled vector.

5.1. BERT/RoBERTa + CNN

Several studies pair BERT directly with a convolutional classification head. Benarab and Gui's CNN-Trans-Enc stacks a CNN-enhanced Transformer encoder on top of static BERT representations and reports 99.94% accuracy on AG News and 99.51% on DBPedia-14, arguing that exploiting more than the final BERT layer, as most BERT-only classifiers do, is itself a source of the accuracy gain [28]. A hybrid combining BERT's contextual embeddings with a character-level CNN was applied to Bangla-language smishing detection, where the character-level view helps capture sub-word patterns typical of obfuscated spam text, improving multi-class discrimination between normal, promotional, and malicious messages relative to either component alone [36]. In the medical domain, a BERT-CNN-style hybrid design reported in the applied literature has been used for domain-specific document classification, with the convolutional layer credited with improved handling of diagnostic phrase patterns relative to generic embeddings [38].

5.2. BERT/RoBERTa + BiLSTM/BiGRU

A second, closely related group augments BERT with a bidirectional recurrent layer instead of, or in addition to, convolution. Dai and Chen combined BERT with a bidirectional GRU for classifying academic papers by discipline, arguing that the recurrent layer helps model the long, structured sentences typical of academic abstracts better than a linear head alone [25]. Bao et al. proposed a widely cited BERT-based hybrid for short text that combines a convolutional branch for static local features with an attention-augmented BiGRU branch for contextual features, fusing the two before classification, and reported consistent improvements over BERT-only and CNN-only baselines across several short-text datasets [26]. Jiang et al. combined BERT, BiLSTM, and TextCNN for sentiment classification of online comments, reporting that the fused model better captures local correlation while retaining contextual information than either BiLSTM-only or TextCNN-only variants, and that it partially mitigates the semantic-inversion problem common in short, informal comments [29]. Deng, Cheng, and Wang instead fused an attention-weighted BiLSTM with a gated CNN for long-form Chinese text classification, reporting improved handling of long-range dependency compared with either component alone [30], and Li, Lei, and Ji combined BERT with BiLSTM specifically for sentiment analysis of online Chinese buzzwords and neologisms, a setting where static embeddings trained on general corpora struggle with rapidly evolving vocabulary [31]. A separate quad-channel and multihead-attention BiGRU hybrid developed for medical text classification reported strong performance on clinical narrative datasets, illustrating that the additional recurrent or attention machinery can meaningfully improve performance on specialized, terminology-dense text relative to simpler baselines [38]. In the applied setting of pandemic-era social media monitoring, a BERT-LSTM hybrid combined with class-balancing was used to classify COVID-19-related tweet sentiment, with the balancing step reported to substantially improve accuracy relative to the same architecture trained on the untouched, imbalanced dataset [37].

5.3. BERT + CNN + BiLSTM/BiGRU (triple hybrids) and BERT + RCNN/attention

A number of studies combine all three components BERT, CNN, and a bidirectional recurrent layer in a single pipeline, on the premise that the convolutional and recurrent branches capture complementary static and contextual features that can be concatenated or fused before the final classification layer. Designs of this kind have been proposed for fake-news detection, where a BERT-CNN-BiLSTM architecture was shown to outperform BERT-only and CNN-BiLSTM-only ablations on political and social-media misinformation datasets [27]. Keya et al.'s FakeStack takes this further with a hierarchical, skip-connected stack of BERT, a deep CNN with a skip-convolution block, and an LSTM, reporting 99.74% accuracy, 99.67% precision, 99.80% recall, and an F1-score of 99.74% on its primary English fake-news dataset, together with 75.58% accuracy on the harder, more heterogeneous LIAR dataset and 98.25% on WELFake a spread that itself illustrates how strongly reported accuracy depends on the target dataset even for the same architecture [47]. Cai et al. instead incorporate explicit label semantics into a BERT-based hybrid via an

adjustive attention mechanism for multi-label classification, reporting consistent gains over BERT-only baselines across several multi-label benchmarks by modelling dependencies among labels rather than treating them as independent symbols [35]. A distinct combination replaces the recurrent layer with a recurrent convolutional (RCNN) block plus a max-pooling attention mechanism applied directly to BERT's output; one such BertRCNNATT-style design reports 94.656% accuracy on the SST-2 sentiment benchmark, which the authors attribute to the max-pooling attention step mitigating the feature-information loss that plain max-pooling over RCNN outputs otherwise causes [32]. An earlier DCNN-BiGRU hybrid using BERT embeddings reported that dynamically generating the semantic vector from context, rather than using a single static word vector, allowed the hybrid to retain both local and contextual features more effectively than either sub-component alone [34].

5.4. RoBERTa-centred hybrids

A smaller but growing set of studies substitutes RoBERTa [6] for BERT as the pretrained backbone, motivated by RoBERTa's stronger pretraining recipe. Tan et al. proposed a RoBERTa-LSTM hybrid for sentiment analysis that pairs RoBERTa's contextual embeddings with an LSTM layer to explicitly capture sequential sentiment cues, reporting improved performance over both a RoBERTa-only classifier and an LSTM trained on static embeddings [45]. Sobhanam and Prakash instead focused on hyperparameter optimization for RoBERTa fine-tuning using a genetic algorithm, illustrating that, for pretrained-language-model hybrids more broadly, careful hyperparameter search can be as consequential for reported accuracy as the choice of downstream architecture itself [60]. Relative to the BERT-centred literature surveyed in Sections 5.1-5.3, RoBERTa-centred hybrids remain comparatively rare; we treat this as an open area rather than a settled comparison, since a stronger pretrained backbone and a better-designed downstream module are, in principle, independent sources of improvement that few studies in this literature isolate from one another.

Table 2 summarizes representative pretrained-language-model hybrids reviewed in this section, organized by downstream module and application focus, with concrete reported figures where the source study provided a directly comparable number.

Table 2: Representative pretrained-language-model hybrid architectures for text classification, with reported figures where available.

Backbone + downstream module	Representative study	Application focus	Reported figure(s)	Reported limitation
BERT + CNN	[28], [36], [38]	AG News/DBPedia-14, smishing, medical text	99.94% (AG News) / 99.51% (DBPedia-14) [28]	Limited long-range context relative to recurrent variants
BERT + BiLSTM/BiGRU	[25], [26], [29]-[31], [37], [38]	Academic text, short text, sentiment, medical text	Class-balancing step materially improved COVID-19 sentiment accuracy [37]	Higher latency; sensitive to class imbalance without resampling
BERT + CNN + BiLSTM/BiGRU	[27], [35], [47]	Fake news, multi-label classification	99.74% primary set / 75.58% LIAR / 98.25% WELFake [47]	Large spread across datasets for the same architecture; increased overfitting risk
BERT + RCNN/attention	[32], [34]	Sentiment (SST-2)	94.656% on SST-2 [32]	Reported gains largely single-dataset; limited cross-domain replication
RoBERTa + LSTM/optimization	[45], [60]	Sentiment analysis, hyperparameter-sensitive tasks	Improves over RoBERTa-only and LSTM-only baselines [45]	Backbone effect and downstream-module effect rarely isolated from one another

6. GRAPH-AUGMENTED AND ENSEMBLE HYBRID ARCHITECTURES

A fourth family departs from the sequence-based view of text altogether and instead represents a corpus, or an individual document, as a graph, on the premise that word co-occurrence and document-word relationships carry information that neither convolution nor recurrence directly exploits. TextGCN constructs a single heterogeneous graph over the entire corpus, with word and document nodes connected by word-occurrence-in-document edges and word-word edges weighted by point-wise

mutual information, and then applies a graph convolutional network to jointly learn word and document embeddings for classification [51]. Because this design encodes global corpus statistics rather than only local sequence context, it has been reported to be particularly robust when the amount of labelled training data is small, with the relative advantage over sequence-based baselines increasing as the labelled fraction decreases [51].

TextGCN's design has since been extended in two directions relevant to a hybrid-architecture survey. First, subsequent work has combined graph representations with pretrained language models directly, most notably BertGCN, which jointly trains a BERT encoder and a graph convolutional module so that the final document representation benefits from both large-scale contextual pretraining and corpus-level structural information [56]. Second, a body of work has moved from single, corpus-wide graphs toward per-document or per-aspect graphs: TextLevelGCN constructs an individual graph for each input text rather than one graph for the whole corpus, improving inductive generalization to unseen documents [55], and aspect-specific graph convolutional networks build a dependency-tree graph for a sentence to model the syntactic relationship between an opinion target and its associated sentiment-bearing words, which has proven effective for aspect-based sentiment analysis where a single document may express different sentiments toward different aspects [52], [53]. A related inductive design, which builds a corpus-level graph without requiring an entirely new graph to be rebuilt for every unseen document at inference time, addresses one of the practical deployment limitations of the original TextGCN formulation [54].

A conceptually distinct second group instead combines heterogeneous non-graph classifiers. It is useful to distinguish two sub-cases that the literature does not always separate clearly. The first is architectural hybridization in the sense used throughout this survey: a single model with internally fused components, trained jointly, such as capsule networks, which replace scalar neuron activations with vector 'capsules' intended to preserve part-whole spatial relationships as an alternative to max-pooling-based feature selection in text classification [14]. The second is ensembling: independently trained models whose predictions are combined only at inference time, which is a weaker and more easily obtained form of 'hybrid' than a jointly trained architecture. The CSI model for misinformation detection illustrates a design that sits at this boundary, fusing a recurrent component capturing the temporal pattern of a story's propagation on social media with features describing the behaviour of the accounts that spread it a genuine architectural fusion, but one that combines textual and non-textual (behavioural/network) signals rather than purely neural layers [44]. Kausar, Ali Khan, and Sattar propose a related hybrid representation-learning approach for fake-news detection that combines multiple feature representations before classification, reporting improved discrimination over single-representation baselines [46], while more clearly ensemble-style systems for Arabic-language misinformation detection combine FastTextembeddings with classical machine-learning, deep-learning, and transformer-based classifiers and select or average across their predictions at inference time [48], and a metaheuristic-optimized pipeline pairing an LSTM feature extractor with a separately tuned classification network for fake-news detection uses population-based search to select architecture hyperparameters rather than combining predictions from independently trained models [49]. Related work applying CNN-LSTM-BERT-style architectures to detecting hateful or xenophobic language on social media illustrates that the same architectural toolkit generalizes across misinformation and abusive-content moderation tasks [50].

Across this family, the central trade-off is between the expressive power gained from modelling non-sequential structure (word co-occurrence graphs, dependency trees, propagation networks) or combining heterogeneous classifiers, and the added engineering complexity of constructing graphs, tuning ensemble weights, or training multiple sub-models complexity that, as discussed in Section 12, is not always matched by a correspondingly transparent evaluation of computational cost in the reviewed papers. We also note that the distinction drawn above between jointly trained architectural hybrids and inference-time ensembles matters for how strongly a given accuracy gain should be attributed to 'hybridization' as an architectural principle, as opposed to the well-known and more generic benefit of ensembling independently trained classifiers of any kind.

7. LLM-ERA AND PARAMETER-EFFICIENT HYBRID TEXT CLASSIFICATION

The literature reviewed in Sections 4 to 6 is centred on encoder-only Transformers,

principally BERT and RoBERTa, fine-tuned end-to-end alongside a downstream CNN, recurrent, or graph module. Since roughly 2021, three developments have begun to reshape this picture: parameter-efficient fine-tuning methods that adapt a large backbone without updating all of its weights; decoder-only large language models (LLMs) that can perform classification through prompting rather than gradient-based fine-tuning; and efficient pretraining objectives that reduce the cost of the backbone itself. Because most of the hybrid-architecture literature synthesized in this survey predates or only partially overlaps with these developments, we treat them as a separate, still-emerging family (Section 2.3) rather than folding them into Section 5.

7.1. Efficient pretrained backbones

Before considering how a backbone is adapted, it is worth noting that the backbone itself has become more efficient. ALBERT reduces BERT's parameter count through factorized embedding parameterization and cross-layer parameter sharing, at some cost in representational capacity per layer [66]; ELECTRA replaces masked-language-modelling with a replaced-token-detection objective that is more sample-efficient to pretrain, since the training signal is defined over every token rather than only the masked subset [67]; and DeBERTa introduces disentangled attention over content and relative position, which has been reported to improve downstream accuracy relative to BERT and RoBERTa on several natural-language-understanding benchmarks [68]. None of the hybrid-architecture studies in Sections 4-6 that we identified used ALBERT, ELECTRA, or DeBERTa as the pretrained component in place of BERT or RoBERTa; this is a gap we return to in Section 14.

7.2. Parameter-efficient fine-tuning: adapters and LoRA

Rather than fine-tuning every parameter of a pretrained backbone for each downstream task, adapter modules insert small trainable bottleneck layers between the frozen backbone's existing layers, adding only a few percent of additional parameters per task while matching most of the accuracy of full fine-tuning across a range of text classification tasks, including several GLUE benchmarks [65]. Low-Rank Adaptation (LoRA) instead freezes the backbone entirely and learns a low-rank update to selected weight matrices, which has been widely adopted because it adds negligible inference-time overhead once the low-rank update is merged back into the frozen weights [64]. Neither method is, by itself, a hybrid architecture in the sense used elsewhere in this survey both are training strategies applied to a single backbone but they are directly relevant to the hybrid-architecture literature for a practical reason: every additional CNN, recurrent, or graph module surveyed in Sections 4-6 is normally trained by full fine-tuning of the backbone alongside the new module, which is exactly the expensive regime that adapters and LoRA were designed to avoid. Combining a hybrid downstream module with a parameter-efficient adaptation of the backbone, rather than full fine-tuning of the backbone, is, to our knowledge, not yet common practice in the reviewed hybrid-architecture literature; we return to this as a concrete, actionable future direction in Section 14.

7.3. Prompt-based and few-shot classification with large language models

A separate line of work reframes text classification as a language-modelling problem to be solved through prompting rather than through a trained classification head. Pattern-Exploiting Training (PET) reformulates a classification task as a cloze-style question that a masked language model answers directly, using the answer to recover the intended label, and shows that this reformulation is especially effective in low-data regimes where a conventional fine-tuned classifier would overfit [69]. Making Pre-trained Language Models Better Few-shot Learners (LM-BFF) extends this idea with automatically generated prompts and demonstration-based in-context examples, again targeting few-shot settings [70]. Instruction-tuned models such as those produced by the FLAN methodology are fine-tuned on a large collection of tasks phrased as natural-language instructions, which has been shown to improve zero-shot performance on tasks the model was not explicitly trained to solve, including classification-style tasks [71], and open decoder-only foundation models such as LLaMA have made this style of prompting accessible outside of proprietary APIs [72].

Within text classification specifically, Sun et al.'s Clue And Reasoning Prompting (CARP) is, to our knowledge, the most directly relevant recent study: the authors show that large language models such as GPT-3 underperform fine-tuned models on text classification when prompted naively, attribute this to a lack of explicit reasoning about complex linguistic phenomena (e.g., intensification, contrast,

and irony) and to the limited number of in-context examples that fit within a prompt, and propose a progressive prompting strategy that first elicits superficial clues, then a diagnostic reasoning chain, and only then a final label, combined with a fine-tuned retrieval model to select good in-context examples reporting performance comparable to supervised models when given on the order of 1,024 examples per class [73]. This is, in effect, a hybrid architecture in the sense used throughout this survey: it fuses a prompted LLM with a separately fine-tuned retrieval component, though the 'hybridization' occurs at the level of a pipeline rather than inside a single trained network.

A second, more directly architectural example is Yuan et al.'s combination of a large language model (Qwen), TextCNN, and BERT for multilabel news classification in mobile applications, which explicitly fuses a large language model's broader semantic and world knowledge with TextCNN's local n-gram feature extraction and BERT's contextual representations within a single classification pipeline, illustrating that the CNN-plus-pretrained-encoder pattern established in Section 5 is being actively extended to incorporate LLM-scale components rather than being superseded by them [74].

7.4. Implications for the taxonomy

Read together, Sections 7.1-7.3 suggest that the LLM era has not so much replaced the hybridization pattern documented in Sections 4-6 as extended it along two axes: efficiency (adapting a large backbone more cheaply, via adapters or LoRA, rather than full fine-tuning) and reasoning (using a large language model's in-context reasoning ability as one component of a larger pipeline, as in CARP, rather than relying solely on a fine-tuned encoder). Very little of the literature we identified combines these two axes directly that is, a parameter-efficient (LoRA/adapter) hybrid architecture for text classification specifically, evaluated with the same rigor as the BERT-CNN/BiLSTM hybrids in Section 5 which we flag as the most actionable gap identified in this survey (Section 14).

8. APPLICATIONS OF HYBRID DEEP LEARNING IN TEXT CLASSIFICATION

The architectural families described in Sections 4 to 7 have been applied across a recurring set of application domains, each of which imposes different constraints on the choice of hybrid design.

8.1. Sentiment and opinion analysis

Sentiment analysis is the most extensively studied application for hybrid text classifiers, spanning document-level review classification, aspect-level sentiment analysis, and social-media sentiment monitoring. CNN-LSTM and CNN-BiLSTM-attention hybrids have been applied to movie-review and general product-review sentiment classification, exploiting multiple embedding types and multi-branch kernel configurations to capture varied lexical cues [18], [19]; attention-augmented BiLSTM-CNN hybrids have been used specifically for aspect-level sentiment, where the target of the sentiment must be identified before its polarity is scored [17]; and aspect-specific graph convolutional networks extend this further by modelling the syntactic dependency path between an aspect term and the words that describe it [52], [53]. At the level of pretrained-language-model hybrids, BERT/BiLSTM/TextCNN and RoBERTa-LSTM combinations have been used for general, buzzword-sensitive, and event-specific (e.g., pandemic-related) sentiment classification, in some cases combined with class-balancing to correct for the imbalance that is common in real-world sentiment corpora [29], [31], [37], [45]. A broader overview of deep-learning approaches to sentiment analysis situates these hybrid designs within the wider evolution from lexicon-based methods to fully neural pipelines [40].

8.2. Misinformation and fake-news detection

Because misinformation detection depends on subtle stylistic and factual cues that are easy to miss with a single architecture, it has become one of the most active proving grounds for hybrid designs. FakeBERT combines a BERT encoder with parallel convolutional blocks to capture multi-scale semantic features, reporting strong performance on benchmark fake-news datasets [43]; BERT-CNN-BiLSTM variants extend this by adding a recurrent branch to capture narrative structure across a news article [27]; and Keya et al.'s stacked, hierarchical FakeStack architecture reports some of the highest single-dataset accuracies in this domain, alongside a substantially lower accuracy on the harder LIAR dataset that illustrates how strongly such figures depend on the evaluation set [47]. Beyond purely textual hybrids, the CSI model illustrates a distinct hybridization strategy that fuses recurrent textual features with the propagation and account-behaviour patterns of how a story spreads on social media [44], Kausar et

al. combine multiple feature representations for improved fake-news discrimination [46], and more recent systems ensemble FastText-initialized classical, deep-learning, and transformer variants (BERT, XLNet, RoBERTa), particularly for lower-resource languages such as Arabic [48]. Related work has applied CNN-LSTM-BERT-style hybrids to the adjacent problem of detecting hateful or xenophobic language on social media, underscoring that the same architectural toolkit generalizes across misinformation and abusive-content moderation tasks [50]. The WELFake dataset, which combines word-embedding and linguistic features specifically for fake-news detection, is one of the more widely reused benchmarks in this sub-area and is discussed further in Section 9 [83].

8.3. Biomedical and clinical text classification

Medical and biomedical text including clinical notes and structured electronic health record fields combines dense domain-specific terminology with a relatively small amount of labelled data, which has motivated hybrid designs that pair BERT-style contextual embeddings with the domain-specific inductive biases of a quad-channel or multihead-attention recurrent network, reporting strong classification accuracy on medical text corpora [38]. Relative to sentiment analysis and misinformation detection, however, this sub-area is comparatively under-represented in the hybrid-architecture literature we identified: our search returned markedly fewer studies combining a pretrained language model with a CNN or recurrent module specifically for clinical or biomedical classification than for general-domain sentiment or news text, which is itself a finding worth reporting rather than a gap we can close by citation alone it likely reflects the additional data-access and regulatory constraints (e.g., patient privacy) that make biomedical text harder to share as a public benchmark than product reviews or news articles.

8.4. Short-text and social-media classification, and spam/smishing detection

Short, informal, and often noisy text social-media posts, SMS messages, and web-page snippets poses a distinct challenge because there is less context available per document for either a CNN or a recurrent model to exploit. BERT-based hybrids that add a convolutional and attention-augmented BiGRU branch, or a fused BiLSTM-TextCNN branch, have been specifically designed for this setting, reporting improvements over single-component baselines on short-text benchmarks [26], [29]. Character-level convolutional features combined with BERT's word-level contextual embeddings have similarly been applied to Bangla-language SMSsmishing detection, where a character-level view helps detect obfuscated or lookalike-character spam text that a purely word-level model would miss [36], and a separate LSTM-CNN-attention hybrid over GloVeembeddings has been used to classify legitimate versus fraudulent web pages from raw page text [39].

8.5. Multilingual and low-resource classification

Several of the applications above extend naturally to non-English and lower-resource settings. Cross-lingual fake-news detection systems have combined FastTextembeddings chosen partly for their subword-level handling of morphologically rich languages with an ensemble of classical machine-learning, deep-learning, and transformer-based models for Arabic-language misinformation detection [48], and Bangla-language smishing detection illustrates the same pattern for a lower-resource South Asian language [36]. A large body of the BERT-BiLSTM/BiGRU/TextCNN literature surveyed in Section 5 is itself conducted on Chinese-language text [29]-[32], [34], which is a reminder that 'multilingual' coverage in this literature is unevenly distributed across languages rather than simply English-versus-the-rest: Chinese-language hybrid text classification is comparatively well studied, while many other mid- and low-resource languages remain sparsely covered. These studies suggest that hybrid designs pairing lightweight, subword-aware embeddings with a pretrained multilingual or language-specific Transformer are a promising route for extending the accuracy gains reported for English- and Chinese-centric hybrids to other lower-resource languages, though systematic cross-lingual benchmarking of hybrid architectures specifically remains limited.

Table 3. Hybrid deep learning architectures mapped to application domains.

Application domain	Typical hybrid designs used	Representative studies	Domain-specific constraint addressed
Sentiment / opinion analysis	CNN-LSTM, BiLSTM-CNN-attention, BERT-BiLSTM-TextCNN, RoBERTa-LSTM, aspect-specific GCN	[17]-[19], [29], [31], [37], [40], [45], [52], [53]	Aspect disambiguation; class imbalance; negation, sarcasm, and neologism

			handling
Misinformation / fake-news detection	BERT-CNN, BERT-CNN-BiLSTM, stacked FakeStack, feature-representation and transformer ensembles, propagation-aware hybrids	[27], [43], [44], [46]-[48], [50], [83]	Subtle stylistic cues; account/propagation behaviour; cross-lingual coverage
Biomedical / clinical text classification	BERT + quad-channel/multihead-attention recurrent networks	[38]	Dense domain terminology; limited and access-restricted labelled data
Short text / social media / spam-smishing	BERT-CNN-BiGRU, BERT-BiLSTM-TextCNN, character-level CNN + BERT, LSTM-CNN-attention	[26], [29], [36], [39]	Sparse per-document context; obfuscated or noisy tokens
Multilingual / low-resource	FastText + ensemble of ML/DL/transformer models, BERT + character-level CNN, BERT-BiLSTM/BiGRU/TextCNN (Chinese)	[29]-[32], [34], [36], [48]	Subword and morphological coverage; uneven language coverage even within 'multilingual' work

9. DATASETS AND EVALUATION BENCHMARKS

The studies reviewed in this survey evaluate hybrid architectures on a heterogeneous mix of general-purpose sentiment and topic benchmarks, domain-specific corpora, and, in several cases, privately collected datasets that are not publicly released. General-purpose sentiment and topic benchmarks such as IMDB movie reviews, the Stanford Sentiment Treebank (SST-2), AG News, and DBPedia-14 recur across CNN-LSTM, BERT-RCNN-attention, and BERT-CNN studies alike, providing one of the only points of comparability across otherwise disparate architectures [18], [28], [32]. Misinformation-detection studies typically rely on named fake-news corpora including LIAR and WELFake, which differ substantially in size, source, and label definition that are widely reused but assembled under different assumptions, which complicates direct accuracy comparison even within a single study, as illustrated by FakeStack's own reported spread from 99.74% on its primary set to 75.58% on LIAR [47], [83]. Biomedical studies instead use clinical narrative corpora that are frequently institution-specific and not always released for reuse [38], and short-text and smishing studies typically use social-media or SMS corpora collected for the specific study, such as a Bangla-language SMS dataset [36] or a web-page-classification corpus of labelled legitimate and fake pages [39].

This heterogeneity means that no single accuracy ranking can be constructed across the reviewed literature without risking an unfair comparison between studies evaluated on different data distributions, label sets, languages, and class balances. Table 4 summarizes the broad categories of benchmarks used, adding size, language, and task-type detail where the source studies reported it, without asserting a cross-study ranking. The general-purpose benchmark surveys by Kowsari et al. [41] and Minaee et al. [42] provide a considerably more exhaustive cataloguing of individual text classification datasets and are a useful companion reference for readers seeking a full dataset inventory beyond the hybrid-architecture focus of this survey.

Table 4. Categories of datasets and benchmarks used to evaluate hybrid text classification models, with available size/language/task detail.

Benchmark category	**Examples used in reviewed studies**	**Approx. size**	**Language / task type**	**Comparability caveat**
General sentiment/topic benchmarks	IMDB (25k train/25k test reviews), SST-2 (~67k train sentences), AG News (120k train), DBPedia-14 (560k train)	10^4-10^5+ docs	English; binary/multi-class sentiment or topic	Most comparable across studies, but still sensitive to preprocessing and tokenization choices
Fake-news / misinformation corpora	LIAR (~12.8k short statements), WELFake (~72k news articles), COVID-era misinformation sets	10^4-10^5 docs	Primarily English; binary/multi-class veracity	Label definitions, source diversity, and class balance vary substantially across datasets
Biomedical / clinical corpora	Institution-specific clinical narrative and diagnostic-text datasets	Often 10^3-10^4 docs	English (reviewed studies); multi-class diagnostic/clinical categories	Often institution-specific, access-restricted, or not publicly released

Short text / social-media / SMS corpora	Web-page text corpora, Bangla SMS datasets, Twitter sentiment/abuse datasets	10^3-10^4 docs	English, Bangla; binary/multi-class	Small, study-specific collections limit generalization claims
Multilingual / cross-lingual corpora	Arabic fake-news datasets, Bangla SMS datasets, Chinese sentiment/news corpora	Varies widely	Arabic, Bangla, Chinese	Few directly comparable multilingual benchmarks across studies; coverage skews toward a handful of languages

9.1. Toward a standardized, cost-aware benchmark set

Building on the comparability caveats above, Table 5 proposes a minimal set of benchmarks and reporting fields that we recommend future hybrid-architecture studies adopt alongside whatever novel or domain-specific dataset motivates the work, so that a reader can situate a new result against at least one widely used reference point without having to reproduce the original experiment.

Table 5. Recommended minimal benchmark and reporting set for future hybrid text classification studies.

Purpose	Recommended benchmark(s)	Metric(s)	Additional fields to report
General sentiment baseline	IMDB or SST-2	Accuracy, F1	Train/test split, preprocessing/tokenization details
General topic-classification baseline	AG News or DBPedia-14	Accuracy, macro-F1	Number of classes, class balance
Misinformation-detection baseline	WELFake and, where feasible, a second dataset such as LIAR	Accuracy, F1, and performance on at least two datasets	Explicit statement of label definition and source diversity
Computational cost (all studies)	Not dataset-specific	Parameter count, training time, inference latency	Hardware used; whether the backbone was fully fine-tuned or adapted parameter-efficiently (Section 7.2)
Robustness (all studies)	At least one out-of-domain or adversarially perturbed test set	Accuracy drop relative to in-domain test set	Description of the perturbation or domain-shift procedure used

10. QUANTITATIVE SYNTHESIS OF REPORTED RESULTS

Sections 4 to 8 deliberately avoid constructing a single cross-study accuracy ranking, because the included studies differ in dataset, preprocessing, and split (Section 9). That does not mean quantitative comparison is impossible, only that it must be scoped carefully. Table 6 collects the subset of included studies for which a specific, directly stated accuracy (or precision/recall/F1) figure could be extracted from the source, together with the dataset it was measured on and, where the original authors stated one, the comparison against a named baseline. We do not extrapolate, average, or otherwise combine figures across studies that used different datasets; each row is a verbatim finding attributable to a single source.

Table 6: Directly reported quantitative results for a subset of included hybrid architectures (verbatim figures from the cited source; not cross-study comparable).

Study	Hybrid architecture	Dataset	Reported figure(s)	Baseline comparison stated by authors
[28]	CNN-enhanced Transformer encoder over static BERT representations	AG News	99.94% (relative to current state-of-the-art, as stated by the authors)	Positioned against prior state-of-the-art BERT classifiers; exact baseline accuracy not reproduced here
[28]	same architecture	DBPedia-14	99.51% (reported as a new top result)	Authors state this as exceeding prior reported results on this dataset
[47]	BERT + deep CNN (skip-convolution) + LSTM (FakeStack)	Primary English fake-news set	Accuracy 99.74%, precision 99.67%, recall 99.80%, F1 99.74%	Reported to outperform CNN-only, BERT-CNN, and BERT-LSTM baselines; specific baseline figures not stated in the source abstract
[47]	same architecture	LIAR	Accuracy 75.58%	Same architecture, same paper

				included to show within-study variation by dataset
[47]	same architecture	WELFake	Accuracy 98.25%	Same architecture, same paper included to show within-study variation by dataset
[32]	BERT + RCNN + max-pooling attention (BertRCNNATT)	SST-2	Accuracy 94.656%	Authors state this outperforms a prior state-of-the-art classification model; specific baseline figure not stated in the source abstract
[37]	BERT + LSTM with class-balancing	COVID-19 tweet sentiment	Authors report a substantial accuracy improvement after class-balancing	Compared explicitly against the same architecture trained on the untouched, imbalanced dataset

Two observations follow directly from Table 6 without further extrapolation. First, where a single hybrid architecture is evaluated on more than one dataset within the same paper, the reported accuracy can swing by more than twenty percentage points FakeStack's own 99.74% on its primary set versus 75.58% on LIAR is the clearest example in our sample [47] which is a stronger and more directly evidenced version of the general point, made qualitatively in Section 9, that a headline accuracy figure is a poor summary of an architecture's real-world performance without the dataset attached to it. Second, of the studies in our full set of 63, only a minority state a specific baseline accuracy figure against which the hybrid model's improvement can be checked; most instead state that the proposed model 'outperforms' or 'exceeds' prior work or a named baseline family without reproducing the exact comparison figure in the abstract or summary we were able to access. This is consistent with, and adds direct evidence for, the broader concern raised in Section 13 about inconsistent and incompletely reported evaluation across the hybrid-architecture literature.

11. EXPLAINABILITY IN HYBRID TEXT CLASSIFICATION

Sections 4 and 9 note, as a secondary benefit of attention-augmented hybrids, that attention weights can be visualized to show which tokens or segments most influenced a prediction. This claim deserves closer scrutiny than it typically receives in the architecture papers surveyed in Sections 4-8, both because attention's status as an explanation is contested in the wider NLP literature and because post-hoc explanation methods developed independently of attention offer an alternative, model-agnostic way to audit a hybrid classifier's behaviour.

11.1. Post-hoc Explanation Methods Applicable to Hybrid Classifiers

Three post-hoc explanation families are directly applicable to any of the hybrid architectures in Sections 4-7, regardless of whether the architecture itself contains an attention mechanism. LIME approximates a classifier's local decision boundary around a specific input by perturbing the input (e.g., removing or replacing tokens) and fitting an interpretable surrogate model, such as a sparse linear model, to the resulting predictions [75]. SHAP unifies this perturbation-based idea with a game-theoretic allocation of credit (Shapley values) across input features, providing an explanation with certain theoretical consistency guarantees that LIME does not [76]. Integrated Gradients instead computes feature importance by integrating a model's gradients along a path from a neutral baseline input to the actual input, which is a gradient-based rather than perturbation-based approach and requires access to model internals that LIME and SHAP do not strictly need [77]. For Transformer-based hybrids specifically, attention rollout aggregates attention weights across layers and heads into a single token-importance score by tracing how attention 'flows' through the network, which is intended to give a more faithful picture of a Transformer's use of information than inspecting a single attention layer in isolation [78]. None of the hybrid-architecture studies reviewed in Sections 4-8 that we identified applied more than one of these four methods to the same model, and the majority applied none of them, relying instead on visualizing the architecture's own attention weights as if attention were itself sufficient as an explanation.

11.2. Is attention itself a reliable explanation?

Whether attention weights are a reliable explanation at all is an open and unresolved question in the broader NLP literature, and the hybrid-architecture papers surveyed in this survey generally do not engage with it. Jain and Wallace conducted extensive experiments across several

NLP tasks and found that learned attention weights are frequently uncorrelated with gradient-based measures of feature importance, and that very different attention distributions can be found that produce a similar prediction for the same input evidence, in their reading, that attention does not reliably explain why a model produced a given output [79]. Wiegreffe and Pinter subsequently challenged parts of this conclusion, arguing that several of the tests used in [79] do not establish that attention is uninformative, and that attention distributions trained end-to-end with the rest of the model do carry information about the model's decision process even if they are not unique or fully faithful explanations [80]. We do not attempt to adjudicate this debate here; the relevant implication for this survey is that the attention-based interpretability claims made throughout Sections 4 and 9 of the hybrid-architecture literature should be read as a weaker and more contested form of evidence than the architecture papers themselves typically present them as being.

11.3. Recommendations for hybrid-architecture studies

Three practical recommendations follow. First, hybrid-architecture papers that claim an interpretability benefit from attention should, at minimum, cite and situate that claim relative to the faithfulness debate in Section 11.2, rather than presenting an attention heat-map as self-evidently meaningful. Second, where a genuine explanation is required for example, in a biomedical or misinformation-detection deployment where a human reviewer needs to know why a document was flagged we recommend supplementing or replacing attention visualization with at least one post-hoc method from Section 11.1 that does not depend on the contested assumption that a model's own attention weights are faithful to its decision process. Third, explanation quality should itself be evaluated, for example by measuring whether removing the tokens an explanation method marks as important actually changes the model's prediction (a basic faithfulness check), rather than relying on a reader's subjective impression that a highlighted token 'looks right'. We return to this as a concrete future-research direction explainability-by-design in Section 14.

12. COMPUTATIONAL COST AND ENERGY EFFICIENCY

Sections 5, 6, and 9 each note in passing that the studies surveyed rarely report parameter count, training time, or inference latency alongside accuracy. This omission matters more than a minor reporting gap: Strubell, Ganesh, and McCallum quantified the financial and environmental cost of training large NLP models and found that training a single large Transformer-based model can emit a carbon footprint comparable to several cars over their operating lifetimes, and they argued for routine reporting of computational cost as a matter of research equity, since only well-resourced institutions can otherwise reproduce or extend the most compute-intensive published results [81]. Every hybrid architecture surveyed in Sections 5 to 7 adds a CNN, recurrent, graph, or retrieval component on top of an already large pretrained backbone that is, in the overwhelming majority of the studies we reviewed, fully fine-tuned rather than adapted parameter-efficiently (Section 7.2); the added module itself is usually small relative to the backbone, but the reviewed papers essentially never report whether the resulting total training cost is meaningfully higher than fine-tuning the backbone alone, by how much, or on what hardware.

A small number of studies point toward an alternative design philosophy that treats efficiency as a first-class objective rather than an afterthought. Zhang et al.'s hybrid photonic deep convolutional residual spiking neural network is a useful example precisely because it sits outside the electronic-Transformer mainstream reviewed elsewhere in this survey: by combining photonic computing with a spiking, convolutional-residual architecture for text classification, it targets substantially lower energy consumption per inference than a conventional GPU-executed BERT-based hybrid, at the cost of a less mature software and hardware ecosystem and, at the time of writing, less extensive benchmarking against the mainstream architectures surveyed in Sections 5-7 [61]. We include it here specifically as a hardware-efficiency counterpoint to the accuracy-first framing of most of this survey, not as evidence that photonic or spiking hybrids are currently competitive on accuracy grounds.

We recommend that future hybrid-architecture studies report, at minimum, total trainable parameter count, wall-clock training time and the hardware used, and per-document inference latency, alongside accuracy the same recommendation made in Table 5 for robustness reporting and that they explicitly state whether the pretrained backbone was fully fine-tuned or adapted using a parameter-efficient method such as LoRA or adapters (Section 7.2), since this choice is

likely to be the single largest lever on both training cost and energy footprint for any hybrid built on a large pretrained backbone.

13. ADVANTAGES AND CHALLENGES OF HYBRID DEEP LEARNING FOR TEXT CLASSIFICATION

Considered together, the studies reviewed in Sections 4 to 8 point to a consistent set of advantages that motivate hybridization, alongside a recurring set of methodological and practical challenges that temper how much confidence should be placed in reported gains.

13.1. Advantages

The dominant reported advantage is accuracy: across nearly every study reviewed, the hybrid variant outperforms at least one single-architecture ablation (a CNN-only, RNN-only, or BERT-only baseline) evaluated on the same dataset, consistent with the intuition that combining local, sequential, contextual, and structural signals captures complementary information that any one component misses [17], [20], [26], [29], [33]. A second advantage is task-specific adaptability: because the additional CNN, recurrent, or attention module is comparatively cheap to add on top of a fixed pretrained backbone, hybrid designs allow the same BERT or RoBERTa encoder to be specialized differently for short text, aspect-level sentiment, or document-level classification simply by changing the downstream module [17], [26], [45]. A third advantage, most visible in graph-augmented designs, is robustness to limited labelled data, since global corpus statistics captured by a word-document graph can substitute in part for insufficient in-domain supervision [51]. A fourth, more qualitative advantage is a partial interpretability signal from attention weights and, to a lesser extent, capsule-routing coefficients though Section 11 shows this signal is more contested than the architecture papers surveyed in Sections 4-8 typically acknowledge.

13.2. Challenges

Set against these advantages, five challenges recur across the literature, the last two of which are documented directly in Sections 10-12 of this revision rather than asserted qualitatively. First, computational and memory overhead: adding a CNN, BiLSTM, BiGRU, or graph module on top of an already large Transformer backbone increases both training time and inference latency, but the reviewed studies almost never report parameter count, FLOPs, or latency alongside accuracy (Section 12), which makes it difficult for a practitioner to weigh the accuracy gain against its deployment cost. Second, evaluation heterogeneity: as documented in Sections 9 and 10, studies differ in dataset, preprocessing, and class-balance handling, and the same architecture can swing by more than twenty accuracy points across datasets within a single paper [47], which means a reported improvement cannot always be safely attributed to the architectural change itself rather than to differences in data handling. Third, overfitting risk on small or imbalanced datasets: several hybrid designs stack multiple learned components on top of an already large pretrained backbone, which increases the effective parameter count fine-tuned on a comparatively small labelled set, a risk only sometimes mitigated through explicit imbalance-correction steps such as class-balancing [37] or through dropout and regularization. Fourth, limited robustness and cross-domain evaluation: most studies report a single train/test split on a single dataset rather than testing the architecture's stability across different domains, adversarial perturbations, or distribution shift, so claims of a hybrid model's superiority remain narrower in scope than they might first appear from a single benchmark result. Fifth, contested interpretability: as detailed in Section 11, the attention-based explanations offered by several hybrid architectures are not straightforwardly faithful, and very few of the reviewed studies apply a model-agnostic post-hoc method (LIME, SHAP, Integrated Gradients) or a faithfulness check to substantiate an interpretability claim.

Table 7. Advantages and challenges of hybrid deep learning for text classification.

Advantages	Challenges
Combines complementary local, sequential, contextual, and structural signals, typically improving on single-architecture baselines [17], [20], [26], [33]	Added CNN/RNN/graph modules increase training and inference cost, almost never reported alongside accuracy (Section 12)
Reuses a fixed pretrained backbone while specializing the downstream module per task (short text, aspect-level, document-level) [17],	Datasets, preprocessing, and class-balance handling vary across studies; the same architecture can swing >20 accuracy points across datasets within one

[26], [45]	paper [47]
Graph-augmented designs improve robustness under limited labelled data by exploiting corpus-level statistics [51]	Stacking multiple learned components increases effective parameter count and overfitting risk on small labelled sets [37]
Attention and capsule-routing weights offer a partial, visualizable interpretability signal [3], [17], [20]	Attention's faithfulness as an explanation is contested [79], [80], and post-hoc, model-agnostic checks (Section 11) are rarely applied

14. FUTURE DIRECTIONS AND OPEN PROBLEMS

Building on the gaps identified above and in Sections 7, 10, 11, and 12, we propose five directions for future work on hybrid deep learning for text classification, in approximate order of how directly actionable we judge each to be given the current state of the literature.

- Parameter-efficient hybrid fine-tuning. Section 7.2 shows that adapters [65] and LoRA [64] are well established for adapting a single pretrained backbone efficiently, yet almost none of the hybrid-architecture studies in Sections 5-6 combine a parameter-efficient adaptation of the backbone with the additional CNN/RNN/graph module, instead fully fine-tuning the backbone alongside the new module. Systematically re-evaluating the BERT-CNN, BERT-BiLSTM/BiGRU, and BERT-RCNN designs of Section 5 with a LoRA- or adapter-adapted backbone in place of full fine-tuning, and reporting the resulting accuracy-versus-cost trade-off directly, is the most concrete and immediately actionable direction we identify.

- Standardized, cost-aware benchmarking. The evaluation heterogeneity documented in Sections 9 and 10 suggests a clear need for a shared benchmark suite, along the lines proposed in Table 5, that fixes dataset, split, and preprocessing across architectural families, and that reports parameter count, training time, and inference latency alongside accuracy (Section 12), so that hybridization's cost-benefit trade-off can be assessed on equal footing across studies rather than compared informally across incompatible experimental setups.

- Explainability-by-design, evaluated rather than asserted. Rather than treating attention weights as an incidental interpretability signal, future hybrid architectures could be designed with faithfulness of explanation as an explicit, measured objective for example, jointly optimizing for classification accuracy and for agreement between attention-based and perturbation-based (LIME/SHAP) or gradient-based (Integrated Gradients) explanations, and reporting a faithfulness metric rather than only a qualitative visualization directly addressing the gap identified in Section 11 between the interpretability claims made in several reviewed papers [3], [17], [20] and the largely unevaluated basis for those claims.

- Robustness and cross-domain / cross-lingual generalization. Extending the promising but still limited multilingual results in Section 8.5 [36], [48] to a systematic cross-lingual and cross-domain evaluation protocol, and testing hybrid architectures under adversarial or naturally shifted text (e.g., dialectal variation, code-switching, or evolving misinformation tactics), would help establish whether the accuracy gains reported on a single benchmark generalize beyond it, and would directly test the robustness gap quantified in Section 10.

- Deeper integration of graph, LLM, and pretrained-language-model representations. Early combinations such as BertGCN [56] and the Qwen-TextCNN-BERT hybrid [74] suggest that jointly learning graph-structural, LLM-scale, and Transformer-contextual representations is more effective than using any one in isolation; extending this integration to aspect-level and multi-label settings, and to lower-resource languages where corpus-level graph statistics or LLM world knowledge may partly compensate for scarce labelled data, remains comparatively under-explored relative to the sequence-only hybrids that dominate the current literature.

15. THREATS TO VALIDITY AND LIMITATIONS OF THIS SURVEY

In the interest of the same transparency this survey asks of the primary studies it reviews (Sections 9, 10, and 13), we state explicitly where our own methodology falls short of a fully systematic review, rather than presenting the PRISMA-informed protocol in Section 3 as equivalent to one.

- Single-pass, tool-assisted screening rather than independent dual review. Title/abstract screening and eligibility assessment (Section 3.3) were conducted as a single pass using general-purpose search and retrieval tooling rather than a dedicated systematic-review platform, and without a second, independent reviewer cross-checking inclusion/exclusion decisions or resolving disagreements, which PRISMA recommends and which reduces the risk of selection bias in a full systematic review.

- Database coverage and reproducibility of the search. Searches were executed through the search interfaces available to the authors rather than through direct, logged API queries against each database, so the exact result sets are not independently re-executable to the same standard as a search string run against a fixed database snapshot; the counts reported in Fig. 2 are therefore best read as an honest approximation of the selection process rather than an exactly reproducible audit trail.

- English-language and well-indexed-venue bias. Non-English-language venues without accessible abstracts were excluded by design (Section 3.2), and the search terms and sources used (arXiv, ACL Anthology, IEEE Xplore, ScienceDirect, SpringerLink, PubMed Central, PLOS ONE, Nature Scientific Reports, Semantic Scholar) skew toward well-indexed international venues; regional or non-English-indexed journals that may contain relevant hybrid-architecture work are likely under-represented.

- Recency and indexing lag. The drop in references from 2022-2023 to 2024-2025 visible in Fig. 3 most likely reflects publication and indexing lag rather than an actual slowdown in hybrid-architecture research; readers should not interpret Fig. 3 as evidence that the field is contracting.

- A small number of references cited by identifier rather than fully verified byline. Table entries and reference-list items marked accordingly (see the reference list) were confirmed via DOI, arXiv ID, or PMC ID and their reported findings, but their full author bylines could not be independently verified within the scope of this revision; we flag rather than silently complete these, and we recommend the authors verify them against the publisher record before camera-ready submission.

- No meta-analytic pooling of effect sizes. Given the dataset, preprocessing, and metric heterogeneity documented in Sections 9 and 10, we deliberately did not pool accuracy figures into a single effect-size estimate per architectural family, since doing so would imply a comparability across studies that Section 9 shows does not hold; the quantitative synthesis in Section 10 is accordingly narrower in scope than a meta-analysis, and should be read as a curated set of verbatim findings rather than an aggregated effect estimate.

- Currency of the LLM-era section. Section 7 reflects the literature available at the time of writing; parameter-efficient and LLM-based approaches to text classification are an active research area, and readers should expect this section in particular to date faster than Sections 4-6.

16. DISCUSSION

Read across the five architectural families and five application domains covered in this survey, hybridization is, in nearly every reviewed case, an empirically motivated response to a specific limitation of a single-architecture baseline: convolution alone cannot model long-range dependencies, recurrence alone is slow and can still lose information over long documents, a linear head on pretrainedembeddings can under-use the full token sequence, a purely sequential view of text cannot capture corpus-level or aspect-level structure, and full fine-tuning of an ever-larger backbone is an increasingly expensive default (Section 7). Framed this way, the choice among the five families in Section 2.3 is less a question of which family is universally best and more a question of which limitation matters most for a given deployment and, as Sections 10 and 12 make concrete, which cost the practitioner is and is not willing to pay to address it.

The quantitative synthesis in Section 10, the explainability analysis in Section 11, and the cost discussion in Section 12 together sharpen a single point that runs through this whole survey: the hybrid-architecture literature is more mature in its accuracy claims than in its accounting of what those claims cost to obtain or how reliably they can be explained. For researchers designing a new hybrid architecture, the practical implication is that

ablation studies against each individual component, explicit reporting of parameter count and latency (Section 12), evaluation on more one dataset (Section 10), and a faithfulness-checked explanation rather than an attention visualization alone (Section 11) would substantially strengthen a paper's contribution relative to the median study reviewed here. For practitioners selecting an architecture for deployment, the taxonomy and comparative tables in Sections 2, 5, 8, 9, and 13 are intended to make the relevant accuracy-cost-robustness-interpretability trade-offs explicit enough to support that choice without requiring an independent literature review of each architectural family.

Before conclusions are drawn, several limitations of the present survey itself should be acknowledged. Screening and eligibility assessment were conducted as a single pass without a second independent reviewer, which increases the risk of selection bias relative to a full systematic review. The search relied on the interfaces available to the authors rather than logged API queries against fixed database snapshots, so the exact result sets are not independently re-executable to the highest standard of reproducibility. The survey is restricted to English-language and well-indexed venues, and a small number of references could be confirmed only by document identifier rather than a fully verified author byline. In addition, no formal meta-analytic pooling of effect sizes was undertaken, precisely because the dataset and metric heterogeneity documented in Sections 9 and 10 would make such pooling unreliable. These constraints are stated so that the scope and strength of the survey's claims remain transparent.

17. CONCLUSION

We return to the four research objectives stated in the Introduction and evaluate the survey against each of them.

Objective O1 asked whether a coherent five-family taxonomy could organise the hybrid deep learning literature for text classification from 2014 to 2025. The taxonomy developed in Section 2.3—CNN-RNN hybrids, attention-augmented hybrids, pretrained-language-model hybrids, graph-augmented/ensemble hybrids, and LLM-era/parameter-efficient hybrids—successfully accommodates the 63 included studies and makes the principal design choices of the field comparable within a single framework.

Objective O2 required a quantitative synthesis rather than purely qualitative description. Section 10 compiles reported accuracy deltas, parameter counts, and dataset dependencies directly from the source papers. The synthesis confirms that hybridization frequently improves on single-architecture baselines, yet it also shows that gains are highly dataset-dependent and that computational overhead is reported far less consistently than accuracy.

Objective O3 concerned explainability and computational cost. Sections 11 and 12 demonstrate that attention visualisation remains the dominant (and often unvalidated) explanation method, while post-hoc techniques such as LIME, SHAP, and Integrated Gradients appear only sporadically. Computational and energy cost are still under-reported, even though many hybrids add layers on top of already large pretrained backbones.

Objective O4 asked for the identification of recurring weaknesses and a forward research agenda. Across the reviewed literature the same problems recur: inconsistent baselines, limited robustness and cross-lingual evaluation, underreported cost, and interpretability claims that are rarely tested for faithfulness. The agenda proposed in Section 14—parameter-efficient hybrid fine-tuning, standardised cost-aware benchmarking, and explainability-by-design—follows directly from these observations.

Overall, hybridization is an empirically well-motivated response to the complementary strengths and weaknesses of individual neural components. At the same time, the literature remains more mature in its accuracy claims than in its accounting of cost, robustness, and the reliability of explanations. Addressing the methodological gaps identified in this survey is a necessary next step if hybrid text classifiers are to become not only more accurate but also more reproducible, interpretable, and practical to deploy.

DECLARATIONS

Funding: No funding for this work
Conflict of interest: The authors declare no conflict of interest.

REFERENCES

[1] Y. Kim, "Convolutional Neural Networks for Sentence Classification," in Proc. 2014 Conf. Empirical Methods in Natural Language

Processing (EMNLP), Doha, Qatar, 2014, pp. 1746-1751.
[2] S. Lai, L. Xu, K. Liu, and J. Zhao, "Recurrent Convolutional Neural Networks for Text Classification," in Proc. 29th AAAI Conf. Artificial Intelligence, 2015, pp. 2267-2273.
[3] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical Attention Networks for Document Classification," in Proc. 2016 Conf. North American Chapter Assoc. Comput. Linguistics: Human Language Technologies (NAACL-HLT), San Diego, CA, 2016, pp. 1480-1489.
[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," in Advances in Neural Information Processing Systems 30 (NeurIPS), 2017, pp. 5998-6008.
[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. 2019 Conf. North American Chapter Assoc. Comput. Linguistics: Human Language Technologies (NAACL-HLT), Minneapolis, MN, 2019, pp. 4171-4186.
[6] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv preprint arXiv:1907.11692, 2019.
[7] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized Autoregressive Pretraining for Language Understanding," in Advances in Neural Information Processing Systems 32 (NeurIPS), 2019.
[8] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep Contextualized Word Representations," in Proc. 2018 Conf. North American Chapter Assoc. Comput. Linguistics: Human Language Technologies (NAACL-HLT), 2018, pp. 2227-2237.
[9] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," arXiv preprint arXiv:1301.3781, 2013.
[10] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global Vectors for Word Representation," in Proc. 2014 Conf. Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1532-1543.
[11] X. Zhang, J. Zhao, and Y. LeCun, "Character-Level Convolutional Networks for Text Classification," in Advances in Neural Information Processing Systems 28 (NeurIPS), 2015, pp. 649-657.
[12] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter," arXiv preprint arXiv:1910.01108, 2019.
[13] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of Tricks for Efficient Text Classification," in Proc. 15th Conf. European Chapter Assoc. Comput. Linguistics (EACL), 2017, pp. 427-431.
[14] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic Routing Between Capsules," in Advances in Neural Information Processing Systems 30 (NeurIPS), 2017, pp. 3856-3866.
[15] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735-1780, 1997.
[16] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation," in Proc. 2014 Conf. Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1724-1734.
[17] Y. Zhu, X. Gao, W. Zhang, S. Liu, and Y. Zhang, "A Bi-Directional LSTM-CNN Model with Attention for Aspect-Level Text Classification," Future Internet, vol. 10, no. 12, p. 116, 2018.
[18] A. Yenter and A. Verma, "Deep CNN-LSTM with Combined Kernels from Multiple Branches for IMDB Review Sentiment Analysis," in Proc. IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conf. (UEMCON), 2017, pp. 540-546.
[19] M. U. Salur and I. Aydin, "A Novel Hybrid Deep Learning Model for Sentiment Classification," IEEE Access, vol. 8, pp. 58080-58093, 2020.
[20] L. Guo, D. Zhang, L. Wang, H. Wang, and B. Cui, "CRAN: A Hybrid CNN-RNN Attention-Based Model for Text Classification," in Conceptual Modeling: 37th Int. Conf. (ER 2018), Xi'an, China, 2018, pp. 571-585.
[21] C. Wang, F. Jiang, and H. Yang, "A Hybrid Framework for Text Modeling with Convolutional RNN," in Proc. 23rd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2017, pp. 2061-2069.
[22] D. Tang, B. Qin, and T. Liu, "Document Modeling with Gated Recurrent Neural Network for Sentiment Classification," in Proc.

2015 Conf. Empirical Methods in Natural Language Processing (EMNLP), 2015, pp. 1422-1432.

[23] J. Y. Lee and F. Dernoncourt, "Sequential Short-Text Classification with Recurrent and Convolutional Neural Networks," arXiv preprint arXiv:1603.03827, 2016.

[24] C. Zhou, C. Sun, Z. Liu, and F. Lau, "A C-LSTM Neural Network for Text Classification," arXiv preprint arXiv:1511.08630, 2015.

[25] J. Dai and C. Chen, "Text Classification System of Academic Papers Based on Hybrid BERT-BiGRU Model," in Proc. 12th Int. Conf. Intelligent Human-Machine Systems and Cybernetics (IHMSC), 2020, pp. 40-44.

[26] T. Bao, N. Ren, R. Luo, B. Wang, G. Shen, and T. Guo, "A BERT-Based Hybrid Short Text Classification Model Incorporating CNN and Attention-Based BiGRU," Journal of Organizational and End User Computing, vol. 33, no. 6, pp. 1-21, 2021.

[27] J. Alghamdi, Y. Lin, and S. Luo, "Modeling Fake News Detection Using BERT-CNN-BiLSTM Architecture," in Proc. IEEE 5th Int. Conf. on Multimedia Information Processing and Retrieval, 2022.

[28] C. E. Benarab and S. Gui, "CNN-Trans-Enc: A CNN-Enhanced Transformer-Encoder on Top of Static BERT Representations for Document Classification," arXiv preprint arXiv:2209.06344, 2022. (Reports 99.94% accuracy on AG News and 99.51% on DBPedia-14.)

[29] X. Jiang, C. Song, Y. Xu, Y. Li, and Y. Peng, "Research on Sentiment Classification for Netizens Based on the BERT-BiLSTM-TextCNN Model," PeerJ Computer Science, vol. 8, p. e1005, 2022.

[30] J. Deng, L. Cheng, and Z. Wang, "Attention-Based BiLSTM Fused CNN with Gating Mechanism Model for Chinese Long Text Classification," Computer Speech & Language, vol. 68, p. 101182, 2021.

[31] X. Li, Y. Lei, and S. Ji, "BERT- and BiLSTM-Based Sentiment Analysis of Online Chinese Buzzwords," Future Internet, vol. 14, no. 11, p. 332, 2022.

[32] F. Li, Y. Huo, and L. Wang, "Research on Chinese Emotion Classification Using BERT-RCNN-ATT," WSEAS Transactions on Communications, vol. 22, 2023.

[33] "A Novel Approach for Text Classification by Combining Pre-Trained BERT Model with CNN Classifier," in Proc. IEEE Conf. (IEEE Xplore document 10392947), 2023. [Cited by document identifier; full author byline not independently confirmed at time of this revision — to be verified against publisher record before camera-ready submission.]

[34] H. Huang et al., "DCNN-BiGRU Text Classification Model Based on BERT Embedding," in Proc. IEEE Int. Conf. Ubiquitous Computing and Communications (IUCC) / Data Science and Computational Intelligence (DSCI) / Smart Computing, Networking and Services (SmartCNS), 2019.

[35] L. Cai, Y. Song, T. Liu, and K. Zhang, "A Hybrid BERT Model That Incorporates Label Semantics via Adjustive Attention for Multi-Label Text Classification," IEEE Access, vol. 8, pp. 152183-152192, 2020.

[36] "Hybrid Machine Learning Model for Detecting Bangla Smishing Text Using BERT and Character-Level CNN," arXiv preprint arXiv:2502.01518, 2025.

[37] "A Hybrid Deep Learning Model for Sentiment Analysis of COVID-19 Tweets with Class Balancing," PubMed Central ID PMC12311151, 2025. [Cited by repository identifier; full author byline to be verified before camera-ready submission.]

[38] "Medical Text Classification Using Hybrid Deep Learning Models with Multihead Attention," PubMed Central ID PMC8486521, 2021. [Cited by repository identifier; full author byline to be verified before camera-ready submission.]

[39] "Research on a Hybrid LSTM-CNN-Attention Model for Text-Based Web Content Classification," arXiv preprint arXiv:2512.18475, 2025.

[40] O. Habimana, Y. Li, R. Li, X. Gu, and G. Yu, "Sentiment Analysis Using Deep Learning Approaches: An Overview," Science China Information Sciences, vol. 63, no. 1, pp. 1-36, 2019.

[41] K. Kowsari, K. JafariMeimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text Classification Algorithms: A Survey," Information, vol. 10, no. 4, p. 150, 2019.

[42] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep Learning-Based Text Classification: A Comprehensive Review," ACM Computing Surveys, vol. 54, no. 3, pp. 62:1-62:40, 2022.

[43] R. K. Kaliyar, A. Goswami, and P. Narang, "FakeBERT: Fake News Detection in Social Media with a BERT-Based Deep Learning

Approach," Multimedia Tools and Applications, vol. 80, no. 8, pp. 11765-11788, 2021.

[44] N. Ruchansky, S. Seo, and Y. Liu, "CSI: A Hybrid Deep Model for Fake News Detection," in Proc. 2017 ACM Conf. Information and Knowledge Management (CIKM), 2017, pp. 797-806.

[45] K. L. Tan, C. P. Lee, K. S. M. Anbananthen, and K. M. Lim, "RoBERTa-LSTM: A Hybrid Model for Sentiment Analysis with Transformer and Recurrent Neural Network," IEEE Access, vol. 10, pp. 21517-21525, 2022.

[46] N. Kausar, A. Ali Khan, and M. Sattar, "Towards Better Representation Learning Using Hybrid Deep Learning Model for Fake News Detection," Social Network Analysis and Mining, vol. 12, no. 1, p. 165, 2022.

[47] A. J. Keya, H. H. Shajeeb, M. S. Rahman, and M. F. Mridha, "FakeStack: Hierarchical Tri-BERT-CNN-LSTM Stacked Model for Effective Fake News Detection," PLOS ONE, vol. 18, no. 12, p. e0294701, 2023. (Reports 99.74% accuracy on the primary dataset, 75.58% on LIAR, and 98.25% on WELFake.)

[48] "Ensemble Based High Performance Deep Learning Models for Fake News Detection," PubMed Central ID PMC11535404, 2024. [Cited by repository identifier; full author byline to be verified before camera-ready submission.]

[49] "A Hybrid Deep Learning Framework for Fake News Detection Using LSTM-CGPNN and Metaheuristic Optimization," Scientific Reports, 2025. [Full author byline to be verified against the publisher record before camera-ready submission.]

[50] J. A. Benitez-Andrades, A. Gonzalez-Jimenez, A. Lopez-Brea, et al., "Detecting Racism and Xenophobia Using Deep Learning Models on Twitter Data: CNN, LSTM and BERT," PeerJ Computer Science, vol. 8, p. e906, 2022.

[51] L. Yao, C. Mao, and Y. Luo, "Graph Convolutional Networks for Text Classification," in Proc. 33rd AAAI Conf. Artificial Intelligence, 2019, pp. 7370-7377.

[52] C. Zhang, Q. Li, and D. Song, "Aspect-Based Sentiment Classification with Aspect-Specific Graph Convolutional Networks," in Proc. 2019 Conf. Empirical Methods in Natural Language Processing and 9th Int. Joint Conf. Natural Language Processing (EMNLP-IJCNLP), 2019.

[53] P. Zhao, L. Hou, and O. Wu, "Modeling Sentiment Dependencies with Graph Convolutional Networks for Aspect-Level Sentiment Classification," Knowledge-Based Systems, 2020.

[54] Y. Zhang, X. Yu, Z. Cui, S. Wu, Z. Wen, and L. Wang, "Every Document Owns Its Structure: Inductive Text Classification via Graph Neural Networks," in Proc. 58th Annual Meeting Assoc. Comput. Linguistics (ACL), 2020.

[55] L. Huang et al., "Text Level Graph Neural Network for Text Classification," in Proc. 2019 Conf. Empirical Methods in Natural Language Processing and 9th Int. Joint Conf. Natural Language Processing (EMNLP-IJCNLP), 2019.

[56] Y. Lin, Y. Meng, X. Sun, Q. Han, K. Kuang, J. Li, and F. Wu, "BertGCN: Transductive Text Classification by Combining GNN and BERT," in Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 2021, pp. 1456-1462.

[57] J. Liu, W.-C. Chang, Y. Wu, and Y. Yang, "Deep Learning for Extreme Multi-Label Text Classification," in Proc. 40th Int. ACM SIGIR Conf. Research and Development in Information Retrieval, 2017, pp. 115-124.

[58] M. Z. Islam, J. Liu, J. Li, L. Liu, and W. Kang, "A Semantics Aware Random Forest for Text Classification," in Proc. 28th ACM Int. Conf. Information and Knowledge Management (CIKM), 2019, pp. 1061-1070.

[59] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785-794.

[60] H. Sobhanam and J. Prakash, "Analysis of Fine Tuning the Hyper Parameters in RoBERTa Model Using Genetic Algorithm for Text Classification," International Journal of Information Technology, vol. 15, no. 7, pp. 3669-3677, 2023.

[61] Y. Zhang et al., "Hybrid Photonic Deep Convolutional Residual Spiking Neural Networks for Text Classification," Optics Express, vol. 31, no. 17, pp. 28489-28502, 2023. (A hardware-efficient hybrid directly relevant to the Green-AI direction discussed in Section 13.)

[62] S. K. Akpatsa, X. Li, and H. Lei, "A Survey and Future Perspectives of Hybrid Deep Learning Models for Text Classification," in Proc. Int. Conf. (Springer), 2021.

[63] "A Survey on Text Classification Using Deep Learning Approaches," in Proc. Int. Conf. (Springer), 2025.
[64] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," arXiv preprint arXiv:2106.09685, 2021.
[65] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-Efficient Transfer Learning for NLP," in Proc. 36th Int. Conf. Machine Learning (ICML), 2019, pp. 2790-2799.
[66] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A Lite BERT for Self-Supervised Learning of Language Representations," in Proc. Int. Conf. Learning Representations (ICLR), 2020.
[67] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "ELECTRA: Pre-Training Text Encoders as Discriminators Rather Than Generators," in Proc. Int. Conf. Learning Representations (ICLR), 2020.
[68] P. He, X. Liu, J. Gao, and W. Chen, "DeBERTa: Decoding-Enhanced BERT with Disentangled Attention," in Proc. Int. Conf. Learning Representations (ICLR), 2021.
[69] T. Schick and H. Schütze, "Exploiting Cloze Questions for Few-Shot Text Classification and Natural Language Inference," in Proc. 16th Conf. European Chapter Assoc. Comput. Linguistics (EACL), 2021, pp. 255-269.
[70] T. Gao, A. Fisch, and D. Chen, "Making Pre-Trained Language Models Better Few-Shot Learners," in Proc. 59th Annual Meeting Assoc. Comput. Linguistics (ACL), 2021, pp. 3816-3830.
[71] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le, "Finetuned Language Models Are Zero-Shot Learners," arXiv preprint arXiv:2109.01652, 2022.
[72] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., "LLaMA: Open and Efficient Foundation Language Models," arXiv preprint arXiv:2302.13971, 2023.
[73] X. Sun, X. Li, J. Li, F. Wu, S. Guo, T. Zhang, and G. Wang, "Text Classification via Large Language Models," in Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, 2023, pp. 8990-9005.
[74] D. Yuan, G. Liang, B. Liu, and S. Liu, "Qwen, TextCNN and BERT Models for Enhanced Multilabel News Classification in Mobile Apps," Scientific Reports, 2025.
[75] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 1135-1144.
[76] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in Advances in Neural Information Processing Systems 30 (NeurIPS), 2017.
[77] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic Attribution for Deep Networks," in Proc. 34th Int. Conf. Machine Learning (ICML), 2017, pp. 3319-3328.
[78] S. Abnar and W. Zuidema, "Quantifying Attention Flow in Transformers," in Proc. 58th Annual Meeting Assoc. Comput. Linguistics (ACL), 2020, pp. 4190-4197.
[79] S. Jain and B. C. Wallace, "Attention Is Not Explanation," in Proc. 2019 Conf. North American Chapter Assoc. Comput. Linguistics: Human Language Technologies (NAACL-HLT), 2019, pp. 3543-3556.
[80] S. Wiegreffe and Y. Pinter, "Attention Is Not Not Explanation," in Proc. 2019 Conf. Empirical Methods in Natural Language Processing and 9th Int. Joint Conf. Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 11-20.
[81] E. Strubell, A. Ganesh, and A. McCallum, "Energy and Policy Considerations for Deep Learning in NLP," in Proc. 57th Annual Meeting Assoc. Comput. Linguistics (ACL), 2019, pp. 3645-3650.
[82] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, et al., "The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews," BMJ, vol. 372, p. n71, 2021.
[83] P. K. Verma, P. Agrawal, I. Amorim, and R. Prodan, "WELFake: Word Embedding Over Linguistic Features for Fake News Detection," IEEE Transactions on Computational Social Systems, vol. 8, no. 4, pp. 881-893, 2021.