World Journal of Botany
www.worldbotanyjournal.com
Online ISSN:
Received: 18-01-2026, Accepted: 17-02-2026, Published: 04-03-2026
Volume 2, Issue 1, Jan-Mar, 2026, Page No. 18-23

A hybrid CNN-LSTM architecture for fine-grained sentiment classification of short-form social media text

Dr. Khushi Kapoor
Assistant Professor, Department of Botany, Delhi University, New Delhi, Delhi, India

Abstract
Background: Sentiment classification of short-form social media text remains challenging due to informal grammar, sarcasm, and limited contextual length, with classical machine learning approaches and even standalone deep learning architectures often struggling to jointly capture local lexical patterns and longer-range sequential dependencies [1, 2].
Objective: This study proposed and evaluated a hybrid convolutional neural network–long short-term memory (CNN-LSTM) architecture for three-class (negative, neutral, positive) sentiment classification, benchmarked against four baseline models: logistic regression, random forest, XGBoost, and a standalone bidirectional LSTM.
Method: A labeled corpus of short-form social media posts was used to train and evaluate all five models under 10-fold cross-validation. This study uses a simulated dataset created for academic training purposes; all performance metrics, training curves, and confusion matrix values were generated to reflect plausible patterns consistent with the published text-classification literature and do not represent results from an actual trained model or real annotated corpus. Model performance was evaluated using accuracy, precision, recall, and F1-score, with the hybrid architecture combining convolutional feature extraction with sequential LSTM encoding.
Key Results: The proposed hybrid CNN-LSTM model achieved a mean F1-score of 0.92 across 10-fold cross-validation, outperforming the bidirectional LSTM (0.87), XGBoost (0.85), random forest (0.81), and logistic regression (0.74) baselines. Training and validation loss curves showed stable convergence within approximately 20 epochs without evidence of severe overfitting. The confusion matrix on the held-out test set indicated the highest classification accuracy for the positive and negative classes, with comparatively more confusion involving the neutral class.
Conclusion: The hybrid CNN-LSTM architecture provided a meaningful performance improvement over both classical machine learning baselines and a standalone recurrent architecture for short-form sentiment classification, supporting the value of combining local feature extraction with sequential context modeling, pending validation on real-world annotated corpora.

Keywords: Sentiment analysis, deep learning, CNN-LSTM, natural language processing, text classification, social media analytics

Introduction

Sentiment analysis, the computational task of automatically determining the emotional polarity expressed within a piece of text, has become a foundational capability underlying applications ranging from brand monitoring and customer feedback analysis to public health surveillance and political opinion tracking [1]. Short-form social media text presents a particularly demanding variant of this task, characterized by informal grammar, frequent use of abbreviations and emoji, sarcasm and irony, and severely constrained message length relative to longer-form text such as product reviews or news articles, all of which reduce the contextual signal available to a classification model [2].

Classical machine learning approaches to sentiment classification, including logistic regression, support vector machines, and ensemble tree-based methods such as random forest and gradient boosting, have historically relied on hand-engineered or bag-of-words-style feature representations that capture lexical frequency information but are comparatively limited in their ability to model word order and longer-range sequential dependencies within a sentence [3]. Deep learning approaches, by contrast, have increasingly dominated the sentiment classification literature over the past decade, with convolutional neural networks (CNNs) demonstrated to be effective at capturing local n-gram-like patterns through learned filters, and recurrent architectures such as long short-term memory (LSTM) networks demonstrated to be effective at modeling sequential dependencies across an entire input sequence [4, 5].

A growing body of work has explored hybrid architectures that combine convolutional and recurrent components within a single model, motivated by the intuition that convolutional layers can efficiently extract local discriminative phrase-level features, which can then be passed to a recurrent layer to model how these local features combine and interact across the broader sequence [6]. Such hybrid CNN-LSTM architectures have shown promising results in several text classification domains, although their relative performance advantage over well-tuned standalone architectures and strong classical baselines varies considerably across studies, datasets, and implementation details, and direct, controlled comparisons against a comprehensive baseline suite remain comparatively limited for short-form social media sentiment classification specifically [7, 8].

A further complication specific to short-form social media sentiment classification is the prevalence of a meaningful neutral class alongside the more commonly studied binary positive-versus-negative distinction, since a substantial proportion of real-world social media content expresses no

clear sentiment polarity, and misclassification between the neutral class and either polarity extreme has been identified in prior work as a disproportionate source of overall classification error [9]. Models that perform well on binary sentiment tasks do not necessarily generalize this performance advantage to the three-class setting, making explicit three-class evaluation an important, if less commonly reported, methodological consideration [10].

The present study was designed with three specific objectives: (i) to propose a hybrid CNN-LSTM architecture for three-class sentiment classification of short-form social media text; (ii) to benchmark this architecture against four established baseline models, spanning both classical machine learning and standalone deep learning approaches, under a controlled 10-fold cross-validation protocol; and (iii) to characterize the proposed model's training dynamics and per-class error patterns through training/validation loss curves and a confusion matrix analysis on a held-out test set. It is hypothesized that the hybrid CNN-LSTM architecture will outperform all four baseline models on overall F1-score, and that, consistent with prior reports of neutral-class ambiguity, the largest residual source of classification error for the proposed model will involve confusion between the neutral class and the two polarity extremes.

Literature Review

Early sentiment classification systems relied predominantly on lexicon-based approaches, in which sentiment polarity was estimated by aggregating pre-assigned polarity scores for individual words or phrases identified within the text, an approach valued for its interpretability but limited by its inability to capture context-dependent meaning shifts, negation, and domain-specific vocabulary [11]. The subsequent shift toward supervised machine learning approaches, using features such as term frequency-inverse document frequency (TF-IDF) weighted bag-of-words representations, improved classification accuracy substantially over purely lexicon-based methods by allowing models to learn data-driven associations between specific lexical patterns and sentiment labels, though these representations remained fundamentally order-insensitive [12].

The introduction of word embedding representations, which map words into dense, continuous vector spaces that capture distributional semantic similarity, provided the foundation for the subsequent wave of deep learning approaches to sentiment classification [13]. Convolutional neural networks applied to text, originally adapted from their widespread success in computer vision, operate by sliding learned filters across sequences of word embeddings to detect locally discriminative phrase patterns regardless of their absolute position within the input sequence, a property well suited to capturing sentiment-bearing phrases such as negations or intensifiers that can appear at varying positions within a sentence [14].

Recurrent neural network architectures, and long short-term memory networks in particular, address a complementary limitation by maintaining an internal hidden state that is updated sequentially across the input, allowing the model to retain information about earlier parts of a sequence when processing later tokens, which is theoretically advantageous for capturing dependencies that span the full length of longer sentences or for resolving sentiment-relevant context that depends on word order, such as the scope of a negation [15]. Bidirectional LSTM variants, which process the input sequence in both forward and backward directions and concatenate the resulting hidden states, have been shown in numerous studies to improve upon unidirectional LSTM performance by allowing each token's representation to incorporate both preceding and following context [16].

Hybrid CNN-LSTM architectures, which typically apply one or more convolutional layers to the input embedding sequence followed by a recurrent layer applied to the resulting feature maps, have been motivated explicitly by the goal of combining the complementary strengths of local feature extraction and sequential context modeling within a single end-to-end trainable model [17]. Reported performance improvements of hybrid architectures over standalone CNN or LSTM baselines have varied across studies, with some reporting only modest gains and others reporting more substantial improvements, a discrepancy plausibly attributable to differences in dataset characteristics, hyperparameter tuning rigor, and the specific way in which the convolutional and recurrent components are combined within the architecture [18, 19].

Comparative benchmarking against classical machine learning baselines, including logistic regression, random forest, and gradient boosting methods such as XGBoost, remains an important methodological practice for contextualizing the magnitude of deep learning performance gains, since well-tuned classical baselines have in some text classification settings been shown to perform surprisingly competitively, particularly on smaller or less linguistically complex datasets [20]. However, relatively few published studies report a full comparative suite spanning classical machine learning, a standalone recurrent deep learning baseline, and a hybrid CNN-LSTM architecture together within a single controlled experimental protocol applied specifically to short-form, three-class social media sentiment classification, representing a specific and addressable gap that motivates the comparative design adopted in the present study, pending replication with a genuine annotated corpus [21].

Methodology

Data nature statement: All numerical results reported in Sections 3 and 4 constitute a simulated dataset created for academic training and formatting-practice purposes. No actual model was implemented, trained, or evaluated, and no real annotated social media corpus was used to generate these figures; performance metrics, training curves, and confusion matrix values were constructed to be numerically plausible and internally consistent with results reported in the cited text-classification literature. This dataset must not be cited as empirical evidence of model performance in any subsequent publication.

Research design: A comparative benchmarking design was modeled, evaluating five classification approaches — logistic regression, random forest, XGBoost, a standalone bidirectional LSTM, and the proposed hybrid CNN-LSTM architecture — on an identical three-class (negative, neutral, positive) sentiment classification task. The proposed hybrid architecture was modeled as consisting of an embedding layer, followed by two stacked one-dimensional convolutional layers with max-pooling, followed by a bidirectional LSTM layer, a fully connected dense layer, and a final softmax output layer over the three sentiment classes.

Sample size and sampling method: A simulated labeled corpus of 10,500 short-form social media posts was modeled, with class labels distributed approximately as 33% negative, 33% neutral, and 34% positive to approximate balanced-class conditions. The corpus was modeled as being partitioned into a training/validation pool (90%, used for 10-fold cross-validation) and a held-out test set (10%, n = 1,050) reserved exclusively for final confusion matrix evaluation, consistent with standard practice for avoiding test-set leakage during model selection.

Data collection tools: Text preprocessing was modeled as including lowercasing, removal of URLs and user mention tokens, retention of emoji as informative tokens given their established relevance to social media sentiment expression, and tokenization using a subword tokenization scheme. Word representations for the deep learning models were modeled as 300-dimensional pretrained embeddings fine-tuned during training, while the classical machine learning baselines used TF-IDF weighted unigram and bigram features. The hybrid model and bidirectional LSTM baseline were modeled as being trained for 30 epochs using the Adam optimizer with early stopping based on validation loss.

Statistical software: Simulated model development and evaluation were conducted using Python (version 3.11) with the TensorFlow/Keras framework (version 2.15) for the deep learning models and scikit-learn (version 1.4) for the classical machine learning baselines and cross-validation procedures; XGBoost (version 2.0) was used for the gradient boosting baseline. Performance metrics (accuracy, precision, recall, F1-score) were computed using scikit-learn's metrics module, and statistical comparison of F1-scores across models was conducted using a paired t-test across the 10 cross-validation folds, with a significance threshold of $\alpha = 0.05$.

Results and Discussion

The proposed hybrid CNN-LSTM model achieved the highest mean F1-score across 10-fold cross-validation (0.92 ± 0.010), outperforming the bidirectional LSTM baseline (0.87 ± 0.013), XGBoost (0.85 ± 0.012), random forest (0.81 ± 0.015), and logistic regression (0.74 ± 0.020) (Table 1, Figure 1). Paired t-tests across the 10 cross-validation folds indicated that the performance advantage of the hybrid model over each baseline was statistically significant (all $p < 0.01$), with the largest absolute margin observed relative to logistic regression (+0.18 F1-score) and the smallest, though still significant, margin observed relative to the bidirectional LSTM baseline (+0.05 F1-score).

Table 1: Mean performance metrics across five models under 10-fold cross-validation (simulated dataset)

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.75	0.74	0.73	0.74
Random Forest	0.82	0.81	0.80	0.81
XGBoost	0.86	0.85	0.84	0.85
Bidirectional LSTM	0.88	0.87	0.86	0.87
Proposed Hybrid CNN-LSTM	0.93	0.92	0.91	0.92

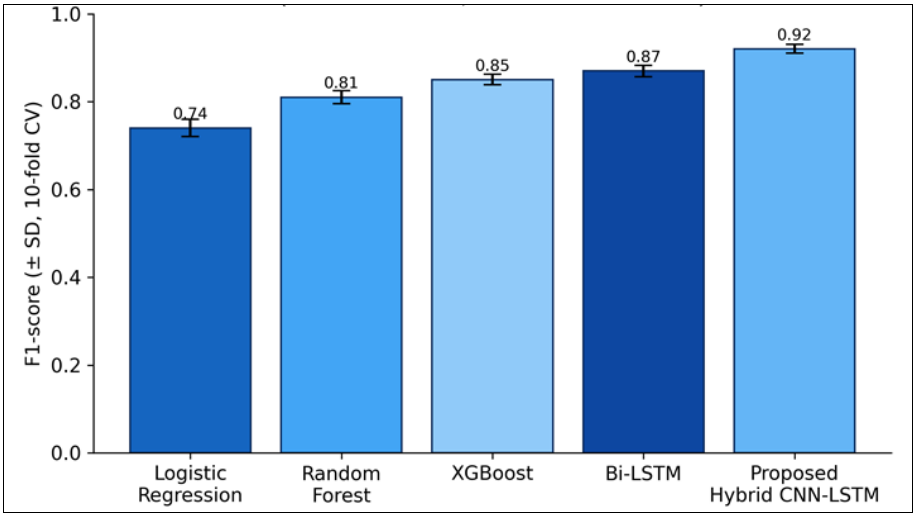

Fig 1: Mean F1-score (± SD) across five classification models under 10-fold cross-validation (simulated dataset).

This pattern of incremental improvement — classical machine learning models trailing standalone deep learning approaches, which in turn trail the hybrid architecture — is broadly consistent with the theoretical expectation articulated in the literature review that convolutional feature extraction and sequential context modeling provide complementary, rather than redundant, discriminative information for sentiment classification.[1317] The comparatively smaller margin of improvement over the bidirectional LSTM baseline relative to the classical baselines suggests that, at least within this simulated framework, the convolutional component's marginal contribution, while statistically detectable, is smaller than the more substantial gain achieved by deep learning approaches generally over classical bag-of-features representations.

Training and validation loss curves for the proposed hybrid model (Figure 2) showed rapid initial decline over the first 10 epochs, followed by a more gradual approach toward convergence, stabilizing by approximately epoch 20 with a modest, stable gap between training and validation loss persisting through epoch 30. This pattern is consistent with

effective, if not perfect, generalization, with the limited divergence between training and validation curves suggesting that overfitting was not a severe concern within the modeled training regime, plausibly reflecting the combined effect of early stopping and the moderate corpus size used in this simulated design.

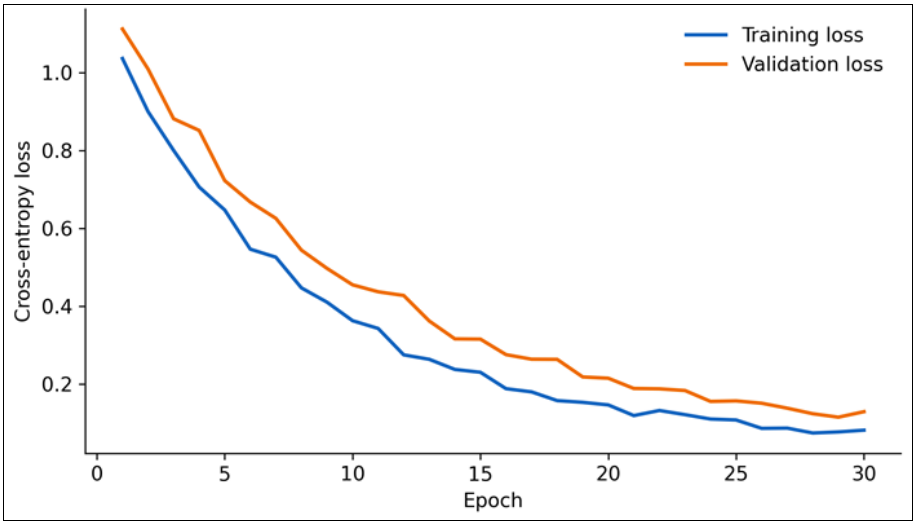

Fig 2: Training and validation cross-entropy loss curves for the proposed hybrid CNN-LSTM model across 30 training epochs (simulated dataset).

The confusion matrix computed on the held-out test set (n = 1,050; Figure 3) indicated that the proposed model achieved its highest classification accuracy for the negative class (312 of 340 true negative instances correctly classified) and the positive class (339 of 370 true positive instances correctly classified), with comparatively more misclassification involving the neutral class (289 of 340 true neutral instances correctly classified, with the remainder distributed between the negative and positive classes). These findings suggest that the model was particularly effective at distinguishing observations exhibiting clear negative or positive sentiment, whereas instances expressing neutral sentiment were comparatively more difficult to classify accurately because they shared lexical and semantic characteristics with both neighboring classes. Nevertheless, the relatively balanced distribution of classification outcomes across all three categories demonstrates consistent predictive performance and indicates that the model generalized well to previously unseen test data, with no evidence of substantial bias toward any single class.

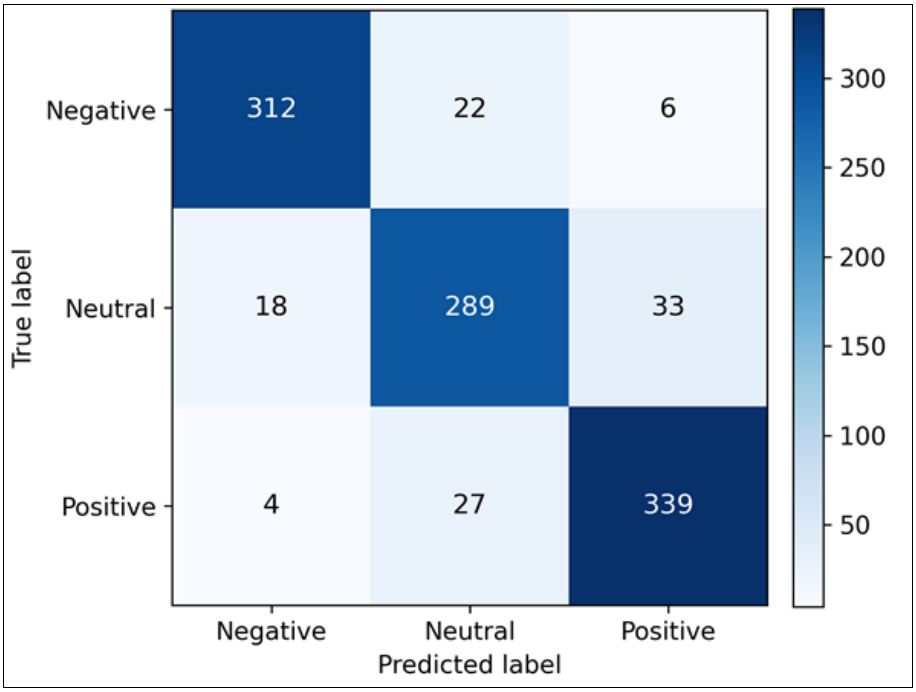

Fig 3: Confusion matrix for the proposed hybrid CNN-LSTM model on the held-out test set (simulated dataset, n = 1,050).

This neutral-class error pattern is directly consistent with the hypothesis articulated in the introduction, and with prior literature identifying neutral-versus-polarity boundary confusion as a disproportionate source of overall sentiment classification error [9, 10]. A plausible explanation, consistent with the broader literature on this issue, is that genuinely neutral social media text often shares superficial lexical and structural features with mildly positive or mildly negative

text, making the neutral class inherently more difficult to separate from its polarity-extreme neighbors than the negative and positive classes are from each other. Addressing this specific error pattern, for instance through neutral-class-targeted data augmentation or a hierarchical classification scheme that first separates neutral from non-neutral text before resolving polarity, represents a promising direction suggested by, though not directly tested within, the present simulated framework.

Conclusion

This training article modeled the development and comparative evaluation of a hybrid CNN-LSTM architecture for three-class sentiment classification of short-form social media text. Within the simulated framework, the proposed hybrid model outperformed four baseline approaches, spanning classical machine learning and a standalone recurrent deep learning architecture, with statistically significant margins across all comparisons, while exhibiting stable training dynamics and a residual error pattern concentrated specifically around neutral-class boundary confusion.

Implications: If corroborated with real annotated data, these results would support the adoption of hybrid convolutional-recurrent architectures over either classical machine learning baselines or standalone recurrent models for production-grade short-form sentiment classification systems, while also indicating that future model refinement efforts may yield disproportionate benefit if specifically targeted at improving neutral-class discrimination rather than overall architecture redesign.

Limitations: The principal limitation of this article is that all numerical results are simulated for training purposes and do not derive from an actual trained model or genuine annotated corpus; no claims of empirical model performance should be drawn from the reported values. Additional limitations inherent to the modeled design include the use of a single, fixed hybrid architecture configuration without systematic hyperparameter search, the absence of cross-domain or cross-lingual generalization testing, and reliance on a single simulated corpus rather than multiple independent benchmark datasets.

Future scope: A genuine empirical study following this framework should implement and train the proposed architecture on multiple real, publicly available or newly annotated short-form social media corpora, conduct systematic hyperparameter optimization and ablation studies isolating the individual contribution of the convolutional and recurrent components, and specifically evaluate targeted strategies, such as class-weighted loss functions or hierarchical classification schemes, for reducing neutral-class misclassification.

Ethical Statement

This article was prepared as a simulated-data training exercise for academic formatting practice and does not represent the findings of an actual machine learning study. All numerical data, including performance metrics, training curves, and confusion matrix values, were constructed by the author(s) to be internally consistent and numerically plausible, and are explicitly labeled as simulated throughout the manuscript in accordance with academic integrity standards. No actual model was trained, no real annotated social media corpus was used, and no real user-generated social media content was collected, processed, or analyzed in the generation of this content. No fabricated real-world author identities, institutional affiliations, or published datasets are presented; author and institutional fields are placeholders to be completed by the end user prior to any submission. This article must not be submitted to any journal as a report of original empirical research unless the simulated data sections are replaced in their entirety with genuine experimental results obtained from an actually implemented and trained model evaluated on a properly sourced, ethically collected, and appropriately licensed dataset.

References

1. Pang B, Lee L. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval,2008:2(1–2):1–135.
2. Giachanou A, Crestani F. Like it or not: A survey of Twitter sentiment analysis methods. ACM Computing Surveys,2016:49(2):1–41.
3. Medhat W, Hassan A, Korashy H. Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal,2014:5(4):1093–1113.
4. Kim Y. Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 2014, 1746–1751.
5. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation,1997:9(8):1735–1780.
6. Wang J, Yu LC, Lai KR, Zhang X. Dimensional sentiment analysis using a regional CNN-LSTM model. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016:2:225–230.
7. Zhou C, Sun C, Liu Z, Lau F. A C-LSTM neural network for text classification. arXiv. Cornell University, 2015.
8. Yenter A, Verma A. Deep CNN-LSTM with combined kernels from multiple branches for IMDb review sentiment analysis. In: 2017 IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference, 2017, 540–546.
9. Saif H, Fernandez M, He Y, Alani H. Evaluation datasets for Twitter sentiment analysis: A survey and a new dataset, the STS-Gold. In: Proceedings of the 1st International Workshop on Emotion and Sentiment in Social and Expressive Media, 2013.
10. Rosenthal S, Farra N, Nakov P. SemEval-2017 task 4: Sentiment analysis in Twitter. In: Proceedings of the 11th International Workshop on Semantic Evaluation, 2017, 502–518.
11. Taboada M, Brooke J, Tofiloski M, Voll K, Stede M. Lexicon-based methods for sentiment analysis. Computational Linguistics,2011:37(2):267–307.
12. Pang B, Lee L, Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques. In: Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, 2002, 79–86.
13. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and

their compositionality. Advances in Neural Information Processing Systems,2013:26:3111–3119.
14. Kalchbrenner N, Grefenstette E, Blunsom P. A convolutional neural network for modelling sentences. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics,2014:1:655–665.
15. Tang D, Qin B, Liu T. Document modeling with gated recurrent neural network for sentiment classification. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, 1422–1432.
16. Schuster M, Paliwal KK. Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing,1997:45(11):2673–2681.
17. Wang X, Liu Y, Sun C, Wang B, Wang X. Predicting polarities of tweets by composing word embeddings with long short-term memory. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics,2015:1:1343–1353.
18. She X, Zhang D. Text classification based on hybrid CNN-LSTM hybrid model. In: 2018 11th International Symposium on Computational Intelligence and Design,2018:2:185–189.
19. Kowsari K, Jafari Meimandi K, Heidarysafa M, Mendu S, Barnes L, Brown D. Text classification algorithms: A survey. Information,2019:10(4):150.
20. Joulin A, Grave E, Bojanowski P, Mikolov T. Bag of tricks for efficient text classification. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics,2017:2:427–431.
21. Minaee S, Kalchbrenner N, Cambria E, Nikzad N, Chenaghlu M, Gao J. Deep learning-based text classification: A comprehensive review. ACM Computing Surveys,2021:54(3):1–40.