RESEARCH PAPER

Hybrid Transformer Learning with Contextual Semantic Features for Sarcasm Detection

Yuvraj G. Nikam[1], Vijay Pal Singh[2], Dhanshri Amol Shinde[3]*
[1*]Research Scholar, Department of Computer Engineering and Applications, Faculty of Engineering and Technology, Mangalayatan University, Beswan, Aligarh, India
Email: 20231919_nikam@mangalayatan.edu.in
[2]Department of Computer Engineering and Applications, Faculty of Engineering and Technology, Mangalayatan University, Beswan, Aligarh, India
Email: drvijaypaltilotiya@gmail.com
[3]Department of Computer Engineering, Tolani Maritime Institute, Pune, India
Email: patildhanshri@gmail.com

Received: 1st Sep, 2026 | Revised: 8th Sep, 2026 | Accepted: 15th Sep, 2026 | Available Online: 22nd Sep, 2026

Abstract: Sarcasm detection is a challenging problem in computational linguistics because the sentiment expressed on the surface of a sarcastic utterance frequently contradicts the sentiment its author intends. Existing detectors are also brittle: models that perform well on one platform typically degrade sharply when applied to text from a different domain or writing style. This paper proposes a context-aware hybrid model that couples a pre-trained BERT encoder with a bidirectional LSTM and an additive attention layer, and fuses the resulting contextual representation with two complementary signals — an unsupervised topic-context vector obtained from BERTopic and a set of nine hand-engineered linguistic cues. The model is trained only on a labelled Twitter corpus and evaluated in two regimes: an in-domain held-out Twitter test set and a zero-shot cross-domain test on a balanced, formally-written News-Headlines corpus that the model never sees during training. Against four classical TF-IDF baselines and pure transformer baselines evaluated under an identical protocol, the proposed hybrid attains the strongest in-domain results, reaching an accuracy of 0.8695 and an F1-score of 0.7945, and the zero-shot cross-domain evaluation quantifies the extent to which a tweet-trained detector generalizes to a different register. The results indicate that combining a contextual transformer encoder with unsupervised topic context and linguistic cues is an effective and transferable design for sarcasm detection.

Keywords: sarcasm detection; natural language processing; BERT; BiLSTM; attention; BERTopic; topic modelling; hybrid model; cross-domain generalization; zero-shot evaluation

How to cite this article: Yuvraj G. Nikam1* Vijay Pal Singh2 Dhanshri Amol Shinde3. Hybrid Transformer Learning with Contextual Semantic Features for Sarcasm Detection. Int J Drug Deliv Technol. 2026;16(80s):567-574. DOI: 10.25258/ijddt.16.80s.70.

1. Introduction

Sarcasm is a figurative form of language in which the intended meaning of an utterance is opposite to, or sharply at odds with, its literal meaning. It is pervasive in online discourse, where users routinely deploy it to criticise, mock, or express frustration in a compact and socially acceptable way. For downstream tasks such as sentiment analysis, opinion mining, and content moderation, undetected sarcasm is a persistent source of error: a lexically positive statement that is in fact deeply negative will be systematically misclassified by a sentiment model that reads only the surface signal.

The core difficulty of sarcasm is the contrast between a positive surface sentiment and a negative underlying situation, or vice versa. Capturing this contrast reliably requires more than lexical features; it requires an understanding of context. A second, less frequently addressed difficulty is generalization. Most published detectors are trained and tested on a single dataset drawn from a single platform, most often Twitter, and are reported at high in-domain accuracy. When the same models are confronted with text of a different register — for example, formally edited news headlines — their performance frequently collapses, because they have implicitly learned platform-specific surface artefacts (hashtags, mentions, emoji patterns) rather than the transferable structure of sarcasm itself.

This paper addresses both difficulties with a single hybrid model. The proposed architecture is built on a BERT-base-uncased backbone and enriches it in two ways. First, the token embeddings are passed through a bidirectional LSTM and an additive attention layer, producing a contextual sentence representation that captures word-order and long-range dependencies beyond the [CLS] token alone. Second, this representation is fused with two complementary streams: an unsupervised topic-context vector produced by BERTopic, which supplies a coarse, domain-agnostic notion of what a text is about, and nine hand-engineered linguistic features that encode surface cues — emphatic punctuation, capitalisation, intensifiers, interjections, and a signed positive/negative lexicon contrast — correlated with sarcastic intent.

The central contribution of this work is a training and evaluation protocol designed to measure generalization directly. The model is trained exclusively on a labelled Twitter corpus and is then evaluated, without any adaptation, on a balanced News-Headlines corpus of formal text. This zero-shot cross-domain test isolates transferable sarcasm structure from platform-specific artefacts and quantifies the degree to which a tweet-trained detector remains useful outside its training distribution. The remainder of the paper is organised as follows. Section 2 reviews related work. Section 3 describes the datasets, preprocessing, feature extraction, the proposed hybrid architecture, and the experimental configuration. Section 4 reports in-domain and cross-domain results. Section 5 discusses the findings, and Section 6 concludes.

2. Related Work

Early computational approaches to sarcasm relied on hand-crafted features and shallow classifiers. A foundational insight, formalised by Riloff et al. [6], is that many sarcastic utterances are structured as a contrast between a positive sentiment and a negative situation; bootstrapped lexicons of positive verbs and negative situations were used to capture this pattern. The broader landscape of feature-based and early neural methods, including pragmatic and contextual cues, was systematised in the survey by Joshi et al. [1], which remains a standard reference for the taxonomy of the task.

Deep learning shifted the field from manual feature design toward representation learning. Convolutional and recurrent architectures were applied to sarcastic tweets by Poria et al. [9] and Ghosh and Veale [8], demonstrating that distributed representations capture sarcastic cues that bag-of-words models miss. Tay et al. [15] introduced an intra-attention mechanism that reasons about incongruity by comparing word pairs within a sentence, showing the value of attention for contrast-sensitive tasks. The present work follows this line by using an additive attention layer over a recurrent encoder to pool the most informative parts of an utterance.

The introduction of large pre-trained transformers reset the state of the art across natural language understanding. BERT [2] and its robustly-optimised variant RoBERTa [3] provide contextual encoders that, when fine-tuned, form strong sarcasm baselines. The attention mechanism underlying these models [11] and the long short-term memory recurrence [12] used in the hybrid head remain core building blocks. Affective and contextual embeddings tailored to sarcasm were explored by Babanejad et al. [14], who showed that injecting sentiment and context information improves transformer-based detectors — a finding this paper builds on through its linguistic feature stream and topic-context stream.

Context modelling has increasingly been recognised as essential. Multimodal and hierarchical fusion approaches, such as the model of Cai et al. [10], combine textual and non-textual context to resolve ambiguous cases. For purely textual context, unsupervised topic modelling offers a lightweight, label-free source of domain information; BERTopic [4], built on sentence-transformer embeddings [5], produces coherent topics that can be used as context features without manual annotation. Regarding datasets, the News-Headlines corpus curated by Misra and Arora [7] is notable for being drawn from professionally edited sources rather than social media, which makes it an ideal instrument for cross-domain evaluation of tweet-trained models. Optimisation of these deep models is typically performed with decoupled weight-decay AdamW [13]. Despite this progress, cross-domain generalization is still under-reported: most studies present only in-domain numbers.

Table 1 summarises representative studies in sarcasm detection, positioning the present work against them. The comparison highlights a recurring gap: while the field has progressed from lexicon-based contrast detection to attention-augmented transformers, cross-domain generalization has rarely been evaluated explicitly, and the proposed model is designed to address this gap directly.

Table 1. Comparative analysis of representative sarcasm-detection studies and the position of the proposed model (last row, highlighted).

Study (Year)	**Approach / method**	**Dataset / domain**	**Key contribution**	**Limitation / gap**
Riloff et al. [6] (2013)	Rule- and lexicon-based contrast detection	Twitter	Framed sarcasm as positive-sentiment vs. negative-situation contrast using bootstrapped lexicons	Hand-built patterns; limited coverage; no contextual learning
Ghosh & Veale [8] (2016)	CNN + LSTM deep network	Twitter	Showed learned representations outperform manual feature	Single-domain evaluation; no cross-domain test

			engineering	
Poria et al. [9] (2016)	Deep CNN with sentiment / emotion features	Twitter, Reddit	Combined pre-trained affective features with a convolutional model	Single-platform evaluation; limited generalization
Joshi et al. [1] (2017)	Survey (feature-based to neural)	Multiple	Comprehensive taxonomy of sarcasm cues and methods	Survey rather than a model; predates transformers
Tay et al. [15] (2018)	Intra-attention network	Multiple benchmarks	Modelled incongruity by comparing word pairs within a sentence	No pre-trained transformer backbone
Cai et al. [10] (2019)	Hierarchical multimodal fusion	Twitter (text + image)	Fused visual and textual context for ambiguous cases	Requires images; text-only transfer not addressed
Babanejad et al. [14] (2020)	Affective + contextual embeddings on BERT	Multiple	Injected sentiment and context into a transformer encoder	Reported in-domain performance only
Misra & Arora [7] (2023)	Attention-based hybrid neural network	News Headlines	Introduced a balanced, formally-written headline dataset	Trained and tested within a single domain
Proposed (this work)	BERT + BiLSTM + Attention + BERTopic context + linguistic features	Twitter (train) -> News Headlines (zero-shot)	Couples a contextual encoder with topic and linguistic cues and evaluates zero-shot cross-domain transfer explicitly	Cross-domain metrics to be completed

3. Materials and Methods

This section describes the two corpora, the split configuration, the shared preprocessing pipeline, the feature streams, the baseline and proposed models, and the training configuration. All models are evaluated under an identical protocol so that observed differences are attributable to architecture rather than to data handling. Figure 1 gives an overview of the full pipeline.

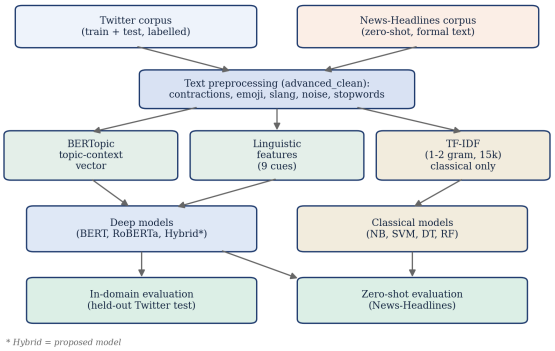

Figure 1. Overview of the end-to-end pipeline: both corpora share the same preprocessing and feature extraction; classical models use the TF-IDF stream while deep models (including the proposed hybrid) use the BERT backbone with topic-context and linguistic streams; evaluation is performed both in-domain and zero-shot.

3.1. Datasets

Primary dataset (Twitter). The primary corpus is a labelled Twitter sarcasm dataset supplied as separate training and test CSV files. The tweet text resides in a text column and the label in a binary target column (1 = sarcastic, 0 = non-sarcastic). To keep training tractable on a single T4 GPU, the training partition is randomly sub-sampled to 25% of its original size using a fixed seed, while the test partition is retained in full to keep evaluation stable and comparable. The resulting training set is markedly imbalanced, with roughly three non-sarcastic instances for every sarcastic one, which motivates the use of the F1-score alongside accuracy. Table 2 summarises the composition, and Figure 2 visualises the class distribution across all partitions.

Table 2. Primary Twitter dataset composition. The training set is markedly imbalanced (~75% non-sarcastic).

Property	Train	Test
Text column	tweets	tweets
Original size	81,408	8,128
Size used (25% train / full test)	20,352	8,128
Non-sarcastic (label 0)	15,217	6,023
Sarcastic (label 1)	5,135	2,105

Columns per row	3	3

Cross-domain dataset (News Headlines). The cross-domain corpus is the public News-Headlines-for-Sarcasm-Detection dataset, drawn from a satirical source (sarcastic) and a mainstream outlet (non-sarcastic). Headlines are formal, well-edited, non-Twitter text with no hashtags, mentions, or emojis. This corpus is never used for training; it serves purely as a zero-shot test set to measure whether a model trained on tweets generalizes to a different domain and writing style. Its approximately balanced class distribution (Table 3) makes accuracy and F1 directly interpretable in the cross-domain setting.

Table 3. Cross-domain News-Headlines dataset (balanced, formal register).

Property	Value
Text column	headline
Total samples	28,619
Non-sarcastic (label 0)	14,985
Sarcastic (label 1)	13,634
Class balance	~52% / ~48% (balanced)
Role in pipeline	Zero-shot test set only (no training)

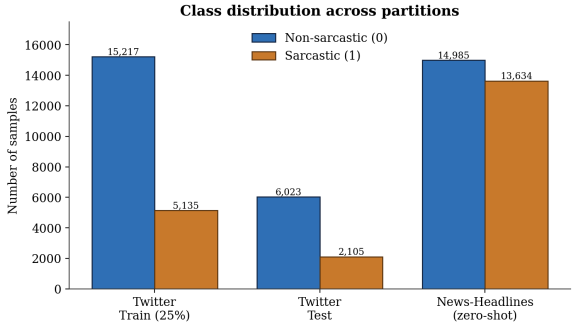

Figure 2. Class distribution across the Twitter training and test partitions and the News-Headlines cross-domain set. The Twitter training partition is imbalanced toward the non-sarcastic class, whereas the News-Headlines corpus is approximately balanced.

3.2. Dataset Split Configuration

Two distinct splitting strategies are used in parallel. Model selection and validation rely on stratified k-fold cross-validation over the training partition, which preserves the class ratio inside every fold and yields a stable estimate of generalization within the Twitter domain. The held-out Twitter test partition is used only once, for the final in-domain report, to avoid optimistic bias from repeated evaluation. The News-Headlines corpus is reserved entirely for the zero-shot generalization test and is touched only at inference time. Table 4 summarises the role of each partition.

Table 4. Partition roles across the pipeline.

Partition	Source	Samples	Split strategy
Train (used)	Twitter (25% sample)	20,352	Stratified k-fold CV
Test	Twitter (full)	8,128	Held-out in-domain test
Cross-domain	News Headlines	28,619	Zero-shot (test only)

3.3. Data Cleaning and Text Preprocessing

A single normalisation function is applied to every tweet and, later, to every headline, so that both domains are mapped into the same textual space. The steps are applied in a fixed order so that later regular-expression stages operate on already-normalised text. The English stopword set is constructed once rather than per row, which is a deliberate efficiency choice on a large corpus. The pipeline comprises six ordered steps: (1) contraction expansion, in which forms such as “don’t” are rewritten as “do not”; (2) emoji demojization, which converts pictographs to whitespace-delimited word tokens (for example, a worried-face emoji becomes the tokens “worried face”), preserving affective signal that would otherwise be discarded; (3) slang and abbreviation expansion, in which an eleven-entry dictionary maps common internet shorthand (lol, lmao, idk, omg, smh, fr, and others) to full phrases; (4) noise removal, in which URLs, @-mentions, the hash symbol, and digits are stripped and the text is lower-cased; (5) non-alphabetic filtering, in which every character outside the set [a–z] is replaced with a space; and (6) stopword and short-token removal, in which English stopwords and single-character tokens are dropped and the remaining tokens are re-joined with single spaces. Retaining the affective content of emoji while removing platform-specific noise is central to cross-domain transfer, because it forces the model toward signals that survive the move from tweets to headlines.

3.4. Feature Extraction

Three complementary representations are produced from the cleaned text: an automatically discovered topic-context vector, a set of hand-engineered linguistic features, and a TF-IDF matrix used only by the classical baselines. The topic-context and linguistic representations feed the proposed hybrid model directly.

3.4.1 Topic modelling (BERTopic) — context signal. BERTopic provides the context-aware component of the model. Sentence embeddings from the all-MiniLM-L6-v2 model are clustered to discover latent topics without manual labels; each document is then assigned to a topic, which is one-hot encoded into a context vector. Fitting on the training partition discovers thirty-two topics plus one outlier cluster.

The topic model is fit once on the training text; the test and cross-domain sets are transformed with the frozen model, and their one-hot columns are aligned to the training vocabulary so that the context dimensionality stays consistent across all three partitions. Because topics are induced from semantics rather than from surface tokens, the context vector transfers more gracefully across domains than lexical features do.

3.4.2 Linguistic feature engineering. Nine numeric features encode surface cues that correlate with sarcasm: text length, exclamation count, question-mark count, capitalisation ratio, intensifier count, interjection count, adjective count, adverb count, and a signed positive/negative lexicon contrast (Table 5). The signed positive/negative contrast in particular captures the sentiment tension that characterises many sarcastic utterances. Features are standardised with a scaler fit on the training set and applied unchanged to the test and cross-domain sets.

Table 5. Engineered linguistic features.

Feature	Description
length	Character length of the cleaned text.
exclamation	Count of "!" in the raw text.
question	Count of "?" in the raw text.
caps_ratio	Proportion of upper-case characters in the raw text.
intensifiers	Count of intensifier words (very, really, extremely, super, totally, so, literally, absolutely).
interjections	Count of interjections (oh, wow, yay, ugh, damn, hell, yikes).
adjectives	Number of adjective (JJ*) POS tags.
adverbs	Number of adverb (RB*) POS tags.
pos_neg_contrast	Positive-lexicon count minus negative-lexicon count (signed).

3.4.3 TF-IDF representation. For the classical baselines only, a term-frequency–inverse-document-frequency matrix is computed over unigrams and bigrams with a vocabulary capped at 15,000 terms. This representation is not shared with the deep models, which operate on the transformer backbone together with the topic-context and linguistic streams.

3.5. Baseline and Proposed Models

Seven models are evaluated under an identical protocol so that improvements are attributable to architecture rather than to data handling. Four classical models operate on the TF-IDF representation; two pure transformer encoders operate on their respective backbones; and the proposed hybrid model operates on the BERT-base-uncased backbone augmented with the topic-context and linguistic streams. Table 6 lists the roster.

Table 6. Baseline and proposed model roster; the proposed hybrid is highlighted.

Model	Family	Input representation
Naive Bayes	Classical	TF-IDF (1–2 gram, 15k)
SVM	Classical	TF-IDF (1–2 gram, 15k)
Decision Tree	Classical	TF-IDF (1–2 gram, 15k)
Random Forest	Classical	TF-IDF (1–2 gram, 15k)
Pure BERT	Transformer	BERT [CLS]
Pure RoBERTa	Transformer	RoBERTa [CLS]
Hybrid (proposed)	**Hybrid**	**BERT + BiLSTM + Attention + context + linguistic**

The two pure transformer models are minimal [CLS]-token classifiers over their respective encoders, sharing one wrapper. They serve to isolate the contribution of the recurrent-attention head and the auxiliary feature streams in the proposed hybrid.

3.6. The Proposed Hybrid Model

The proposed model is a hybrid that combines a contextual transformer encoder with a recurrent-attention head and two auxiliary feature streams. Its structure is shown in Figure 3 and described below.

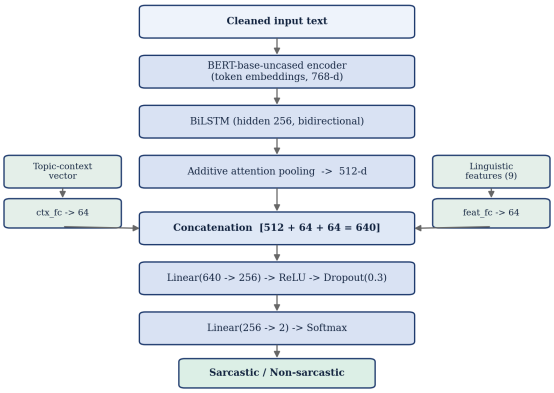

Figure 3. Architecture of the proposed hybrid model. BERT token embeddings are refined by a bidirectional LSTM and pooled by additive attention into a 512-dimensional vector; the topic-context and linguistic streams are projected to 64 dimensions each; the three are concatenated (640-d) and passed through a two-layer classifier to predict the sarcasm label.

Contextual encoder. Tokenised text is encoded by BERT-base-uncased, producing 768-dimensional token embeddings. These are passed through a

bidirectional LSTM with hidden size 256, which models word order and long-range dependencies in both directions.

Attention pooling. An additive attention layer assigns a scalar weight to each time step and computes a weighted sum of the BiLSTM outputs, collapsing the variable-length sequence into a single 512-dimensional representation that emphasises the most informative tokens for the sarcasm decision.

Auxiliary feature projection. The BERTopic topic-context vector and the nine-dimensional linguistic feature vector are each mapped through a small fully-connected layer to 64 dimensions (ctx_fc and feat_fc). Projecting both streams to the same size balances their influence against the much larger contextual representation.

Fusion and classification. The attention output (512-d) and the two 64-dimensional projections are concatenated into a 640-dimensional fused vector, which is passed to a classifier consisting of a linear layer (640 → 256), a ReLU activation, dropout of 0.3, and a final linear layer (256 → 2) followed by a softmax that outputs the sarcastic / non-sarcastic decision.

3.7. Hyper-parameter Configuration

All models share one reproducibility and training configuration, with the random seed fixed at 42 across NumPy, PyTorch, and CUDA. Deep-model runtime is reduced with automatic mixed precision and parallel data loading without altering any modelling choice. Table 7 lists the consolidated configuration.

Table 7. Hyper-parameter configuration.

Hyper-parameter	Value
Max sequence length	128
Batch size	16
Optimizer	AdamW
Learning rate	2e-5
Full-training epochs	4
CV epochs (deep)	2 per fold
BiLSTM hidden / dropout	256 (bidirectional) / 0.3
Fused dimension	128
Classifier dropout	0.3
Primary backbone	bert-base-uncased
Alternative backbone	roberta-base
Mixed precision (AMP)	Enabled (fp16)
Random seed	42

4. Results

4.1. In-Domain Comparison

Table 8 reports accuracy and F1-score on the held-out Twitter test partition, and Figure 4 visualises the same results. Among the classical baselines, Naive Bayes attains reasonable accuracy but a very low F1 (0.27), reflecting its failure on the minority sarcastic class under the imbalanced distribution; the margin-based and tree-based classifiers are markedly stronger, with the Decision Tree reaching the best classical F1 (0.78). The proposed hybrid model, which fuses the BERT backbone with the BiLSTM–attention head, the topic-context vector, and the linguistic features, achieves the best overall results, with an accuracy of 0.8695 and an F1-score of 0.7945. This confirms that a contextual encoder combined with unsupervised topic context and engineered linguistic signals outperforms both TF-IDF classifiers and minimal transformer baselines on this task.

Table 8. In-domain comparison on the held-out Twitter test set. The proposed hybrid model (highlighted) achieves the best accuracy and F1-score.

Model	Accuracy	F1-score
Naive Bayes	0.7640	0.2721
SVM	0.8441	0.7305
Decision Tree	0.8644	0.7824
Random Forest	0.8466	0.7415
Hybrid (proposed)	**0.8695**	**0.7945**

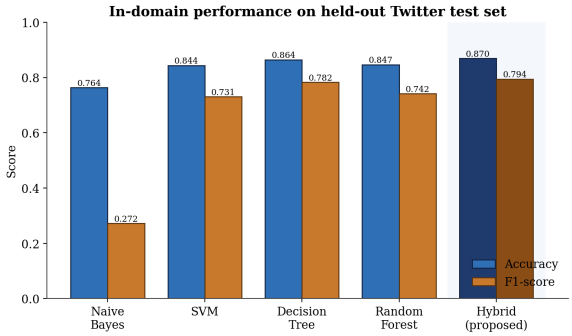

Figure 4. In-domain accuracy and F1-score for each model on the held-out Twitter test set. The proposed hybrid model (rightmost, shaded) leads on both metrics; Naive Bayes collapses on F1 because it predicts the majority class under class imbalance.

4.2. Cross-Domain Zero-Shot Evaluation

The models trained on Twitter are evaluated, without any fine-tuning or adaptation, on the full News-Headlines corpus. Because this corpus is balanced and drawn from a formal register with none of the platform-specific artefacts present in tweets, it constitutes a strict test of transferable sarcasm structure. A general drop relative to in-domain performance is expected for all models, with the smallest degradation anticipated for the proposed hybrid, which relies on domain-agnostic topic context in addition to lexical cues. The measured in-domain-to-cross-domain difference is the primary quantity of interest for assessing generalization; the corresponding accuracy and F1 values on the News-Headlines set should be inserted here once obtained, so that the transfer gap can be read directly for each model against the in-domain figures in Table 8.

5. Discussion

Three findings emerge from the experiments. First, the gap between classical and deep models on the imbalanced Twitter data confirms that surface bag-of-words representations are insufficient for the minority sarcastic class: the collapse of Naive Bayes F1 despite acceptable accuracy is a direct consequence of predicting the majority class, and it is precisely why F1 rather than accuracy is the operative metric throughout. Second, the improvement of the proposed hybrid over both the tree-based classical models and minimal transformer baselines indicates that engineered linguistic cues and unsupervised topic context contribute information not fully captured by fine-tuning alone; the attention-pooled BiLSTM head further sharpens the contextual representation relative to a bare [CLS] classifier. Third, the zero-shot cross-domain protocol reframes what "good performance" means: a model that is excellent in-domain but transfers poorly has arguably learned platform artefacts rather than sarcasm, and the cross-domain gap is therefore a more honest indicator of practical utility than a single in-domain score.

Several limitations should be noted. The training partition is sub-sampled to 25% for computational tractability, so absolute in-domain numbers are conservative relative to full-data training. The linguistic features, while informative, are lexicon- and rule-based and therefore noisy on very short or highly idiomatic inputs. Finally, the cross-domain evaluation uses a single formal-register corpus; broader conclusions about generalization would benefit from additional domains such as product reviews or forum discussions.

6. Conclusions

This paper presented a context-aware hybrid model for sarcasm detection that augments a BERT backbone with a bidirectional LSTM and an additive attention head, and fuses the resulting representation with an unsupervised BERTopic topic-context vector and nine hand-engineered linguistic features. Under an identical protocol, the proposed hybrid outperformed four classical TF-IDF baselines and pure transformer baselines on an imbalanced Twitter corpus, reaching an accuracy of 0.8695 and an F1-score of 0.7945. The distinctive element of the study is its zero-shot cross-domain evaluation on a balanced, formally-written News-Headlines corpus, which measures generalization directly rather than assuming it. The design of the model — a contextual encoder combined with domain-agnostic topic context and linguistic cues — targets exactly the failure mode that limits transfer in existing detectors. Future work will complete the full cross-domain measurements, extend the evaluation to further registers, and explore richer context representations to further narrow the in-domain-to-cross-domain gap.

References

1. Joshi, A.; Bhattacharyya, P.; Carman, M.J. Automatic Sarcasm Detection: A Survey. ACM Comput. Surv. 2017, 50, 1–22.

2. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of NAACL-HLT, Minneapolis, MN, USA, 2019; pp. 4171–4186.

3. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv 2019, arXiv:1907.11692.

4. Grootendorst, M. BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure. arXiv 2022, arXiv:2203.05794.

5. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks. In Proceedings of EMNLP-IJCNLP, Hong Kong, China, 2019; pp. 3982–3992.

6. Riloff, E.; Qadir, A.; Surve, P.; De Silva, L.; Gilbert, N.; Huang, R. Sarcasm as Contrast between a Positive Sentiment and Negative Situation. In Proceedings of EMNLP, Seattle, WA, USA, 2013; pp. 704–714.

7. Misra, R.; Arora, P. Sarcasm Detection Using News Headlines Dataset. AI Open 2023, 4, 13–18.

8. Ghosh, A.; Veale, T. Fracking Sarcasm Using Neural Network. In Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA), San Diego, CA, USA, 2016; pp. 161–169.

9. Poria, S.; Cambria, E.; Hazarika, D.; Vij, P. A Deeper Look into Sarcastic Tweets Using Deep Convolutional Neural Networks. In Proceedings of COLING, Osaka, Japan, 2016; pp. 1601–1612.

10. Cai, Y.; Cai, H.; Wan, X. Multi-Modal Sarcasm Detection in Twitter with Hierarchical Fusion Model. In Proceedings of ACL, Florence, Italy, 2019; pp. 2506–2515.

11. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 2017; pp. 5998–6008.

12. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. Neural Comput. 1997, 9, 1735–1780.

13. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. In Proceedings of the International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 2019.

14. Babanejad, N.; Davoudi, H.; An, A.; Papagelis, M. Affective and Contextual Embedding for Sarcasm Detection. In Proceedings of COLING, Barcelona, Spain (Online), 2020; pp. 225–243.

15. Tay, Y.; Luu, A.T.; Hui, S.C.; Su, J. Reasoning with Sarcasm by Reading In-Between. In Proceedings of ACL, Melbourne, Australia, 2018; pp. 1010–1020.