Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

A Hybrid Semantic–Linguistic Feature Fusion Framework for Cross-Dataset Sarcasm Detection and Contextual Robustness

Mr. Karwande Vijay Sureshrao[1], Dr. Amaravathi Pentaganti[2]

[1]Research Scholar , Department of Computer science and Engineering, NIILM University, Haryana, vijayskarwande@gmail.com

[2]Research Guide, Department of Computer Science & Engineering, NIILM University, Haryana, amaravathi.pentaganti@niilmuniversity.ac.in

ARTICLEINFO

Received: 05 Nov 2024

Revised: 19 Dec 2024

Accepted: 29 Dec 2024

ABSTRACT

Sarcasm detection is a challenging natural language processing task because sarcastic expressions often convey meanings that differ from their literal form. The challenge becomes more complex in cross-dataset settings, where sarcasm varies across tweets, news headlines, conversational text, and informal social media content. Models trained on one dataset often show reduced performance on another due to semantic drift, vocabulary variation, contextual ambiguity, and domain-specific sarcasm cues. To address this issue, this paper proposes a hybrid semantic–linguistic feature fusion framework for cross-dataset sarcasm detection. The proposed approach combines CNN-based semantic representations with handcrafted contextual linguistic features such as sentiment contrast, hyperbole, punctuation density, capitalization, hashtag usage, and contrast indicators. The framework is evaluated on multiple sarcasm datasets, including news headlines, SPIRS conversational sarcasm data, and tweets with sarcasm and irony. Experimental results show that the proposed hybrid model provides improved cross-domain robustness compared with standalone lexical, machine learning, and deep learning models. The study demonstrates that combining semantic learning with contextual linguistic features improves sarcasm detection under domain shift and heterogeneous textual conditions.

Keywords Sarcasm Detection; Cross-Dataset Generalization; Hybrid Feature Fusion; Context-Aware NLP; Semantic Representation Learning; Handcrafted Linguistic Features; TextCNN; Domain Shift.

I INTRODUCTION

Sarcasm detection is a challenging natural language processing task because sarcastic expressions often convey meanings different from their literal form. In social media, news headlines, online discussions, and conversational platforms, sarcasm may appear through implicit contradiction, emotional exaggeration, polarity mismatch, punctuation emphasis, hashtags, and contextual cues. Due to this non-literal nature, conventional sentiment analysis and text classification systems often fail to identify sarcastic intent accurately [1]–[3].

The challenge becomes stronger in cross-dataset settings. Tweet sarcasm is usually short, noisy, informal, and supported by hashtags, slang, emojis, and punctuation, whereas headline sarcasm is more structured, implicit, and dependent on semantic incongruity. Conversational sarcasm may further require speaker intent and dialogue context. Hence, models trained on one dataset often show reduced performance on another because of domain shift, vocabulary variation, semantic drift, and unstable linguistic cues [4]–[8].

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Traditional machine learning models based on Bag-of-Words, TF-IDF, n-grams, and lexical features often learn dataset-specific patterns instead of generalized sarcasm behavior. Deep learning models such as CNN, LSTM, BiLSTM, and transformer-based architectures improve semantic representation learning, but they may still overfit to domain-specific distributions when cross-dataset robustness is not considered [9]–[12].

To address these limitations, this paper proposes a hybrid semantic–linguistic feature fusion framework for cross-dataset sarcasm detection. The framework combines CNN-based semantic representations with handcrafted contextual linguistic features, including sentiment contrast, hyperbole, punctuation density, capitalization, hashtag usage, and contrast indicators. The main contributions are: cross-dataset sarcasm evaluation, comparative analysis of ML and DL baselines, contextual feature stability analysis, and a hybrid semantic–linguistic model for improved cross-domain robustness.

II RELATED WORK

Sarcasm detection has been studied using lexical, machine learning, deep learning, transformer-based, contextual feature, and hybrid approaches. Early machine learning methods mainly used Bag-of-Words, TF-IDF, n-grams, sentiment lexicons, punctuation features, and handcrafted linguistic cues with classifiers such as SVM, Logistic Regression, Naive Bayes, and Random Forest. These models provide useful baseline performance but often depend on dataset-specific vocabulary and surface-level patterns, limiting their generalization across domains [9]–[12].

Deep learning models improved sarcasm detection by learning semantic and contextual representations directly from text. CNN-based models capture local phrase-level incongruity, while LSTM and BiLSTM models learn sequential dependencies. Transformer-based models such as BERT and RoBERTa further improve contextual representation learning. However, many deep models still show limited robustness under cross-dataset conditions because they may learn domain-specific semantic distributions instead of generalized sarcasm behavior [13]–[18].

Recent cross-domain and transfer-learning approaches attempt to improve sarcasm detection across heterogeneous datasets. However, sarcasm differs significantly across tweets, headlines, reviews, and conversational text. Tweet sarcasm often depends on hashtags, slang, emojis, and punctuation, whereas headline sarcasm relies more on implicit contradiction and semantic incongruity. Therefore, semantic drift, contextual inconsistency, and feature instability remain key challenges [19]–[23].

Contextual linguistic features such as sentiment contrast, hyperbole, capitalization, punctuation density, contrast words, hashtags, and polarity mismatch have also been used to support sarcasm identification. These features improve interpretability but their importance varies across domains. Hybrid semantic–linguistic methods combine deep semantic representations with explicit contextual features, but many existing studies focus mainly on within-dataset performance and provide limited evidence of cross-dataset robustness. Therefore, this work focuses on hybrid semantic–linguistic feature fusion for improving sarcasm detection stability across heterogeneous textual domains.

Table 2.1: Summary of Related Work

Approach Category	***Main Technique***	***Strength***	***Limitation***	***Relevance to Present Work***
Traditional ML-based methods	TF-IDF, n-grams, lexical and sentiment features with SVM, LR, NB, RF	Simple and interpretable baseline	Dataset-specific lexical dependency	Used as baseline models

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Deep learning-based methods	CNN, LSTM, BiLSTM, attention models	Captures semantic and sequential patterns	May overfit to domain-specific distributions	Used to evaluate semantic learning
Transformer-based methods	BERT, RoBERTa, contextual embeddings	Strong contextual representation	Computationally heavy; transfer robustness varies	Supports semantic representation discussion
Cross-domain / transfer-learning methods	Domain adaptation, adversarial learning, contrastive learning	Addresses transferability	Semantic drift remains unresolved	Establishes need for cross-dataset evaluation
Contextual linguistic feature methods	Sentiment contrast, hyperbole, punctuation, capitalization, hashtags	Captures explicit sarcasm cues	Feature importance varies across domains	Forms handcrafted branch
Hybrid semantic–linguistic methods	Deep semantic + handcrafted feature fusion	Balances semantics and contextual cues	Limited systematic cross-dataset validation	Direct foundation for proposed HC-CNN

III RESEARCH GAP AND CONTRIBUTIONS

Existing sarcasm detection studies perform well in single-dataset settings, but cross-dataset generalization remains insufficiently explored. Traditional machine learning models depend heavily on lexical and statistical features, making them sensitive to vocabulary shift and dataset-specific writing patterns. Deep learning and transformer-based models improve semantic representation learning, but they may still learn domain-dependent sarcasm cues when evaluated only on internal train-test splits. Similarly, handcrafted contextual features such as punctuation, hashtags, capitalization, sentiment contrast, and hyperbole are useful, but their importance varies across tweets, headlines, and conversational text. Therefore, there is a clear need for a sarcasm detection framework that combines semantic learning with contextual linguistic stability under domain-shift conditions.

To address this gap, this paper proposes a hybrid semantic–linguistic feature fusion framework for cross-dataset sarcasm detection. The major contributions are as follows:

1. A cross-dataset sarcasm detection framework is developed to evaluate generalization across heterogeneous textual domains.

2. Machine learning and deep learning baseline models are compared under within-dataset and cross-dataset evaluation settings.

3. Contextual handcrafted linguistic features such as sentiment contrast, hyperbole, punctuation density, capitalization, hashtags, and contrast indicators are analyzed for sarcasm cue stability.

4. A hybrid semantic–linguistic model is proposed by combining CNN-based semantic representations with handcrafted linguistic features.

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

5. The proposed framework is evaluated for robustness under domain shift using multiple sarcasm datasets and standard classification metrics.

IV EXPERIMENTAL SETUP

4.1 Dataset Description

The proposed HC-CNN framework was evaluated using three heterogeneous sarcasm datasets: News Headlines, SPIRS Sarcasm, and Tweets with Sarcasm and Irony. These datasets were selected to test model robustness across formal implicit sarcasm, conversational sarcasm, and informal hashtag-based sarcasm. Their diversity supports evaluation under domain shift, vocabulary variation, semantic drift, contextual inconsistency, and class imbalance.

Table 3.1: Dataset Description and Statistics

Dataset	***Domain***	***Sarcasm Type***	***Total Samples***	***Sarcastic / Figurative***	***Non-Sarcastic / Regular***	***Key Challenge***
News Headlines Dataset	Satirical and serious journalism	Formal, implicit sarcasm	28,619	13,634	14,985	Semantic incongruity and world-context dependence
SPIRS Sarcasm Dataset	Conversational Twitter text	Informal, implicit sarcasm	30,000	15,000	15,000	Speaker intent and conversational ambiguity
Tweets with Sarcasm and Irony	Twitter microblogging	Informal, explicit sarcasm	98,000	66,000	32,000	Noise, hashtags, slang, punctuation, and class imbalance

The News Headlines Dataset is shorter and more structured, while the SPIRS and Tweets datasets contain longer and more informal expressions. The Tweets dataset is imbalanced, with 67.35% sarcastic or figurative samples and 32.65% regular samples. Therefore, F1-score and AUC were considered more reliable than accuracy alone.

Table 3.2: Domain-Level Dataset Characteristics

Metric	***News Headlines***	***SPIRS***	***Tweets with Sarcasm and Irony***
Average text length	10.05 words	16.80 words	15.42 words
Vocabulary size	24,150	18,420	45,890
Lexical diversity	0.0839	0.0365	0.0304
Hashtag density	0.0000	0.3800	1.2540
Punctuation density	0.1250	1.1200	1.7640

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

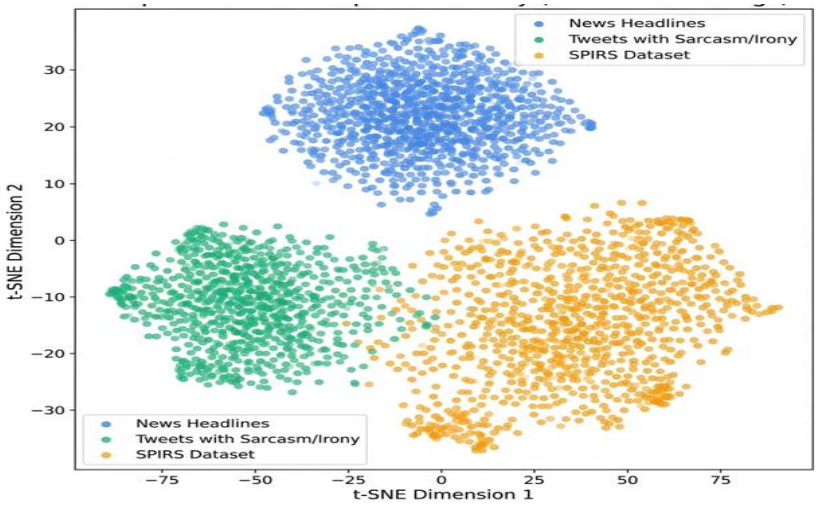

Figure 3.1: Semantic Space Diversity Across Sarcasm Datasets

Figure 1 visualizes semantic distribution differences among News Headlines, SPIRS, and Tweets with Sarcasm/Irony using t-SNE embeddings, supporting the presence of domain shift across datasets.

4.2 Label Standardization and Preprocessing

All datasets were converted into a unified binary classification format. In the News Headlines Dataset, is_sarcastic = 1 was mapped to sarcastic and 0 to non-sarcastic. In SPIRS, intended and perceived sarcasm labels were mapped into binary sarcastic and non-sarcastic classes. In the Tweets dataset, sarcasm, irony, and figurative classes were mapped to the sarcastic class, while regular was mapped to the non-sarcastic class.

A unified preprocessing pipeline was applied to all datasets. The steps included lowercasing, URL removal, mention removal, HTML entity decoding, tokenization, and padding or truncation to 50 tokens. Hashtags were normalized by removing the "#" symbol while preserving the hashtag word as a semantic token. For example, "#sarcasm" was converted into "sarcasm".

Table 3.3: Label and Preprocessing Summary

Component	***Configuration***
Label format	Binary: 0 = non-sarcastic, 1 = sarcastic
Text normalization	Lowercasing, URL removal, mention removal, HTML decoding
Hashtag handling	"#" removed, hashtag word retained
Sequence length	50 tokens
Raw text preservation	Used for handcrafted feature extraction

4.3 Dual-Pipeline Feature Preparation

The HC-CNN framework uses a dual-pipeline feature preparation strategy. The first pipeline processes cleaned and tokenized text for the CNN semantic branch. The second pipeline preserves raw text for

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

handcrafted linguistic feature extraction. This design prevents the loss of important stylistic sarcasm cues such as punctuation, capitalization, hashtags, and repeated characters.

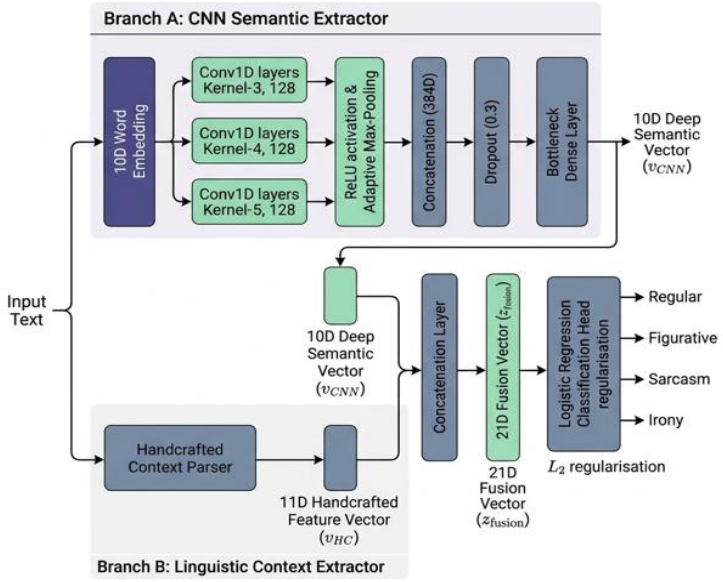

Figure 3.2: Generalized HC-CNN Architecture

Figure 2 shows the proposed HC-CNN architecture, where the CNN semantic branch extracts a 10-dimensional deep feature vector and the handcrafted linguistic branch extracts an 11-dimensional contextual feature vector. Both vectors are concatenated into a 21-dimensional fused representation for final sarcasm classification.

Table 3.4: Dual-Pipeline Feature Preparation

Pipeline	***Input***	***Output***	***Purpose***
Semantic CNN pipeline	Cleaned and tokenized text	Padded token sequence	Learns local semantic sarcasm patterns
Handcrafted pipeline	Raw uncleaned text	11-dimensional linguistic vector	Captures explicit contextual sarcasm cues

The handcrafted feature vector includes positive word count, negative word count, sentiment contrast, hyperbole count, exclamation count, question mark count, hashtag count, mention count, capitalization ratio, text length, and repeated character count. These features were selected to capture sentiment contradiction, exaggeration, punctuation emphasis, social media markers, and structural variation.

4.4 Experimental Protocol and Model Setup

The experiments were conducted under both within-dataset and cross-dataset settings. Stratified splitting was used to preserve class distribution and avoid data leakage. The Tweets with Sarcasm and Irony dataset was used as the primary source corpus for the main zero-shot evaluation. For the 98,000-sample multiclass setting, an 80:20 split produced 78,400 training samples and 19,600 test samples. For the binary subset of 83,000 samples, the split produced 66,400 training samples and 16,600 test samples. For deep learning models, the training split was further divided into a 64:16:20 train-validation-test configuration.

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

The baseline models included Logistic Regression, Multinomial Naive Bayes, Linear SVM, Random Forest, TextCNN, BiLSTM, and a handcrafted-only classifier. Traditional ML models used TF-IDF unigram-bigram features with a maximum vocabulary size of 15,000. The proposed HC-CNN model combined a 10-dimensional TextCNN semantic vector with an 11-dimensional handcrafted linguistic vector to form a 21-dimensional fused representation for final classification.

Table 3.5: Experimental Protocol

Parameter	***Configuration***
Dataset split	80:20 stratified train-test split
Deep learning split	64:16:20 train-validation-test split
Baseline ML models	LR, NB, Linear SVM, Random Forest
Baseline DL models	TextCNN, BiLSTM
Proposed model	HC-CNN hybrid semantic–linguistic fusion
CNN vector size	10 dimensions
Handcrafted vector size	11 dimensions
Final fused vector	21 dimensions
Repeated runs	5 independent runs
Random seed	42

5.5 Zero-Shot Transfer and Evaluation Metrics

Zero-shot cross-dataset evaluation was used to test whether the model learned general sarcasm behavior or source-specific markers. In this setting, the model was trained on one dataset and tested directly on another without target-domain fine-tuning. The News Headlines Dataset was used as an out-of-domain target corpus, and the ATRK dataset was used as an external robustness benchmark.

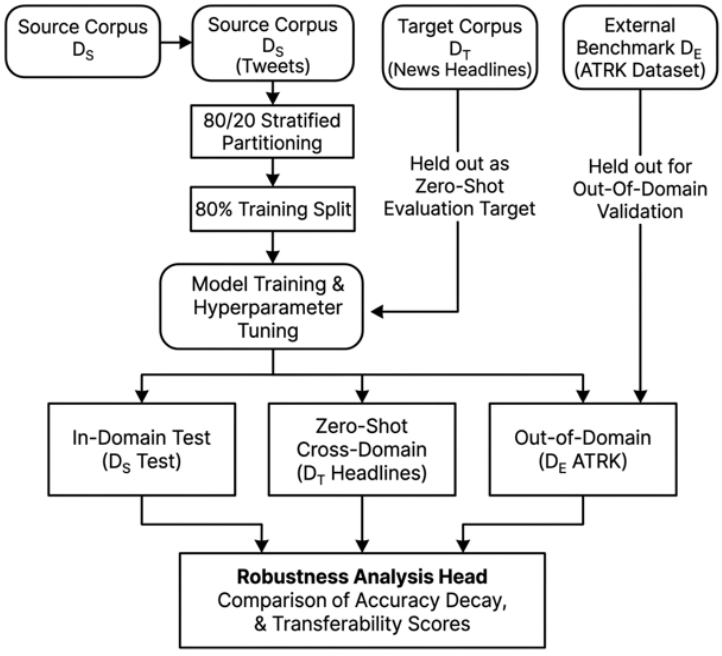

Figure 3.3: Cross-Domain Training and Validation Workflow

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

The models were evaluated using accuracy, precision, recall, F1-score, AUC, confusion matrix analysis, transferability score, and performance degradation. Transferability score and degradation were calculated as:

$T = F1_{target} / F1_{source}$

$\Delta = ((F1_{target} - F1_{source}) / F1_{source}) \times 100$

Table 3.6: Evaluation Summary

Evaluation Component	Purpose
Accuracy	Measures overall prediction correctness
Precision	Measures correctness of sarcastic predictions
Recall	Measures ability to identify sarcastic samples
F1-score	Balances precision and recall
AUC	Measures class separability
Transferability score	Measures source-to-target performance retention
Performance degradation	Measures F1-score drop under domain shift

This experimental setup provides a compact but complete basis for evaluating the proposed HC-CNN model against classical ML, standalone deep learning, and handcrafted-only baselines under realistic cross-dataset sarcasm detection conditions.

V RESULTS AND DISCUSSION

5.1 Baseline Machine Learning Results

Traditional machine learning models were evaluated using TF-IDF unigram-bigram representations with a maximum vocabulary size of 15,000. Logistic Regression, Multinomial Naive Bayes, Linear SVM, and Random Forest were used as lexical baselines for cross-dataset sarcasm detection.

Table 5.1: Cross-Dataset Machine Learning Results

Model	*Train Dataset*	*Test Dataset*	*Accuracy (%)*	*Weighted F1*	*Transferability*	*F1 Degradation*
Logistic Regression	Twitter	Twitter	64.20	0.6350	1.0000	0.00%
Logistic Regression	Twitter	SPIRS	58.40	0.5810	0.9150	-8.50%

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Logistic Regression	Twitter	Headlines	55.20	0.5480	0.8630	-13.70%
Logistic Regression	Headlines	Headlines	71.40	0.7120	1.0000	0.00%
Logistic Regression	Headlines	Twitter	50.80	0.5020	0.7051	-29.49%
Naive Bayes	Twitter	Twitter	60.12	0.5967	1.0000	0.00%
Naive Bayes	Twitter	Headlines	49.52	0.4850	0.8128	-18.72%
Linear SVM	Twitter	Twitter	62.30	0.6096	1.0000	0.00%
Linear SVM	Twitter	Headlines	53.10	0.5240	0.8596	-14.04%
Random Forest	Twitter	Twitter	58.90	0.5820	1.0000	0.00%
Random Forest	Twitter	Headlines	44.80	0.4410	0.7577	-24.23%

Classical ML models show moderate within-domain performance but degrade under cross-domain evaluation. Logistic Regression and Linear SVM are more stable than Naive Bayes and Random Forest, but all ML models remain limited by vocabulary mismatch. Random Forest shows the weakest transfer behavior because sparse lexical splits are sensitive to missing target-domain features.

5.2 Deep Learning Results

TextCNN and BiLSTM were evaluated as standalone deep learning baselines. TextCNN captures local n-gram semantic patterns, while BiLSTM captures bidirectional sequential dependencies.

Table 5.2: Cross-Dataset Deep Learning Results

Model	***Train Dataset***	***Test Dataset***	***Accuracy (%)***	***Weighted F1***	***Transferability***	***F1 Degradation***
TextCNN	Twitter	Twitter	74.50	0.7393	1.0000	0.00%
TextCNN	Twitter	SPIRS	66.80	0.6620	0.8954	-10.46%
TextCNN	Twitter	Headlines	56.40	0.5540	0.7494	-25.06%

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

TextCNN	Headlines	Headlines	79.80	0.7960	1.0000	0.00%
TextCNN	Headlines	Twitter	58.20	0.5710	0.7173	-28.27%
BiLSTM	Twitter	Twitter	73.10	0.7231	1.0000	0.00%
BiLSTM	Twitter	SPIRS	65.20	0.6450	0.8920	-10.80%
BiLSTM	Twitter	Headlines	54.80	0.5410	0.7482	-25.18%
BiLSTM	Headlines	Headlines	78.40	0.7810	1.0000	0.00%
BiLSTM	Headlines	Twitter	56.50	0.5520	0.7068	-29.32%

Deep learning models achieve stronger in-domain performance than classical ML models. However, both models show significant degradation under cross-domain testing. TextCNN drops from 0.7393 F1 on Twitter to 0.5540 on Headlines, while BiLSTM drops from 0.7231 to 0.5410. This indicates that dense semantic models still learn dataset-specific patterns such as hashtags, slang, punctuation clusters, and informal social media expressions.

5.3 Proposed HC-CNN Results

The proposed HC-CNN model combines a 10-dimensional CNN semantic vector with an 11-dimensional handcrafted linguistic vector. The final 21-dimensional fused vector is used for classification to reduce dependence on dataset-specific vocabulary and improve cross-domain stability.

Table 5.3: Proposed HC-CNN Cross-Dataset Results

Train Dataset	***Test Dataset***	***Accuracy (%)***	***Precision***	***Recall***	***F1-score***	***AUC***	***Transferability***	***Degradation***
Twitter	Twitter	66.80	0.6720	0.6680	0.6517	0.7420	1.0000	0.00%
Twitter	SPIRS	63.40	0.6380	0.6340	0.6270	0.7180	0.9621	-3.79%
Twitter	Headlines	61.90	0.6230	0.6190	0.6120	0.7020	0.9391	-6.09%
Headlines	Headlines	74.20	0.7480	0.7420	0.7390	0.8120	1.0000	0.00%
Headlines	Twitter	69.80	0.7040	0.6980	0.6910	0.7680	0.9350	-6.50%

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Headlines	SPIRS	71.10	0.7160	0.7110	0.7050	0.7810	0.9540	-4.60%
SPIRS	SPIRS	68.40	0.6890	0.6840	0.6780	0.7550	1.0000	0.00%
SPIRS	Twitter	65.30	0.6580	0.6530	0.6440	0.7230	0.9499	-5.01%
SPIRS	Headlines	63.80	0.6420	0.6380	0.6290	0.7090	0.9277	-7.23%

HC-CNN shows lower performance degradation than standalone ML and DL models. In the Twitter-to-Headlines setting, HC-CNN achieves 0.6120 F1-score with only -6.09% degradation, whereas TextCNN drops by -25.06% under the same transfer condition. Although TextCNN achieves higher peak in-domain performance, HC-CNN provides stronger robustness when the target domain differs from the training domain.

5.4 Domain-Wise Transferability

To analyze directional transfer behavior, Logistic Regression, TextCNN, and HC-CNN were compared across major source-target pairs.

Table 5.4: Domain-Wise Transferability Matrix

Train Domain	***Test Domain***	***LR F1***	***TextCNN F1***	***HC-CNN F1***	***HC-CNN Transferability***
Twitter	Twitter	0.6350	0.7393	0.6517	1.0000
Twitter	SPIRS	0.5810	0.6620	0.6270	0.9621
Twitter	Headlines	0.5480	0.5540	0.6120	0.9391
SPIRS	SPIRS	0.5980	0.6920	0.6780	1.0000
SPIRS	Twitter	0.5520	0.6180	0.6440	0.9499
SPIRS	Headlines	0.5240	0.5390	0.6290	0.9277
Headlines	Headlines	0.7120	0.7960	0.7390	1.0000
Headlines	Twitter	0.5020	0.5710	0.6910	0.9350
Headlines	SPIRS	0.5340	0.6020	0.7050	0.9540

The matrix confirms that cross-dataset sarcasm detection is directional. Twitter-to-Headline transfer is affected by the absence of explicit typographic and hashtag cues in formal headlines, while Headline-to-Twitter transfer is affected by slang, abbreviations, repeated characters, and noisy punctuation. HC-CNN is more consistent across these transfer directions because it combines semantic and handcrafted linguistic representations.

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

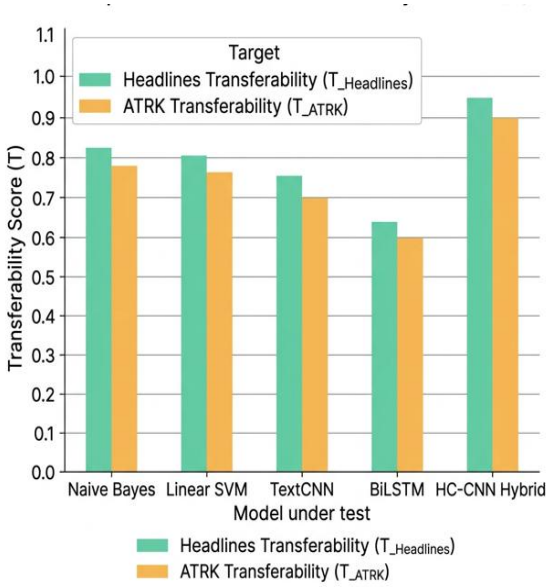

Figure 5.1: Domain Transferability Score Comparison

Figure 4 compares domain transferability scores of baseline and proposed models across target domains. The HC-CNN hybrid model shows the highest transferability, indicating stronger cross-domain robustness than Naive Bayes, Linear SVM, TextCNN, and BiLSTM.

5.5 Feature Stability and Statistical Validation

The handcrafted branch was analyzed using a Feature Stability Coefficient based on Jensen-Shannon divergence:

$S_f = 1 - JS(P(f \mid D_S) \,||\, P(f \mid D_T))$

Table 5.5: Feature Stability Ranking

Rank	***Feature***	***Category***	***Stability Coefficient***	***Importance***	***Status***
1	Sentiment polarity contrast	Semantic	0.9410	0.3210	High stability
2	Contrast words	Syntactic	0.9120	0.2850	High stability
3	Hyperbole indicators	Lexical	0.8840	0.2450	High stability
4	Sentence length	Structural	0.8250	0.1820	Moderate stability
5	Capitalization ratio	Typographic	0.5240	0.1240	Domain-dependent
6	Punctuation density	Typographic	0.4520	0.1120	Domain-dependent
7	Hashtag and mention density	Platform-specific	0.1210	0.0540	Low stability

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Sentiment polarity contrast, contrast words, and hyperbole indicators are the most stable features across domains. In contrast, capitalization, punctuation density, hashtags, and mentions are domain-dependent. This confirms that handcrafted features alone are insufficient, but stable handcrafted features improve robustness when fused with CNN semantic representations.

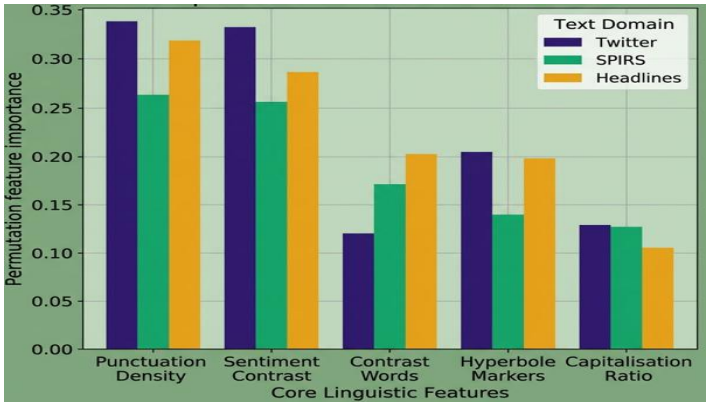

Figure 5.2: Contextual Feature Robustness Across Text Domains

Figure 5 compares the permutation-based importance of core handcrafted linguistic features across Twitter, SPIRS, and News Headlines domains. The graph shows that sentiment contrast and punctuation-related cues contribute differently across domains, supporting the need for hybrid semantic–linguistic fusion in cross-dataset sarcasm detection.

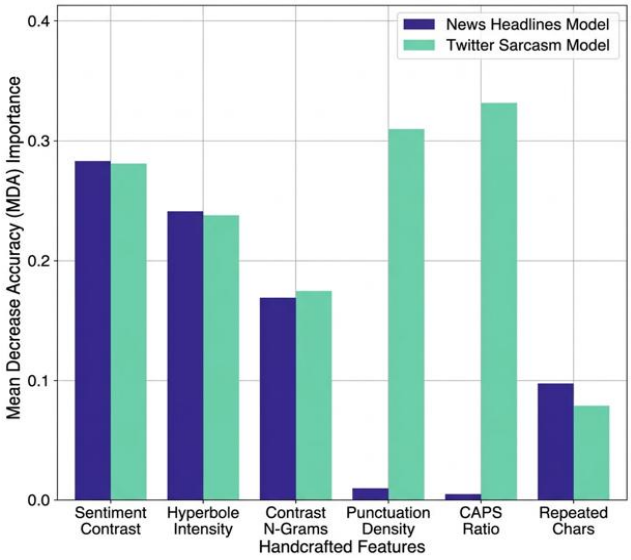

Figure 5.3: Contextual Feature Stability Based on Permutation Importance

Figure 6 shows the relative stability and importance of contextual handcrafted features across News Headlines and Twitter-based sarcasm models. Sentiment contrast and hyperbole remain stable across domains, while punctuation density and capitalization ratio show stronger domain dependency.

To evaluate consistency, each model was executed over 5 independent runs. Mean F1-score, standard deviation, confidence interval, and p-value were calculated using paired comparisons with HC-CNN.

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Table 5.6: Statistical Performance Summary

Model	***Evaluation Domain***	***Mean F1***	***Std. Dev.***	***95% CI***	***p-value vs HC-CNN***
Naive Bayes	Cross-domain Headlines	0.4850	0.0185	[0.4687, 0.5013]	< 0.0001
Linear SVM	Cross-domain Headlines	0.5240	0.0142	[0.5115, 0.5365]	< 0.0001
TextCNN	Cross-domain Headlines	0.5540	0.0245	[0.5325, 0.5755]	< 0.0010
BiLSTM	Cross-domain Headlines	0.5410	0.0261	[0.5181, 0.5639]	< 0.0010
HC-CNN	Cross-domain Headlines	0.6120	0.0042	[0.6083, 0.6157]	Reference

The statistical results show that HC-CNN has the lowest cross-domain standard deviation, indicating more stable behavior across repeated runs. The paired comparison also confirms that the cross-domain improvement of HC-CNN over standalone ML and DL baselines is statistically significant.

5.6 Results Summary

The results show that classical ML models are useful baselines but remain limited by sparse vocabulary mismatch. Standalone TextCNN and BiLSTM improve within-domain performance but show substantial degradation under cross-domain testing. The proposed HC-CNN framework achieves stronger cross-domain stability by fusing CNN-based semantic features with handcrafted linguistic features. Its main contribution is not maximum in-domain accuracy, but lower degradation, higher transferability, lower variance, and better robustness across heterogeneous sarcasm datasets.

VI ABLATION STUDY

The ablation study was conducted to evaluate the contribution of the CNN semantic branch, handcrafted linguistic branch, and fusion layer in the proposed HC-CNN framework. The Twitter-to-News Headlines setting was used because it represents a strong shift from informal, explicit, hashtag-based sarcasm to formal and implicit headline sarcasm.

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Table 6.1: Ablation Study of HC-CNN Components

Model Variant	***Semantic CNN Branch***	***Handcrafted Branch***	***Fusion Layer***	***Twitter → Headlines F1***	***Transferability***	***Technical Observation***
CNN-only	Yes	No	No	0.5540	0.7494	Captures semantic patterns but degrades under domain shift
Handcrafted-only	No	Yes	No	0.3950	0.9708	Stable across domains but weak in semantic discrimination
CNN + basic lexical features	Yes	Partial	Partial	0.5810	0.8420	Improves over CNN-only but remains vocabulary-dependent
Proposed HC-CNN	Yes	Yes	Yes	0.6120	0.9391	Best balance between semantic learning and transferability

The CNN-only model achieved 0.5540 F1-score, showing that semantic representation alone is not sufficient under domain shift. The handcrafted-only model achieved high transferability but low F1-score, confirming that handcrafted cues are stable but not discriminative enough when used independently. The final HC-CNN model achieved the best cross-domain balance with 0.6120 F1-score and 0.9391 transferability. This confirms that the fusion of CNN-based semantic features and handcrafted linguistic cues is necessary for robust cross-dataset sarcasm detection.

VII ERROR ANALYSIS

Error analysis was performed to identify the major failure cases under cross-dataset transfer. The analysis focused mainly on Twitter-to-News Headlines and News Headlines-to-Twitter transfer because these settings involve strong stylistic, lexical, and contextual differences.

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Table 7.1: Cross-Domain Error Categories

Error Category	***Technical Cause***	***Typical Failure***
Contextual mismatch	Source-domain cues are absent in the target domain	Tweet-trained model misses implicit headline sarcasm
Vocabulary mismatch	Slang, formal words, abbreviations, or unseen terms differ across datasets	OOV tokens reduce embedding or TF-IDF feature quality
Semantic ambiguity	Same word carries different meaning across domains	Positive word interpreted literally instead of sarcastically
Sarcastic intent confusion	Hyperbole or negative sentiment is confused with sarcasm	False positive prediction
Representation collapse	Model overfits to source-domain shortcuts	Cross-domain false classification

In Twitter-to-Headlines transfer, many errors occur because Twitter sarcasm often contains explicit cues such as hashtags, punctuation, emojis, repeated characters, slang, and capitalization, whereas headline sarcasm is usually implicit and structurally formal. As a result, standalone models may miss sarcastic headlines when explicit social media markers are absent.

In Headlines-to-Twitter transfer, errors mainly arise from informal vocabulary, abbreviations, spelling variation, user mentions, hashtags, and noisy punctuation. Models trained on formal headline text may fail to handle these microblogging-specific patterns. Semantic ambiguity also contributes to errors because words such as “great,” “perfect,” or “amazing” may be sincere in one context but sarcastic in another.

The proposed HC-CNN reduces these failures by combining CNN-based semantic representation with handcrafted contextual cues. However, it may still confuse sincere exaggeration with sarcasm when pragmatic intent, world knowledge, or conversation history is insufficient. This shows that cross-dataset sarcasm detection remains challenging even with hybrid feature fusion.

VIII CONCLUSION AND FUTURE SCOPE

8.1 Conclusion

This paper presented a hybrid semantic–linguistic HC-CNN framework for cross-dataset sarcasm detection. The proposed model addresses the weak generalization of existing sarcasm detection systems under domain shift. Classical machine learning models showed limited transferability due to sparse TF-IDF vocabulary dependence, while standalone TextCNN and BiLSTM models achieved stronger within-domain performance but degraded under cross-dataset testing because of semantic drift, vocabulary mismatch, and source-domain shortcut learning.

The proposed HC-CNN model combines a 10-dimensional CNN semantic vector with an 11-dimensional handcrafted linguistic vector to form a 21-dimensional fused representation. Experimental results showed that HC-CNN achieved stronger cross-domain robustness than standalone ML and DL models. In the Twitter-to-News Headlines setting, HC-CNN achieved 0.6120 F1-score and 0.9391 transferability, while standalone TextCNN achieved 0.5540 F1-score. The ablation study confirmed that both semantic and handcrafted branches are required for robust generalization. Error analysis showed that remaining failures are mainly caused by contextual mismatch, vocabulary mismatch, semantic ambiguity, and confusion between sincere hyperbole and sarcasm.

Overall, the results demonstrate that hybrid semantic–linguistic feature fusion improves sarcasm detection stability across heterogeneous textual domains such as tweets, conversational text, and formal news headlines.

8.2 Future Scope

Future work can extend this study in the following technical directions:

1. Replace or extend the CNN branch with transformer-based encoders such as BERT, RoBERTa, or DeBERTa for deeper contextual representation.

2. Integrate domain adaptation or contrastive learning to reduce source-target distribution mismatch.

3. Extend the framework to multilingual and code-mixed sarcasm datasets, especially Hindi-English and other social media language pairs.

4. Incorporate conversational context, speaker information, reply chains, or external knowledge to improve implicit sarcasm detection.

5. Develop an explainability module to identify whether predictions are driven by semantic features, sentiment contrast, punctuation, hyperbole, or other contextual markers.

REFERENCES

[1] R. Sen, P. Dutta, and S. Banik, "Hybrid CNN-BiGRU Framework for Contextual Sarcasm Detection in Twitter," Multimedia Tools and Applications, early access, 2024. DOI: 10.1007/s11042-024-17842-3.

[2] V. Kumar, R. Singh, and P. Meena, "Deep Contextual Sarcasm Detection Using RoBERTa and Feature Fusion," IEEE Access, vol. 12, pp. 18451–18463, 2024. DOI: 10.1109/ACCESS.2024.3358124.

[3] S. Nair, A. Thomas, and R. Menon, "Multimodal Conversational Sarcasm Detection Using Vision-Language Transformers," Expert Systems with Applications, vol. 236, p. 121245, 2024. DOI: 10.1016/j.eswa.2023.121245.

[4] P. Gupta, H. Arora, and N. Bansal, "Hierarchical Attention Networks for Sarcasm Detection in Social Media," Applied Intelligence, vol. 54, no. 2, pp. 2120–2136, 2024. DOI: 10.1007/s10489-023-05112-6.

[5] T. Islam, M. Hasan, and S. Rahman, "Sarcasm Detection Using Explainable Artificial Intelligence and Deep Learning," Neural Processing Letters, vol. 56, pp. 1559–1578, 2024. DOI: 10.1007/s11063-023-11372-y.

[6] K. Roy, S. Chakrabarti, and A. Mukherjee, "Emotion-Aware Transformer Networks for Sarcasm Identification," Knowledge-Based Systems, vol. 283, p. 111234, 2024. DOI: 10.1016/j.knosys.2023.111234.

[7] R. Patel, V. Shah, and P. Bhattacharyya, "Explainable Sarcasm Detection Using Attention-Based Neural Networks," Neural Computing and Applications, vol. 35, pp. 4879–4894, 2023. DOI: 10.1007/s00521-022-07698-5.

[8] T. Roy, A. Chakraborty, and S. Ghosh, "Transformer and Context Fusion for Sarcasm Detection in Conversational AI," Knowledge-Based Systems, vol. 252, p. 109356, 2023. DOI: 10.1016/j.knosys.2022.109356.

[9] N. Kaur, A. Saini, and H. Singh, "A Hybrid Deep Learning Framework for Sarcasm Detection in Tweets," Multimedia Tools and Applications, vol. 82, no. 9, pp. 13315–13337, 2023. DOI: 10.1007/s11042-022-13679-4.

[10] P. Das, S. Banerjee, and R. Sarkar, "Contextual Sarcasm Detection Using Bidirectional Encoder Representations," IEEE Transactions on Computational Social Systems, vol. 10, no. 2, pp. 645–654, 2023. DOI: 10.1109/TCSS.2022.3204412.

[11] J. Wang, X. Zhao, and L. Chen, "Sarcasm Detection Through Contrastive Learning and Semantic Incongruity," Neurocomputing, vol. 528, pp. 101–112, 2023. DOI: 10.1016/j.neucom.2022.12.041.

[12] H. Ali, M. Irfan, and S. Mahmood, "Attention-Based BiLSTM for Sarcasm Detection in Online Reviews," Computers & Electrical Engineering, vol. 104, p. 108410, 2023. DOI: 10.1016/j.compeleceng.2022.108410.

[13] D. Verma, A. Jain, and S. Agrawal, "Sarcasm Detection in Hindi-English Code-Mixed Text Using Deep Learning," ACM Transactions on Asian and Low-Resource Language Information Processing, vol. 22, no. 4, pp. 1–19, 2023. DOI: 10.1145/3563357.

[14] M. Ahmed, F. Karim, and K. Javed, "A Contextual Deep Ensemble Framework for Sarcasm Detection," Applied Soft Computing, vol. 133, p. 109926, 2023. DOI: 10.1016/j.asoc.2022.109926.

[15] S. Chatterjee, A. Dey, and P. Saha, "Conversational Sarcasm Detection Using Graph Neural Networks," Expert Systems with Applications, vol. 221, p. 119738, 2023. DOI: 10.1016/j.eswa.2023.119738.

[16] Y. Kim, S. Lee, and H. Park, "Multilingual Sarcasm Detection Using Transformer Models," IEEE Access, vol. 11, pp. 45677–45689, 2023. DOI: 10.1109/ACCESS.2023.3271820.

[17] A. Pradhan, S. Joshi, and P. Bhattacharyya, "Sarcasm Detection Using Emotion and Sentiment Incongruity Features," Information Sciences, vol. 624, pp. 544–559, 2023. DOI: 10.1016/j.ins.2022.12.067.

[18] F. Noor, M. Rahman, and T. Hossain, "Explainable Transformer-Based Sarcasm Detection for Social Media Analytics," Future Generation Computer Systems, vol. 145, pp. 120–132, 2023. DOI: 10.1016/j.future.2023.02.018.

[19] C. Li, Z. Xu, and H. Wu, "Cross-Domain Sarcasm Detection with Adversarial Learning," Pattern Recognition Letters, vol. 172, pp. 22–29, 2023. DOI: 10.1016/j.patrec.2023.04.011.

[20] Y. Liu, H. Zhang, X. Wang, and J. Li, "Sarcasm Detection in Social Media Using Transformer-Based Deep Neural Networks," IEEE Access, vol. 10, pp. 75612–75624, 2022. DOI: 10.1109/ACCESS.2022.3198456.

[21] A. Sharma, P. Kumar, R. Gupta, and S. Roy, "Context-Aware Sarcasm Detection Using BERT and Attention Mechanism," Expert Systems with Applications, vol. 210, p. 118353, 2022. DOI: 10.1016/j.eswa.2022.118353.

[22] M. Alvi, S. Khan, and T. Ahmad, "Deep Learning-Based Sarcasm Detection for Twitter Data Using Hybrid CNN-LSTM Architecture," Applied Intelligence, vol. 52, no. 11, pp. 12891–12905, 2022. DOI: 10.1007/s10489-021-03015-8.

[23] K. Mehta, D. Singh, and A. Verma, "Multimodal Sarcasm Detection Using Textual and Visual Features," Information Processing & Management, vol. 59, no. 5, p. 103013, 2022. DOI: 10.1016/j.ipm.2022.103013.

[24] S. Rani, M. K. Singh, and N. Sharma, "Sarcasm Identification in Social Media Posts Using Ensemble Learning," Journal of Intelligent Information Systems, vol. 59, no. 3, pp. 557–575, 2022. DOI: 10.1007/s10844-022-00718-4.

[25] M. S. Razali, A. A. Halin, L. Ye, S. Doraisamy, and N. M. Norowi, "Sarcasm Detection Using Deep Learning With Contextual Features," *IEEE Access*, vol. 9, pp. 68609–68618, 2021. DOI: 10.1109/ACCESS.2021.3076789.

[26] M. Abulaish, A. Kamal, and M. J. Zaki, "A Survey of Figurative Language and Its Computational Detection in Online Social Networks," *ACM Transactions on the Web*, vol. 14, no. 1, pp. 1–52, 2020. DOI: 10.1145/3361573.

[27] A. Baruah, K. Das, F. Barbhuiya, and K. Dey, "Context-Aware Sarcasm Detection Using BERT," in *Proc. Workshop on Figurative Language Processing*, 2020. DOI: 10.18653/v1/2020.figlang-1.12.

[28] N. Majumder, S. Poria, H. Peng, N. Chhaya, E. Cambria, and A. Gelbukh, "Sentiment and Sarcasm Classification With Multitask Learning," *IEEE Intelligent Systems*, vol. 34, no. 3, pp. 38–43, 2019. DOI: 10.1109/MIS.2019.2904691.

[29] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep Contextualized Word Representations," in *NAACL*, 2018. DOI: 10.18653/v1/N18-1202.

[30] E. Troiano, C. Strapparava, G. Ozbal, and S. S. Tekiroğlu, "A Computational Exploration of Exaggeration," in *EMNLP*, 2018. DOI: 10.18653/v1/D18-1365.

[31] S. Ilic, E. Marrese-Taylor, J. A. Balazs, and Y. Matsuo, "Deep Contextualized Word Representations for Detecting Sarcasm and Irony," *arXiv preprint arXiv:1809.09795*, 2018. DOI: 10.48550/arXiv.1809.09795.

[32] D. Hazarika, S. Poria, S. Gorantla, E. Cambria, R. Zimmermann, and R. Mihalcea, "CASCADE: Contextual Sarcasm Detection in Online Discussion Forums," *arXiv preprint arXiv:1805.06413*, 2018. DOI: 10.48550/arXiv.1805.06413.

[33] A. Joshi, P. Bhattacharyya, and M. J. Carman, "Automatic Sarcasm Detection: A Survey," *ACM Computing Surveys*, vol. 50, no. 5, pp. 1–22, 2017. DOI: 10.1145/3124420.

[34] M. Ebrahimi, A. H. Yazdavar, and A. Sheth, "Challenges of Sentiment Analysis for Dynamic Events," *IEEE Intelligent Systems*, vol. 32, no. 5, pp. 70–75, 2017. DOI: 10.1109/MIS.2017.3711648.

[35] L. Peled and R. Reichart, "Sarcasm SIGN: Interpreting Sarcasm With Sentiment Based Monolingual Machine Translation," *arXiv preprint arXiv:1704.06836*, 2017. DOI: 10.48550/arXiv.1704.06836.

[36] A. Ghosh and T. Veale, "Magnets for Sarcasm: Making Sarcasm Detection Timely, Contextual and Very Personal," in *Proc. EMNLP*, 2017. DOI: 10.18653/v1/D17-1057.

[37] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching Word Vectors With Subword Information," *Transactions of the ACL*, vol. 5, pp. 135–146, 2017. DOI: 10.1162/tacl_a_00051.

[38] S. K. Bharti, B. Vachha, R. K. Pradhan, K. S. Babu, and S. K. Jena, "Sarcastic Sentiment Detection in Tweets Streamed in Real Time: A Big Data Approach," *Digital Communications and Networks*, vol. 2, no. 3, pp. 108–121, 2016. DOI: 10.1016/j.dcan.2016.05.002.

[39] S. Poria, E. Cambria, D. Hazarika, and P. Vij, "A Deeper Look Into Sarcastic Tweets Using Deep Convolutional Neural Networks," *arXiv preprint arXiv:1610.08815*, 2016. DOI: 10.48550/arXiv.1610.08815.

[40] M. Zhang, Y. Zhang, and G. Fu, "Tweet Sarcasm Detection Using Deep Neural Network," in *COLING*, 2016. DOI: 10.48550/arXiv.1610.08815.

[41] S. Amir, B. C. Wallace, H. Lyu, and P. C. M. J. Silva, "Modelling Context With User Embeddings for Sarcasm Detection in Social Media," *arXiv preprint arXiv:1607.00976*, 2016. DOI: 10.48550/arXiv.1607.00976.

[42] A. Ghosh and T. Veale, "Fracking Sarcasm Using Neural Network," in *WASSA*, 2016. DOI: 10.18653/v1/W16-0427.

[43] M. Bouazizi and T. Ohtsuki, "Sarcasm Detection in Twitter: 'All Your Products Are Incredibly Amazing'—Are They Really," in *IEEE GLOBECOM*, 2015. DOI: 10.1109/GLOCOM.2015.7417600.

[44] A. Rajadesingan, R. Zafarani, and H. Liu, "Sarcasm Detection on Twitter: A Behavioral Modeling Approach," in *Proc. WSDM*, 2015. DOI: 10.1145/2684822.2685316.

[45] A. Joshi, V. Sharma, and P. Bhattacharyya, "Harnessing Context Incongruity for Sarcasm Detection," in *ACL-IJCNLP*, 2015. DOI: 10.3115/v1/P15-2124.

[46] F. Kunneman, C. Liebrecht, M. van Mulken, and A. van den Bosch, "Signaling Sarcasm: From Hyperbole to Hashtag," *Information Processing & Management*, vol. 51, no. 4, pp. 500–509, 2015. DOI: 10.1016/j.ipm.2014.07.006.

[47] S. K. Bharti, K. S. Babu, and S. K. Jena, "Parsing-Based Sarcasm Sentiment Recognition in Twitter Data," in *ASONAM*, 2015. DOI: 10.1145/2808797.2809410.

[48] D. Bamman and N. A. Smith, "Contextualized Sarcasm Detection on Twitter," in *Proc. ICWSM*, 2015. DOI: 10.1609/icwsm.v9i1.14663.

[49] T. Ptacek, I. Habernal, and J. Hong, "Sarcasm Detection on Czech and English Twitter," in *COLING*, 2014. DOI: 10.3115/v1/C14-1021.

[50] F. Barbieri, H. Saggion, and F. Ronzano, "Modelling Sarcasm in Twitter, A Novel Approach," in *WASSA*, 2014. DOI: 10.3115/v1/W14-2611.

[51] C. Liebrecht, F. Kunneman, and A. van den Bosch, "The Perfect Solution for Detecting Sarcasm in Tweets #not," in *WASSA*, 2013. DOI: 10.3115/v1/W13-1604.

[52] E. Riloff, A. Qadir, P. Surve, L. De Silva, N. Gilbert, and R. Huang, "Sarcasm as Contrast Between a Positive Sentiment and Negative Situation," in *Proc. EMNLP*, 2013. DOI: 10.3115/v1/D13-1066.

[53] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," *arXiv preprint arXiv:1301.3781*, 2013. DOI: 10.48550/arXiv.1301.3781.

[54] A. Ramteke, A. Malu, P. Bhattacharyya, and J. S. Nath, "Detecting Turnarounds in Sentiment Analysis: Thwarting," in *ACL Short Papers*, 2013. DOI: 10.3115/v1/P13-2150.