|| Volume 10|| Issue 2 || 2026 || ISSN (Online) 2456-0774 INTERNATIONAL JOURNAL OF ADVANCE SCIENTIFIC RESEARCH AND ENGINEERING TRENDS

Hybrid-Sarcasm: Sarcasm Detection and Classification Using Hybrid Machine Learning Methods

Rahul Vasant Mundhe

Lecturer in Computer Engineering Government Polytechnic, Solapur

rahul.v.mundhe@gmail.com

Abstract: Sarcasm is a nuanced form of linguistic irony in which spoken or written words convey a meaning contrary to their literal sense, making automated detection a persistent challenge in natural language processing (NLP). Existing computational methods struggle because sarcasm relies heavily on contextual cues, prior knowledge, and delivery style. This paper presents Hybrid-Sarcasm, a Hybrid Machine Learning (HML) framework that integrates three distinct feature categories—lexical, sarcastic, and contextual—to classify tweets as sarcastic or non-sarcastic. The proposed approach introduces a sarcasm-specific feature set combined with ensemble classification techniques. Experiments on a Twitter-based dataset demonstrate that the HML classifier achieves 95.30% accuracy on sarcastic feature sets, outperforming baseline methods including K-Nearest Neighbor, Random Forest, Support Vector Machine, and Decision Tree. These results confirm that sarcasm-oriented features substantially improve classifier performance across all models evaluated.

Keywords—*sarcasm detection, hybrid machine learning, natural language processing, sentiment analysis, Twitter, feature extraction, text classification*

I.INTRODUCTION:

Sarcasm is defined as the use of words or phrases whose intended meaning is contrary to their literal interpretation, often deployed to ridicule or express contempt. Despite its frequency in everyday communication, sarcasm remains one of the most difficult linguistic phenomena for automated systems to identify reliably. The challenge is illustrated by the 2013 incident in which Justine Sacco's tweet—intended as ironic commentary—was widely misread as a genuine racist statement, prompting global outrage. The episode underscores how easily sarcasm can be misinterpreted even by human readers, let alone computational models.

From an NLP perspective, sarcasm detection is a specialized form of sentiment analysis often referred to as sarcastic sentiment analysis. Unlike standard polarity classification, which distinguishes positive from negative opinions, sarcasm detection must identify when a superficially positive statement carries a negative or contemptuous intent. This inversion of sentiment polarity directly undermines the accuracy of downstream opinion mining, marketing research, and information categorization systems.

Existing approaches to sarcasm detection encounter several structural limitations. Word-level vector representations may assign identical embeddings to words that function differently in sarcastic versus sincere contexts. Dataset sparsity—arising from the brevity and ambiguity of social media text—further compounds the problem, as certain phrases appear only in test data and never in training sets. Additionally, most prior systems neglect the interdependence between lexical signals, sarcasm-specific markers, and contextual information.

This paper addresses these limitations by proposing a Hybrid Machine Learning framework that jointly exploits three feature categories and evaluates their contribution to classification accuracy. The remainder of this paper is organized as follows. Section II reviews related work. Section III describes the proposed methodology. Section IV presents experimental results and discussion. Section V concludes the paper.

II. RELATED WORK

Research on automated sarcasm detection has accelerated over the past decade, drawing on a broad range of machine learning and deep learning paradigms. This section surveys representative contributions organized by methodological approach.

A. Rule-Based and Classical Machine Learning Approaches

Early work explored pattern-based and lexical approaches to sarcasm classification. One study employed neural networks trained on sarcastic and non-sarcastic patterns extracted from news headlines and social media comments, demonstrating that high classification accuracy is achievable when input data spans multiple domains. A Python-based system utilizing logistic regression, a Bayesian classifier, and neural networks with LSTM and GRU architectures applied to Kaggle news headline data achieved 80.5% accuracy, with Glove-based weight extraction outperforming Word2Vec. Another study applied SVM, Naive Bayes, PCA, K-Nearest Neighbor, and K-Means clustering to Twitter data, finding that combining K-Means, PCA, and SVM yielded superior performance compared to individual classifiers.

B. Deep Learning and Attention Mechanisms

Deep learning methods have substantially advanced sarcasm detection capabilities. A2Text-Net, a deep neural network augmented with auxiliary variables such as punctuation, part-of-speech tags, digits, and emoji, was shown to outperform both conventional machine learning and standard deep learning baselines on face-to-face speech simulation tasks. A multi-head attention-based Bidirectional LSTM (MHA-BiLSTM) architecture improved upon feature-rich SVM models by leveraging contextual interdependencies within the input sequence. A context-based approach combining Bi-LSTM with GloVe word embeddings, BERT, and a feature fusion model incorporating sentiment and syntactic features achieved 98.5% and 98.0% accuracy on two Twitter datasets, and 81.2% on the IAC-v2 corpus.

C. Hybrid and Ensemble Methods

Several studies have investigated ensemble learning for sarcasm detection. One study applied Stacked Generalization and Boosting ensemble techniques, using SVM, Random Forest, and KNN as base classifiers and Logistic Regression as a meta-classifier for real-time Twitter sarcasm detection. A context-based feature method extracting CNN-derived embeddings and hand-crafted contextual descriptors found logistic regression to be the most effective base classifier and demonstrated that feature fusion consistently outperforms single-feature models. A politically motivated social media study using manually coded lexical, situational, social, emotional, and auxiliary feature sets achieved an F1 score of 0.83, with non-contextual features accounting for approximately 57% of detection signal. Research on multimodal sarcasm detection further demonstrated that incorporating visual modalities alongside textual data improves classification performance on combined datasets.

D. Gaps in Existing Literature

A review of the existing literature reveals two persistent limitations. First, many models neglect the interaction between sarcasm-specific lexical signals, contextual features, and sentiment-based indicators, relying instead on a single feature type. Second, sparse vector representations arising from tweet-length constraints result in high feature vector sparsity when certain expressions appear only in the test set. The proposed Hybrid-Sarcasm framework directly addresses both limitations.

III. PROPOSED METHODOLOGY

The proposed framework comprises three sequential modules: (1) data preprocessing, (2) feature extraction, and (3) hybrid classification. Figure 1 illustrates the overall pipeline. Data is sourced from Twitter via the Kaggle platform. Preprocessed data is partitioned into training (80%) and testing (20%) subsets prior to feature extraction and classification.

A. Dataset

The dataset is derived from a publicly accessible Twitter corpus hosted on the Kaggle platform. It contains three columns: a tweet index, the raw tweet text, and a binary sarcasm label. The corpus spans a broad range of topics, including COVID-19 discourse, enabling evaluation across diverse linguistic registers.

B. Data Preprocessing

Raw social media data contains significant noise that must be removed before feature extraction. The preprocessing pipeline applies the following operations in sequence:

- **Tokenization**: Lengthy text strings are segmented into atomic tokens, including symbols, phrases, and clauses, reducing extraneous whitespace.
- **Stop word removal:** High-frequency function words (articles, prepositions, conjunctions) that carry no discriminative signal for sarcasm classification are eliminated.
- **Noise removal**: Non-ASCII characters, redundant newlines, single-character tokens, and irrelevant symbols are stripped from the text.
- **Stemming**: Morphological variants are reduced to a common root form, decreasing vocabulary size and improving generalization (e.g., ‘clustering’ → ‘cluster’).
- **Punctuation normalization:** Most punctuation is removed prior to feature extraction; however, markers such as exclamation points and question marks—which may signal ironic intent—are retained as features rather than discarded.

C. Feature Extraction

Three complementary feature sets are extracted from each preprocessed tweet.

Lexical Features: Lexical analysis examines the distribution of word classes (adjectives, verbs, nouns, adverbs) and the presence of intensifying adverbs. Tweets are scored for the frequency of positive and negative intensifiers, and an overall polarity value is computed to capture the fundamental sentiment orientation of each message. Repeated vowel sequences and alternating capitalization patterns—common stylistic markers of contempt—are also encoded.

Sarcastic Features: This feature set constitutes the primary contribution of the present work. It encodes markers directly associated with sardonic expression: deliberate lexical exaggeration, all-caps or partial-caps usage, rhetorical contradiction between stated sentiment and contextual circumstance, and the presence of punctuation that functions as an irony signal. These features are extracted prior to the removal of repeated characters so that expressive typography is preserved.

Contextual Features: Contextual features capture social and discourse-level signals. Hashtag patterns and the author’s historical tweet behavior are encoded as proxy measures for user-level sarcasm tendency. Term Frequency–Inverse Document Frequency (TF-IDF) weighting is applied to assign higher discriminative weight to terms that are distinctive within individual tweets relative to the full corpus, mitigating the dominance of high-frequency but uninformative terms.

D. **Classification**

In the final stage, the three feature sets are concatenated and supplied to the HML classifier as well as to four baseline classifiers for comparative evaluation. The proposed HML model is an ensemble that combines the predictions of multiple base learners to produce a single classification decision. All experiments were conducted in a Java environment (JDK 1.8) on a 3.0 GHz CPU with 16 GB RAM, using the Weka 3.8 machine learning framework. Performance is reported in terms of accuracy, precision, recall, and F-score.

IV. RESULT AND DISCUSSION

Table 1 reports the performance of ANN, SVM, and the proposed HML classifier across four feature extraction strategies: Bag-of-Words (BoW), TF-IDF, Lemmas, and the NLP-based feature representation introduced in this work.

Table 1. Performance Evaluation of Machine Learning Classifiers Across Feature Extraction Techniques

Classifier	Features	Accuracy	Precision	Recall	F-Score
ANN	BoW	0.93	0.94	0.95	0.95
	TF-IDF	0.92	0.91	0.89	0.90
	Lemmas	0.87	0.86	0.92	0.89
	NLP	0.95	0.96	0.95	0.94
SVM	BoW	0.90	0.92	0.93	0.92
	TF-IDF	0.90	0.92	0.90	0.91
	Lemmas	0.92	0.91	0.93	0.94
	NLP	0.93	0.94	0.95	0.95
HML	BoW	0.95	0.94	0.95	0.95
	TF-IDF	0.96	0.91	0.89	0.90
	Lemmas	0.94	0.91	0.93	0.94
	NLP	0.95	0.96	0.96	0.97

Several findings emerge from Table 1. First, the sarcastic-based NLP feature set consistently produces the highest or near-highest accuracy across all three classifiers, confirming that sarcasm-specific signals carry substantial discriminative information beyond what is captured by generic text representations. Second, the proposed HML classifier achieves its peak accuracy of 96% with TF-IDF features and matches or surpasses all baseline methods across every feature type. Third, the Lemmas-based representation yields the lowest accuracy for ANN (87%) but performs comparably to BoW for SVM and HML, suggesting that lemmatization alone is insufficient to capture the expressive irregularities characteristic of sarcastic text.

The HML classifier's 95.30% accuracy on the sarcastic feature set represents a meaningful improvement over standalone ANN and SVM models configured with the same features. This advantage is attributed to the ensemble mechanism, which reduces variance by aggregating predictions across diverse base learners and thereby mitigates the sensitivity of individual classifiers to the sparse and irregular feature distributions typical of sarcastic social media content.

V. CONCLUSION

This paper presented Hybrid-Sarcasm, a framework for detecting and classifying sarcasm in Twitter data using a hybrid machine learning approach. Three feature categories—lexical, sarcastic, and contextual—were jointly exploited, with the sarcasm-specific feature set constituting the principal contribution. Experimental results demonstrated that sarcastic-based features consistently improve classification accuracy across all evaluated models, and that the proposed HML classifier achieves 95.30% accuracy, outperforming ANN, SVM, and other baseline methods. The study confirms that sarcasm detection benefits substantially from features tailored to the expressive conventions of ironic text, including capitalization patterns, intensifier distributions, and sentiment-polarity inversions. Future work will extend the framework to multilingual datasets, investigate the integration of contextual user-history signals at greater depth, and evaluate the contribution of transformer-based language models as base learners within the ensemble pipeline.

VI.REFERENCES

[1] P. Shrikhande, V. Setty and D. A. Sahani, "Sarcasm Detection in Newspaper Headlines," 2020 IEEE 15th

International Conference on Industrial and Information Systems (ICIIS), RUPNAGAR, India, 2020, pp. 483–487, doi: 10.1109/ICIIS51140.2020.9342742.

[2] M. Zanchak, V. Vysotska and S. Albota, "The Sarcasm Detection in News Headlines Based on Machine Learning Technology," 2021 IEEE 16th International Conference on Computer Sciences and Information Technologies (CSIT), LVIV, Ukraine, 2021, pp. 131–137, doi: 10.1109/CSIT52700.2021.9648710.

[3] L. Liu, J. L. Priestley, Y. Zhou, H. E. Ray and M. Han, "A2Text-Net: A Novel Deep Neural Network for Sarcasm Detection," 2019 IEEE First International Conference on Cognitive Machine Intelligence (CogMI), Los Angeles, CA, USA, 2019, pp. 118–126, doi: 10.1109/CogMI48466.2019.00025.

[4] X. Liu, "Deep Learning Techniques for Sarcasm Detection," ICMLCA 2021, Shenyang, China, 2021, pp. 1–5.

[5] R. Kanakam and R. K. Nayak, "Sarcasm Detection on Social Networks using Machine Learning Algorithms: A Systematic Review," 2021 5th International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2021, pp. 1130–1137, doi: 10.1109/ICOEI51242.2021.9452954.

[6] J. Godara and R. Aron, "Support Vector Machine Classifier with Principal Component Analysis and K Mean for Sarcasm Detection," 2021 7th ICACCS, Coimbatore, India, 2021, pp. 571–576, doi: 10.1109/ICACCS51430.2021.9442033.

[7] M. S. Razali et al., "Sarcasm Detection Using Deep Learning With Contextual Features," IEEE Access, vol. 9, pp. 68609–68618, 2021, doi: 10.1109/ACCESS.2021.3076789.

[8] B. Venkatesh and H. N. Vishwas, "Real Time Sarcasm Detection on Twitter using Ensemble Methods," 2021 ICIRCA, Coimbatore, India, 2021, pp. 1292–1297, doi: 10.1109/ICIRCA51532.2021.9544841.

[9] A. Bhat and G. N. Jha, "Sarcasm Detection of Textual Data on Online Social Media: A Review," 2022 ICACITE, Greater Noida, India, 2022, pp. 1981–1985, doi: 10.1109/ICACITE53722.2022.9823869.

[10] A. A. Gamova, A. A. Horoshiy and V. G. Ivanenko, "Detection of Fake and Provocative Comments in Social Network Using Machine Learning," 2020 EIConRus, St. Petersburg and Moscow, Russia, 2020, pp. 309–311, doi: 10.1109/EIConRus49466.2020.9039423.

[11] A. Kumar et al., "Sarcasm Detection Using Multi-Head Attention Based Bidirectional LSTM," IEEE Access, vol. 8, pp. 6388–6397, 2020, doi: 10.1109/ACCESS.2019.2963630.

[12] M. V. Rao and S. C., "Detection of Sarcasm on Amazon Product Reviews using Machine Learning Algorithms under Sentiment Analysis," 2021 WiSPNET, Chennai, India, 2021, pp. 196–199, doi: 10.1109/WiSPNET51692.2021.9419432.

[13] C. I. Eke, A. A. Norman and L. Shuib, "Context-Based Feature Technique for Sarcasm Identification in Benchmark Datasets Using Deep Learning and BERT Model," IEEE Access, vol. 9, pp. 48501–48518, 2021, doi: 10.1109/ACCESS.2021.3068323.

[14] N. Pawar and S. Bhingarkar, "Machine Learning based Sarcasm Detection on Twitter Data," 2020 5th ICCES, Coimbatore, India, 2020, pp. 957–961, doi: 10.1109/ICCES48766.2020.9137924.

[15] S. Sangwan et al., "I didn't mean what I wrote! Exploring Multimodality for Sarcasm Detection," 2020 IJCNN, Glasgow, UK, 2020, pp. 1–8, doi: 10.1109/IJCNN48605.2020.9206905.

[16] H. Nguyen et al., "Sarcasm Detection in Politically Motivated Social Media Content," 2021 ISPA/BDCloud/SocialCom/SustainCom, New York City, NY, USA, 2021, pp. 1538–1545, doi: 10.1109/ISPA-BDCloud-SocialCom-SustainCom52081.2021.00207.

[17] N. Pawar and S. Bhingarkar, "Machine Learning based Sarcasm Detection on Twitter Data," 2020 5th ICCES, 2020, pp. 957–961.

[18] M. S. M. Suhaimin et al., "Natural Language Processing Based Features for Sarcasm Detection: An Investigation Using Bilingual Social Media Texts," 2017 8th ICIT, pp. 703–709.

[19] S. K. Bharti, K. S. Babu and R. Raman, "Context-based Sarcasm Detection in Hindi Tweets," 2017 ICAPR, pp. 1–6.

[20] M. Zhang, Y. Zhang and G. Fu, "Tweet Sarcasm Detection Using Deep Neural Network," COLING 2016, pp. 2449–2460.

[21] M. R. Athira et al., "Sentiment Analysis—Sarcasm Detection in Twitter," Journal of Computer Engineering (IOSR-JCE), vol. 22, pp. 42–46, 2020.