Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 47 No. 03 (2026)

Detecting Sarcasm Using Commonsense Reasoning Capabilities of Comet and Representation Power of Bert

Obarakpo Ogechukwu Peace, Oyelami Olufemi Moses, Popoola Solomon Olugbemiga

Department of Computer Science Bowen University Osun, Nigeria.

Abstract: Sarcasm poses significant challenges for natural language processing systems due to its reliance on implicit meaning and contextual cues. While models like BERT have advanced sarcasm detection, they still face difficulties with non-literal interpretations. This paper introduces a hybrid model that combines BERT's contextual representations with COMET's commonsense reasoning. The approach involved fine-tuning BERT on sarcasm datasets and integrating commonsense inferences via a feature integration layer. Evaluated on three key datasets, the BERT–COMET model demonstrated superior performance in recognizing subtle sarcastic expressions and contradictions, highlighting that incorporating commonsense knowledge can enhance detection accuracy and reduce errors.

Keywords: Commonsense knowledge, knowledge graphs, sarcasm detection, sentiment analysis.

1. Introduction

Sarcasm detection as a type of sentiment analysis has undergone significant evolution over the years. Early efforts primarily relied on rule-based systems and manually crafted linguistic cues to identify sarcastic content in text. As natural language processing (NLP) technologies progressed, researchers began to leverage machine learning and statistical methods to capture more nuanced forms of expression. From machine learning models, hybrid models emerged, combining rule-based and machine learning models and later transitioning to deep learning models. However, sarcasm remains a particularly challenging aspect of language due to its reliance on implicit meanings, contextual cues, and commonsense reasoning. The increasing prevalence of user-generated content on social media has further necessitated the need for accurate sarcasm detection, given its impact in sentiment analysis, opinion mining, and dialogue systems [1]. Misreading sarcasm can result in flawed sentiment assessments, distorted opinions, and poor chatbot interactions, impacting user experience. Furthermore, content moderation may misinterpret sarcastic comments, leading to unnecessary censorship of humor and irony. Improved understanding of sarcasm could enhance user engagement, making online interactions feel more genuine and connected.

Previous studies have approached sarcasm detection using various techniques including syntactic patterns, sentiment polarity shifts, and contextual embeddings. For instance, models have been trained to detect incongruities between expected and actual sentiment within sentences [2]. While approaches using deep learning and transformer-based models like BERT have yielded notable improvements, limitations persist particularly in capturing the commonsense knowledge necessary to understand the implied meanings behind sarcastic expressions. As sarcasm often involves a contrast between literal and intended meaning, models that lack real-world knowledge and reasoning capabilities frequently fail to interpret such subtleties. This misinterpretation can have severe implications including inaccurate sentiment classification in customer feedback systems, misleading conclusions in public opinion mining, and ineffective conversational agents in customer service or virtual assistants.

Chowdhury and Chaturvedi in [3] tried to solve the problem by integrating commonsense knowledge to detect sarcasm. They also highlighted significant limitations in current approaches to performing sarcasm detection using commonsense knowledge, particularly with COMET-based models, their work inclusive. First, COMET's generic inferences lack adaptation to sarcasm-specific cues such as contradictions or ironic intent, limiting their

effectiveness. Second, graph convolutional network (GCN)-based fusion methods used, failed to meaningfully integrate commonsense with textual data as saliency tests reveal minimal reliance on COMET's outputs. Third, the absence of a filtering mechanism for irrelevant or noisy COMET inferences often degrades performance, especially in non-sarcastic text. To address these gaps in [3], this study proposes an approach that combines the language representation power of BERT with the commonsense reasoning capabilities of COMET (Commonsense Transformers). The rationale stems from the observation that sarcasm frequently depends on background knowledge and real-world expectations which traditional models often lack. By integrating COMET, which generates plausible inferences based on everyday events and social norms with BERT's contextual understanding, the study aims to bridge the gap between surface-level text analysis and deeper semantic comprehension.

2. Related Works

Current approaches to modeling sarcasm in sentiment analysis systems typically employ one of three strategies [4]. The first involves training deep learning models on large corpora of labeled sarcastic utterances, allowing the system to learn patterns of polarity inversion. The second approach incorporates contextual features such as user history, topic consistency, and situational knowledge to identify probable sentiment reversals [5,6]. The third strategy focuses on detecting linguistic markers commonly associated with sarcasm including specific punctuation patterns, emoji usage, and stylistic devices like hyperbole. While each method shows promise, many studies suggest that combining multiple techniques may help address the complex and multifaceted nature of sarcasm [1].

Recent advancements in NLP have focused on multimodal approaches, contextual analysis, and deep learning techniques to enhance detection accuracy. Multimodal methods leverage a combination of data types such as text, audio, and visual cues, to capture the diverse ways sarcasm is conveyed [7,8]. For example, integrating textual data with speech intonation or facial expressions has significantly improved models' ability to understand and classify sarcastic remarks in multimedia contexts like videos and social media posts [9]. Contextual analysis, both historical and situational, plays a pivotal role in sarcasm detection [10]. By analyzing the broader conversational or situational context in which a statement is made, models can better infer whether sarcasm is present. This involves leveraging attention mechanisms and transformer-based architectures like BERT and Generative Pre-trained Transformer (GPT) to capture subtle dependencies in dialogue or written exchanges [11].

Deep learning techniques, particularly in the form of convolutional neural networks (CNNs), recurrent neural networks (RNNs) and transformers, have further enhanced the ability to detect sarcasm by enabling models to learn complex patterns and relationships within data [12,13]. Pre-trained language models fine-tuned for sarcasm detection are now a key area of research as they combine general language understanding with task-specific insights [14].

Mane and Khatavkar, in [15] proposed a semigraph-based method to detect sarcasm by analyzing text structure and calculating sarcastic polarity scores. While this approach achieved strong performance (F1-score: 0.83), its reliance on document-level patterns makes it struggle with short-form social media text where sarcasm often depends on external context rather than internal document structure. Similarly, Wang in [16] integrated commonsense knowledge using a heterogeneous graph attention network (HGAT), demonstrating improved accuracy on Reddit and Internet Argument Corpus datasets. However, this method's dependency on pre-existing knowledge graphs limits its adaptability to evolving or culture-specific sarcasm forms such as memes and regional slang.

Anbarasu, et al. in [2] introduced Conditional Random Fields (CRF) with a modified Expectation Maximization (EM) algorithm to address sarcasm detection in social media content. The methodology leverages CRF's ability to model sequential dependencies in text while incorporating the modified EM algorithm to handle the latent semantic relationships characteristic of sarcastic expressions. Pokhriyal and Jain in [17] proposed an unsupervised method combining probabilistic distributions with sentiment lexicons, but faced challenges with sarcastic expressions in dynamic social media contexts. Alaramma et al. in [18] explored an ensemble classifier using incongruity features but did not assess its robustness across domains, raising questions about generalizability beyond the initial training data.

Chen, et al. in [19] used different approaches in sentiment analysis which include machine learning (ML), deep learning (DL), and hybrid approaches based on Kitchenham's framework. They found out that ML models effectively managed explicit negation but struggled with implicit shifts, while DL models like BERT excelled in understanding context. hybrid models achieved the highest accuracy (97%) by leveraging both ML and DL strengths. The study emphasizes that explicit negation is better addressed than implicit one with Word2Vec and BERT embeddings outperforming traditional methods in feature engineering.

Gregory et al. in [20] fine-tuned RoBERTa, BERT, LSTM and a transformer ensembled made up of BERT-base-uncased, RoBERTa-base, XLNet-base-cased, RoBERTa-large, and ALBERT-base-v2model on sarcasm datasets. The results show that transformer ensembled model achieved a precision score of 0.758, recall as 0.767 and f1-macro as 0.756.

Earlier systems relied on supervised learning methods while deep learning has shifted focus to neural architectures with RNNs, particularly LSTMs, used for sequential detection despite their limitations. Hybrid approaches are also gaining popularity. Jing in [21] combined pragmatic features with BERT embeddings, and multitask learning frameworks. The model was trained to recognize both sentiment and sarcasm at the same time. This approach helps improve the model's ability to interpret complex language nuances effectively. Recent studies have also integrated diverse architectures. For instance, Qiu, et al. in [22] combined GNNs and transformers, and ensemble methods that enhance robustness across models.

Bosselut, et al. in [23] presented COMET, a generative framework for commonsense knowledge bases (KBs), fine-tuned a pre-trained transformer (GPT) to create subject-relation-object triples that generate commonsense inferences. Liu, et al. in [24] introduced PrimeNet, structuring commonsense knowledge into three layers and achieving a 92.4% human acceptance rate. Nguyen, et al. in [25] discussed the Ascent framework for knowledge extraction, outperforming previous methods in informativeness. Wang, et al. in [26] presented ECoK for emotional reasoning, integrating emotional events into the ATOMIC format. Chowdhury and Chaturved in [27] explored the use of COMET for sarcasm detection, showing limited success. Ghosal, et al. in [28] developed COSMIC for emotion recognition in conversations on multiple conversational datasets with weighted-F1 scores of 65.28 on IEMOCAP and 65.21 on MELD, and Macro-F1 of 51.05 on DailyDialog, outperforming prior models at the time. Puri, et al. in [29] introduced a commonsense-based text mining approach linking legal texts to public sentiment with promising results.

3. Methods

The proposed model builds on the framework depicted in Figure 1.

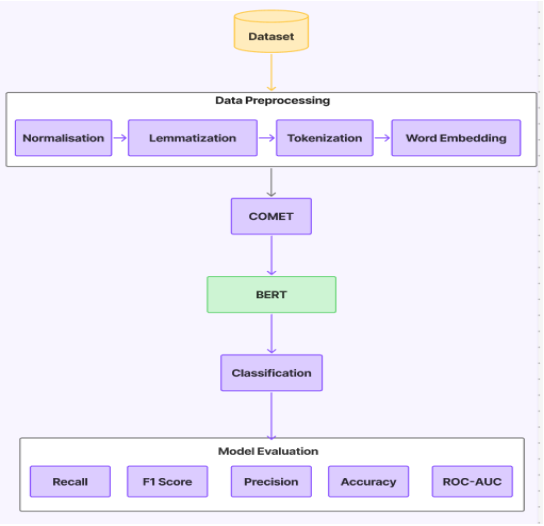

Fig. 1. Model Methodology

The proposed model builds on the framework depicted in Figure 1.

Prior to the datasets being utilized effectively by BERT algorithm, it had to go through a series of preprocessing steps to convert the raw data into a clean, structured, and analyzable state. Normalization was carried out and the techniques include normalizing the text by bringing all the text to lowercase to maintain consistency along with the removal of punctuation, special characters, and numbers if they were not conveying important information. Lemmatization was also performed to reduce words to their base forms and ensure they were dictionary words. Tokenisation was then used for splitting sentences into individual words, a necessary process for transformer-based models like BERT. Further preprocessing included correcting common misspellings and eliminating very brief or unimportant text segments. For text data gathered from Reddit sites, further cleaning was necessary to remove markup, URLs, user tags, or duplicate material. These preprocessing steps are distinct from word embedding, as they focus solely on improving text quality rather than representation. The actual word embeddings were generated automatically within COMET as seen in Eqn. 1, which converts cleaned tokens into contextual vector representations for classification. The internal workflow of the COMET model for generating commonsense knowledge started with the input data, structured as a tuple consisting of subject phrase (s) and a relation (r). These input data were first tokenized into smaller units (ti). Eqn. 1 is the sequence of vectors (X(0)) that represented the input text with both meaning and order captured.

$$X(0)=[E[CLS]+P0,E1+P1,\ldots,En+Pn \quad \text{.....eqn. 1}$$

Each token was converted into a dense numerical vector (Ei) using an embedding layer. Since word order is important, we also added positional encodings (Pi) to help the model understand the sequence of the tokens. The input text (x) was also passed into COMET, a model trained to generate commonsense inferences. Eqn. 2 is the assigned weights given to each commonsense inferences.

$$\alpha i = \frac{\exp(u^T \tanh(W_a Z_i))}{\sum_{j=1}^{K} \exp(u^T \tanh(W_a Z_i))} \quad \text{.....eqn. 2}$$

Each of these commonsense inferences (ki) was converted into a vector (zi) using the ϕ(ki) function which converts each generated knowledge triple into a vector. Since not all generated knowledge is equally important, an attention mechanism assigns weights (αi) to them, giving more importance to the most relevant ones. Where u is the trainable parameter vector used for attention scoring, Wa is the weight matrix used in the attention mechanism. Finally, these weighted vectors - (αi) were combined into a single commonsense knowledge embedding as in eqn. 3(g).

$$g = \sum_{i=1} \alpha i z_i \quad \text{.....eqn. 3}$$

At inference time, the model was given an event and a relation token (xEffect or xIntent), and it autoregressively generated the most likely continuation of the sequence. Rather than performing logical reasoning, COMET learns to mimic statistical patterns observed in its training data. The inferences it produces are not derived from explicit logic or symbolic rules, but rather from statistical associations encoded during pretraining on curated knowledge triples. Next, the token embeddings in eqn. 1 and the commonsense embedding in eqn. 3 were fed into BERT (Bidirectional Encoder Representations from Transformers).

Inside BERT, a stack of transformer layers applies self-attention. This means every token can interact with other tokens in the sentence to understand the context. At the core of BERT is the Transformer encoder, composed of layers of multi-head self-attention mechanisms and position-wise feedforward neural networks. In the multi-head attention module (MHAttn) , each head projects the input into query (Q), key (K), and value (V) vectors and computes attention scores Attn(Q,K,V) using the scaled dot-product of queries and keys leading to eqn. 4:

$$\text{Attn}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad \text{.....eqn. 4}$$

These scores determined the weighted sum (Wh,Wg) of the values, allowing each token to attend to all others in the sequence and capture deep contextual relationships. The attention outputs were followed by layer normalization (LN) to stabilize training and a feedforward network (FFN(x)) that applied non-linear transformations. Another layer normalization followed this step to ensure consistency before passing to the next

layer. Eqn. 5 is the feedforward network (FFN(x)) which induces non-linearity and increases the model's representational capacity.

$$FFN(x) = \max(0, xW1 + b1)W2 + b2 \text{.....eqn. 5}$$

Eqn. 6 is the Layer Normalization (LN) which is employed after both the multi-head attention and feedforward sublayers to stabilize training and ensure consistent feature scaling.

$$LN = -\sum_{i=1}^{n} \log f_i y_i(x; \theta) \text{.....eqn. 6}$$

Eqn. 7 (F(ℓ)) applies multi-head self-attention and normalizes the result using LN.

$$F(\ell)=LN(H(\ell-1)+MHAttn(Q,K,V)),\ldots\ldots \quad \text{eqn. 7}$$

Eqn. 8 is the hidden state h(ℓ) which denotes the output representation of the ℓ-th Transformer layer, obtained through residual connections, followed by normalization, allowing effective gradient flow and deep contextual representation learning.

$$h(\ell)=LN(F(\ell)+FFN(F(\ell))) \text{.....eqn. 8}$$

The encoded representation, particularly the one corresponding to the [CLS] token h, is passed through a linear layer followed by a softmax activation function which is qn. 9 to produce output probabilities for tasks like classification.

$$\tilde{y}=softmax(Wch + bc) \text{.....eqn. 9}$$

The model makes a guess (prediction) and we measured how wrong it was which is loss (L) with eqn. 10. Where, C is the number of classes:

$$L= -\sum_{c=1}^{C} y_c \log \tilde{y}_c + \lambda ||\Theta||_2^2 \text{ .....eqn. 10}$$

Eqn. 11 shows how to adjust each weight to reduce the error. Adam (Θ) then applied these adjustments smartly using momentum and adaptive learning rates.

$$\Theta \leftarrow AdamUpdate(\Theta, \nabla\Theta L) \text{.....eqn. 11}$$

The parameters Θ were updated, and the model performed a little better on the next round. This process undergoes many iterations (epochs) until the model learns well.The commonsense embedding is fused with BERT's internal representation so that the model does not just rely on surface-level text, but also on commonsense reasoning. After several layers, BERT produces the final representation of the whole input, summarized in the [CLS] token vector, which acts like the "summary meaning" of the sentence. The model is then evaluated using different metrics such as accuracy, precision, recall and F1 score. Together, these metrics provide a comprehensive view of the model's effectiveness, allowing for a better understanding of its strengths and weaknesses.

4. Results

In this section, we present the implementation details of our proposed model, followed by the baseline systems used for comparison and the evaluation metrics employed across the experiments.

Experiment Setup: Our sarcasm detection model was implemented using PyTorch, a widely adopted deep learning framework for natural language processing tasks. All experiments were conducted using the HuggingFace Transformers library from which we loaded the BERT encoder and the COMET commonsense inference generator. To generate external commonsense knowledge, we utilized the pre-trained COMET-ATOMIC model, which produced inferential triples such as xIntent, xReact, xEffect, and xReason. These generated sequences were encoded using BERT and integrated with the contextual representation of the input text through a fusion layer. We empirically examined multiple fusion configurations to establish the most effective combination of contextual and commonsense embeddings. A series of fusion dimensionalities and attention weights were tested to determine the optimal configuration. Unless otherwise stated, the embedding size for both BERT and COMET-derived representations is 768 dimensions. All experiments were executed on an NVIDIA GeForce RTX GPU

environment, enabling efficient batch-level inference and training cycles. Training was performed for up to 20 epochs, with early stopping applied to prevent overfitting. The Adam optimizer was used with a learning rate of 2e-5, and dropout was set to 0.3 across all trainable layers. Batch size was fixed at 16, while the COMET-generated sequences were truncated or padded to maintain a consistent maximum length. The classifier consists of a fully connected layer with ReLU activation, followed by a softmax output for binary classification (sarcastic vs. non-sarcastic).

Baseline: To ensure a fair comparison and to highlight the contribution of integrating commonsense reasoning, we compared our model with several strong baseline systems by using the same datasets and evaluation metrics widely used in sarcasm detection. DistilBERT + FFNN + COMET (Chowdhury & Chaturvedi, 2021), and GCN (Chowdhury & Chaturvedi, 2021) providing a competitive knowledge-enhanced baseline including commonsense elements. The aforementioned baselines were modified to fit the binary classification scenario because our goal was limited to sarcasm detection. To set a lower bound on knowledge integration performance, commonsense embeddings were simply appended to the textual embeddings without fusion optimization for COMET-related baselines.
Metrics for Evaluation: We employed many evaluation measures that are often used in binary classification and sarcasm detection to provide a thorough and trustworthy evaluation of model performance across all datasets. For the sarcasm detection job, the following standard metrics were employed: Accuracy, Precision, Recall, F1-Score and ROU-AUC. Given the nature of sarcasm where class imbalance and misclassification of minority sarcastic samples are typical, the F1-Score serves as the most trustworthy metric of success.

The results obtained are presented in Table 1 below:

Table 1: Results Of Dataset

	News Headlines	English Sarcasm Dataset	Reddit Dataset
Accuracy	0.8558	0.9546	0.7680
Precision	0.8102	0.9706	0.3458
Recall	0.8813	0.8302	0.2624
F1 Score	0.8443	0.8949	0.2984
ROC-AUC	0.8715	0.9755	0.5929

The results of this study reveal important insights into how the integration of BERT and COMET influences sarcasm detection across datasets of varying complexity, structure, and noise levels. The overall performance trends demonstrate that the BERT+COMET model excels when sarcasm is expressed through clear linguistic cues or when data is clean, structured, and well-annotated, as seen prominently in the English Sarcasm Dataset. With an accuracy of 0.95, precision of 0.97, F1-score of 0.89, and ROC-AUC of 0.98, the model shows exceptional ability to differentiate between sarcastic and non-sarcastic statements in such controlled environments. The high precision indicates that the model rarely misclassifies non-sarcastic texts as sarcastic, while the very high AUC suggests the model maintains a strong balance between sensitivity and specificity across decision thresholds. This superior performance aligns with the theoretical expectation that BERT's deep contextual understanding combined with COMET's commonsense inferencewould particularly benefit datasets where sarcasm relies on subtle semantic contradictions rather than messy conversational constructs. In essence, the English sarcasm

dataset provides the ideal conditions - clean sentences, consistent labeling, and moderate linguistic context, allowing the hybrid model to demonstrate its full capacity.

A second pattern emerges from the results obtained on the news headlines dataset where the model also performed strongly but with slightly lower scores than in the English sarcasm dataset. Achieving an accuracy of 85.58%, F1-score of 0.84, and ROC-AUC of 0.87, the model demonstrates that structured and formal textual environments allow it to capture sarcasm with considerable reliability. News headlines often express sarcasm through irony, exaggeration, and contrast between expected and actual scenarios- patterns that BERT is naturally good at identifying. COMET's inferential knowledge further aids by supplying likely causes, intents, or reactions that help the model sense implicit contradictions. The balanced behaviour between precision and recall in this dataset indicates that the model is not only detecting sarcasm effectively but doing so with relatively few false positives and false negatives. However, the confusion matrix for this dataset reveals a recurring issue: the model tends to perform better at identifying non-sarcastic headlines than sarcastic ones, a consequence primarily of class imbalance and limited linguistic variation in sarcastic headlines. Nevertheless, the model's performance here reaffirms the value of integrating commonsense reasoning into transformer-based architectures for text where sarcasm is lexically or contextually constrained.

In contrast, the results from the Reddit dataset expose one of the critical limitations of the BERT+COMET framework: its difficulty in handling noisy, user-generated, and highly informal data. The architecture struggles significantly, earning an accuracy of 76.80%, precision of only 0.34, recall of 0.26, F1-score of 0.29, and a weak ROC-AUC of 0.59. These metrics collectively show that the model frequently mislabels sarcastic statements as non-sarcastic (high false negatives) and also incorrectly flags many non-sarcastic statements as sarcastic (false positives). The confusion matrix confirms this imbalance - 104 sarcastic instances were missed, while 70 non-sarcastic posts were wrongly flagged. Reddit conversations often depend heavily on cultural references, inside jokes, discourse history, and implicit context, all of which are difficult for both BERT and COMET to interpret without additional conversational or multimodal cues. Moreover, Reddit posts contain slang, humour, code-switching, metaphor, and unreliable labeling, all of which introduce noise that reduces the reliability of COMET's commonsense generation. Since COMET draws from general commonsense rather than internet-specific social discourse, much of its inferred knowledge fails to align with the informal communication found on Reddit. This mismatch explains the low precision, weak recall, and near-random ROC-AUC score.

Table 2 shows the results comparing the BERT+COMET model to prior baselines such as DistilBERT+FFNN+COMET (Chowdhury & Chaturvedi, 2021), Qwen2.5 (Yang et al. 2024), BERT+GGNN (Qin et al. 2025), and GCN architectures (Chowdhury & Chaturvedi, 2021).

Table 2: Comparision Of Different Models

	News Headlines	SemEval 2015 English Sarcasm dataset	Irony FigLang 2020 (Reddit)
Baseline - DistilBERT + FFNN + COMET [27]	96.13%	69.09%	67.95%
GCN [27]	96.16%	67.88%	67.50%
BERT+GGNN [22]	90.80%	86.40%	83.20%

Qwen2.5 [35]	87.2%	83.2% [36]	81.6%
BERT + COMET	85.58%	95.46% [36]	76.80%

From the table, an interesting pattern emerges. The classical baselines achieve consistent performance in structured datasets- around 96% accuracy in News Headlines, but struggle more with noisy or domain-specific datasets, especially Reddit. In contrast, the BERT+COMET model behaves differently, underperforming the baselines in the News Headlines dataset but strongly outperforming them in the English Sarcasm and Reddit datasets. This indicates that the hybrid model is far more sensitive to datasets where commonsense reasoning is beneficial, particularly in situations where sarcasm is subtle, multi-layered, or reliant on violation of expected human behaviour. The superior performance on the English Sarcasm dataset suggests that COMET successfully boosts BERT's ability to identify contextual inconsistencies and sentiment inversions. Meanwhile, the improvement over baselines in the Reddit dataset (despite overall low performance) suggests that the model captures slightly more nuanced cues than simpler architectures, even if the noise obscures the benefits.

Overall, these results show that integrating commonsense reasoning with contextual language modelling improves sarcasm detection, but the degree of improvement is heavily dependent on the nature and quality of the dataset. Clean and semantically rich datasets benefit the most, while noisy, user-generated data requires additional strategies such as context-aware modeling, conversational history incorporation, or advanced preprocessing. The model's sensitivity to sarcasm in structured environments demonstrates its potential, but the variability in performance across datasets highlights the need for future work addressing noise reduction, domain adaptation, and enhanced contextual grounding..

5. Discussion

Sarcasm detection relies on context, social knowledge, cultural norms, and commonsense reasoning, which traditional computational models struggle to capture. This study addressed the challenge of detecting sarcasm in natural language processing (NLP) systems by combining BERT and COMET. The research proposed a mix of contextual strength of BERT and commonsense reasoning abilities of COMET to improve sarcasm detection performance. The hybrid model showed better results compared to basic approaches especially when sarcasm is implied without clear cues like punctuation, exaggeration, or hashtags. The study also contributed to the understanding of how commonsense knowledge plays a role in NLP, as COMET can generate inferences about intentions, causes, and effects. However, it sometimes produces noisy or irrelevant inferences which is a complementary effect with BERT.

The study also contributes to applied sentiment analysis and opinion mining by addressing sarcasm, a linguistic phenomenon that often distorts the true sentiment of a statement, leading to misclassification in conventional sentiment analysis systems. By enabling models to identify sarcastic expressions and underlying semantic incongruity, the study shows an improvement in the accuracy of sentiment polarity detection beyond literal word usage. To address its limitations, future work should explore incorporating multimodal signals such as emojis, images, and conversational context, which often provide additional cues for detecting sarcasm beyond plain text. Expanding training data using cross-domain datasets and low-resource augmentation techniques may also improve generalizability across diverse linguistic styles and platforms. Finally, integrating more advanced commonsense reasoning frameworks or dynamically updated knowledge bases could help the model better interpret evolving slang, cultural references, and implicit meanings.

Refrences:

[1] S. Alaramma, A. Habu, B. Ya'u, and A. Gamsha, "Sentiment analysis of sarcasm detection in social media," Gadau Journal of Pure and Allied Sciences, vol. 2, pp. 76–82, 2023, doi: 10.54117/gjpas.v2i1.72.

[2] S. Anbarasu, S. Karthik, and P. Anandhakumar, "CRF-MEM: Conditional Random Field Model Based Modified Expectation Maximization Algorithm for Sarcasm Detection in Social Media," Journal of Internet Technology, 2023, doi: 10.53106/160792642023012401005.
[3] S. B. R. Chowdhury and S. Chaturvedi, "Does Commonsense Help in Detecting Sarcasm?," arXiv, 2021, arXiv:2109.08588.
[4] A. Anouska, M. Agarwal, P. K. Mallick, and N. Padhy, "Beyond Words: Contextual Sarcasm Detection in News Texts using Advanced Models," in Proc. ESIC, 2024, doi: 10.1109/esic60604.2024.10481639.
[5] A. Baruah, K. Das, F. Barbhuiya, and K. Dey, "Context-aware sarcasm detection using BERT," in Proc. 2nd Workshop on Figurative Language Processing, pp. 83–87, Jul. 2020.
[6] A. Bhat and G. N. Jha, "Sarcasm Detection of Textual Data on Online Social Media: A Review," in Proc. ICACITE, 2022, pp. 1981–1985, doi: 10.1109/icacite53722.2022.9823869.
[7] Y. Bengong and X. Ji, "Multi-modal sarcasm detection based on emotion perception and cross-modality attention fusion," 2024, doi: 10.3233/jifs-233163.
[8] M. Jia, C. Xie, and L. Jing, "Debiasing Multimodal Sarcasm Detection with Contrastive Learning," arXiv, 2023.
[9] R. Das and T. D. Singh, "Multimodal sentiment analysis: a survey of methods, trends, and challenges," ACM Computing Surveys, vol. 55, no. 13s, pp. 1–38, 2023.
[10] N. A. Helal, A. Hassan, N. L. Badr, and Y. M. Afify, "A contextual-based approach for sarcasm detection," Scientific Reports, vol. 14, no. 1, p. 15415, 2024.
[11] P. Dubey, P. Dubey, and P. N. Bokoro, "Unpacking Sarcasm: A Contextual and Transformer-Based Approach for Improved Detection," Computers, vol. 14, no. 3, p. 95, 2025.
[12] M. M. Krishna and J. Vankara, "Detection of sarcasm using bi-directional RNN based deep learning model," J. Adv. Res. Appl. Sci., vol. 31, no. 2, pp. 352–362, 2023.
[13] R. Jamil et al., "Detecting sarcasm in multi-domain datasets using CNN and LSTM," PeerJ Computer Science, vol. 7, p. e645, 2021.
[14] R. Kanakam and R. K. Nayak, "Sarcasm detection on social networks using machine learning algorithms: A systematic review," in Proc. ICOEI, 2021, pp. 1130–1137.
[15] S. Mane and V. Khatavkar, "Polarity Based Sarcasm Detection Using Semigraph," arXiv preprint arXiv:2304.01424, 2023.
[16] W. mingqian, "Commonsense Knowledge-Powered Heterogeneous Graph Attention Networks for Semi-Supervised Short Text Classification," Expert Systems with Applications, vol. 229, 2023.
[17] H. Pokhriyal & G. Jain, "Sarcasm Detection via Sentiment and Emotion Analysis of News Headlines Using Optimized Simulation of Weibull Entropy Distribution," Journal of Intelligent & Fuzzy Systems, 2023.
[18] S. K. Alaramma, A. A. Habu, B. I. Ya'u, and A. G. Madaki, "Sentiment Analysis of Sarcasm Detection in Social Media," Gadau Journal of Pure and Allied Sciences, vol. 2, no. 1, pp. 1–12, 2023.
[19] W. Chen, F. Lin, G. Li, B. Liu "A Survey of Automatic Sarcasm Detection: Fundamental Theories, Methods, and Opportunities," Neurocomputing, vol. 567, pp. 127428, 2024.
[20] H. Gregory, S. Li, P. Mohammadi, N. Tarn, R. Draelos, & C. Rudin. A transformer approach to contextual sarcasm detection in Twitter. In Proc. of Second Workshop on Figurative Language Processing (pp. 270–275). Association for Computational Linguistics. ttps://doi.org/10.18653/v1/2020.figlang-1.37
[21] D. Jing, "A BERT-Based with Fuzzy Logic Sentimental Classifier for Sarcasm Detection," 2024, doi: 10.1109/icaace61206.2024.10548550.
[22] Z. Qin, Q. Luo, Z. Zang, and H. Fu, "Detecting sarcasm in user-generated content integrating transformers and gated graph neural networks," PeerJ Computer Science, vol. 11, p. e2817, Apr. 2025, doi: 10.7717/peerj-cs.2817'
[23] A. Bosselut, H. Rashkin, M. Sap, C. Malaviya, A. Çelikyilmaz, and Y. Choi, "COMET: Commonsense transformers for automatic knowledge graph construction," in Proc. 57th Annu. Meeting Assoc. Comput. Linguistics (ACL), Florence, Italy, 2019, pp. 4762–4779.

[24] Q. Liu, S. Han, E. Cambria, Y. Li, and K. Kwok, "PrimeNet: A framework for commonsense knowledge representation and reasoning based on conceptual primitives," Cognitive Computation, vol. 16, no. 6, pp. 3429–3456, 2024, doi: 10.1007/s12559-024-10345-6.

[25] T.-P. Nguyen, S. Razniewski, and G. Weikum, "Advanced Semantics for Commonsense Knowledge Extraction," Proceedings of the Web Conference (WWW), pp. 2522–2532, 2020.

[26] Z. Wang et al., "ECoK: Emotional Commonsense Knowledge Graph for Emotional Reasoning," in Findings of the Association for Computational Linguistics (ACL), 2024.

[27] S. B. R. Chowdhury and S. Chaturvedi. Does Commonsense help in detecting Sarcasm?. arXiv preprint arXiv:2109.08588, 2021

[28] D. Ghosal, N. Majumder, A. Gelbukh, R. Mihalcea, and S. Poria, "COSMIC: COmmonSense Knowledge for Emotion Identification in Conversations," in Findings of the Association for Computational Linguistics: EMNLP, pp. 2470–2481, 2020.

[29] M. Puri, A. S. Varde, and G. de Melo, "Commonsense-Based Text Mining for Linking Legal Texts with Public Sentiment," IEEE Intelligent Systems, vol. 38, no. 2, pp. 68–76, 2023.

[30] R. Misra and P. Arora, "Sarcasm Detection using News Headlines Dataset," AI Open, vol. 4, pp. 13–18, 2023, doi: 10.1016/j.aiopen.2023.01.001.

[31] I. Abu Farha, S. V. Oprea, S. R. Wilson, and W. Magdy, "SemEval-2022 Task 6: iSarcasmEval, Intended Sarcasm Detection in English and Arabic," in Proc. 16th Int. Workshop on Semantic Evaluation (SemEval-2022), Seattle, United States, 2022, pp. 802–814, doi: 10.18653/v1/2022.semeval-1.111

[32] D. Ghosh, A. Vajpayee, and S. Muresan, "A Report on the 2020 Sarcasm Detection Shared Task," in Proc. 2nd Workshop on Figurative Language Processing (FigLang 2020), Seattle, United States, 2020, pp. 1–11, doi: 10.18653/v1/P17.

[33] S. Castro et al., "Towards Multimodal Sarcasm Detection," 2019, doi: 10.48550/arXiv.1906.01815.

[34] Z. Yu, D. Jin, X. Wang, Y. Li, L. Wang, and J. Dang, "Commonsense knowledge enhanced sentiment dependency graph for sarcasm detection," in Proc. Int. Joint Conf. Artificial Intelligence (IJCAI), Macao, China, 2023, pp. 2423–2431, doi: 10.24963/ijcai.2023/269.

[35] A. Yang et al., "Qwen2.5 technical report," arXiv preprint arXiv:2412.15115, 2024, doi: 10.48550/arXiv.2412.15115.

[36] 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. Conf. Empirical Methods in Natural Language Processing (EMNLP), Seattle, WA, USA, 2013, pp. 704–714.