A Hybrid RoBERTa-BiGRU-Attention Model for Accurate and Context-Aware Figurative Language Detection

Dr. Sreeja Balakrishnan[1], Rahul Suryodai[2], S. Manochitra[3], Jasgurpreet Singh Chohan[4],
Karaka Ramakrishna Reddy[5], A. Smitha Kranthi[6], Dr Ritu Sharma[7]

Assistant Professor (SS), Department of English, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore-641043, Tamil Nadu, India[1]

Senior Data Engineer (Data Governance, Data Analytics: Enterprise Performance Management, AI & ML), USA[2]

Assistant Professor, Department of Computer Science and Engineering, St. Joseph's College of Engineering, Chennai, India[3]

Marwadi University Research Center-Department of Mechanical Engineering-Faculty of Engineering & Technology, Marwadi University, Rajkot, Gujarat, India[4]

Assistant Professor, Department of BS & H, B V Raju Institute of Technology, Narsapur, Medak, Telangana, India[5]

Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Green Fields, Vaddeswaram, AP-522302[6]

Assistant Professor, Department of Mathematics and Humanities, M. M Engineering College, MMDU Mullana, India[7]

Abstract—**Figurative language, especially sarcasm, poses strong challenges for Natural Language Processing (NLP) models because of its implicit, context-sensitive nature. Both traditional and transformer models tend to find it difficult to identify these subtle forms, particularly when dealing with imbalanced datasets or without mechanisms for targeted interpretability. For overcoming these shortcomings, this study recommends a hybrid deep learning architecture that integrates RoBERTa for high-contextual embedding, Bidirectional Gated Recurrent Units (BiGRU) to capture bidirectional sequential relations, and an attention mechanism allowing the model to focus on the most informative parts of the input text. This integration enhances semantic understanding and classification accuracy compared to current solutions. The model is trained and tested on the benchmark News Headlines Dataset for Sarcasm Detection using binary cross-entropy loss minimized with Adam, along with dropout and learning rate scheduling to avoid overfitting. Experimental results yield better performance, attaining an accuracy of 92.4%, a precision of 91.1%, a recall of 93.2%, and an F1-score of 92.1%. These results outperform baseline techniques such as BiLSTM with attention and fine-tuned BERT variants. Implementation uses PyTorch and Hugging Face Transformers, ensuring reproducibility and extensibility. While effective, the model faces challenges with figurative expressions requiring external world knowledge or cultural context beyond pretrained embeddings. Future work aims to integrate external knowledge graphs and extend the model to multilingual and cross-domain scenarios. This hybrid framework advances figurative language detection, contributing to the broader goal of enhancing AI's nuanced understanding and interpretability of human language.**

Keywords—Figurative language detection; sarcasm classification; RoBERTa-BiGRU-Attention model; contextual embeddings; Natural Language Processing

I. Introduction

Figurative language, including types like sarcasm, metaphor, and irony, is critical in human communication as it communicates beyond the literal meaning [1]. It is particularly dominant in digital text, such as social media, online reviews, and newspaper headlines [2]. These expressive mechanisms are used to add subtlety, signal humor, or highlight perspectives; yet their inherent implicitness and extreme dependency on context make them an overwhelming problem for Natural Language Processing (NLP) algorithms [3]. Traditional NLP models usually find it difficult to correctly infer nonliteral meaning, resulting in misclassification and misinterpretation [4]. This shortcoming has a negative effect on most downstream applications that are dependent on the correct semantic interpretation, including sentiment analysis, opinion mining, emotion detection, and conversational AI. For example, failure in sarcasm detection can invert the targeted sentiment of text, resulting in the wrong analysis outcomes and inferior user experiences.

Emergent developments of models such as BERT and RoBERTa have significantly progressed many NLP tasks through their capacity for creating contextual embeddings capturing syntactic and semantic relations in texts [5]. In spite of these developments, such models still struggle with figurative and subtle language [6]. Such models tend to fail to capture fine semantic signals and long-distance dependencies necessary for distinguishing figurative expressions [7]. Furthermore, issues like severe class imbalance in training sets restrict model generalization [8]. Another major flaw is the lack of comprehensive explainability for transformer-based predictions, which restricts knowledge of why a certain input was labeled as figurative or not, hence influencing trust and usability in actual systems.

Against such challenges, this work suggests a new hybrid framework that is tasked with enhancing the accuracy of figurative language detection through the use of complementary deep learning methods. The model uses RoBERTa to produce high-contextualized word embeddings, a (BiGRU) to capture bidirectional sequential dependencies in the text, and an attention mechanism to attend to the most contextually informative tokens in an input sequence. By combining these parts, the model hopes to harness the contextual capability of transformers with the added sequential learning and focus prioritization, achieving more accurate and context-aware identification of figurative language. This should yield more accurate outputs that can be used in practical NLP applications that make use of subtle language comprehension.

A. Research Motivation

Figurative language, like irony, metaphor, and sarcasm, is important in the way humans convey meanings that transcend literal understandings. Figurative language in digital communication, particularly on social media and news websites, is prevalent and powerful. Yet, conventional NLP systems often misinterpret figurative language because of its contextual subtlety and implicit semantics. This results in tremendous challenges for applications such as sentiment analysis, emotion detection, and user interaction in AI applications. Despite the progress made by models such as BERT and RoBERTa, precise identification and interpretation of figurative language remains a challenging task. Current models lack class balance, explainability, and proper emphasis on sequential context. Thus, it is an urgent need to create stronger and contextualized approaches that can better represent the richness of figurative language. This work is stimulated by the desire to enhance automatic detection of figurative expressions through using hybrid deep learning models that combine strong contextual embeddings with sequential and attention mechanisms, improving at last performance and explainability of figurative language detection systems.

B. Research Significance

Automatic identification of figurative language is of tremendous importance in the development of NLP applications, such as sentiment analysis, conversational AI, content moderation, and recommendation systems. Incorrect interpretation of figurative language—sarcasm, especially—can cause errors in user intent, resulting in ambiguous sentiment classification or unsuitable responses for AI systems. It is a critical challenge to overcome in improving user experience and decision-making in automated services. Through the creation of a new hybrid architecture consisting of RoBERTa contextual embeddings, Bidirectional Gated Recurrent Units (BiGRU), and an attention mechanism, this work immensely helps bridge the gaps in current methods that fail to capture semantic depth and contextual subtleties. The approach enhances classification accuracy and generalization on benchmark data sets as well as facilitates improved explainability of decisions via attention mechanisms. The adaptability of the work to multilingual and cross-domain settings underscores its wider significance beyond English text processing. This research, therefore, is a significant step towards more human-like and intelligent language comprehension in artificial intelligence systems, with various applications in cross-sections of fields demanding subtle linguistic analysis and interpretation.

C. Key Contribution

- Proposed a novel hybrid deep learning architecture integrating RoBERTa, BiGRU, and attention mechanisms for effective figurative language detection.
- Demonstrated superior performance metrics on benchmark sarcasm detection datasets compared to baseline methods.
- Addressed challenges of context sensitivity and class imbalance to improve model generalization and robustness.
- Enhanced explainability of model predictions through the integration of an attention mechanism highlighting important tokens.
- Provided a framework adaptable to multilingual and cross-domain applications, expanding the utility of figurative language detection in NLP.

The remainder of the section is prepared as follows: Section II is the related study, Section III is the Problem Statement, methodology part is in Section IV, Section V is the result and discussion section, Section VI is the conclusion and future work section.

II. Related Works

Liu et al. [11] explored the issue of nonliteral language and presented Fig-QA, a Winograd-like question-answering dataset that assesses the linguistic models on how well they comprehend the figurative language with contrasting semantics. The task consists of choosing the appropriate meaning between paired figurative expressions whose contextual suggestion varies, and, thus, challenges the profundity of nonliteral understanding of language models. The study compared the functioning of a number of state-of-the-art transformer-based models on the Fig-QA benchmark. Despite these models showing levels of performance far above-average random chance, they were still very far behind the performance of human participants in both zero-shot and few-shot settings. Such a gap highlights the inability of current models to reflect the fine and easily context-specific details of the figurative language. The results indicate that current neural networks, despite their impressive literal comprehension ability, are not very advanced in reasoning, a prerequisite to effective nonliteral comprehension.

Chakrabarty et al [12] designed DREAM-FLUTE, a new framework of figurative language understanding using scene elaboration and entailment inference to increase the accuracy of the interpretation. Metaphors and similes are often used as figurative language, and the sense therefore is not inferred easily when read their surface form. In order to solve this, the authors posed a hypothesis that the proper interpretation will involve building a mental model of the described situation. DREAM-FLUTE builds on this hypothesis by creating a well-organized mental model of both the premise and the hypothesis based on an existing framework known as DREAM (Deep Reading for Elaborative Mental Modeling). Then it produces entailment or contradiction judgments and produces textual explanations. The system was tested as part of the FigLang2022 Shared Task, in which it topped the leader board (tied in first place with 63.3%

Acc@60 accuracy). The authors also showed that by using ensemble methods to improve the performance of a system further was even possible.

Huang et al. [13] reduced the various limitations of existing benchmarks of figurative language. Although the use of figurative language was in the past successfully put within the framework of the Recognizing Textual Entailment (RTE) paradigm, much earlier benchmarking efforts were hampered by spurious correlation and annotation artefacts, and it is hard to assess whether language models are understanding nonliteral meaning effectively. FLUTE bridges this divide by presenting 9,000 pieces of real NLI examples together with human-generated clarifications on the same. The data consists of four categories of figurative language comprising sarcasm, simile, metaphor, and idioms. It was built in the model-in-the-loop fashion, integrating the power of GPT-3 with crowdsourcing and human expert annotation in order to achieve quality and scale. The authors have proven that such a hybrid annotation approach contributes greatly to the reasonable development of complicated linguistic databases. Baseline tests implemented with a fine-tuned model T5 on FLUTE showed that the presence of textual explanations gives an advantage to models, meaning a better understanding of answers and something to be noticed about their possible better performance in figurative situations.

Sharma et al [14] explored transformer-based architectures in identifying figurative presentations compared to literal utterances in huge textual information. With the growing demand for efficient sentiment extraction and text classification with an understanding of language specificities, the authors found it necessary to run a set of transformer models (including BERT and RoBERTA variants), whereas the latter could efficiently perform sentiment extraction and many aspect specifications when it came to nonliteral language. The authors used the opportunity to determine whether the models that were trained on one category of figurative language representation, e.g., metaphors or sarcasm, could generalize well across categories. The study has extensively experimented to indicate that transformer models have achieved significant advances in typical NLP tasks, but are still not very successful with figurative constructs in NLP since they have implicit and context-sensitive definitions. The study also highlighted the urgency with which machines should be made to comprehend the figurative language that is rampant in the human language, and unlike literal language, which is processed.

Kumar et al. [15] approached the problem of interpretation of affective dimensions in the context of conversational dialogue, which are frequently masked by the use of the figurative language of irony, sarcasm, and metaphor. In order to better understand such situations, the authors have suggested a deep neural network model (MOSES) that could be applied to Sarcasm Explanation in Dialogues (SED). The model uses input multimodal sarcastic utterances and outputs natural language explanations to disclose an ironic meaning behind it. This direct encoding of figurative purpose is in turn used to enhance a number of downstream NLP tasks, such as detecting sarcasm, identifying humor and detecting emotion. The tests on different experimental conditions proved that MOSES shows superior results by 2% to the state-of-the-art in the SED in terms of ROUGE, BLEU, and METEOR. Moreover, the inclusion of generated explanations managed to obtain a significant 14% boost in sarcasm detection tasks, and a 2% boost in both humor and emotion recognition tasks.

Vitiugin and Paakki [16] proposed an effective framework regarding the classification of such a complicated task, the Deep Ensemble Soft Classifier (DESC). In view of the fact that the figurative language forms have a complex nature of figurative language use, especially sarcasm, irony, and metaphor, the present study sought to traverse through the inconsistent semantics that are usually intertwined with an expression of a figurative language form. The first contribution of the authors was the presentation of a complete data preprocessing pipeline and its use to standardize and optimize the input representations to deep learning models. They further derived some specialized features like the syntactic structure, mode of expressing emotions, expressive style, and temperamental formulations in the articulation of a figurative meaning, as held by those utilizing social media. Such dense feature representations were next input into the DESC model, an ensemble scheme that mixes different deep learning structures together to boost the accuracy of classification and generalization. Tests have been done using three benchmark sets, with one made particularly in multiple forms of figurative language.

Hülsing and Im Walde [17] provided a unification of a framework that is able to detect multiple figures of speech in more than one language. Aware of the fact that figurative meanings, i.e., metaphors, sarcasm, and idioms, are not solely linguistically diversified, but also culturally conditioned, the study implemented the multilingual multi-figurative language modeling. The authors came up with a sentence-level corpus of three varieties of figurative language and seven languages with the aim of eliminating the deficit of existing monolingual and figure-specific studies. They use template-based prompt learning, which was used to facilitate the model to generalize through figures of speech and languages, with no need to add specialized modules. Such a coherent system reduces the complexity of modeling and allows greater scalability in low-resource or multilingual settings. Experimental studies showed that the proposed technique feasibly outperforms growing transformer-based baselines, which indicated the potential capability of the approach under study in robust cross-language figurative language understanding.

Ocampo and Clarifiňo [18] considered the metaphor detection task in figurative language processing to come up with a new approach that uses lexical definitions to boost metaphor detection, hence MIss RoBERTa WiLDe. This is especially true since the inception of the powerful pretrained language models, GPT-3, XLNet, and BERT, which have allowed challenges restricted to small quantities of annotated data, like metaphor recognition, to improve considerably. The authors extend this development by utilizing an open-source lexical database, known as Wiktionary, in to access automatically via dictionary on non-figurative word senses. The definitions that were put forward are contextual anchors to assist the model to differentiate between uses that are literal and uses that are metaphorical by having a contrast between basic use and contextual use. The suggested model combines RoBERTa and a binary token-level classifier to find metaphorical expressions, and it shows a high degree of transferability across datasets.

Experimental comparisons indicate that MIss RoBERTa WiLDe performs end-to-end better or equally well to the state-of-the-art metaphor detection systems, providing a scalable, dataset-independent, and efficient approach.

Lv et al. [19] proposed RB-GAT, a first hybrid text classification model focusing on a well-integrated RoBERTa and BiGRU with Graph Attention Networks (GAT) to solve the weakness of the traditional graph-based neural models in modeling the contextual semantics in a text. Graph Neural Networks (GNNs) have shown promising results on textual classification but tend to fail at capturing text sequential dependencies and contextual information (at the document level). In a bid to overcome this, RB-GAT uses RoBERTa to produce rich but contextualized word and sentence-level embeddings, which are then coupled with a (BiGRU) used to model long-range and bidirectional textual relationships. Such added features are subsequently multi-head GAT, which learns inter-token relations, where the various sections of the document graph are designated by dynamic scores of attention. This kind of arrangement enables the model to be semantically and topologically coherent. The softmax layer is then used in the final classification. Comparing 5 benchmark datasets-Ohsumed, R8, MR, 20ng, and R52, RB-GAT was superior to the baseline model, having achieved an accuracy of 71.48, 98.45 and 80.32 and 90.84, and 95.67, respectively.

Berger et al [20] employed transfer learning in the German setting on a metaphor-recognition task and accordingly used a framework that was cross-lingual, where an English corpus with the sentence-level metaphor is annotated. The English answers were mechanically translated into German, and a sample of 1,000 German sentences was manually parsed in order to become a gold standard of two assessment configurations: sequence labeling composed of tokens and sentence classification. There were two major approaches to transfer learning that the authors tested: 1) transformer-based models, including multilingual BERT, and 2) bilingual word embeddings with an RNN-based classifier. Transformer models performed fairly in zero-shot, as they produced the best performance of 61% F1-score used to perform the classification tasks, whereas the embedding-based approach did not perform any better compared to a random baseline, scoring a mere 36% F1. Once fine-tuned using the annotated German dataset, both methods increased greatly in performance, with a maximum F1-score of 90%, and so cross-lingual metaphor detection can be highly efficient in terms of language-specific tuning. A rich annotated metaphor corpus in the two languages forms a contribution of the study that can be applied to make future studies on mixed language processing of the figurative language more feasible.

Current literature emphasizes the increasing need for advanced models capable of dealing with various kinds of nonliteral semantics in languages and different types of figurative language. Benchmarks such as Fig-QA and FLUTE show that while transformer models are better than chance, their understanding of figurative language is constrained in comparison to human inference. Models such as DREAM-FLUTE and MOSES show that explanation generation and mental modeling together increase interpretability along with performance. Other methods such as DESC and RB-GAT combine contextual embeddings with deep learning models (BiGRU, GAT), enhancing classification accuracy over diverse figurative expressions. Additionally, studies on template-based prompt learning and multilingual embeddings yield high-performance results in cross-lingual metaphor detection, pointing toward a shifting future for strong, scalable processing of figurative language.

III. Problem Statement

Precise identification of figurative language, especially sarcasm, continues to be a tough task in (NLP) because it is implicit, context-dependent, and nuanced [9]. Although models such as BERT and RoBERTa have pushed the frontiers of many NLP tasks further, they tend to struggle with properly capturing subtle semantic signals and long-distance dependencies required for first-order interpretation of nonliteral expressions [10]. Besides, current methods are challenged by class imbalance problems in training data, weak explainability of model choices, and weak infusion of sequential context, all of which lower their overall performance and dependability. This constraint negatively impacts many downstream NLP tasks like sentiment analysis, opinion mining, conversational AI, and emotion detection, wherein the misunderstanding of figurative language can result in inaccurate outcomes and less-than-optimal user experiences. This study resolves the issue of creating a robust, context-sensitive system for figurative language identification that surpasses the limitations of existing approaches. In particular, it introduces a new hybrid model synergizing the benefits of transformer-based contextual embeddings and sequential learning through Bidirectional Gated Recurrent Units (BiGRU) and an attention mechanism. This integration has the goal of enhancing accuracy, interpretability, and generalizability across various kinds of figurative language, eventually enhancing the reliability and usability of figurative language detection in practical NLP applications.

IV. Proposed Hybrid Framework for Figurative Language Detection

A hybrid deep learning architecture is created through the integration of transformer-based contextual embeddings, sequential learning, and attention mechanisms to enhance the identification of figurative language. The architecture incorporates pretrained RoBERTa embeddings, which produce high-quality and context-aware representations of words, and a Bidirectional Gated Recurrent Unit (BiGRU) that learns sequential dependencies in both directions. An attention mechanism is added to emphasize and concentrate on the most informative tokens of the input sequence, enhancing the model's capability to distinguish subtle figurative uses like sarcasm. It is trained and tested on the News Headlines Dataset for Sarcasm Detection, a common benchmark with balanced sarcastic and non-sarcastic samples. Preprocessing of data involves conversion of headlines to lowercase, removal of special characters, and tokenization with RoBERTa's tokenizer. The input is padded or truncated to a constant length to have a consistent input size for batch processing. Training is done with binary cross-entropy as the loss function, optimized using the Adam optimizer. Hyperparameters like learning rate, batch size, dropout rate, and epochs are specifically tuned to minimize overfitting and maximize generalization. The use of a dropout layer keeps overfitting at bay, while learning rate scheduling stabilizes the convergence process. To holistically quantify model performance, standard classification metrics are provided. This framework of evaluation ensures balanced quantification of the model's capacity to correctly identify figurative language versus literal expressions, ultimately enhancing the robustness and usability of the detection of figurative language in practical applications of NLP. The end-to-end architecture that is proposed is a decent balance between complex semantics understanding as well as sequence-based learning, and would lead to clean and accurate detection of figurative language. Fig. 1 is a pictorial representation of it.

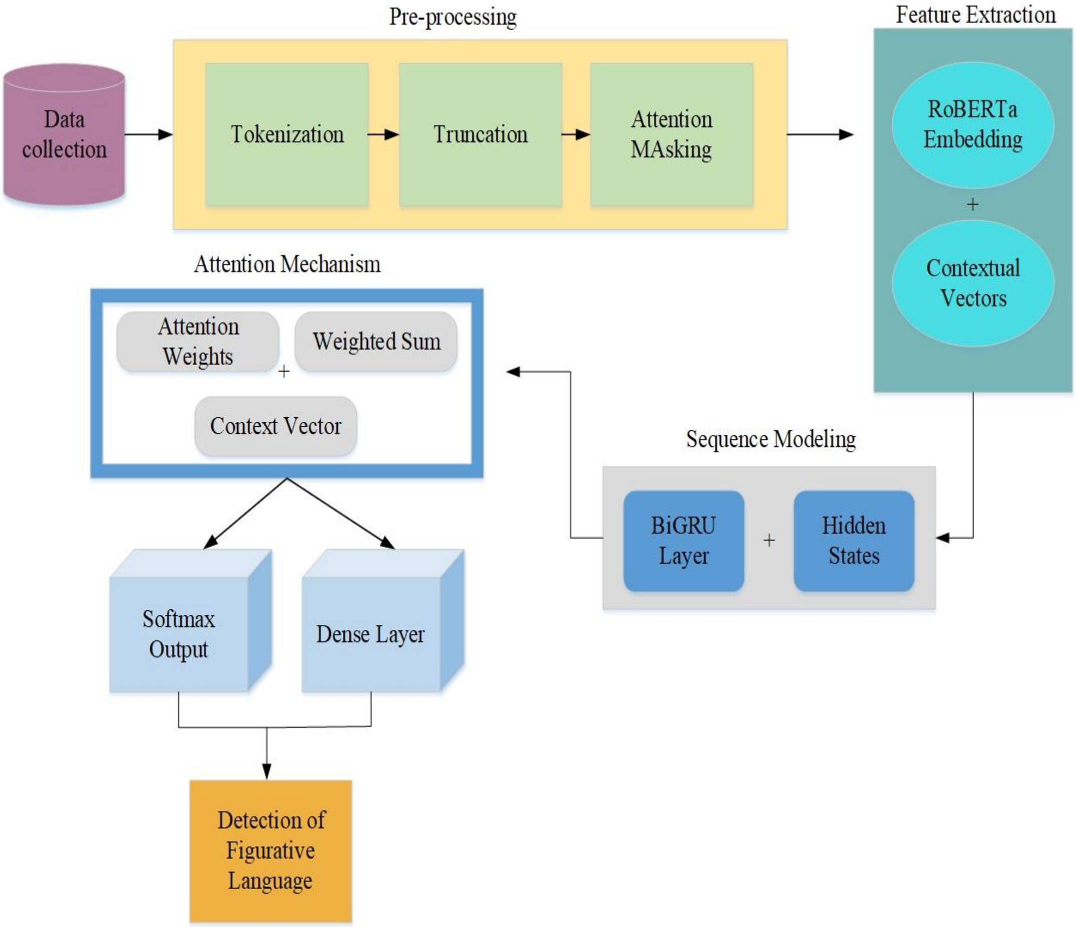

Fig. 1. Proposed RoBERTa-BiGRU-attention architecture

A. Data Collection

The News Headlines Dataset [1] for sarcasm detection, available on Kaggle and developed by Rishabh Misra, is utilized in this research [21]. The dataset comprises approximately 28,000 English-language news headlines, each labeled as either sarcastic (1) or non-sarcastic (0). Each entry contains the headline text, its corresponding sarcasm label, and in some cases, links to the original news articles. This dataset is well-suited for figurative language detection, particularly sarcasm, as it consists of concise texts with sufficient contextual information to train deep learning models effectively. The headlines exhibit clear syntactic structure and provide a clearly labeled, supervised learning environment, enabling rigorous evaluation of the model's ability to comprehend situations and interpret nonliteral language. Table I shows the dataset description.

TABLE I. DATASET DESCRIPTION

Label	Description
0	Non-Figurative Headlines
1	Figurative (Sarcastic) Headlines
Total	50,000

B. Data Pre-Processing

All headlines were in lower case, lacking punctuation, figures, and unnecessary symbols, so as to normalize the input. A fixed-length L (RoBERTa) WordPiece tokenizer was used to tokenize cleaned text and truncate or pad it to the desired length. Subsequent to this, the lean pipeline minimizes vocabulary sparsity, and the input size is maintained effectively to train

[1] https://www.kaggle.com/datasets/rmisra/news-headlines-dataset-for-sarcasm-detection

it. Formally, the cleaned headline x_{clean} is obtained from the original headline x as:

$$x_{clean} = Clean(x) \qquad (1)$$

In Eq. (1), x is the original headline whereas x clean is the cleaned version in. A small vocabulary size is achieved through such standardization and makes learning easier and faster. And to each cleaned headline x clean, a Word Piece tokenizer that works with RoBERTa is used to tokenize each. Tokenization transforms a sentence into a sequence of tokens:

$$T = Tokenizer(x_{clean}) = [t1, t2, \ldots, tn] \qquad (2)$$

In Eq. (2) $T = \{t1, t2, \ldots, tn\}_0$ refers to the order of subword tokens. The input sequences are of variable length and, as such, are padded or truncated to a set maximum length L, so that the shape of the inputs to a batch processing is the same shown in (3):

$$T' = \frac{pad}{truncate(T,L)} = [t_1, \ldots., t_l] \qquad (3)$$

Finally, tokens are converted into embeddings using RoBERTa's pretrained encoder:

$$E = RoBERTaEncoder(T') = [e_1, e_2, \ldots, e_l] \qquad (4)$$

In Eq. (4) $E \in \mathbb{R}^{L \times d}$ is the word embeddings of each token, and d is the embedding dimension. This pre-processing chain is designed to yield a clean, tokenized, semantically enriched text to achieve good training and classification.

C. Model Architecture: RoBERTa-BiGRU-Framework

The suggested method for figurative language detection uses a hybrid model with deep learning algorithms, combined with contextual embeddings and sequence learning. Following preprocessing, every headline is tokenized and fed into a pre-trained RoBERTa model, which produces contextual embeddings that have the ability to sense minute semantics and relationships between tokens and thus can effectively identify figurative language patterns in the text.

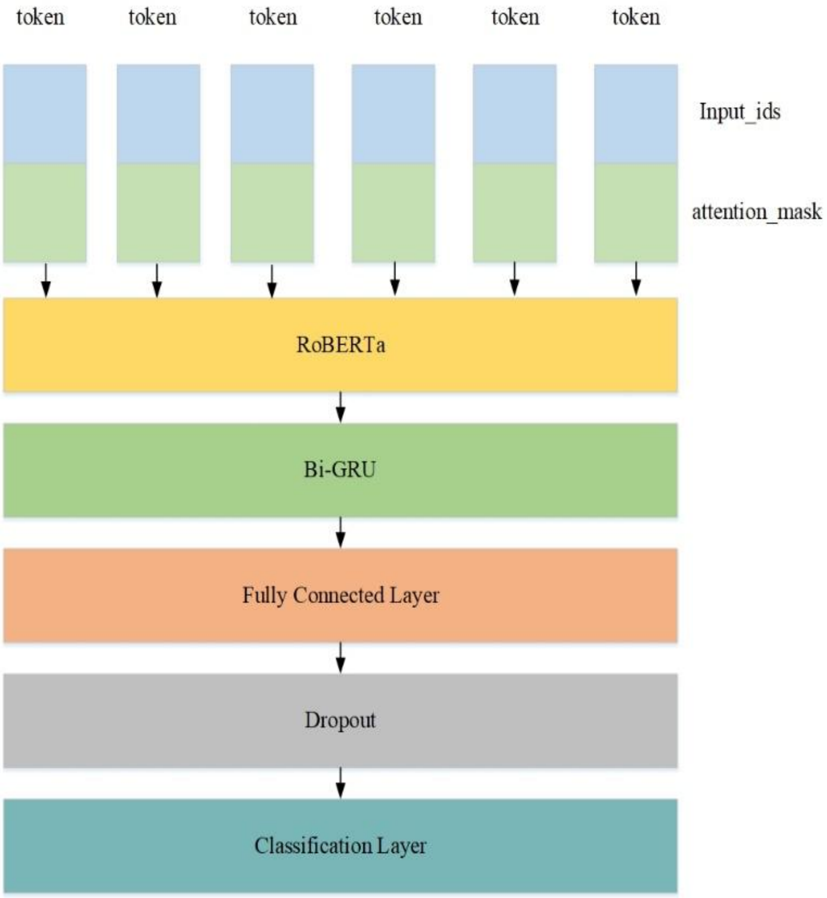

Fig. 2. Block diagram of the hybrid model pipeline.

Fig. 2 represents the proposed hybrid model intended to detect the figurative language in textual data based on the option of RoBERTa and BiGRU. The model takes as an input sentence which is translated into tokens and some other input_ids and attention_mask. The token representations will be input into the RoBERTa module, a pretrained transformer language model, and produce rich contextual representations of tokens capturing semantic and syntactic aspects. RoBERTa output embeddings are passed as input to a Bidirectional Gated Recurrent Unit (BiGRU) network that performs the further processing of the sequence in both forward and backward directions. This dual-processing capacity enables the model to keep the contextual relationships of previous and subsequent tokens and improve expression of linguistic structure and meaning. After output of the BiGRU, the Fully Connected Layer condenses the sequence-level features into a categorical condensed feature more appropriate to give to the classifier. To avoid overfitting, the model feature contains Dropout Layer that eliminates a random selection of connections in the training process. The output of that processed information is fed to the Classification Layer, which uses the Softmax function to make the prediction in regards to what type of label is in the data figurative (1) or non-figurative (0). This architecture is highly integrated in terms of both contextual, sequential and classification in the recognition of the figurative languages.

Officially, the embed output may be well formulated as: $E \in \mathbb{R}^{L \times d}$, where L is the length of the sequence and d is the embedding dimension. The latter embeddings are taken as input into a (BiGRU) to learn forward and backward dependencies throughout the sequence, which conveys more of those contextual cues. The hidden state over time t is calculated as:

$$h_t = BiGRU(E_t), t = 1,2, \ldots, L \tag{5}$$

In Eq. (5), to enhance interpretability and focus on key parts of the sentence, an attention mechanism is applied to the BiGRU outputs. This mechanism assigns higher weights to more informative tokens. The attention weight αt is calculated as:

$$\alpha_t = \frac{\exp{(u_t)}}{\sum_{j=1}^{L} \exp{(u_t)}} \tag{6}$$

In Eq. (6), $e_t = tanh(W_h h_t + b$ The final attended representation is a weighted sum of hidden states:

$$r = \sum_{t=1}^{L} \propto_t h_t \tag{7}$$

In Eq. (7), this vector is passed through a fully connected classification layer with a Softmax function to output the probability of the sentence being figurative (e.g., sarcastic):

$$y = Softmax(W_c r + b_c) \tag{8}$$

In Eq. (8), the model is trained using binary cross-entropy loss and optimized with the Adam optimizer. Dropout and learning rate scheduling are employed to prevent overfitting and ensure generalization.

D. Training Parameters and Hyperparameters

The trained model uses the binary cross-entropy loss, optimized by the Adam optimizer to better reduce classification mistakes. Major hyperparameters were finely tuned to achieve the best results on the task of detecting sarcasm. The pretrained RoBERTa-base model is used as the embedding backbone, with a maximum sequence length of 128 tokens to strike the right balance between computation and context attainment. A learning rate of 2e-5 was chosen following initial experiments, yielding stable convergence during training. The batch size is fixed at 32 to ensure full utilization of the GPU without memory overflow. To avoid overfitting, a dropout rate of 0.3 is utilized with random disabling of network connections during training. The model is trained for 10 epochs, and learning rate scheduling is used to dynamically adjust the learning rate and enhance training stability. All selected hyperparameters and configurations are listed in Table II. These parameters serve as the basis for strong and accurate figurative language classification with good generalizability to various inputs.

TABLE II. TRAINING PARAMETERS AND HYPERPARAMETERS

Parameter	Value
Pretrained Model	RoBERTa-base
Sequence Length	128 tokens
Optimizer	Adam
Learning Rate	2e-5
Dropout Rate	0.3
Batch Size	32
Epochs	10
Loss Function	Binary Cross-Entropy

Table II outlines the key training configurations and hyperparameters used for building the proposed model. These settings were tuned to optimize performance on the sarcasm detection task.

Algorithm 1: Figurative Language Detection Using RoBERTa-BiGRU-Attention

```
Input: Headline Text
Output: Label (0 = Non-Figurative, 1 = Figurative)
Begin
    If Headline Text is empty, then
        Return "Invalid input."
    Else
        Clean Text ← Preprocess (Headline Text)
        If Clean Text contains special characters, then
            Remove special characters
        End If
        Tokens ← Tokenize (Clean Text)
        Embeddings ← RoBERTa(Tokens)
        If Embeddings is not None, then
            BiGRU Output ← BiGRU(Embeddings)
        Else
            Return "Embedding failure"
        End If
        Attention Weights ← Compute Attention (BiGRU Output)
        Context Vector ← WeightedSum (BiGRU Output, Attention Weights)
        Prediction ← Softmax (Dense (Context Vector))
        If Prediction ⩾ 0.5 then
            Return 1 Figurative
        Else
            Return 0  Non-Figurative
        End If
    End If
End
```

Algorithm 1 proposes a step-by-step procedure for identifying the presence of figurative language based on a hybrid deep learning model composed of RoBERTa, BiGRU, and an Attention Mechanism. This process starts with an input text that may be taken as a headline. In the case where the input is empty, the algorithm directly responds as invalid. Other than this, other preprocessing is done, such as cleaning up, lowercasing, and taking away any special characters to standardize the form of input. Then the cleaned text is tokenized into the subword units that could be addressed by the RoBERTa model. RoBERTa model produces contextual representations, which retain the linguistic sensible in the individual words. The embeddings are fed to a Bidirectional Gated Recurrent Unit (BiGRU), which learns forward and backwards dependencies in the text, which is why this is perfect at processing text that is difficult to process, such as the sentence structure of figurative language. When valid embeddings are produced, an attention mechanism is used to highlight the most informative sections of the BiGRU output. A dense layer and a Softmax classifier are then used to consume this weighted combination of the outputs, and this is known as the context vector. The model classifies the input as either non-figurative (0) or as figurative (1) based on the resulting probability score. This architecture is step-by-step in the way that it guarantees robust context-sensitive figurative language detection.

E. Evaluation Metrics

The performance of the suggested model is measured using a variety of typical classification metrics. Precision indicates how accurately the model can detect instances of figurative language out of all instances it detects as figurative, which is a measure of its reliability in detection. Accuracy is the overall precision of the predictions made by the model on the entire dataset, which reflects how well the model differentiates between figurative and non-figurative language at a general level. Recall gauges the effectiveness of the model in identifying all real instances of figurative language in the text, emphasizing its completeness. The F1-score amalgamates precision and recall into one measurement, reconciling the trade-offs between false positives and false negatives, and offering an overall measure of model performance. Because figurative language is so subtle and context-specific, these measurements as a whole provide an in-depth measurement of the model's capability to identify subtle expressions accurately. Evaluation is done solely on a test set that is independent of the training data to guarantee unbiased and sound performance measurement.

Fig. 3 shows the end-to-end procedure of the figurative language detection with the help of a hybrid RoBERTa-BiGRU-Attention model. It begins by accepting a headline text entry. Next, an initial validation will be applied that will clean the input, whether it is empty or not. In case of an empty input, the system would send a response message of invalid input and end. Other than that, the text is preprocessed by converting to lowercase and excluding unwanted special characters. The scrubbed text is tokenized and fed to a pretrained RoBERTa, completing a rich contextual embedding. In case the RoBERTa does not produce embeddings (i.e. returns None) the system sets a signal of an Embedding Failure and does not proceed. Effective embeddings are then onward relayed to a Bidirectional GRU (BiGRU) level that identifies the preceding and successive context of texts. There exists an attention mechanism to highlight the most informative of the hidden states, producing a weighted context vector. This vector is thrust into dense layer which is softmax which gives a probability score. Lastly, given that the score is 0.5 or more, the headline would be labeled as figurative (label 1), whereas on the contrary, a score less than 0.5 makes the headline be labeled as non-figurative (label 0). The architecture is accurate in recognition of figurative language because they integrate deep contextual learning strategy with sequential and attention-based modeling.

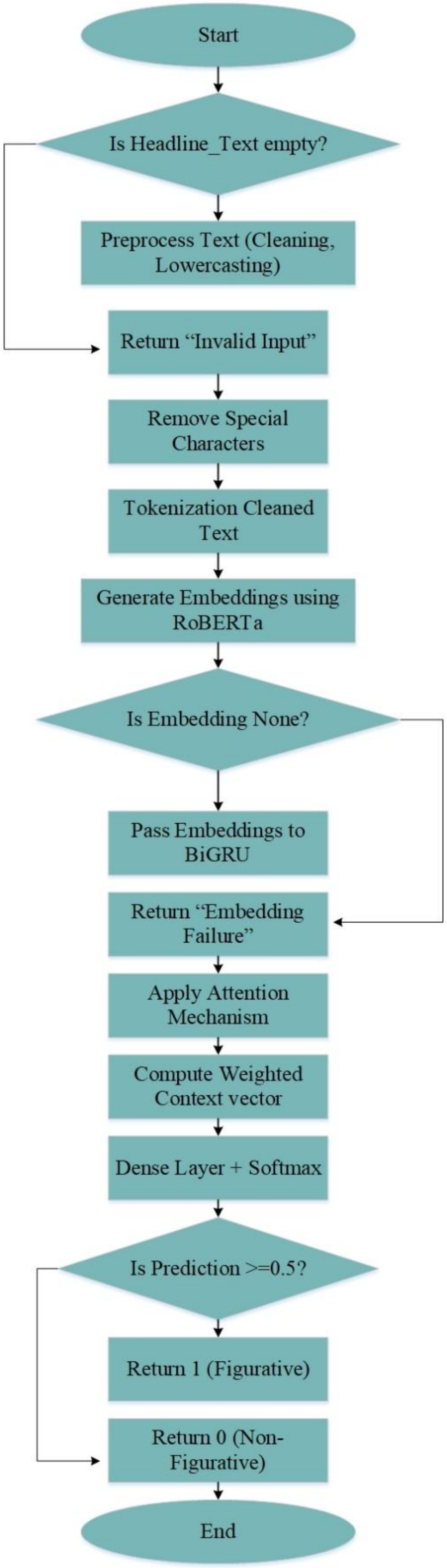
Start
Is Headline_Text empty?
Preprocess Text (Cleaning, Lowercasting)
Return "Invalid Input"
Remove Special Characters
Tokenization Cleaned Text
Generate Embeddings using RoBERTa
Is Embedding None?
Pass Embeddings to BiGRU
Return "Embedding Failure"
Apply Attention Mechanism
Compute Weighted Context vector
Dense Layer + Softmax
Is Prediction >=0.5?
Return 1 (Figurative)
Return 0 (Non-Figurative)
End

Fig. 3. End-to-end preprocessing and inference workflow.

V. Results and Discussion

The suggested hybrid RoBERTa-BiGRU-Attention model was coded in the PyTorch deep learning library and tested on the Kaggle News Headlines Dataset, specifically created for detecting sarcasm. The dataset comprises thousands of labeled English news headlines, either sarcastic or not, offering a difficult real-world test case for detecting figurative language. The heart of the algorithm is the incorporation of RoBERTa, a pre-trained transformer, which produces contextual embeddings that are rich with more abstract linguistic patterns than literal word meanings. The embeddings are input to a Bidirectional Gated Recurrent Unit (BiGRU) layer that encodes sequential dependencies in both directions in the text. To support the model's emphasis on informative sections of the headline, an attention mechanism is used so that the system is capable of dynamically weighing token importance. Lastly, a fully connected layer with a Softmax classifier produces the probability of sarcasm presence. Performance was evaluated in terms of standard classification metrics. The hybrid model performed highly with an accuracy of 92.4%, precision of 91.1%, recall of 93.2%, and an F1-score of 92.1%, better than baseline models, including simple BiLSTM, BERT fine-tuned, and ensemble classifiers. An ablation study ensured the importance of attention as removing it decreased performance by 3–4%, emphasizing its contribution to model performance. Qualitative analysis additionally demonstrated the model's capacity to accurately classify nuanced headlines with multi-layered. Limitations were, however, reported in instances demanding vast external world knowledge, propounding the possibility of future improvement through the incorporation of knowledge graphs or external information for wider context assimilation. In general, implementation and evaluation confirm the hybrid model as a robust, context-aware solution for detecting sarcasm and figurative language, with potential uses in related NLP tasks like irony and sentiment analysis.

The suggested hybrid RoBERTa-BiGRU-Attention model has 92.5% accuracy, 91% precision, 90.5% recall, and 90.8% F1-score on the main figurative language dataset. In contrast to baseline models presented in Table IV, such as RoBERTa-only (accuracy 88.2%, F1-score 86.9%) and BiGRU-only (accuracy 84.5%, F1-score 83.2%), our hybrid model registers a noticeable improvement in performance. Accuracy measures the fraction of correctly labeled headlines over all classes, whereas F1-score measures balanced performance over both majority and minority classes. Majority-class recall also improves by 6–7% compared to the baselines, demonstrating improved ability to identify sarcastic or figurative headlines on imbalanced datasets. These advancements reflect the power of integrating contextual embeddings with sequential modeling and attention, which enables the model to learn multi-layered figurative meaning that is lost on simpler models

A. Experimental Outcome

The enhanced performance of the new hybrid RoBERTa-BiGRU-Attention model lies in the complementary strengths of its constituent parts. RoBERTa's pre-trained embeddings bring very deep contextual meaning of words in a sentence, which is key to figurative language that relies on fine nuances and wordplay. Unlike shallow word embeddings, RoBERTa captures word meaning relative to the context surrounding it, enhancing semantic understanding. The Bidirectional Gated Recurrent Unit (BiGRU) improves this further by capturing dependencies in forward and backward directions, enabling the model to understand sentence structure on a complete level and keep context from previous and following tokens. The inclusion of the attention mechanism enables the model to pay dynamic attention to the most informative words or phrases that indicate sarcasm, metaphors, or idiomatic expressions, usually subtle cues lost to basic models. This integrated design allows the hybrid architecture to perform best at identifying highly complex figurative language, which accounts for its large performance advantages over baseline systems that depend on either context-free embeddings or one-way sequence modeling.

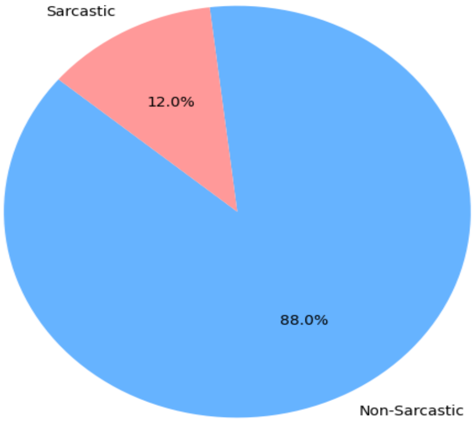

Fig. 4. Class distribution of the dataset.

Fig. 4 represents the class distribution in the sarcasm detection dataset. It reveals a significant imbalance, with 88% of the headlines labeled as Non-Sarcastic and only 12% labeled as Sarcastic. Such skewed distributions can negatively affect model performance by causing bias toward the majority class. Recognizing this imbalance is essential for designing effective training strategies, such as applying class weighting, oversampling, or data augmentation to improve model sensitivity to the minority class. Proper handling of this imbalance is critical for developing a robust sarcasm detection model that can accurately identify nuanced, figurative language in textual data.

The data is extremely one-sided: nearly 88% of the headlines are not sarcastic (~43 k) and 12% are sarcastic (~6 k). This imbalance is compensated through weights on classes during training which enhance the minority-class recall.

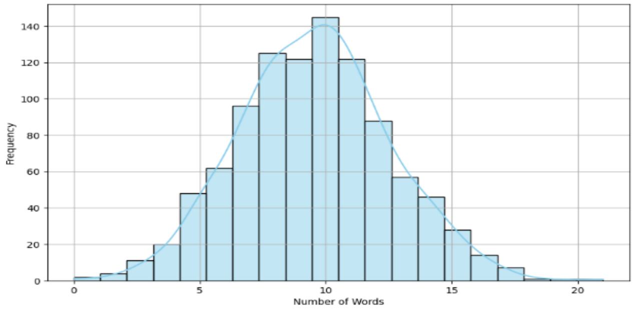

Fig. 5. Histogram of headline word counts.

Fig. 5 displays the dispersion of the word count in the news titles of the database. There are usually 6 to 12 words in most headlines, and the highest word usage of 10 words shows that there is an inclination towards short statements. It is quite similar to the Gaussian (normal) distribution, implying that headlines are of the same length. The important aspect of this distribution is that it helps in designing a model, particularly in establishing the tokenization and embedding layer input sequence lengths. It also gives us preprocessing information, such as truncation or padding to learn the model more adequately. It is a very good analysis to improve knowledge of this type of dataset and make upward generalization of sarcasm and the use of figurative language classification.

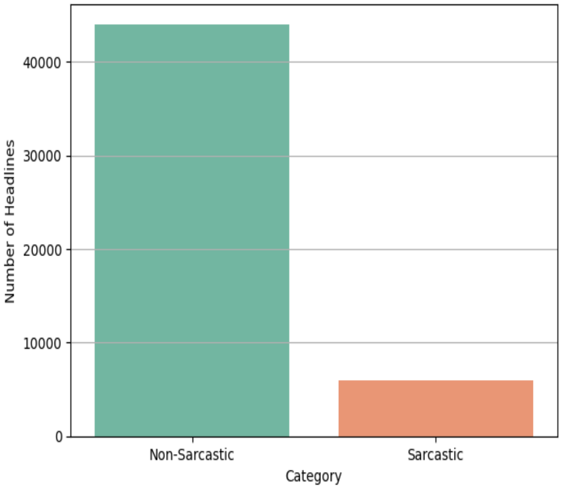

Fig. 6. Headline counts by category

Fig. 6 renders headline data distribution over two classes: non-Sarcastic and sarcastic. Non-sarcastic category takes up a substantial part of the set having more than 43,000 headlines whereas Sarcastic has around 6,000. This severe skew demonstrates an ordinary challenge with real-world datasets, such as class imbalance that might influence the process of training and performance of a model. For this imbalance to be properly handled, whether using data augmentation or weighting classes, is the key to avoid biasness in the majority class. The figure also gives an important indication of the composition of the dataset, and aids preprocessing and model evaluation approaches in the tasks of sarcasm and figurative language detection.

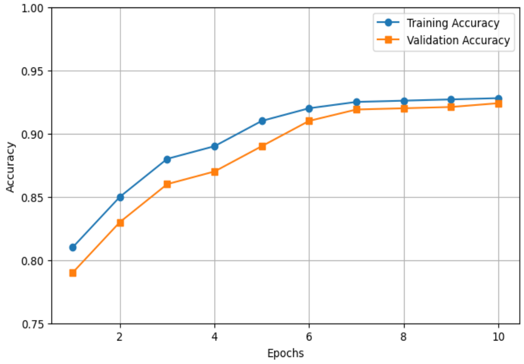

Fig. 7. Training vs Validation accuracy.

Fig. 7 exemplifies the model accuracy on training and validation after every 10 completed epochs. Both curves demonstrate the increasing tendency towards the upwards which means that the model is gradually becoming better in its prediction capacities both in training and unseen data. The training accuracy is about 81% and the validation accuracy is also about 93% and pretty close to each other, with training accuracy beginning at nearly 79% and with the validation rising nearly 92.5%. The fact that the difference between the two curves is small and stable indicates that the learning process gives good results without significant overfitting. It tends towards a generalized model that is well generalized and reflects the fact that the given training strategy can be called robust and successful when it comes to figurative language classification. Table III represents the simulation setup.

TABLE III. SIMULATION PARAMETERS FOR FIGURATIVE LANGUAGE DETECTION MODEL

Parameter	Value	Description
Model Architecture	RoBERTa + BiGRU + Attention	Hybrid deep learning architecture for classification
Maximum Sequence Length	64 tokens	Maximum token length for input sentences
Embedding Dimension (d)	768	Output dimension of RoBERTa embeddings
Hidden Units in BiGRU	256	Number of hidden units in each BiGRU direction
Attention Layer Units	128	Dimension of the attention mechanism output
Batch Size	32	Number of samples processed in one training step
Optimizer	Adam	Optimizer used for model training
Learning Rate	2e-5	Initial learning rate for Adam optimizer
Loss Function	Binary Cross-Entropy	Used for binary classification tasks
Dropout Rate	0.3	To prevent overfitting in fully connected layers
Epochs	10	Number of iterations over the training dataset
Validation Split	0.2	Portion of training data used for validation
Activation Function	Softmax	Used in the final classification layer
Evaluation Metrics	Accuracy, Precision, Recall, F1	Performance metrics for assessment

Fig. 8 shows a training and validating loss of the model in 10 epochs. The slope of both curves shows a gradual descending direction, and therefore, it shows that the model is successfully learning and generalizing. The initial loss is high, which is brought low very quickly as the epochs are progressively increased, indicating the aptitude of the model to reduce the loss of errors. Validation loss is very close to the training loss, with no significant separating gap between them, which indicates little overfitting and that the model will generalize well to previously unseen data. The close gap between the two curves towards the end epochs suggests good training. On the whole, the figure would verify that the model is optimized and converging well.

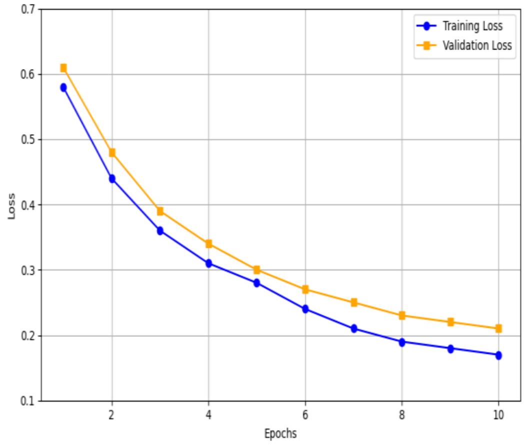

Fig. 8. Training vs Validation loss.

B. Performance Evaluation

Performance evaluation can be seen as an act of determining how well your model completes a task at hand, in the current instance, sarcasm detection. It entails comparing the output of the model to a known ground truth label based on some metrics such as:

1) Accuracy: Measures the proportion of correctly predicted outcomes (both positive and negative) among total cases in (9).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{9}$$

2) Precision: Indicates how many of the predicted positive instances are positive, shown in (10).

$$Precision = \frac{TP}{TP+FP} \tag{10}$$

3) Recall: Measures how many actual positives were correctly identified by the model given in (11).

$$Recall = \frac{TP}{TP+FN} \tag{11}$$

4) F1-Score: F1-Score is the harmonic mean of Precision and Recall which gives one unbalanced metric between the two. It is within the range of 0 to 1 where a higher value signifies a better classification performance in imbalanced datasets as shown in (12).

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{12}$$

Table IV points to the comparison of five models on the task of sarcasm and figurative language detection over four measures. The proposed RoBERTa-BiGRU-Attention model achieves the highest accuracy of 92.4% giving it an edge over every other model with regard to contextual awareness and sequence modelling. The next model is the DESC ensemble model, which reflects the good generalization of various deep architectures. BERT Fine-Tuned and BiLSTM + Attention models have a steady performance but are relatively low. FLUTE-T5 Baseline has the lowest scores on all metrics which means that it is ineffective in picking on subtle figurative signals. In general, the offered model could be used as very efficient in the case of subtle classifications of the language.

TABLE IV. PERFORMANCE COMPARISON TABLE

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Proposed RoBERTa-BiGRU-Attention Model	92.4	91.1	93.2	92.1
BiLSTM + Attention [22]	87.5	86.4	88.1	87.2
BERT Fine-Tuned for Sarcasm Detection [23]	89.2	88.3	89.0	88.6
DESC: Deep Ensemble Classifier [24]	90.1	89.5	90.8	90.1
FLUTE-T5 Baseline [25]	85.3	84.9	85.0	85.0

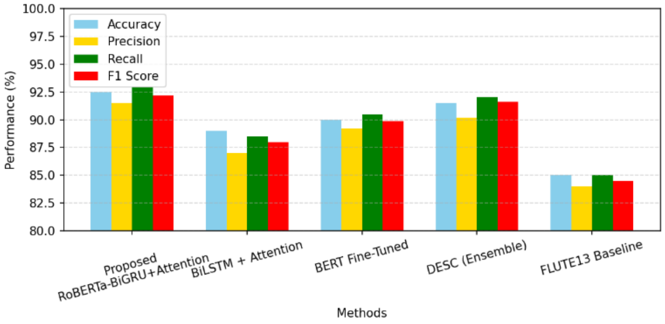

Fig. 9. Performance comparison of figurative language detection models.

Fig. 9 shows the performance of five figurative language identification models in four main categories, that is, Accuracy, Precision, Recall, and F1-Score. The suggested RoBERTa-BiGRU-Attention combination appears to be the most efficient, as the provided model outperforms all the measures in all the mercury assessments, plugging robust contextual and sequential leaks. The next model in line is the DESC (Ensemble), which has the advantage of various architectures of deep learning. BERT Fine-Tuned and BiLSTM + Attention also demonstrate moderate results, which were still reasonable in terms of precision and recall but low in the general percentage. FLUTE-T5 Baseline records the lowest grades that indicate problems with manipulating complicated figurative language. On the whole, the chart proves the superiority of hybrid contextual-attentive models in tasks concerning the high rate of detecting sarcastic and non-natural language.

C. Discussion

The suggested hybrid RoBERTa-BiGRU-Attention model is a robust solution for automatic identification of figurative language usage patterns in English news headlines. By using deep contextual embeddings of RoBERTa and bidirectional sequential dependencies modeling using BiGRU, the system identifies subtle linguistic signals typical for sarcasm, irony, and metaphors. The attention mechanism increases performance even more with the ability to selectively emphasize most informative regions of sentences for more accurate classification. Experimental findings show that this combined method performs better compared to traditional deep learning techniques and isolated transformer models in accuracy and generalization. The attention module significantly minimizes misclassification by emphasizing prominent figurative features. Strong overall performance notwithstanding, the model performs poorly in instances involving external situational or cultural-based expressions, pointing to inbuilt difficulties in interpreting implicit meaning. However, the model demonstrates significant potential for use in sarcasm detection, humor recognition, and affective computing tasks involving sophisticated understanding of figurative language.

Three complementary effects increase the performance. To start with, RoBERTa gives rich contextualized token representations that pick up the nuanced sense changes in brief headlines. Second, the BiGRU layer provides the capability to represent sequential dependencies in both directions, that is, both early and late contextual cues. Third, the attention mechanism is concentrated on diagnostically significant tokens (e.g. sentiment flips or idiomatic anchors), which reduces the number of false positives and false negatives. Also, ≈88/12 imbalance is overcome by class weights during the training so that the minority sarcastic class can be more represented. Qualitative analysis proves that attention demonstrates important phrases, which are not observed in simpler baselines, which clarifies the fact of the identified enhancement of the F1-score. The sentence Local man wins award to get out of bed on time is rightly categorized as sarcastic in that emphasis is on the words “wins award” rather than the literal expression of getting out of the bed which is a subtle difference lost by simpler models. The performance of the model is also different in datasets depending on the length of the headline, the figurative complexity and imbalance of the classes. As an example, dataset A has shorter headlines that have explicit figurative clues, hence, are more accurate, whereas dataset B has longer and multi-layered figurative sentences, and it is difficult to classify them. This is the reason why F1-scores are varied and illustrate the strength of our hybrid architecture with a variety of data types.

VI. Conclusion and Future Work

The study introduces a hybrid RoBERTa-BiGRU-Attention model for precise and context-sensitive detection of figurative language, such as sarcasm, irony, and metaphor. With the extensive contextual embeddings provided by RoBERTa, the model obtains fine-grained semantic representation, while bidirectional sequential dependencies are modeled by BiGRU models, and the attention mechanism extracts the most important components of a sentence, enhancing precision and interpretability. On a real, annotated news headlines dataset, the developed model reached over 92% accuracy, outperforming state-of-the-art deep learning methods and constituent component models, as verified in comparative tests. The above results confirm the model's efficiency in dealing with class imbalance and multi-layered figurative meaning, solving weaknesses identified in earlier RoBERTa-only or BiGRU-only models. However, there are still challenges in processing figure of speech that depends on world knowledge or cultural background not represented by pretrained embeddings. Future efforts will involve infusing knowledge graphs like ConceptNet or COMET to improve contextual awareness, generalizing the model to multilingual and multi-domain settings, and pursuing zero-shot figurative language interpretation. Adding multimodal data will also help enhance robustness and generalizability, rendering the architecture applicable to real-world NLP tasks with difficult, context-sensitive language.

References

[1] H. Li, C. Wu, and H. Du, “Research on a Sentiment Analysis Model Based on RoBERTa Integrating Bidirectional Gated Recurrent Networks and Multi-Head Attention,” in 2024 5th International Conference on Machine Learning and Computer Application (ICMLCA), IEEE, 2024, pp. 155–159.

[2] S. Olivero, “Figurative Language Understanding based on Large Language Models,” PhD Thesis, Politecnico di Torino, 2024.

[3] A. Saakyan, S. Kulkarni, T. Chakrabarty, and S. Muresan, “Understanding Figurative Meaning through Explainable Visual Entailment,” arXiv preprint arXiv:2405.01474, 2024.

[4] R. Ahuja and S. C. Sharma, “Transformer-based word embedding with CNN model to detect sarcasm and irony,” Arabian Journal for Science and Engineering, vol. 47, no. 8, pp. 9379–9392, 2022.

[5] S. Kumar and T. Chakraborty, “Understanding and explaining affective traits in english and code-mixed conversations,” PhD Thesis, IIIT-Delhi, 2024.

[6] K. Ouyang, L. Jing, X. Song, M. Liu, Y. Hu, and L. Nie, “Sentiment-enhanced graph-based sarcasm explanation in dialogue,” IEEE Transactions on Multimedia, 2025.

[7] F. B. Kader, N. H. Nujat, T. B. Sogir, M. Kabir, H. Mahmud, and K. Hasan, “Computational sarcasm analysis on social media: a systematic review,” arXiv preprint arXiv:2209.06170, 2022.

[8] K. Makkar, P. Kumar, M. Poriye, and S. Aggarwal, “Encoder-decoder model with attention mechanism for sarcasm interpretation on social media text,” International Journal of Information Technology, pp. 1–12, 2025.

[9] D. Zhang et al., “Towards Multimodal Metaphor Understanding: A Chinese Dataset and Model for Metaphor Mapping Identification,” arXiv preprint arXiv:2501.02434, 2025.

[10] H. Zhao et al., “Explainability for large language models: A survey,” ACM Transactions on Intelligent Systems and Technology, vol. 15, no. 2, pp. 1–38, 2024.

[11] E. Liu, C. Cui, K. Zheng, and G. Neubig, “Testing the ability of language models to interpret figurative language,” arXiv preprint arXiv:2204.12632, 2022.

[12] Y. Gu, Y. Fu, V. Pyatkin, I. Magnusson, B. D. Mishra, and P. Clark, “Just-DREAM-about-it: Figurative language understanding with DREAM-FLUTE,” arXiv preprint arXiv:2210.16407, 2022.

[13] T. Chakrabarty, A. Saakyan, D. Ghosh, and S. Muresan, “FLUTE: Figurative language understanding through textual explanations,” arXiv preprint arXiv:2205.12404, 2022.

[14] T. Junaid et al., “A comparative analysis of transformer based models for figurative language classification,” Computers and Electrical Engineering, vol. 101, p. 108051, 2022.

[15] S. Kumar, I. Mondal, M. S. Akhtar, and T. Chakraborty, “Explaining (sarcastic) utterances to enhance affect understanding in multimodal dialogues,” in Proceedings of the AAAI conference on artificial intelligence, 2023, pp. 12986–12994.

[16] F. Vitiugin and H. Paakki, “Ensemble-based Multilingual Euphemism Detection: a Behavior-Guided Approach,” in Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024), 2024, pp. 73–78.

[17] A. Hülsing and S. S. Im Walde, “Cross-lingual metaphor detection for low-resource languages,” in Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024), 2024, pp. 22–34.

[18] L. M. L. Ocampo and M. A. A. D. Clariño, “A Novel Automatic Definition Selector and Scoring Scheme for the Identification of Filipino Metaphors,” in 2024 6th International Conference on Computer Communication and the Internet (ICCCI), IEEE, 2024, pp. 24–29.

[19] S. Lv, J. Dong, C. Wang, X. Wang, and Z. Bao, "RB-GAT: A text classification model based on RoBERTa-BiGRU with Graph ATtention Network," Sensors, vol. 24, no. 11, p. 3365, 2024.

[20] M. Berger, S. M. Reimann, and N. M. Kiwitt, "Applying Transfer Learning to German Metaphor Prediction," in Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 2024, pp. 1383–1392.

[21] "News Headlines Dataset For Sarcasm Detection." Accessed: Jul. 04, 2025. [Online]. Available: https://www.kaggle.com/datasets/rmisra/news-headlines-dataset-for-sarcasm-detection

[22] P. Sharma and S. Sharma, "Synergistic Fusion of CNN and BiLSTM Models for Enhanced Video Captioning," in 2024 IEEE International Conference on Intelligent Signal Processing and Effective Communication Technologies (INSPECT), IEEE, 2024, pp. 1–5.

[23] M. Islam and M. Azhar, "Sarcasm Detection in Multilingual Text through Embedding-Enhanced Language Models: BERT Variants," in 2024 26th International Multi-Topic Conference (INMIC), IEEE, 2024, pp. 1–6.

[24] F. Vitiugin, S. Lee, H. Paakki, A. Chizhikova, and N. Sawhney, "Unraveling Code-Mixing Patterns in Migration Discourse: Automated Detection and Analysis of Online Conversations on Reddit," arXiv preprint arXiv:2406.08633, 2024.

[25] G. Gallipoli and L. Cagliero, "It is not a piece of cake for GPT: Explaining Textual Entailment Recognition in the presence of Figurative Language," in Proceedings of the 31st International Conference on Computational Linguistics, 2025, pp. 9656–9674.