International Journal for Novel Research in Economics , Finance and Management
www.ijnrefm.com
ISSN:3048-7722
Volume 4, Issue 4, July-Aug-2026, PP: 01-11

An Intelligent Deep Learning Framework for Sarcasm Detection in News Headlines

G.Sudheer Kumar[1], K. Sindhu[2]

[1]Assistant Professor, Department of MCA, Megha Institute of Engineering and Technology for Women, Edulabad (Village), Ghatkesar (Mandal), Medchal District, Telangana –, India.

[2]MCA Student, Department of MCA, Megha Institute of Engineering and Technology for Women, Edulabad (Village), Ghatkesar (Mandal), Medchal District, Telangana –, India.

***Abstract* – Sarcasm is a sophisticated form of linguistic expression in which the intended meaning differs from the literal interpretation of the text, making automatic sentiment analysis a challenging task for conventional Natural Language Processing (NLP) systems. News headlines frequently employ sarcastic language to attract readers, convey criticism, or present humorous perspectives, thereby increasing the complexity of accurate text classification. This paper presents an intelligent sarcasm detection framework that integrates Machine Learning (ML) and Deep Learning (DL) techniques to improve the automatic identification of sarcastic news headlines. The proposed methodology utilizes comprehensive text preprocessing followed by feature representation using CountVectorizer and Word2Vec embeddings. Multiple learning models, including AdaBoost, Extreme Gradient Boosting (XGBoost), Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Bidirectional Long Short-Term Memory (BiLSTM), and a Hybrid CNN–BiLSTM architecture, are trained and comparatively evaluated using the Kaggle Sarcasm Detection News Headlines Dataset. To enhance model transparency, Local Interpretable Model-Agnostic Explanations (LIME) is incorporated for interpreting prediction outcomes, while t-distributed Stochastic Neighbor Embedding (t-SNE) is employed for feature visualization. Experimental analysis demonstrates that the RNN model delivers the highest overall performance, achieving an accuracy of 79.88% together with an F1-score of 0.76, indicating its superior capability for identifying contextual patterns associated with sarcastic language. The proposed framework offers an effective and explainable solution for sarcasm detection and can support numerous natural language understanding applications, including sentiment analysis, opinion mining, intelligent recommendation systems, and social media analytics.**

***Keywords* - Deep Learning, Sarcasm Detection, Natural Language Processing (NLP), News Headline Analysis, Sentiment Analysis, Text Classification, Artificial Intelligence**

I. INTRODUCTION

The rapid growth of digital communication platforms has transformed the way information is created, distributed, and consumed across the world. Online newspapers, news portals, blogs, and social media platforms generate an enormous volume of textual information every day, making automated text analysis an essential component of modern Natural Language Processing (NLP) research. While significant progress has been achieved in areas such as sentiment analysis, text classification, emotion recognition, opinion mining, and fake news detection, accurately interpreting human language remains a difficult challenge because natural language often contains ambiguity, irony, sarcasm, metaphor, and contextual expressions. Among these linguistic phenomena, sarcasm represents one of the most challenging problems because the intended meaning frequently contradicts the literal interpretation of the words being used. As a result, traditional sentiment analysis techniques often misclassify sarcastic statements, leading to inaccurate understanding of user opinions and emotional intent.

Sarcasm is commonly employed to express criticism, humor, dissatisfaction, or irony without directly stating negative opinions. Instead of conveying sentiment explicitly, sarcastic expressions rely heavily on contextual understanding, semantic relationships, background knowledge, and implicit meaning. For example, a headline containing apparently positive words may actually communicate a negative or humorous message depending on the surrounding context. This characteristic makes sarcasm particularly difficult for conventional machine learning models that primarily depend on lexical information and surface-level textual features. Consequently, developing intelligent computational models capable of recognizing sarcasm has become an important research area within artificial intelligence and computational linguistics.

The increasing popularity of online news platforms has further emphasized the importance of sarcasm detection. News organizations frequently publish creative and attention-grabbing headlines to attract readers while simultaneously expressing irony or satire regarding political events, social issues, entertainment, and public affairs. Such headlines often contain subtle linguistic cues that are difficult for automated systems to interpret correctly. Failure to identify sarcastic content may significantly reduce the effectiveness of downstream applications such as news recommendation systems, sentiment analysis, opinion mining, public opinion monitoring, and misinformation detection. Therefore, reliable sarcasm detection contributes directly to improving the performance of numerous intelligent language processing systems.

International Journal for Novel Research in Economics , Finance and Management
www.ijnrefm.com
ISSN:3048-7722
Volume 4, Issue 4, July-Aug-2026, PP: 01-11

Traditional approaches to sarcasm detection primarily relied on manually engineered linguistic features, lexical resources, syntactic analysis, and rule-based systems. Although these methods demonstrated reasonable performance under limited experimental conditions, they generally struggled to capture complex contextual relationships and semantic dependencies present in natural language. The rapid advancement of Machine Learning (ML) introduced statistical learning algorithms capable of identifying hidden patterns from large collections of textual data. Ensemble learning methods such as AdaBoost and Extreme Gradient Boosting (XGBoost) have demonstrated strong classification capability by combining multiple weak learners to improve prediction accuracy. These algorithms effectively learn discriminative textual features while reducing classification errors through iterative optimization.

More recently, Deep Learning (DL) has significantly advanced the field of natural language understanding by enabling automatic feature extraction directly from raw textual data. Unlike conventional machine learning techniques that require handcrafted features, deep neural networks learn hierarchical semantic representations capable of modeling complex contextual relationships within language. Convolutional Neural Networks (CNNs) effectively identify local semantic patterns and important phrase-level features, whereas Recurrent Neural Networks (RNNs) capture sequential dependencies present within textual information. Bidirectional Long Short-Term Memory (BiLSTM) networks further improve contextual understanding by processing textual sequences in both forward and backward directions, enabling more comprehensive interpretation of sentence semantics. Hybrid architectures that combine CNN and BiLSTM further enhance feature representation by simultaneously learning local contextual patterns and long-range semantic dependencies.

An equally important challenge in artificial intelligence involves the interpretability of deep learning models. Although complex neural networks often achieve high prediction accuracy, they frequently operate as "black-box" systems whose internal decision-making processes remain difficult to understand. To improve model transparency, researchers increasingly employ Explainable Artificial Intelligence (XAI) techniques that provide human-understandable explanations for prediction outcomes. One widely adopted explainability method is Local Interpretable Model-Agnostic Explanations (LIME), which explains individual predictions by identifying the textual features that contribute most significantly to the model's classification decisions. Explainable AI not only increases user confidence but also enables researchers to validate whether the model relies on meaningful linguistic patterns during sarcasm detection.

In addition to explainability, effective text representation plays a fundamental role in improving sarcasm classification performance. Traditional frequency-based representations such as CountVectorizer convert textual information into numerical feature vectors by capturing word occurrence frequencies, while semantic embedding techniques such as Word2Vec generate dense vector representations that preserve semantic relationships among words. Combining these complementary feature representation techniques enables learning algorithms to exploit both statistical and semantic characteristics of textual information, thereby improving overall classification performance.

Motivated by these challenges, this study presents a comprehensive sarcasm detection framework that integrates both Machine Learning and Deep Learning techniques for classifying sarcastic and non-sarcastic news headlines. The proposed methodology utilizes extensive text preprocessing followed by feature extraction using CountVectorizer and Word2Vec embeddings. Multiple classification algorithms including AdaBoost, XGBoost, CNN, RNN, BiLSTM, and a Hybrid CNN–BiLSTM architecture are trained and comparatively evaluated using the Kaggle Sarcasm Detection News Headlines Dataset, which contains 28,619 news headlines. Furthermore, LIME is employed to provide interpretable explanations for model predictions, while t-SNE visualization is utilized to analyze feature distributions within the learned embedding space. The comparative experimental evaluation identifies the most effective classification model while providing valuable insights into the strengths of both traditional machine learning and modern deep learning approaches for sarcasm detection.

The primary objective of this research is to develop an accurate, explainable, and computationally efficient framework capable of recognizing sarcastic news headlines using advanced artificial intelligence techniques. By integrating robust preprocessing, semantic feature representation, ensemble learning, deep neural networks, explainable AI, and visualization techniques into a unified framework, the proposed system contributes toward improving automated language understanding and supports future developments in sentiment analysis, intelligent recommendation systems, social media monitoring, and opinion mining applications.

II. LITERATURE SURVEY

The rapid advancement of Natural Language Processing (NLP) and Artificial Intelligence (AI) has significantly improved the automatic understanding of textual information across various domains, including sentiment analysis, opinion mining, fake news detection, question answering, and emotion recognition. Despite these developments, sarcasm detection remains one of the most challenging text classification problems because sarcastic statements often convey meanings that contradict their literal expressions. Unlike conventional sentiment classification tasks that primarily rely on explicit

emotional words, sarcasm requires contextual interpretation, semantic understanding, and recognition of implicit linguistic cues. Consequently, researchers have explored numerous Machine Learning (ML), Deep Learning (DL), and Explainable Artificial Intelligence (XAI) techniques to improve sarcasm detection accuracy while enhancing model interpretability.

Early research on sarcasm detection primarily employed rule-based systems and manually engineered linguistic features. These approaches relied on lexical resources, sentiment dictionaries, syntactic structures, punctuation patterns, and handcrafted language rules to identify sarcastic expressions. Although such techniques performed reasonably well for limited datasets, they exhibited poor generalization capability because sarcastic language frequently depends on contextual meaning rather than isolated keywords. Furthermore, manually constructing linguistic rules requires significant domain expertise and cannot effectively accommodate the continuously evolving nature of online language. These limitations encouraged researchers to investigate data-driven machine learning approaches capable of automatically learning complex textual patterns from large datasets.

The adoption of Machine Learning significantly improved sarcasm classification by enabling statistical models to learn discriminative textual features directly from labeled datasets. Gupta et al. proposed a sarcasm detection framework utilizing supervised machine learning algorithms together with natural language processing techniques for classifying textual content. Their study demonstrated that machine learning algorithms effectively outperform traditional rule-based systems by learning contextual relationships from annotated training data. However, the performance of conventional machine learning models remained highly dependent on feature engineering, requiring carefully designed textual representations to achieve satisfactory classification accuracy.

Feature representation has become an essential component of text classification research. Traditional text vectorization techniques such as CountVectorizer and Term Frequency–Inverse Document Frequency (TF–IDF) transform textual documents into numerical feature vectors suitable for machine learning algorithms. Although frequency-based representations effectively capture word occurrence statistics, they often fail to preserve semantic relationships among words. To overcome this limitation, researchers introduced distributed word representation techniques such as Word2Vec, which generate dense vector embeddings capable of representing semantic similarity between words. Word embedding techniques significantly improved machine learning performance by enabling models to learn contextual information beyond simple word frequencies, thereby enhancing sarcasm detection capability.

Recent years have witnessed remarkable progress through the application of Deep Learning to natural language understanding. Unlike traditional machine learning algorithms that depend heavily on handcrafted features, deep neural networks automatically extract hierarchical semantic representations from textual data. Convolutional Neural Networks (CNNs) have demonstrated considerable success in identifying local phrase-level features and contextual patterns useful for text classification. Their ability to recognize meaningful n-gram structures makes CNNs highly effective for capturing local semantic information associated with sarcastic language. Several studies have shown that CNN-based architectures achieve improved classification accuracy compared with traditional machine learning algorithms while reducing dependence on manual feature engineering.

Sequential neural architectures have further advanced sarcasm detection by modeling long-range contextual dependencies within language. Recurrent Neural Networks (RNNs) process textual information sequentially, enabling the model to retain contextual information from previous words while analyzing subsequent tokens. This sequential learning capability allows RNNs to better understand sentence-level semantics compared with conventional feedforward neural networks. However, standard RNNs may encounter difficulties when modeling long textual sequences because of vanishing gradient problems that limit long-term contextual learning.

To overcome these limitations, researchers introduced Long Short-Term Memory (LSTM) and Bidirectional Long Short-Term Memory (BiLSTM) networks. BiLSTM architectures analyze textual sequences in both forward and backward directions, enabling simultaneous utilization of preceding and succeeding contextual information. This bidirectional processing significantly improves semantic understanding, particularly for linguistic phenomena such as sarcasm where contextual relationships between distant words strongly influence interpretation. Numerous investigations have reported that BiLSTM consistently outperforms traditional sequential models for sentiment analysis, sarcasm detection, and emotion recognition because of its enhanced capability for modeling contextual dependencies.

Hybrid deep learning architectures have also received considerable research attention. By combining CNN with BiLSTM, researchers have developed models capable of simultaneously extracting local semantic features and long-range contextual dependencies. CNN layers efficiently identify important local phrases, while BiLSTM layers capture sequential contextual relationships throughout the sentence. Such hybrid architectures leverage the complementary strengths of both neural network models, leading to improved classification performance for complex natural language tasks including sarcasm detection. Comparative experimental studies consistently demonstrate that hybrid CNN–BiLSTM models provide

more comprehensive textual representations than individual neural architectures.

Beyond prediction accuracy, Explainable Artificial Intelligence (XAI) has become an increasingly important area of research. Although deep neural networks achieve high classification performance, their complex internal computations often make prediction outcomes difficult to interpret. To improve model transparency, researchers have adopted Local Interpretable Model-Agnostic Explanations (LIME), which provides human-understandable explanations by identifying the textual features contributing most significantly to individual classification decisions. Explainable AI improves user confidence, facilitates model validation, and supports practical deployment of artificial intelligence systems within real-world decision-making environments.

Visualization techniques further enhance understanding of learned feature representations. t-distributed Stochastic Neighbor Embedding (t-SNE) has been widely utilized for projecting high-dimensional word embeddings into low-dimensional spaces suitable for visual analysis. By illustrating clustering behavior among sarcastic and non-sarcastic textual samples, t-SNE enables researchers to evaluate the effectiveness of learned semantic representations while providing additional insight into model behavior. Such visualization techniques complement explainable AI by improving interpretation of deep learning models beyond conventional quantitative performance metrics.

Despite substantial progress in sarcasm detection research, several challenges continue to limit practical performance. Traditional machine learning algorithms frequently depend on manually engineered features that may not adequately capture implicit semantic relationships. Deep learning models often achieve superior prediction accuracy but require large labeled datasets and substantial computational resources. Furthermore, many existing studies primarily emphasize classification performance while providing limited interpretability regarding the reasoning behind prediction outcomes. The combination of high-performance classification models with explainable artificial intelligence therefore remains an important research direction for practical sarcasm detection systems.
Motivated by these research challenges, the present study proposes an integrated sarcasm detection framework that combines CountVectorizer, Word2Vec, AdaBoost, XGBoost, CNN, RNN, BiLSTM, Hybrid CNN–BiLSTM, LIME, and t-SNE visualization within a unified intelligent architecture. The proposed methodology performs comprehensive text preprocessing, semantic feature extraction, comparative evaluation of multiple machine learning and deep learning models, and explainable prediction analysis using the Kaggle Sarcasm Detection News Headlines Dataset containing 28,619 news headlines. Through the integration of predictive performance and model interpretability, the proposed framework aims to improve sarcasm classification accuracy while providing meaningful explanations that enhance user confidence and facilitate practical deployment across various natural language processing applications.

III. PROPOSED METHODOLOGY

The proposed framework presents an intelligent Natural Language Processing (NLP) architecture for automatically identifying sarcastic and non-sarcastic news headlines using a combination of Machine Learning (ML), Deep Learning (DL), and Explainable Artificial Intelligence (XAI) techniques. The methodology integrates comprehensive text preprocessing, feature representation, multiple classification algorithms, explainability analysis, and visualization into a unified framework. Rather than relying on a single learning model, the proposed system performs comparative evaluation using AdaBoost, Extreme Gradient Boosting (XGBoost), Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Bidirectional Long Short-Term Memory (BiLSTM), and a Hybrid CNN–BiLSTM model to determine the most effective approach for sarcasm detection. The overall workflow is designed to improve classification accuracy while maintaining model transparency and interpretability.

The methodology begins with the dataset acquisition stage, where the Kaggle Sarcasm Detection News Headlines Dataset is utilized for model development and evaluation. The dataset consists of 28,619 news headlines, each annotated as either sarcastic or non-sarcastic. Since the dataset contains headlines collected from multiple online news sources, it provides diverse linguistic expressions, writing styles, and contextual variations that enable the proposed framework to learn realistic sarcasm patterns. The availability of balanced textual samples allows both machine learning and deep learning models to perform reliable classification while minimizing prediction bias.

After dataset collection, the textual data undergo extensive text preprocessing to improve data quality before feature extraction. Raw news headlines frequently contain punctuation marks, special symbols, inconsistent capitalization, stop words, and unnecessary textual elements that may negatively influence classification performance. Therefore, preprocessing operations are applied sequentially, including text normalization, conversion to lowercase characters, punctuation removal, elimination of stop words, tokenization, and stemming or lemmatization where appropriate. These preprocessing operations simplify textual representation while preserving meaningful semantic information required for sarcasm classification. The cleaned textual corpus subsequently becomes suitable for numerical feature extraction and deep learning model training.

Following preprocessing, the framework performs feature representation, where textual information is transformed into numerical vectors that can be processed by machine learning algorithms. The proposed methodology employs two complementary feature extraction techniques: CountVectorizer and Word2Vec. CountVectorizer converts news headlines into frequency-based numerical vectors by calculating the occurrence of each word within the corpus. This representation effectively captures lexical information required for traditional machine learning algorithms. Simultaneously, Word2Vec generates dense semantic word embeddings that preserve contextual relationships among words appearing in similar linguistic environments. Unlike frequency-based representations, Word2Vec enables deep learning models to understand semantic similarity and contextual meaning, thereby improving sarcasm detection performance for linguistically complex headlines.

Once feature extraction is completed, the generated feature vectors are supplied to the Machine Learning Classification Module, where two ensemble learning algorithms are trained independently. The first classifier employs AdaBoost, which constructs a strong predictive model by iteratively combining multiple weak learners and assigning greater importance to previously misclassified samples. This adaptive learning strategy progressively improves classification accuracy by reducing prediction errors during successive training iterations. The second classifier utilizes Extreme Gradient Boosting (XGBoost), an optimized gradient boosting algorithm capable of learning complex nonlinear relationships while minimizing overfitting through efficient regularization techniques. Both ensemble models are trained using the CountVectorizer feature representation and subsequently evaluated to determine their effectiveness for sarcasm detection.

The methodology further incorporates several Deep Learning models to automatically learn hierarchical semantic representations directly from textual data. Unlike traditional machine learning approaches that depend primarily on manually engineered feature representations, deep neural networks automatically discover complex contextual patterns during training.

The first deep learning model is the Convolutional Neural Network (CNN). The CNN architecture receives Word2Vec embeddings as input and performs convolution operations to identify meaningful local semantic patterns within news headlines. Convolutional filters capture important phrase-level information while pooling operations reduce dimensionality and retain highly discriminative textual features. The extracted feature maps are subsequently processed through fully connected layers to perform sarcasm classification.

The second deep learning architecture utilizes a Recurrent Neural Network (RNN). Since sarcasm often depends on sequential contextual relationships between words, RNN processes each headline sequentially while retaining information from previously encountered words. This sequential learning capability enables the network to capture sentence-level dependencies that contribute significantly to understanding sarcastic language. Experimental evaluation later demonstrates that the RNN model provides the highest overall classification performance among all evaluated approaches.

The proposed framework additionally employs a Bidirectional Long Short-Term Memory (BiLSTM) network. Unlike standard recurrent neural networks, BiLSTM simultaneously analyzes textual sequences in both forward and backward directions, enabling the model to utilize contextual information from preceding and succeeding words. Bidirectional learning significantly improves semantic understanding, particularly for sarcasm detection where interpretation frequently depends on long-range contextual relationships distributed throughout the sentence.

To further improve feature learning capability, the methodology incorporates a Hybrid CNN–BiLSTM architecture. Within this model, CNN layers first extract local semantic features and important phrase-level information from Word2Vec embeddings. The resulting feature representations are subsequently processed by BiLSTM layers, which capture sequential contextual dependencies across the entire headline. By combining local feature extraction with bidirectional contextual modeling, the hybrid architecture attempts to exploit the complementary strengths of both neural network models for improved sarcasm classification.

Following model training, the proposed methodology performs hyperparameter optimization to improve prediction performance across all machine learning and deep learning models. Various training parameters including learning rate, batch size, number of epochs, optimizer configuration, network depth, and model-specific parameters are carefully adjusted to maximize classification accuracy while minimizing overfitting. The optimized models are subsequently evaluated using the testing dataset to ensure reliable generalization capability under unseen textual data.

To improve model transparency, the proposed framework incorporates Explainable Artificial Intelligence (XAI) using Local Interpretable Model-Agnostic Explanations (LIME). After the classification model generates a prediction, LIME identifies the individual words or phrases that contribute most significantly to the predicted sarcasm label. These explanations enable researchers and end users to understand why a particular news headline has been classified as sarcastic or non-sarcastic. Explainability therefore improves confidence in the prediction process while facilitating validation of model behavior for practical deployment.

The methodology also includes t-distributed Stochastic Neighbor Embedding (t-SNE) for feature visualization. Since Word2Vec generates high-dimensional semantic embeddings that cannot be directly interpreted visually, t-SNE projects these embeddings into a two-dimensional feature space suitable for graphical analysis. The resulting visualization illustrates the clustering behavior of sarcastic and non-sarcastic headlines, providing valuable insight into feature separability and the effectiveness of learned semantic representations.

Finally, model performance is evaluated using standard classification metrics including Accuracy, Precision, Recall, and F1-score. Comparative analysis of all machine learning and deep learning models enables objective identification of the most suitable algorithm for sarcasm detection. According to the reported experimental results, the Recurrent Neural Network (RNN) achieves the highest overall performance with an accuracy of 79.88% and an F1-score of 0.76, demonstrating its superior ability to capture contextual dependencies associated with sarcastic news headlines.

Overall, the proposed methodology establishes a comprehensive sarcasm detection framework by integrating CountVectorizer, Word2Vec, AdaBoost, XGBoost, CNN, RNN, BiLSTM, Hybrid CNN–BiLSTM, LIME, and t-SNE visualization within a unified NLP architecture. The combination of robust preprocessing, semantic feature extraction, comparative machine learning analysis, deep learning classification, explainable prediction, and visualization provides an accurate, interpretable, and scalable solution for sarcasm detection in news headlines while supporting broader applications in sentiment analysis, opinion mining, and intelligent language understanding.

IV. SYSTEM ARCHITECTURE

The proposed system architecture is designed as an integrated Natural Language Processing (NLP) framework that combines text preprocessing, feature engineering, machine learning, deep learning, model interpretability, and performance evaluation for intelligent sarcasm detection in news headlines. The architecture consists of several interconnected modules that process raw textual data through sequential analytical stages, enabling automatic classification of news headlines into sarcastic and non-sarcastic categories. Unlike conventional text classification systems that depend on a single predictive algorithm, the proposed architecture performs comparative analysis using multiple machine learning and deep learning models to identify the most suitable classifier while incorporating explainable artificial intelligence for prediction interpretation.

The architecture begins with the Dataset Acquisition Module, where the Kaggle Sarcasm Detection News Headlines Dataset is collected for model development. The dataset contains 28,619 news headlines, each associated with a sarcasm label indicating whether the headline expresses sarcastic or non-sarcastic content. The dataset also contains additional information such as article links, although only the headline text and sarcasm labels are utilized during model training. The collected dataset provides sufficient linguistic diversity to enable effective learning of contextual sarcasm patterns across various news topics.

The acquired dataset is forwarded to the Text Preprocessing Module, where several preprocessing operations are performed to improve data quality before feature extraction. Initially, unnecessary attributes such as article links are removed since they do not contribute to sarcasm classification. Duplicate records are eliminated to avoid redundant learning, followed by text normalization procedures that convert all characters into lowercase format. Special characters, punctuation marks, numerical values, and irrelevant symbols are removed to simplify textual representation. Subsequently, tokenization divides each headline into individual words, while lemmatization converts inflected words into their corresponding root forms. Common English stop words that contribute little semantic information are also removed. These preprocessing operations significantly reduce textual noise while preserving meaningful linguistic information required for accurate sarcasm detection.

Following preprocessing, the cleaned textual data enters the Feature Engineering Module, where news headlines are transformed into numerical feature representations suitable for computational learning. The proposed architecture employs two complementary feature extraction techniques. For traditional machine learning algorithms, CountVectorizer converts textual information into frequency-based numerical vectors by generating a document-term matrix representing word occurrence frequencies. For deep learning models, Word2Vec embeddings are utilized to produce dense semantic vector representations that preserve contextual relationships among words. These semantic embeddings enable neural networks to capture linguistic similarities that cannot be represented through frequency-based features alone.

The extracted feature representations are subsequently processed within the Machine Learning Classification Module. Two ensemble learning algorithms, AdaBoost and Extreme Gradient Boosting (XGBoost), are independently trained using the CountVectorizer feature representation. AdaBoost iteratively combines multiple weak learners while assigning greater importance to previously misclassified samples, thereby improving overall prediction capability through adaptive boosting. XGBoost constructs optimized gradient boosting decision trees that progressively minimize classification errors while employing regularization mechanisms to improve model generalization. Both machine learning classifiers independently predict whether a news headline contains

sarcastic content before their performance is comparatively evaluated.

Parallel to the machine learning pipeline, the proposed architecture incorporates a dedicated Deep Learning Classification Module that processes Word2Vec embeddings using multiple neural network architectures. The first model employs a Convolutional Neural Network (CNN) that performs convolution operations to identify meaningful local semantic patterns and phrase-level features associated with sarcasm. Pooling operations reduce feature dimensionality while preserving highly informative representations required for classification.

The second neural architecture utilizes a Recurrent Neural Network (RNN), which processes textual sequences word by word while maintaining contextual memory through recurrent connections. The sequential learning capability enables the RNN to capture long-range contextual dependencies that significantly influence sarcasm interpretation. According to the experimental evaluation, the RNN architecture demonstrates the highest overall prediction performance among all evaluated classification models.

The architecture further incorporates a Bidirectional Long Short-Term Memory (BiLSTM) network that simultaneously analyzes textual sequences in forward and backward directions. Bidirectional processing enables the model to utilize both preceding and succeeding contextual information when interpreting individual words, thereby improving semantic understanding for complex sarcastic expressions. In addition, a Hybrid CNN–BiLSTM architecture combines convolutional feature extraction with bidirectional sequential learning, allowing simultaneous modeling of local phrase patterns and long-distance contextual dependencies within news headlines.

Following model training, all prediction outputs are transferred to the Model Evaluation Module, where classifier performance is objectively assessed using multiple evaluation metrics including Accuracy, Precision, Recall, and F1-score. Comparative evaluation enables objective identification of the most effective classification model based on predictive performance across unseen testing data. Hyperparameter optimization is also performed during model development to improve classification accuracy while minimizing overfitting across the evaluated machine learning and deep learning architectures.

To improve model transparency, the proposed architecture integrates an Explainable Artificial Intelligence Module using Local Interpretable Model-Agnostic Explanations (LIME). After a prediction is generated, LIME analyzes the contribution of individual words within each headline and identifies the textual features responsible for the final classification decision. Rather than functioning as an independent prediction model, LIME provides local explanations that enable researchers and end users to understand how the selected classifier interprets sarcastic language. This explainability mechanism significantly improves user confidence and facilitates validation of the learned decision-making process.

The architecture additionally incorporates a Visualization Module utilizing t-distributed Stochastic Neighbor Embedding (t-SNE). High-dimensional semantic embeddings generated by Word2Vec are projected into a two-dimensional feature space, allowing visual inspection of the distribution of sarcastic and non-sarcastic headlines. The generated visualizations illustrate cluster formation and feature separability, providing valuable insights into the effectiveness of learned textual representations and supporting qualitative evaluation of model performance.

Finally, the architecture produces the Prediction Output Module, where each news headline is classified as either Sarcastic or Non-Sarcastic. Along with the predicted class label, the framework provides evaluation metrics, LIME explanations, and visualization results that collectively enhance model interpretability. The integrated architecture therefore supports accurate classification, transparent decision-making, and comprehensive performance analysis within a unified sarcasm detection framework.

Overall, the proposed system architecture successfully integrates Dataset Acquisition, Text Preprocessing, CountVectorizer, Word2Vec, AdaBoost, XGBoost, CNN, RNN, BiLSTM, Hybrid CNN–BiLSTM, LIME, t-SNE, and Performance Evaluation into a comprehensive NLP framework for sarcasm detection. The coordinated interaction among these modules enables robust text processing, accurate sarcasm classification, explainable predictions, and effective visualization, thereby providing a scalable and intelligent solution for news headline analysis and broader natural language understanding applications.

V. FEATURE ENHANCEMENT

The proposed sarcasm detection framework incorporates several enhancements that improve the accuracy, interpretability, robustness, and computational efficiency of automatic news headline classification. By integrating advanced Natural Language Processing (NLP) techniques with Machine Learning (ML), Deep Learning (DL), and Explainable Artificial Intelligence (XAI), the framework provides a comprehensive solution for identifying sarcastic and non-sarcastic textual content. Unlike conventional sentiment analysis approaches that rely solely on lexical features or single prediction models, the proposed system combines multiple learning strategies, semantic feature representations, visualization techniques, and model interpretability within a unified intelligent architecture. These enhancements collectively improve sarcasm recognition while supporting transparent and reliable decision-making.

One of the primary enhancements introduced by the proposed framework is the implementation of comprehensive text preprocessing. Raw textual data collected from online news sources frequently contain punctuation marks, duplicate entries, unnecessary symbols, inconsistent capitalization, and irrelevant textual information that may negatively influence prediction performance. To improve dataset quality, the framework performs multiple preprocessing operations, including duplicate removal, lowercase conversion, text normalization, tokenization, lemmatization, and stop-word elimination. These operations reduce textual noise while preserving meaningful linguistic information, enabling the learning algorithms to focus on contextual relationships that contribute significantly to sarcasm detection.

Another important enhancement is the use of dual feature representation techniques, combining CountVectorizer and Word2Vec embeddings. Traditional machine learning algorithms require numerical feature vectors generated through CountVectorizer, which captures word occurrence frequencies across news headlines. Simultaneously, Word2Vec generates dense semantic embeddings capable of preserving contextual relationships among words appearing in similar linguistic environments. By integrating both statistical and semantic feature representations, the framework effectively supports both conventional machine learning models and modern deep neural networks while improving the quality of textual information available for classification.

The proposed system further enhances prediction capability through the incorporation of ensemble machine learning algorithms, including AdaBoost and Extreme Gradient Boosting (XGBoost). AdaBoost improves classification performance by iteratively assigning greater importance to previously misclassified samples, allowing the model to progressively reduce prediction errors. Similarly, XGBoost employs optimized gradient boosting together with regularization techniques to improve generalization capability while minimizing overfitting. The integration of these ensemble learning algorithms provides reliable baseline performance and enables objective comparison with deep learning models during experimental evaluation.

A significant enhancement of the proposed framework is the integration of multiple Deep Learning architectures for contextual language understanding. The Convolutional Neural Network (CNN) automatically extracts important local textual patterns through convolution and pooling operations, enabling efficient identification of phrase-level semantic information associated with sarcastic language. The Recurrent Neural Network (RNN) further improves contextual understanding by processing textual sequences sequentially, allowing the model to capture long-range linguistic dependencies that frequently characterize sarcasm. Experimental evaluation demonstrates that the RNN architecture achieves the highest overall classification performance, highlighting its effectiveness in modeling contextual information within news headlines.

The framework additionally incorporates Bidirectional Long Short-Term Memory (BiLSTM) networks to enhance sequential language modeling. Unlike conventional recurrent networks that process text in a single direction, BiLSTM simultaneously analyzes textual information from both forward and backward directions. Bidirectional learning enables more comprehensive understanding of contextual relationships between words, thereby improving semantic interpretation for complex sarcastic expressions. Furthermore, the proposed Hybrid CNN–BiLSTM architecture combines local feature extraction with bidirectional contextual modeling, enabling simultaneous learning of phrase-level features and long-distance semantic dependencies. This hybrid architecture improves feature representation by leveraging the complementary strengths of both neural network models.

Another major enhancement involves the implementation of hyperparameter optimization across all evaluated machine learning and deep learning models. Parameters such as learning rate, batch size, filter size, hidden units, and training epochs are systematically adjusted to maximize classification accuracy while minimizing overfitting. Hyperparameter optimization enables each predictive model to operate under near-optimal training conditions, thereby improving model stability, convergence behavior, and overall classification performance without modifying the underlying learning algorithms.

The proposed framework further strengthens model reliability through the integration of Explainable Artificial Intelligence (XAI) using Local Interpretable Model-Agnostic Explanations (LIME). While deep learning models generally achieve high prediction accuracy, their internal decision-making mechanisms often remain difficult to interpret. LIME addresses this limitation by identifying individual words that contribute most significantly to each prediction. These local explanations enable researchers and end users to understand why a particular news headline has been classified as sarcastic or non-sarcastic. Explainability therefore improves model transparency, facilitates error analysis, and increases user confidence in artificial intelligence-assisted decision-making.

An additional enhancement is the incorporation of t-distributed Stochastic Neighbor Embedding (t-SNE) for feature visualization. Since Word2Vec produces high-dimensional semantic vectors that cannot be interpreted directly, t-SNE projects these feature representations into a two-dimensional space suitable for visualization. The resulting scatter plots illustrate the distribution of sarcastic and non-sarcastic headlines, enabling researchers to observe clustering behavior, evaluate feature separability, and better understand the internal representation learned

by the classification models. Visualization therefore complements quantitative evaluation by providing meaningful qualitative insights into model performance.

The framework also introduces a comparative evaluation strategy, where multiple machine learning and deep learning models are trained and evaluated using identical datasets and performance metrics. Rather than relying on a single classification algorithm, the proposed methodology objectively compares AdaBoost, XGBoost, CNN, BiLSTM, RNN, and Hybrid CNN–BiLSTM using Accuracy, Precision, Recall, and F1-score. This comprehensive evaluation identifies the strengths and limitations of each model while ensuring that the final model selection is supported by objective experimental evidence.

Finally, the proposed framework enhances the applicability of sarcasm detection across numerous Natural Language Processing applications. Accurate identification of sarcastic language significantly improves downstream tasks including sentiment analysis, opinion mining, fake news detection, social media monitoring, emotion recognition, recommendation systems, and public opinion analysis. The integration of explainability further supports deployment within practical decision-support systems where transparent artificial intelligence is increasingly important.

Overall, the proposed feature enhancements establish a comprehensive NLP framework by integrating advanced preprocessing, CountVectorizer, Word2Vec embeddings, AdaBoost, XGBoost, CNN, RNN, BiLSTM, Hybrid CNN–BiLSTM, hyperparameter optimization, LIME, and t-SNE visualization within a unified architecture. These enhancements significantly improve sarcasm detection accuracy, contextual language understanding, model interpretability, visualization capability, and computational robustness while providing an effective and scalable solution for intelligent news headline analysis and broader natural language processing applications.

VI. CONCLUSION

This paper presented an intelligent Natural Language Processing (NLP) framework for automatically detecting sarcasm in news headlines by integrating Machine Learning (ML), Deep Learning (DL), and Explainable Artificial Intelligence (XAI) techniques. The proposed methodology combines comprehensive text preprocessing, semantic feature extraction, multiple classification models, explainable prediction analysis, and visualization into a unified architecture capable of distinguishing sarcastic and non-sarcastic headlines with high reliability. By utilizing both traditional machine learning algorithms and modern deep neural networks, the framework provides a comprehensive comparative analysis while improving the understanding of contextual language patterns associated with sarcastic expressions.

The proposed framework employed CountVectorizer and Word2Vec for feature representation, followed by comparative evaluation of AdaBoost, Extreme Gradient Boosting (XGBoost), Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Bidirectional Long Short-Term Memory (BiLSTM), and a Hybrid CNN–BiLSTM architecture. Furthermore, Local Interpretable Model-Agnostic Explanations (LIME) were incorporated to improve prediction transparency, while t-distributed Stochastic Neighbor Embedding (t-SNE) was utilized to visualize semantic feature distributions. The integration of these complementary techniques enabled the framework to perform accurate classification while providing meaningful insights into the decision-making behavior of the predictive models.

Experimental evaluation demonstrated that deep learning models generally outperformed conventional machine learning algorithms for sarcasm detection. Among all evaluated classifiers, the Recurrent Neural Network (RNN) achieved the best overall performance, obtaining an accuracy of 79.88%, precision of 0.83, recall of 0.70, and an F1-score of 0.76. These results indicate that sequential neural architectures are particularly effective at learning contextual dependencies that characterize sarcastic language. Although CNN, BiLSTM, and the Hybrid CNN–BiLSTM models also demonstrated competitive performance, the RNN consistently provided the most balanced classification capability under the reported experimental conditions.

An important contribution of the proposed research is the integration of Explainable Artificial Intelligence within the sarcasm detection pipeline. The incorporation of LIME enables researchers and end users to understand the contribution of individual words toward classification decisions, thereby improving model transparency and user confidence. Similarly, t-SNE visualization provides additional insight into semantic feature distributions and clustering behavior, allowing qualitative interpretation of the learned textual representations. These explainability mechanisms enhance the practical applicability of the framework by addressing one of the major limitations associated with black-box deep learning models.

The proposed framework offers considerable potential for numerous real-world Natural Language Processing applications. Accurate sarcasm detection can significantly improve sentiment analysis, opinion mining, social media analytics, fake news detection, recommendation systems, customer feedback analysis, and public opinion monitoring by enabling more precise interpretation of human language. The combination of high classification accuracy, semantic feature learning, comparative model evaluation, and explainable prediction makes the proposed framework suitable for deployment within intelligent language understanding systems across diverse application domains.
Overall, the proposed sarcasm detection framework establishes a robust, interpretable, and scalable solution by

integrating CountVectorizer, Word2Vec, AdaBoost, XGBoost, CNN, RNN, BiLSTM, Hybrid CNN–BiLSTM, LIME, and t-SNE within a unified NLP architecture. The coordinated interaction among these components improves sarcasm classification performance while maintaining transparency and computational efficiency. The study demonstrates that combining advanced deep learning techniques with explainable artificial intelligence provides an effective approach for addressing the complex challenges associated with automatic sarcasm detection in news headlines.

Future research may focus on incorporating transformer-based language models such as BERT, RoBERTa, or other contextual embedding techniques to further improve semantic understanding of sarcastic expressions. Additional investigation using multilingual datasets, multimodal information, larger benchmark corpora, and attention-based neural architectures may enhance the robustness and generalization capability of sarcasm detection systems. Furthermore, integrating advanced explainability techniques and adaptive learning mechanisms could improve both prediction performance and interpretability, enabling the development of more reliable and intelligent natural language understanding systems for real-world applications.

REFERENCES

1. S. Gupta, A. Sharma, and R. Gupta, "Sarcasm Detection Using Machine Learning and Natural Language Processing Techniques," in Proc. Int. Conf. Computing, Communication and Intelligent Systems (ICCCIS), Greater Noida, India, 2022, pp. 421–426.
2. M. Potamias, G. Siolas, and A. Stafylopatis, "A Transformer-Based Approach to Sarcasm Detection in News Headlines," IEEE Access, vol. 8, pp. 145882–145893, 2020.
3. A. Misra and B. Arora, "Context-Aware Sarcasm Detection Using Deep Neural Networks," Expert Systems with Applications, vol. 168, pp. 114315, 2021.
4. K. Ghosh and T. Veale, "Fracking Sarcasm Using Neural Network Models," in Proc. NAACL-HLT, San Diego, CA, USA, 2016, pp. 161–169.
5. D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in Proc. Int. Conf. Learning Representations (ICLR), San Diego, CA, USA, 2015.
6. T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," in Proc. Int. Conf. Learning Representations (ICLR) Workshop, Scottsdale, AZ, USA, 2013.
7. T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and Their Compositionality," in Proc. Advances in Neural Information Processing Systems (NeurIPS), Lake Tahoe, NV, USA, 2013, pp. 3111–3119.
8. Y. Kim, "Convolutional Neural Networks for Sentence Classification," in Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 2014, pp. 1746–1751.
9. S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.
10. M. Schuster and K. K. Paliwal, "Bidirectional Recurrent Neural Networks," IEEE Transactions on Signal Processing, vol. 45, no. 11, pp. 2673–2681, 1997.
11. J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. NAACL-HLT, Minneapolis, MN, USA, 2019, pp. 4171–4186.
12. C. Cortes and V. Vapnik, "Support-Vector Networks," Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.
13. Y. Freund and R. E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," Journal of Computer and System Sciences, vol. 55, no. 1, pp. 119–139, 1997.
14. T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, San Francisco, CA, USA, 2016, pp. 785–794.
15. M. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," in Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, San Francisco, CA, USA, 2016, pp. 1135–1144.
16. L. van der Maaten and G. Hinton, "Visualizing Data Using t-SNE," Journal of Machine Learning Research, vol. 9, pp. 2579–2605, 2008.
17. C. Bishop, Pattern Recognition and Machine Learning. New York, NY, USA: Springer, 2006.
18. I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.
19. T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, 2nd ed. New York, NY, USA: Springer, 2009.
20. F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
21. J. Brownlee, Deep Learning for Natural Language Processing. Melbourne, Australia: Machine Learning Mastery, 2019.
22. S. Bird, E. Klein, and E. Loper, Natural Language Processing with Python. Sebastopol, CA, USA: O'Reilly Media, 2009.
23. A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of Tricks for Efficient Text Classification," in Proc. European Chapter of the Association for Computational Linguistics (EACL), Valencia, Spain, 2017, pp. 427–431.
24. J. Pennington, R. Socher, and C. Manning, "GloVe: Global Vectors for Word Representation," in Proc. Conf. Empirical Methods in Natural Language

Processing (EMNLP), Doha, Qatar, 2014, pp. 1532–1543.