International Journal of Scientific World, 11 (2) (2025) 71-78

International Journal of Scientific World

Website: www.sciencepubco.com/index.php/IJSW
https://doi.org/10.14419/t1xnaq87
Review paper

Deep Learning for Natural Language Processing: A Review of Models and Applications

Renjbar Sh. Othman [1, 2] *, Ibrahim Mahmood Ibrahim [3]

[1] *Duhok Polytechnic University, Technical Institute of Amedi, Department of Information Technology, Duhok, Kurdistan Region, Iraq.*
[2] *Akre University for Applied Sciences, Technical College of Informatics-Akre, Department of Information Technology, Akre, Kurdistan Region, Iraq.*
[3] *Akre University for Applied Sciences, Technical College of Informatics-Akre, Department of Computer Networks and Information Security, Kurdistan Region, Iraq.*
**Corresponding author E-mail: renjbar.othman@dpu.edu.krd*

Received: April 24, 2025, Accepted: June 15, 2025, Published: August 29, 2025

Abstract

This review provides a critical analysis of the transformative impact of deep learning on the advancement of Natural Language Processing (NLP). With the increasing volume of unstructured textual data, traditional rule-based and statistical methodologies have demonstrated limitations in effectively capturing the intricacies of human language. In contrast, deep learning models, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and transformer-based architectures such as BERT and GPT, have significantly enhanced NLP capabilities by facilitating context-aware, scalable, and highly accurate language comprehension. The primary objective of this review is to deliver a comprehensive synthesis of deep learning architectures utilized in essential NLP tasks, including sentiment analysis, text classification, machine translation, and question answering. Additionally, it examines their evolution, key applications, and comparative performance across various domains. By reviewing recent literature from 2021 to 2025, this analysis also emphasizes hybrid models, multimodal systems, and adaptations for low-resource environments. The goal is to identify emerging trends, challenges (e.g., interpretability, computational cost), and future directions, including data augmentation, self-supervised learning, and cross-domain generalization, ultimately-ly guiding researchers towards the development of more adaptive and trustworthy NLP systems.

Keywords: *Deep Learning; Natural Language Processing; Transformer Models; Neural Network Architectures; Sentiment Analysis; Text Classification; Multimodal and Hybrid Models.*

1. Introduction

Natural Language Processing (NLP), a key subdomain of Artificial Intelligence (AI), bridges the gap between human communication and machine understanding. With the exponential growth of digital text data and the increasing demand for intelligent systems capable of interpreting and generating human language, NLP has become indispensable across numerous sectors, including healthcare, finance, education, and e-commerce[1]. Over the last decade, the advent and advancement of deep learning have significantly transformed the landscape of NLP, enabling breakthroughs in a range of complex tasks such as sentiment analysis, machine translation, question answering, and text classification [2], [3].

Traditional NLP techniques heavily relied on manual feature engineering and rule-based systems. These methods, though foundational, struggled with scalability and adaptability to the linguistic complexity inherent in natural language[4]. Deep learning models, in contrast, have demonstrated the ability to automatically learn hierarchical feature representations from raw text data, drastically improving performance and generalizability across tasks [1], [5].

Deep neural networks (DNNs), including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM), and more recently, Transformer-based models like BERT and GPT, have pushed the boundaries of what machines can achieve in language understanding [6], [2], [7]. These architectures excel at capturing syntactic and semantic nuances, making them suitable for both low-level linguistic tasks (e.g., part-of-speech tagging, named entity recognition) and high-level applications (e.g., machine reading comprehension, dialogue systems) [1], [8].

Despite their success, deep learning models in NLP are not without challenges. Issues such as adversarial robustness, interpretability, computational cost, and data dependency continue to pose significant hurdles [9], [10]. Moreover, developing systems that can understand context, sarcasm, and ambiguity remains an open area. The growing interest in contextual language models, like BERT and ELMo, has led to notable advancements in this regard, showing that pre-training on massive corpora followed by fine-tuning can yield state-of-the-art results on multiple NLP benchmarks [3], [8].

This review paper endeavors to deliver a thorough overview of the contemporary landscape of deep learning models in Natural Language Processing (NLP), emphasizing their architectures, applications, and the challenges they encounter. We examine foundational models and trace their evolution into sophisticated systems adept at managing multimodal and multilingual tasks. Furthermore, we underscore promising avenues for future research, encompassing model interpretability, robustness, and generalization towards Artificial General Intelligence (AGI).
The remainder of this review paper is structured as follows: Section 2 provides an overview of foundational theories in NLP and deep learning architectures relevant to the field. Section 3 presents a comprehensive literature review, analyzing recent studies and key developments. Section 4 discusses the comparative evaluation of deep learning models, highlighting statistical trends, performance metrics, and notable insights. Finally, Section 5 concludes the review by summarizing key findings and outlining future research directions in deep learning-based NLP.

2. Background theory

2.1. Natural language processing (NLP)

Natural Language Processing (NLP) is a core area of artificial intelligence (AI) that focuses on enabling machines to understand, interpret, and generate human language. Traditionally, NLP relied on rule-based systems and symbolic reasoning, which were limited in handling linguistic ambiguity and context sensitivity [11], [12]. Over the years, statistical and machine learning methods, such as Naïve Bayes, SVMs, and Hidden Markov Models (HMM), have become widely adopted to process language more flexibly [12]. However, with the rise of big data and powerful computational resources, particularly GPUs, deep learning emerged as a transformative force in NLP, making it possible to model complex syntactic and semantic structures directly from raw data [12], [13].

2.2. Deep learning in NLP

Deep learning refers to neural network models with multiple layers capable of learning hierarchical representations from large datasets. Unlike traditional models that rely on hand-crafted features, deep learning enables automated feature extraction, making it highly suitable for unstructured text data[13].
Key advantages of deep learning in NLP include:

- Ability to capture context and meaning.
- Scalability to massive datasets.
- State-of-the-art performance in a wide range of tasks, such as text classification, machine translation, sentiment analysis, and question answering [12], [13].

2.3. Neural network architectures

Several deep learning architectures are central to modern NLP tasks:

2.3.1. Feedforward neural networks (FNNs)

The simplest form of neural networks, FNNs, are used mainly in tasks where the context is not sequential. They are composed of input, hidden, and output layers, with non-linear activation functions like ReLU and Sigmoid [13].

2.3.2. Convolutional neural networks (CNNs)

CNNs are particularly effective for extracting local features and patterns from text, such as phrases or key terms in sentiment analysis or sentence classification [14], [15].

2.3.3. Recurrent neural networks (RNNs) and variants

RNNs, including LSTMs and GRUs, are tailored for sequential data. They maintain memory of previous states, making them useful for tasks like text generation, machine translation, and speech recognition. Bidirectional RNNs further improve performance by processing sequences in both directions [14], [12], [16].

2.3.4. Transformer models

Transformers, especially models like BERT and GPT, represent a significant leap in NLP. They use self-attention mechanisms to model long-range dependencies and context in a sequence without relying on recurrence, resulting in superior performance across nearly all NLP benchmarks [13], [17], [18].

2.4 Word embedding techniques

Word embeddings are vector representations of words that capture semantic relationships. They form the foundation of modern deep learning models by converting words into dense, continuous vector spaces:

- Static embeddings: Word2Vec, GloVe, and FastText provide fixed embeddings for words regardless of context [14], [17].
- Contextual embeddings: Models like ELMo, BERT, and GPT produce embeddings based on surrounding context, offering more accurate representations of word meaning [14], [16].

2.5. Applications of deep learning in NLP

1) Text Classification
Deep learning enables effective categorization of documents, emails, and social media posts. CNNs and RNNs are commonly used for spam detection, topic labeling, and content moderation [16], [15], [7].
2) Sentiment Analysis
Sentiment analysis benefits greatly from contextual embeddings and deep models like LSTM and BERT, especially in multilingual or domain-specific contexts [11], [19].
3) Machine Translation
Transformers and attention-based models dominate this area. BERT and its multilingual variants facilitate high-quality translation even between low-resource languages [12], [16].
4) Question Answering (QA)
QA systems leverage both extractive and generative models using deep architectures. Models like BERT and GPT have pushed the boundaries of QA by enabling machines to infer answers from unstructured texts [7], [19].
5) Named Entity Recognition (NER) and Information Extraction
NER is significantly improved with deep learning by reducing reliance on hand-crafted rules. CNNs, Bi-LSTMs, and CRFs are often combined to enhance performance [12], [15].

2.6. Challenges and future directions

Despite remarkable progress, deep learning in NLP continues to face key challenges, including limited labeled data in domain-specific tasks, a lack of model interpretability, and persistent biases in large language models [11], [17]. Ensuring fairness, transparency, and robustness in AI-driven NLP systems remains an ongoing concern [12]. Additionally, integrating multimodal data (text, image, audio) and managing computational efficiency are pressing issues [16]. Promising research directions include self-supervised learning, transfer learning, and model compression, which aim to enhance generalization and reduce resource consumption [13], [16]. These innovations will shape the next generation of adaptive and trustworthy NLP systems.

3. Literature review

Recent advancements in deep learning have significantly revolutionized Natural Language Processing (NLP), enabling models to more adeptly capture context, semantics, and complex language patterns. A growing body of research explores the application of deep learning architectures such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory networks (LSTMs), and transformers in tackling fundamental NLP tasks, including sentiment analysis, question answering, and text classification. This section provides a comprehensive review of key studies conducted from 2021 to 2025, emphasizing the methodologies utilized, performance metrics attained, and practical insights gained. The analysis highlights the transition from traditional approaches to more sophisticated hybrid and transformer-based models:
Xiong et al. (2024) investigated how deep neural networks (DNNs) integrating multimodal information can enhance intelligent question answering (QA) systems. The research asked whether combining multimodal inputs (text, audio, visual) using techniques like word embedding, attention mechanisms, and graph neural networks (GNN) improves semantic understanding and performance in QA. The hypothesis was that such integration would yield higher accuracy and user satisfaction than traditional models. The study used a DNN model trained on QA corpora, manipulating input modalities and measuring precision, recall, and user satisfaction. Data were analyzed through loss function trends and comparative metrics. Results showed significantly higher precision, recall, and satisfaction scores using the proposed method. The findings imply that multimodal integration improves QA robustness and semantic comprehension without overclaiming universal effectiveness [20].
Chen (2023) explored how neural language models have evolved and impacted NLP, especially with the rise of deep learning. The central research question examined how these models, particularly those based on neural networks, enhance language modeling tasks like word prediction and sentence generation. This topic is compelling because it addresses core challenges in enabling machines to understand and generate human language. The hypothesis was that neural models outperform traditional statistical approaches in accuracy and contextual understanding. Using comparative evaluations across models like RNNs, CNNs, and LSTMs, the study measured perplexity, precision, and translation quality. Results showed that neural models, especially those with attention and dropout mechanisms, significantly improved performance, with LSTMs yielding better results than CNNs in most tasks. The findings underscored that while neural approaches greatly enhance NLP, challenges like model evaluation and generalization remain [21].
Sruthi et al. (2024) explored how deep learning models can enhance sentiment analysis by evaluating different architectures like CNNs, RNNs, and Transformers. The central question was whether deep neural networks outperform traditional approaches in accurately detecting sentiment in text data, which is crucial for applications in business intelligence and user feedback analysis. The hypothesis was that deep learning models, especially transformer-based ones, would yield higher accuracy due to their ability to capture contextual nuances. The team trained and tested models on labeled datasets, using metrics such as accuracy, precision, recall, and F1 score to assess performance. Results showed the transformer model achieved the highest accuracy (94.3%), significantly outperforming CNNs and RNNs. These findings suggested that deep learning offers a promising path for improving sentiment classification, though model interpretability and data imbalance remain challenges [22].
Liu and Wang (2024) examined how combining LSTM and CNN models could improve sentiment analysis by better handling both contextual flow and local features in text. They posed whether this hybrid model would outperform standalone deep learning approaches, a relevant question due to the complex, varied nature of online language. They hypothesized that the LSTM-CNN combination would yield higher classification accuracy and generalizability. Using datasets from six domains, they trained the model with Word2Vec embeddings and tested different configurations, evaluating precision, recall, and F1 scores. Results showed the hybrid model achieved an F1 score of 0.9279, outperforming LSTM (0.9154) and CNN (0.9092). This suggested that integrating temporal and spatial features enhances model effectiveness, though computational cost remains a limitation [23].
Geddama et al. (2024) investigated how various deep learning models, CNNs, RNNs, LSTMs, and Transformers like BERT perform in classifying news articles. The study asked whether combining models such as BERT and LSTM could improve accuracy, which is relevant given the need for fast, reliable categorization in digital journalism. The authors hypothesized that hybrid models would outperform

standalone ones. They used the BBC News dataset, applying preprocessing, tokenization, TF-IDF, and Word2Vec, then trained models using cross-entropy loss and Adam optimizer. They compared performance based on accuracy. BERT+LSTM achieved the highest accuracy (98.4%), significantly outperforming CNN (90%) and even BERT alone (96.5%). These results suggest hybrid models offer a strong approach to handling complex language patterns in real-world text classification without overstating generalizability [24].

Singh et al. (2023) explored how various deep learning models, including RNNs, LSTMs, Transformers, GANs, and VAEs, perform in generating natural language text. They asked which models are most effective for producing fluent, coherent text, a key question given the growing demand for human-like responses in applications like chatbots and summarization. The authors hypothesized that newer architectures like Transformers and hybrid models would outperform traditional RNNs in quality and consistency. Their study compared models using datasets like Yelp, Amazon, and WebText, evaluating performance with metrics such as perplexity, BLEU, and ROUGE scores. Results showed Transformer-based models (e.g., GPT-2, CTRL) had the lowest perplexity and highest human evaluation scores, outperforming RNNs and LSTMs. The findings emphasized that while traditional models still have merit, advanced architectures yield more natural and context-aware outputs, though challenges like interpretability and data demands persist [25].

Cao (2023) studied the effectiveness of deep learning models, specifically CNN and LSTM, for analyzing sentiment in social media posts. The research aimed to determine whether these models could reliably detect emotional tone in diverse, informal online text, an important question due to the noisy and dynamic nature of social media language. The hypothesis was that CNN and LSTM, with word embeddings and proper tuning, would yield strong performance. Using a dataset of 10,000 labeled posts from platforms like Twitter and Facebook, the team applied text cleaning, tokenization, and feature representation, then trained both models. Performance was measured using accuracy, precision, recall, F1-score, and AUC. Results showed both models performed well, with high accuracy and balanced metrics. The study concluded that deep learning offers a promising path for sentiment analysis, though challenges like cross-cultural context and multimodal data remain areas for future work [26].

Prathyakshini and Shetty (2024) examined the effectiveness of deep learning models for text classification across various domains and datasets. They explored whether newer models like CNNs, RNNs, and BERT outperform traditional techniques such as SVM and Naive Bayes in handling large-scale, unstructured text. The hypothesis was that deep learning would provide better accuracy due to its ability to capture semantic patterns without manual feature engineering. They applied preprocessing steps like tokenization, stemming, and word embedding, then trained models on datasets including news articles, tweets, and medical texts. Performance was measured using accuracy, precision, recall, and F1 score. Results showed that models like SVM with RBF and Set-CNN reached up to 98% and 95% accuracy, respectively, while simpler models like GNB performed significantly worse. The study suggested that deep learning models, especially when optimized, are better suited for complex classification tasks, though challenges remain with interpretability and data quality [27].

Vijaya Lakshmi and Murugesh (2023) reviewed advancements in deep learning techniques for sentiment analysis, asking whether models like LSTM, CNN, and Transformers can better capture emotional cues in text compared to traditional approaches. They hypothesized that deep learning, especially when paired with attention and transfer learning, offers superior performance in context-rich environments. Their analysis spanned models applied across datasets such as tweets and reviews, measuring accuracy, efficiency, and robustness. Results showed models like BERT and CNN-RNN hybrids achieved high accuracy but required considerable resources. Significant gaps remain in cross-domain generalization, interpretability, and multimodal integration. The study emphasized that while deep learning has enhanced sentiment detection, future work must balance accuracy with explainability and accessibility [28].

Qiu et al. (2021) introduced EasyTransfer, a scalable deep transfer learning platform tailored for NLP tasks. They explored whether a unified toolkit could simplify and accelerate training, fine-tuning, and deployment of large pre-trained models in real-world applications. The team hypothesized that combining model and data parallelism with custom algorithms would enhance efficiency and performance. Using tasks like paraphrase detection, sentiment analysis, and recommendation systems, they measured accuracy, AUC, and training time. The platform showed up to 4.4× training speedups and competitive accuracy (e.g., 89.4% with MetaKD vs. 86.7% with TinyBERT). The results supported the toolkit's practical value while acknowledging challenges like scalability and memory limits in ultra-large models[29].

Mohanraj et al. (2024) developed a hybrid deep learning framework to automate the evaluation of handwritten answer scripts, addressing the tedious nature of manual grading. They asked whether combining NLP, OCR, OpenCV, and DCCNN could deliver accurate, scalable assessments for varied answer types, including diagrams and equations. The hypothesis was that this ensemble system would outperform existing models in precision and reduce evaluator bias. They used over 5,000 scanned scripts and applied multiple stages: OCR for text extraction, NLP for semantic matching, and DCCNN for equation analysis. The model achieved up to 95% accuracy in mathematical tasks and 93% in sentence similarity. Results showed significantly lower cross-entropy loss compared to previous methods, suggesting the framework's efficiency and near-human grading reliability, though future enhancements for complex rubrics are still needed [30].

Purohit et al. (2025) conducted a large-scale study to analyze media sentiment across 143,000 news articles from 15 U.S. outlets between 2000 and 2017. They asked whether deep learning techniques like LSTM, BERT, and GPT could uncover patterns of bias and emotional tone over time timely question given ongoing concerns about media influence. They hypothesized that NLP models would reveal consistent sentiment trends tied to events, authorship, and editorial stance. Using web-scraped data, they applied extensive preprocessing and sentiment labeling via lexicons and ML classifiers. Metrics such as average sentiment score, standard deviation, and model accuracy (BERT: 89%) were analyzed. BERT outperformed GPT (86%) and LSTM (82%), with notable variation in sentiment by source and author. The findings suggested a slight positive bias overall, with sentiment shifting in response to sociopolitical events, showing how NLP can surface complex media dynamics without overgeneralizing [31].

Sheik et al (2024) examined whether neural data augmentation could reduce annotation costs and improve model performance in the legal overruling task. They hypothesized that small deep learning models trained on synthetically generated data could rival large language models. Using just 100 annotated samples from a legal corpus, they generated over 41,000 augmented sentences with GPT-3, GPT-2, BERT variants, and Word2Vec, then trained BiLSTM, BiGRU, and ConvBiLSTM classifiers. They measured accuracy, F1-score, and inference time, showing that BiGRU achieved 96% accuracy and outperformed GPT-3's 74% in a few-shot setting. Results confirmed data augmentation significantly boosted performance, especially for resource-efficient student models, suggesting a promising approach for domain-specific NLP with limited data [32].

Manzoor et al (2025) tackled the challenge of detecting intrinsic plagiarism in Urdu, a low-resource language, by asking whether a stylometry-based machine learning framework could outperform deep and large language models. They hypothesized that handcrafted linguistic features across six textual levels would be more effective than relying on complex neural networks. Using a custom Urdu dataset of 5,436 documents, they extracted 43 stylometry features and tested multiple classifiers under two conditions: general and topic-based (e.g., moral lessons, national events). Accuracy, precision, recall, and F1 scores were used to evaluate performance. Random Forest and XGBoost emerged as top performers, with up to 99% accuracy, significantly surpassing deep models like BiLSTM and even GPT-2. The study suggested that in resource-limited contexts, well-engineered features and classic models can deliver strong, interpretable results [33].

Islam et al (2024) evaluated how well NLP and CNN models detect disaster types from social media data. They asked whether language models like BERT variants and image-based CNNs could effectively classify disaster-related posts—an important question for improving real-time emergency responses. The hypothesis was that advanced NLP models would offer better accuracy and speed in identifying disaster contexts. Using the CrisisMMD dataset, they tested four NLP models and two CNN models, measuring accuracy, precision, recall, and F1 scores. All NLP models achieved 94% accuracy, with DistilBERT-Base-Uncased showing the fastest runtime. DenseNet slightly outperformed Inception v3 in CNN tasks, with 78% accuracy. The findings suggested NLP models are more accurate and efficient than CNNs for disaster classification, especially useful for rapid response, though results may vary by disaster type and data quality [34].

Tüfekci and Kösesoy (2024) investigated whether deep learning models could accurately identify an author's biological gender from Turkish news texts. They asked if models like LSTM or CNN could outperform traditional classifiers such as Naive Bayes and Random Forest, a question of interest in linguistics, marketing, and media studies. They hypothesized that deep learning would achieve better accuracy due to its ability to capture complex language patterns. Using a newly curated dataset of 43,292 balanced articles (IAG-TNKU), they tested four models through tenfold cross-validation and measured accuracy, precision, recall, and F1 score. LSTM emerged as the top performer with 88.51% accuracy, significantly higher than traditional methods. These results showed that advanced neural models can effectively support gender-related text analysis, though broader language and media diversity should be explored further [35].

Ba Alawi and Bozkurt (2024) explored which embedding methods work best for deep learning-based sentiment analysis in Turkish, a language known for its morphological complexity. They asked whether combining character-level and word-level embeddings would yield more accurate sentiment detection than using standard embeddings alone. The authors hypothesized that hybrid embedding techniques would outperform single approaches. Using Twitter (THED) and hotel review (HRD) datasets, they tested six deep learning models (e.g., LSTM, CNN, CNN-BiLSTM) across several embeddings (Word2Vec, GloVe, FastText, COE, CIE, and hybrids). Accuracy and F1-score were the main evaluation metrics. CNN-BiLSTM with hybrid CIE-WE achieved an F1-score of 0.8392 and 95.43% accuracy on HRD, outperforming others. The study concluded that hybrid embeddings improve sentiment classification, especially in morphologically rich languages, while noting trade-offs in training time and complexity [36].

Wang et al. (2025) explored whether deep learning models, particularly transformer-based architectures like GPT and T5, could improve de novo drug generation for targeting EGFR mutations in non-small cell lung cancer. They hypothesized that enhancing GPT with novel techniques (RoPE, DeepNorm, GEGLU) and integrating transfer learning in a T5-based encoder-decoder model (T5MolGe) would generate valid, diverse, and target-specific molecules. The study trained models on the GuacaMol and ChEMBL datasets using SMILES sequences, testing performance through metrics like validity, uniqueness, and Tanimoto similarity. Results showed T5MolGe achieved the highest similarity (0.963) and validity (0.989), outperforming other models in conditional generation. These findings highlighted the value of combining structural guidance and language models for efficient molecule generation, though further testing on synthesis and efficacy is needed [37].

Rana et al. (2025) investigated whether a hybrid deep learning model, BERT-BiGRU-Senti-GCN, could improve sentiment analysis of e-commerce reviews by better capturing contextual and emotional nuances. They hypothesized that integrating BERT for embeddings, BiGRU for sequence understanding, and GCNs enriched with SentiWordNet would outperform existing models. The team used three annotated datasets (employee, SemEval-2014, and Amazon reviews), applying preprocessing, feature extraction, and classification with GCNs, followed by pattern-based sentiment ranking using finite automata. Performance was measured via accuracy, precision, and recall. Their model achieved 93.35% accuracy, outperforming baselines like BERT+GNN and DGEDT. The findings suggest this architecture can meaningfully enhance sentiment analysis in diverse e-commerce feedback scenarios, though scalability and sarcasm detection remain areas for future work [38].

Raza et al. (2024) explored whether machine learning and deep learning models could effectively detect implied threats in text, a subtle and complex challenge in social media analysis. They hypothesized that deep models, especially BiLSTM and DNN, would outperform traditional ML models using lexical and linguistic features. They created a synthetic dataset with human-in-the-loop validation, then applied preprocessing, TF-IDF, and encoded vector representations. Models were tested for accuracy, precision, recall, and F1-score. BiLSTM achieved the highest F1-score (91.61%), outperforming logistic regression (77.13%) and classical ensemble models, which tended to overfit on linguistic features. Their results highlighted the promise of deep learning for nuanced threat detection while acknowledging limits in interpretability and generalization [39].

Martín-Noguerol et al. (2024) examined whether NLP-based deep learning models could distinguish high-grade gliomas (HGG) from brain metastases using MRI radiology reports. The study addressed the challenge of differentiating these conditions, which often appear similar on scans. They hypothesized that models like CNN, BiLSTM, and RadBERT could classify reports accurately even without explicit diagnostic terms. The team trained models on 185 annotated MRI reports (with biased terms removed), using metrics like precision, sensitivity, F1-score, and AUC for evaluation. CNN outperformed other models with an F1-score of 87.23% and an AUC of 87.45%. The findings suggested CNN could assist radiologists in challenging diagnoses, though the study acknowledged limits like dataset size and language translation effects [40].

Table 1 provides a comprehensive comparison of modern deep learning models employed in various NLP tasks, highlighting key innovations, dataset applications, and significant findings from numerous studies. The diverse range of techniques demonstrates the adaptability of deep learning in tackling both domain-specific and general NLP challenges.

4. Comparison and discussion

Table 1: Comparative Overview of Deep Learning Models and Their Applications in NLP (2021–2025): Objectives, Techniques, and Key Outcomes

Ref. No.	Task / Objective	Models / Techniques	Dataset	Key Results	Notable Insight
[20]	Multimodal QA systems	DNN + Word Embedding + GNN + Attention	QA Corpora	↑ Precision, recall, satisfaction	Multimodal inputs enhance semantic understanding
[21]	Neural language modeling	RNN, CNN, LSTM (with attention/dropout)	Multiple NLP benchmarks	LSTM > CNN in word prediction	Neural models better generalize context than statistical ones
[22]	Sentiment analysis	CNN, RNN, Transformer	Labeled sentiment datasets	Transformer: 94.3% accuracy	Transformers are best at context handling
[23]	Hybrid model for sentiment	CNN + LSTM	6 domains	F1: 0.9279 > LSTM/CNN alone	Hybrid temporal-spatial features outperform individual models
[24]	News classification	CNN, RNN, BERT, LSTM	BBC News	BERT+LSTM: 98.4% accuracy	Hybrid models outperform single ones

[25]	Text generation	GPT-2, CTRL, RNN, LSTM	Yelp, Amazon, WebText	GPT-2 lowest perplexity	Transformers generate the most fluent text
[26]	Social media sentiment	CNN, LSTM	Twitter, Facebook	High AUC & F1	Deep learning is effective despite informal language
[27]	Cross-domain text classification	CNN, RNN, BERT vs SVM/Naive Bayes	Mixed (news, tweets, medical)	Set-CNN: 95–98% accuracy	Deep models outperform traditional ones in semantics
[28]	Sentiment analysis	LSTM, CNN, BERT + attention/transfer learning	Tweets & reviews	High accuracy, needs more resources	Accuracy ↔ Interpretability trade-off
[29]	Transfer learning toolkit	EasyTransfer (with MetaKD, TinyBERT)	Paraphrase, Sentiment, Recsys	Up to 4.4× speedup	Toolkits simplify scaling and deployment
[30]	Grading handwritten answer scripts	DCCNN + NLP + OCR + OpenCV	5,000 scanned scripts	Accuracy: up to 95%	The hybrid model is nearly as reliable as human grading
[31]	Media sentiment bias over 2 decades	LSTM, BERT, GPT	143k news articles	BERT: 89% accuracy	NLP reveals long-term sentiment trends
[32]	Legal overruling with a few samples	GPT-3, GPT-2, BiGRU, BiLSTM, ConvBiLSTM	100 real + 41k synthetic	BiGRU: 96%, GPT-3: 74%	Data augmentation boosts small model performance
[33]	Plagiarism detection in Urdu	Stylometry + ML/DL (XGBoost, BiLSTM, GPT-2)	5,436 Urdu docs	XGBoost: 99% > GPT-2	Classical methods can outperform DL in low-resource settings
[34]	Disaster detection from social media	BERT variants, CNNs	CrisisMMD	NLP: 94%, CNN: 78%	NLP > Vision models for textual emergency data
[35]	Gender detection from news text	LSTM, CNN, RF, Naive Bayes	IAG-TNKU (43k articles)	LSTM: 88.51%	DL captures subtle linguistic gender cues
[36]	Turkish sentiment analysis	CNN-BiLSTM + Word/Char embeddings	THED & HRD	F1: 0.8392, Acc: 95.43%	Hybrid embeddings improve morphology-rich language handling
[37]	Drug molecule generation	GPT, T5MolGe (with RoPE, DeepNorm)	SMILES from ChEMBL, GuacaMol	Tanimoto sim: 0.963	Transformers excel in generative bio-NLP
[38]	E-commerce sentiment	BERT + BiGRU + SentiGCN	Amazon, SemEval, Employee reviews	Acc: 93.35%	GCN+sentiment resources boost emotion detection
[39]	Implied threat detection	BiLSTM, DNN, Logistic Regression	Synthetic + real text data	BiLSTM F1: 91.61%	Deep models outperform lexical ML in nuance
[40]	Glioma vs. Metastasis via reports	CNN, BiLSTM, RadBERT	185 MRI radiology reports	CNN F1: 87.23%	NLP assists in subtle diagnostic text classification

A central trend in the literature is the growing dominance of hybrid and transformer-based architectures. For example, Geddama et al. (2024) demonstrated that combining BERT and LSTM achieved an impressive 98.4% accuracy in news classification, outperforming individual models. Similarly, Liu and Wang (2024) showed that their CNN-LSTM hybrid for sentiment analysis achieved an F1-score of 0.9279, significantly higher than either model alone. These findings affirm that integrating spatial and temporal learning capacities enhances model robustness across domains.

Transformers consistently outperformed traditional RNNs and CNNs in text generation and classification. Singh et al. (2023) highlighted the superiority of GPT-2 and CTRL in generating coherent and context-rich text, while Sruthi et al. (2024) reported that transformer models yielded 94.3% accuracy in sentiment classification, surpassing CNNs and RNNs. These results solidify the position of transformer architectures as state-of-the-art solutions for context-aware tasks.

Multimodal and domain-specific adaptations of deep learning also proved highly effective. Xiong et al. (2024) integrated text, audio, and visual inputs using DNNs and attention mechanisms to improve QA systems, leading to enhanced precision and user satisfaction. In the legal domain, Sheik et al. (2024) demonstrated that data augmentation with GPT-generated sentences empowered smaller models like BiGRU to outperform GPT-3, reaching 96% accuracy. This highlights the potential of synthetic data in resource-constrained NLP applications.

Cross-linguistic and low-resource settings received substantial attention. Manzoor et al. (2025) found that traditional models like XGBoost surpassed deep models in Urdu plagiarism detection, achieving 99% accuracy, suggesting that in certain scenarios, handcrafted linguistic features are still highly valuable. Similarly, Ba Alawi and Bozkurt (2024) emphasized the utility of hybrid embeddings in Turkish sentiment analysis, where CNN-BiLSTM with combined embeddings achieved 95.43% accuracy.

Several studies tackled socially relevant tasks. Islam et al. (2024) and Raza et al. (2024) evaluated NLP models for disaster detection and threat identification, respectively. Both reported that deep learning models like BERT and BiLSTM significantly outperformed classical techniques in capturing nuanced textual cues, with BiLSTM achieving an F1-score of 91.61% in threat detection.

Furthermore, domain adaptation was effectively demonstrated in biomedical NLP by Martín-Noguerol et al. (2024), where CNN models successfully differentiated between gliomas and metastases using textual radiology reports. This reinforces the adaptability of deep learning for complex diagnostic tasks in healthcare.

Therefore, underscores that deep learning models, particularly transformers, hybrids, and domain-tailored frameworks, offer superior performance across a broad spectrum of NLP tasks. However, challenges such as interpretability, computational cost, and cross-domain generalization persist. Continued research into efficient architectures and data augmentation strategies, as evidenced by Qiu et al. (2021) and Rana et al. (2025), will be essential for expanding the impact of NLP across real-world applications.

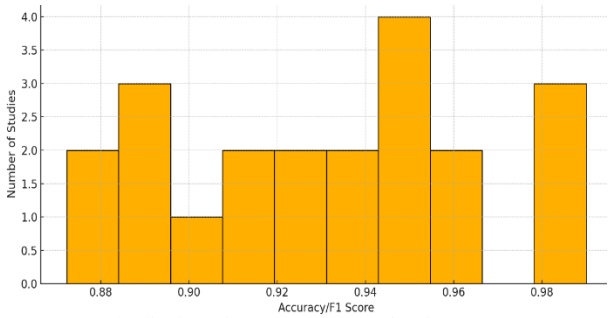

Fig. 1: Distribution of Reported Accuray/Performance Scores.

Figure 1 presents a histogram summarizing the accuracy and F1-scores reported across 21 NLP studies, as detailed in Table 1 of the reviewed paper. The histogram is segmented into bins representing different performance ranges, with most models achieving scores between 0.94 and 0.96, indicating a strong trend toward high-performing models in recent research. Notably, a significant number of studies also cluster in the 0.98–1.0 range, showcasing the effectiveness of advanced architectures such as transformers, hybrids, and optimized transfer learning techniques. The tail ends of the distribution, particularly scores around 0.88 to 0.90, are less populated, suggesting that most recent deep learning approaches have surpassed earlier performance baselines. Overall, this distribution reflects a growing maturity in deep learning-based NLP, with models consistently achieving high predictive power, especially when fine-tuned on domain-specific tasks or supported by hybrid frameworks. The histogram reinforces the conclusion that deep learning continues to push the performance boundaries across various NLP applications.

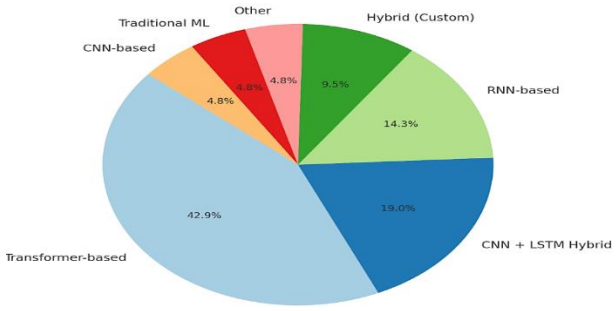

Fig. 2: Proportional Use of Deep Learning Model Categories in NLP Research.

Figure 2: The pie chart illustrates the distribution of deep learning model categories used across selected NLP studies. Transformer-based models dominate with 42.9%, reflecting their superior performance and widespread adoption. CNN + LSTM hybrids follow with 19%, effectively capturing spatial and temporal features. RNN-based models contribute 14.3%, showing continued relevance in sequential tasks. Custom hybrid models make up 9.5%, used in domain-specific solutions. CNN-only, traditional ML, and other models each represent 4.8%. This distribution highlights the growing preference for transformer and hybrid architectures in modern NLP research.

5. Conclusion

Deep learning has significantly advanced the field of Natural Language Processing (NLP), achieving remarkable accuracy across a variety of applications. Transformer-based models have become predominant, representing 42.9% of research utilization, while hybrid architectures such as CNN-LSTM account for 19%. Empirical studies indicate performance peaks: BERT combined with LSTM has attained 98.4% accuracy in news classification, whereas hybrid models have achieved F1-scores exceeding 0.92 in sentiment analysis. This underscores the enhanced context-awareness and adaptability of deep learning models, particularly in specialized and multilingual contexts. However, challenges remain, including interpretability, substantial computational demands, and performance limitations in low-resource environments. Future research should focus on developing efficient architectures (e.g., model compression), self-supervised learning techniques, and promoting fairness in artificial intelligence. Data augmentation, especially through synthetic inputs, holds potential for alleviating data scarcity. As deep learning continues to evolve, its integration with multimodal inputs, transfer learning frameworks such as EasyTransfer, and explainable AI will be essential for creating adaptive, generalizable, and reliable NLP systems. Ongoing advancements in this domain will not only enhance machine comprehension of human language but also drive innovations across sectors such as healthcare, law, and education.

References

[1] A. Torfi, R. A. Shirvani, Y. Keneshloo, N. Tavaf, and E. A. Fox, “Natural Language Processing Advancements By Deep Learning: A Survey,” pp. 1–23, 2020, [Online]. Available: http://arxiv.org/abs/2003.01200.

[2] A. Raj, R. Jindal, A. K. Singh, and A. Pal, “A Study of Recent Advancements in Deep Learning for Natural Language Processing,” *Proc. - 2023 IEEE World Conf. Appl. Intell. Comput. AIC 2023*, pp. 300–306, 2023, https://doi.org/10.1109/AIC57670.2023.10263979.

[3] K. MOHAMAD and K. M. KARAOĞLAN, “Enhancing Deep Learning-Based Sentiment Analysis Using Static and Contextual Language Models,” *Bitlis Eren Üniversitesi Fen Bilim. Derg.*, vol. 12, no. 3, pp. 712–724, 2023, https://doi.org/10.17798/bitlisfen.1288561.

[4] P. R. Kshirsagar, D. H. Reddy, M. Dhingra, D. Dhabliya, and A. Gupta, “A Review on Application of Deep Learning in Natural Language Processing,” *Proc. 5th Int. Conf. Contemp. Comput. Informatics, IC3I 2022*, pp. 1834–1840, 2022, https://doi.org/10.1109/IC3I56241.2022.10073309.

[5] E. O. Arkhangelskaya and S. I. Nikolenko, “Deep Learning for Natural Language Processing: A Survey,” *J. Math. Sci. (United States)*, vol. 273, no. 4, pp. 533–582, 2023, https://doi.org/10.1007/s10958-023-06519-6.

[6] Z. Wang and Z. Zhang, “Research Convey on Text Classification Method based on Deep Learning,” *2022 7th Int. Conf. Intell. Comput. Signal Process. ICSP 2022*, pp. 285–288, 2022, https://doi.org/10.1109/ICSP54964.2022.9778518.

[7] M. Haroon, "Deep Learning Based Question Answering System: A Review," 2019, https://doi.org/10.20944/preprints202312.1739.v1.

[8] R. Rejimoan, B. Gnanapriya, and J. S. Jayasudha, "A Comprehensive Review on Deep Learning Approaches for Question Answering and Machine Reading Comprehension in NLP," *2nd Ed. IEEE Delhi Sect. Own. Conf. DELCON 2023 - Proc.*, pp. 1–6, 2023, https://doi.org/10.1109/DELCON57910.2023.10127327.

[9] B. Alshemali and J. Kalita, "Improving the Reliability of Deep Neural Networks in NLP: A Review," *Knowledge-Based Syst.*, vol. 191, p. 105210, 2020, https://doi.org/10.1016/j.knosys.2019.105210.

[10] J. Liu, X. Chu, Y. Wang, and M. Wang, "Deep Text Retrieval Models based on DNN, CNN, RNN and Transformer: A review," *Proc. 2022 8th IEEE Int. Conf. Cloud Comput. Intell. Syst. CCIS 2022*, pp. 391–400, 2022, https://doi.org/10.1109/CCIS57298.2022.10016379.

[11] Y. Chen, H. Wang, K. Yu, and R. Zhou, "Artificial Intelligence Methods in Natural Language Processing: A Comprehensive Review," *Highlights Sci. Eng. Technol.*, vol. 85, pp. 545–550, 2024, https://doi.org/10.54097/vfwgas09.

[12] D. W. Otter, J. R. Medina, and J. K. Kalita, "A Survey of the Usages of Deep Learning for Natural Language Processing," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 32, no. 2, pp. 604–624, 2021, https://doi.org/10.1109/TNNLS.2020.2979670.

[13] Z. Yang, "Deep Learning Applications in Natural Language Processing and Optimization Strategies," pp. 1–5. https://doi.org/10.70767/jmec.v1i2.257.

[14] D. S. Asudani, N. K. Nagwani, and P. Singh, *Impact of word embedding models on text analytics in deep learning environment: a review*, vol. 56, no. 9. Springer Netherlands, 2023. https://doi.org/10.1007/s10462-023-10419-1.

[15] Muhammad Zulqarnain *et al.*, "Text Classification Using Deep Learning Models: A Comparative Review," *Cloud Comput. Data Sci.*, pp. 80–96, 2023, https://doi.org/10.37256/ccds.5120243528.

[16] M. Gupta, S. K. Verma, and P. Jain, "Detailed Study of Deep Learning Models for Natural Language Processing," *Proc. - IEEE 2020 2nd Int. Conf. Adv. Comput. Commun. Control Networking, ICACCCN 2020*, pp. 249–253, 2020, https://doi.org/10.1109/ICACCCN51052.2020.9362989.

[17] S. Singh, N. Zaidi, and A. Singh, "Deep learning for natural language understanding: A review of recent advances," *Int. J. Appl. Res.*, vol. 4, no. 10, pp. 310–314, 2018, https://doi.org/10.22271/allresearch.2018.v4.i10d.11459.

[18] Z. Xu, "Research on Deep Learning in Natural Language Processing," *Adv. Comput. Commun.*, vol. 4, no. 3, pp. 196–200, 2023, https://doi.org/10.26855/acc.2023.06.018.

[19] J. Liu *et al.*, "Application of Deep Learning-Based Natural Language Processing in Multilingual Sentiment Analysis," *Mediterr. J. Basic Appl. Sci.*, vol. 08, no. 02, pp. 243–260, 2024, https://doi.org/10.46382/MJBAS.2024.8219.

[20] Z. Xiong, L. Zeng, Y. Wu, J. Li, X. Yuan, and B. Mo, "Application of Deep Neural Networks Integrating Multimodal Information in Intelligent Question Answering Systems," *Proc. - 2024 3rd Int. Conf. Artif. Intell. Auton. Robot Syst. AIARS 2024*, pp. 693–698, 2024, https://doi.org/10.1109/AIARS63200.2024.00131.

[21] Z. Chen, "Neural Language Models in Natural Language Processing," *Proc. - 2023 2nd Int. Conf. Data Anal. Comput. Artif. Intell. ICDACAI 2023*, pp. 521–524, 2023, https://doi.org/10.1109/ICDACAI59742.2023.00104.

[22] S. Sruthi, V. G. Trinath, V. Jayanth, V. P. Balaji, T. Singh, and A. Mandal, "Natural Language Processing for Sentiment Analysis with Deep Learning," *2024 3rd Int. Conf. Innov. Technol. INOCON 2024*, pp. 1–6, 2024, https://doi.org/10.1109/INOCON60754.2024.10511769.

[23] Q. Liu and X. Wang, "The application of deep Learning-Based natural language processing models in sentiment analysis," *2024 5th Int. Conf. Electron. Commun. Artif. Intell.*, pp. 686–689, 2024, https://doi.org/10.1109/ICECAI62591.2024.10674860.

[24] G. Geddam, G. Dharmaraju, G. P. Kumar, M. Babu Ketha, and A. Lakshmanarao, "Exploring Deep Learning Approaches for News Classification with CNNs, RNNs and Transformers," *2024 1st Int. Conf. Innov. Commun. Electr. Comput. Eng. ICICEC 2024*, pp. 1–5, 2024, https://doi.org/10.1109/ICICEC62498.2024.10808249.

[25] B. Singh, A. Kumar, S. Kaur, S. Shekhar, and G. Singh, "Exploring the Effectiveness of Various Deep Learning Techniques for Text Generation in Natural Language Processing," *2023 Int. Conf. Adv. Comput. Commun. Inf. Technol. ICAICCIT 2023*, pp. 70–75, 2023, https://doi.org/10.1109/ICAICCIT60255.2023.10466068.

[26] L. Cao, "Sentiment Analysis of Social Media Text Based on Deep Learning," *3rd IEEE Int. Conf. Mob. Networks Wirel. Commun. ICMNWC 2023*, pp. 1–5, 2023, https://doi.org/10.1109/ICMNWC60182.2023.10435901.

[27] Prathyakshini and J. Shetty, "DeepText: Pioneering the Future of Text Classification with Innovative Deep Learning Techniques," *5th Int. Conf. Electron. Sustain. Commun. Syst. ICESC 2024 - Proc.*, no. Icesc, pp. 911–917, 2024, https://doi.org/10.1109/ICESC60852.2024.10689751.

[28] E. al. P. Vijaya Lakshmi, "Advances in Sentiment Analysis in Deep Learning Models and Techniques," *Int. J. Recent Innov. Trends Comput. Commun.*, vol. 11, no. 9, pp. 474–482, 2023, https://doi.org/10.17762/ijritcc.v11i9.8831.

[29] M. Qiu *et al.*, "EasyTransfer: A Simple and Scalable Deep Transfer Learning Platform for NLP Applications," *Int. Conf. Inf. Knowl. Manag. Proc.*, pp. 4075–4084, 2021, https://doi.org/10.1145/3459637.3481911.

[30] M. G, N. R.K, M. M, and S. V, "An enhanced framework for smart automated evaluations of answer scripts using NLP and deep learning methods," *Multimed. Tools Appl.*, 2024, https://doi.org/10.1007/s11042-024-19182-z.

[31] S. Purohit *et al.*, "Analyzing two decades of media sentiments: NLP and deep learning insights into news bias and trends," *Iran J. Comput. Sci.*, 2025, https://doi.org/10.1007/s42044-025-00235-x.

[32] R. Sheik, K. P. Siva Sundara, and S. J. Nirmala, "Neural Data Augmentation for Legal Overruling Task: Small Deep Learning Models vs. Large Language Models," *Neural Process. Lett.*, vol. 56, no. 2, pp. 1–21, 2024, https://doi.org/10.1007/s11063-024-11574-4.

[33] M. F. Manzoor, M. S. Farooq, and A. Abid, *Stylometry-driven framework for Urdu intrinsic plagiarism detection: a comprehensive analysis using machine learning, deep learning, and large language models*, vol. 37, no. 9. Springer London, 2025. https://doi.org/10.1007/s00521-024-10966-w.

[34] M. A. Islam, F. Rabbi, and N. U. I. Hossain, "Performance evaluation of NLP and CNN models for disaster detection using social media data," *Soc. Netw. Anal. Min.*, vol. 14, no. 1, pp. 1–17, 2024, https://doi.org/10.1007/s13278-024-01374-y.

[35] P. Tüfekci and M. Bektaş Kösesoy, "Biological gender identification in Turkish news text using deep learning models," *Multimed. Tools Appl.*, vol. 83, no. 17, pp. 50669–50689, 2024, https://doi.org/10.1007/s11042-023-17622-w.

[36] A. Ba Alawi and F. Bozkurt, "Performance Analysis of Embedding Methods for Deep Learning-Based Turkish Sentiment Analysis Models," *Arab. J. Sci. Eng.*, no. X, 2024, https://doi.org/10.1007/s13369-024-09360-4.

[37] Y. Wang, M. Guo, X. Chen, and D. Ai, "Screening of multi deep learning-based de novo molecular generation models and their application for specific target molecular generation," *Sci. Rep.*, vol. 15, no. 1, pp. 1–15, 2025, https://doi.org/10.1038/s41598-025-86840-z.

[38] M. R. R. Rana, A. Nawaz, S. U. Rehman, M. A. Abid, M. Garayevi, and J. Kajanová, "BERT-BiGRU-Senti-GCN: An Advanced NLP Framework for Analyzing Customer Sentiments in E-Commerce," *Int. J. Comput. Intell. Syst.*, vol. 18, no. 1, pp. 1–18, 2025, https://doi.org/10.1007/s44196-025-00747-1.

[39] M. O. Raza *et al.*, "Reading Between the Lines: Machine Learning Ensemble and Deep Learning for Implied Threat Detection in Textual Data," *Int. J. Comput. Intell. Syst.*, vol. 17, no. 1, 2024, https://doi.org/10.1007/s44196-024-00580-y.

[40] T. Martín-Noguerol, P. López-Úbeda, A. Pons-Escoda, and A. Luna, "Natural language processing deep learning models for the differential between high-grade gliomas and metastasis: what if the key is how we report them?," *Eur. Radiol.*, vol. 34, no. 3, pp. 2113–2120, 2024, https://doi.org/10.1007/s00330-023-10202-4.