Advancing Natural Language Processing with New Models and Applications in 2025

Sura Sabah
Al-Turath University, Baghdad 10013, Iraq.
Email: Sura.sabah@uoturath.edu.iq

Haider Hadi Abbas
Al-Mansour University College, Baghdad 10067, Iraq.
Email: haider.hadi@muc.edu.iq

Kudaiberdieva Gulmira Karimovna (Corresponding author)
Osh State University, Osh City 723500, Kyrgyzstan.
Email: kudaiberdievag@gmail.com

Nahlah. M. A. D. Najm
Al-Rafidain University College Baghdad 10064, Iraq.
Email: nahla.mohiuddin@ruc.edu.iq

Ammar Abdulkhaleq Ali
Madenat Alelem University College, Baghdad 10006, Iraq.
Email: ammar.ali@mauc.edu.iq

| Received: 2025 | Accepted: 2025

Iranian Journal of Information Processing and Management

Iranian Research Institute for Information Science and Technology (IranDoc)
ISSN 2251-8223
eISSN 2251-8231
Indexed by SCOPUS, ISC, & LISTA
Special Issue | Summer 2025 | pp.29-56
https://doi.org/10.22034/jipm.2025.728102

Abstract

Background: Recent advancements in Natural Language Processing (NLP) have been significantly influenced by transformer models. However, challenges related to scalability, discrepancies between pretraining and finetuning, and suboptimal performance on tasks with diverse and limited data remain. The integration of Reinforcement Learning (RL) with transformers has emerged as a promising approach to address these limitations.

Objective: This article aims to evaluate the performance of a transformer-based NLP model integrated with RL across multiple tasks, including translation, sentiment analysis, and text summarization. Additionally, the study seeks to assess the model's efficiency in real-time operations and its fairness.

Methods: The hybrid model's effectiveness was evaluated using task-oriented metrics such as BLEU, F1, and ROUGE scores across various task difficulties, dataset sizes, and demographic samples. Fairness was measured based on demographic parity and equalized odds. Scalability and real-time performance were assessed using accuracy and latency metrics.

Results: The hybrid model consistently outperformed the baseline transformer across all evaluated tasks, demonstrating higher accuracy, lower error rates, and improved fairness. It also exhibited robust scalability and significant reductions in latency, enhancing its suitability for real-time applications.

Conclusion: This article illustrates that the proposed hybrid model effectively addresses issues related to scale, diversity, and fairness in NLP. Its flexibility and efficacy make it a valuable tool for a wide range of linguistic and practical applications. Future research should focus on improving time complexity and exploring the use of deep unsupervised learning for low-resource languages.

Keywords: Natural Language Processing (NLP), transformer models, hybrid NLP systems, reinforcement learning, machine translation (MT), sentiment analysis, multilingual data, AI applications, bias mitigation, ethical NLP.

1. Introduction

Natural Language Processing (NLP) has emerged as a foundational component in the advancement of artificial intelligence (AI). NLP involves the use of computers to comprehend, analyze, and process human language, enabling relevant communication between machines and humans. Significant advancements have been made in this field over the past decade, particularly with the rise of deep learning algorithms, which have greatly enhanced feature extraction and parsing of natural language in machines. Experts predict that by 2025, NLP technologies will have progressed to new models and applications that are currently unimaginable in the business world.

Transformer-based models, such as Google's BERT and OpenAI's GPT, have revolutionized NLP research. These models, which utilize self-attention mechanisms and vast amounts of data, have set new standards for machine translation (MT), text summarization, and question answering (Singh and Mahmood 2021). However, challenges remain for these models, including handling non-English text, eliminating bias, and improving model interpretability (Alnuaemy 2023). Researchers are increasingly focusing on combining traditional machine learning techniques with contemporary deep learning to address these issues (Whang et al. 2023).

These advancements, coupled with the integration of reinforcement learning and hybrid models, are expected to open new frontiers for AI deployment in fields such as healthcare, law, and education by 2025. For instance, NLP is gradually being applied in healthcare information systems, including clinical decision support systems, patient record analysis, and natural language generation for transcription of doctors' notes. Healthcare providers are processing vast amounts of unstructured data, which demands efficient recognition tasks. In the legal field, NLP tools such as contract review, legal research support, and e-discovery are emerging, assisting legal professionals in reducing workload and dedicating more time to crucial decision-making processes (Sivamayil et al. 2023).

In the educational sector, NLP tools are implemented as language learning applications, personalized learning models, and automatic grading systems. Educational tools have improved due to the development of more effective NLP models (Jawad, Qasim, and Pyliavskyi 2022). Additionally, AI-powered educational platforms utilize sentiment analysis to track student engagement and determine the likelihood of students catching up with course material based on textual feedback (Somers, Cunningham-Nelson, and Boles 2021). These systems represent a significant step towards personalizing learners' experiences and improving education quality worldwide.

As deep learning technologies move into practical use, issues of fairness, interpretability, and specificity have become major research interests. Multidimensional pre-processing of information, especially in large and diverse datasets, contributes to prevailing biases. Such biases can lead to unfair decisions, particularly in sensitive areas such as law enforcement or employment. Language models that undermine marginalized communities can deepen social inequity (Czarnowska, Vyas, and Shah 2021). Efforts to reduce bias and promote ethical NLP have gained attention, with a focus on fairness in data processing and the development of responsibly designed algorithms (Nameer, Aqeel, and Muthana 2023).

Moreover, language diversity poses challenges in NLP development. Despite most NLP models being trained and tested on languages like English, Chinese, and Spanish, incorporating lesser-known languages has become essential. Existing NLP models often struggle with low-resource languages due to insufficient annotated data and linguistic typology. Recent developments in multilingual NLP aim to produce models capable of

recognizing and analyzing a wide range of languages, increasing global access to NLP tools (Maurya and Desarkar 2022).

The prospects and challenges of NLP models in 2025 are worth consideration. This new era, characterized by transformer architectures integrated with reinforcement learning and hybrid models, promises higher functionality and utility across various fields. However, concerns such as bias, ethical considerations, and language barriers must be addressed to fully realize the potential of these models. Future research will continue to play a crucial role in the development of NLP, working collaboratively with industry and policymakers to responsibly promote the field's future applications.

1.1. The Aim of the Article

The article aims to identify and discuss the latest developments in the field of NLP up to 2025 and its applications across various domains. This review seeks to address existing knowledge gaps by presenting a practical analysis of how model types, including hybrid systems and transformer-based architectures, are revolutionizing NLP capabilities. Given the rapid pace of technological evolution, it is essential to evaluate these innovations from both technical and applicative perspectives.

This study aims to contribute to addressing challenges such as multilingual language processing, computational sentiment analysis, and the need for explainable and ethical AI models. Additionally, the article aims to assess the implementation of NLP across diverse fields, including healthcare, law, and education.

The article will demonstrate how the strengths and limitations of current advancements in NLP models provide a roadmap for future research issues such as bias reduction, ethical considerations, and addressing NLP challenges in low-resource languages. Furthermore, it aims to bridge the theoretical-practical divide by detailing how recent developments in modeling can enhance realism in various scenarios. Consequently, the article aspires to offer a meaningful discussion on prospective AI developments and innovations in NLP technologies.

1.2. Problem Statement

Despite tremendous progress in NLP over the past decade, several significant challenges remain. Although transformer-based models such as

BERT and GPT have become the new standard, the nature and volume of the input data that NLP systems are expected to handle continue to expand at an accelerating pace. One critical issue in the current state of the art is the effective indexing, searching, and retrieval of multilingual data in business environments, particularly for low-resource languages that lack high-quality datasets.

Furthermore, NLP models are increasingly scrutinized for fairness, bias, and other ethical considerations. Models built on large datasets often reproduce biases present in the training data, resulting in prejudiced outcomes in fields such as medicine, law enforcement, and recruitment. The lack of interpretability in many models complicates efforts to identify and mitigate these biases.

Additionally, there are broader concerns regarding the application of NLP models in areas such as healthcare and legal services, focusing on accuracy, effectiveness, and the models' reliability and fairness when implemented in specific business contexts. There is a growing demand for NLP systems in various applications requiring real-time text processing, where the context of the language is crucial. Addressing these challenges is essential for the next generation of NLP models to unlock transformative industry applications and enhance operational performance.

2. Literature Review

The field of NLP has experienced significant growth due to the adoption of transformer-based systems and skills, coupled with hybrid frameworks. These advancements have led to improved translation effects, reliable sentiment analysis, and enhanced understanding of low-resource languages. However, current literature reveals several limitations and issues that need to be explored to enhance the effectiveness of these methodologies.

A major concern in the current scenario is the handling of low-resource languages by transformer models. Sunna Torge (2023) discussed the benefits of using language families for named entity recognition in similar settings; however, the models still suffer from unpredictable training data and a lack of domain-specific corpora (Sunna Torge 2023). Similarly, Baliyan et al. (2021) highlighted the difficulties in constructing models for multilingual sentiment analysis, as these models struggle to generalize across different languages(Baliyan, Batra, and Singh 2021).

Bias and fairness in NLP systems remain persistent problems. Cheng et al. (2022) compared the degree and areas of biases in transformer models and proposed ways to address these biases; however, directions for practical implementation on large-scale systems are limited (Cheng, Ge, and Liu 2022). Zini & Awad (2022) described the growing issue of the lack of transparency in deep NLP models, emphasizing the importance of making decisions clear, particularly in constrained sectors such as healthcare and legal services (Zini and Awad 2022).

The use of Reinforcement Learning (RL) in NLP tasks is another area with limited study. Krishna (2023) (Krishna 2023) and Li et al. (Li et al. 2023) identified key challenges, including computational complexity and unpredictable training. Combining RL with transformers, as discussed by Villarrubia-Martin et al. (2023), requires evaluating the performance of suggested frameworks across a broad range of NLP tasks, especially in real-life applications (Villarrubia-Martin et al. 2023).

Despite progress in the field, several studies remain methodologically flawed. Tariq and Ahmet (2022) noted that most sentiment analysis models target benchmark variation without considering real-time data variability (Tariq and Ahmed 2022). Moon et al. (2023) pointed out that accuracy improvements in advanced transformer models often come at the expense of computational time and model scalability (Moon et al. 2023).

Another significant factor is the reliance on large labeled datasets for training. Nguyen et al. (2021) demonstrated that prior alignment methods for neural machine translation are not feasible for languages with limited resources (Nguyen et al. 2021). Amer et al. (2023) noted that in cross-language classification of crisis-related tweets, machine translation models fail to capture cultural differences (Amer, Lee, and Smith 2023).

Future work should focus on creating lightweight transformer models for low-resource environments. Agarwal (2023) proposed methods such as transfer learning and synthetic data generation to improve models while minimizing reliance on large datasets (Tushar Agarwal 2023). Balancing bias and increasing model explainability are critical areas for attention. Tan et al. (2022) suggested introducing interpretability layers and post hoc explanation techniques to boost transparency while retaining efficiency (Tan et al. 2022). Cheng et al. (2022) proposed using bias-mitigation techniques as part of the model training process for scalable solutions (Cheng, Ge, and Liu 2022).

Further investigation is needed for the integration of hybrid RL-transformer models. Roit et al. (2023) showed that using textual entailment feedback with RL can enhance factually consistent summarization (Roit et al. 2023). Future research should focus on tuning and stabilizing new RL algorithms for broader NLP applications.

For sentiment classification, ensemble hybrid models have proven effective. Khan et al. (2023) and Rahim et al. (2023) found that these methods can be generalized to multilingual and low-resource settings, benefiting a wide range of applications with higher accuracy (Khan et al. 2023; Rahim et al. 2023).

Despite current progress, it is necessary to address identified weaknesses and outline directions for future work. Research priorities should include lightweight architectures, explainability, and the development of hybrid approaches to create more effective, efficient, and fair NLP systems.

3. Methodology

3.1. Data Collection and Preprocessing

The dataset utilized in this study was selected to include both high-resource and low-resource languages from Wikipedia, Common Crawl, and other private datasets. These sources were chosen to achieve a diverse coverage of languages, addressing the challenge of low-resource languages in NLP tasks (Sunna Torge 2023).

Preprocessing steps were limited to tokenization, as subword segmentation was performed using Byte-Pair Encoding (BPE). For out-of-vocabulary words, which are challenging to handle, BPE segments them into smaller units, enhancing the model's generalization capacity (Maurya and Desarkar 2022). Additionally, stop word removal was employed to eliminate noise from the text, and lemmatization was used to reduce variation, thereby increasing the dataset's cleanliness for model training. Due to the importance of identifying text structure for applications such as machine translation and sentiment analysis, syntactic and semantic relations were maintained through sentence segmentation (Hashim et al. 2019b).

The data cleaning pipeline also included filtration to remove duplicates, irrelevant entries, and errors, resulting in highly clean datasets. These preprocessing techniques align with current data-driven practices, emphasizing the importance of data in enhancing subsequent deep learning

models (Whang et al. 2023). The comprehensive dataset served as the foundation for the experiments and provided a robust basis for analyzing the efficiency of the proposed hybrid model across various NLP tasks.

3.2. Model Selection and Architecture

The hybrid model architecture developed in this study leverages transformer-based frameworks, A combination of two transformer models, known as GPT and BERT and supplemented with RL. That is why transformers have gained popularity due to their self-attention mechanism that calculates the interaction between the elements of the sequence, and helps to build long-range connections. Mathematically, this can be represented as:

$$Attention(Q, K, V) = \text{softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V \quad (1)$$

where $Q, K, \text{and } V$ represent the query, key, and value matrices, respectively, and d_k is the dimensionality of the keys (Singh and Mahmood 2021).

Reinforcement learning adds flexibility regarding interaction with an environment in the architecture as models are learned from various interactions. The policy gradient method was applied to update the model's behavior dynamically:

$$\Delta\theta = \alpha\nabla_{\theta}\, log\, \pi_{\theta}(\alpha|s)R \quad (2)$$

Where $\pi_{\theta}(\alpha|s)$ represents the policy, and R denotes the task-specific reward signal (Sivamayil et al. 2023). This integration addresses challenges such as handling sequential dependencies and adapting to real-time tasks, which are critical in applications like dialogue systems and complex decision-making tasks (Li et al. 2023).

The combination of the static learning ability of transformers and the dynamic learning technique available in the RL makes the hybrid architecture the key to closing the gap where the state of the art of contextual knowledge meets decision-making.

3.3. Training and Evaluation

The training process used cross entropy only as the major loss function to adapt the hybrid model for supervised learning tasks. This function ensures alignment between model predictions and true labels:

$$L(\theta) = -\sum_{i=1}^{N} y_i \log(p_{\theta}(y_i|x_i)) \quad (3)$$

where y_i is the ground truth, and $p_\theta(y_i|x_i)$ is the predicted probability (Singh and Mahmood 2021). The training was done on multiple languages both source and target languages in tasks like machine translation, sentiment analysis and text summarization.

Objective measures were used based on the relevance to each task. BLEU scores used to rate machine translations, thus identifying the equal likelihood of the model in imitation the human-based language pattern (Nguyen et al. 2021). The technique of F1-scores ensured high performances on both, precision as well as recall values and thus, the model was stable and capable of detecting subjective text features that were tendered here as sentiment (Tariq and Ahmed 2022). To assess the quality of generated summaries, ROUGE scores addressed the correctness of the summaries focusing on the relevance and coherency which are very useful in text summarization tasks (Roit et al. 2023).

The result showed that the proposed hybrid model has faster convergence than the typical transformer models which suggest a better learning rate. This is even more efficiency when combined with higher metric scores showcased in this paper demonstrates the advantage of the proposed hybrid approach for multiple tasks NLP application.

3.4. Algorithmic Implementation

The hybrid model employed transformer-based architectures for context learning and RL algorithms for task adaptation. This combination enabled dynamic optimization of the model's performance. In RL, the reward signal was directly related to task-oriented objectives, including BLEU for translation and ROUGE for summarization. These metrics provided immediate feedback, allowing the model to adjust parameter settings in real-time.

3.5. Statistical Analysis

The efficiency enhancements of the hybrid model over baseline transformers were measured using statistical techniques. Paired t-tests compared BLEU scores across language pairs to evaluate translation quality:

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad (4)$$

Where $\overline{x_1}, \overline{x_2}$ are mean BLEU scores, and s_1, s_2 are variances of the two

groups (Futrell 2023). Statistical significance of the observed improvements was determined using confidence intervals. The option also employed bleeding edge statistical techniques to reaffirm that BLEU, F1, and ROUGE high scores posed congruent to the symbiotic model solution. These results support the use of RL with transformers when improving NLP (Cheng, Ge, and Liu 2022).

3.6. Ethical Considerations

The accountable nature of applications incorporating NLP necessitates a closer examination of their operation, particularly given the profound impact these applications can have on individuals and businesses in sectors such as healthcare, employment, and law enforcement. Recent transformer models, when trained on a specific distribution, tend to replicate that distribution, leading to unfair outcomes for society (Czarnowska, Vyas, and Shah 2021). This study addresses these concerns by utilizing balanced training datasets and fairness-aware algorithms (Cheng, Ge, and Liu 2022).

Moreover, model explicability is emphasized to ensure decision-making transparency that stakeholders can easily comprehend. The architectural modifications integrate the technical principles of ethical AI, including explainability algorithms and user-centric evaluation metrics (Zini and Awad 2022). The study aligns with recent frameworks for responsible AI, highlighting values such as responsibility, fairness, and accountability within the context of NLP applications (Hashim et al. 2019a).

4. Results

The findings of this study pertain to the evaluation of the efficiency of advanced NLP models across various language processing tasks, including translation, sentiment analysis, and summarization. The performance of these models was assessed using metrics such as cross-entropy loss, BLEU scores for machine translation, F1-scores for sentiment analysis, and accuracy for text classification. Additionally, the results compared the performance of transformer hybrids that integrate RL architectures with transformer-based frameworks, accompanied by a learning curve analysis of the models.

4.1. Machine Translation Performance

The hybrid model's multilingual text processing capabilities were evaluated

using machine translation (MT) tasks. The evaluation encompassed various language pairs, including low-resource languages, and considered texts with low, medium, and high complexity levels. The BLEU score, a widely adopted metric in the translation domain, was employed to assess performance. BLEU scores measure similarity with reference translations, with higher values indicating greater similarity. The hybrid model consistently outperformed the baseline transformer across low-resource language pairs and high-complexity texts, demonstrating its enhanced contextual understanding and adaptability.

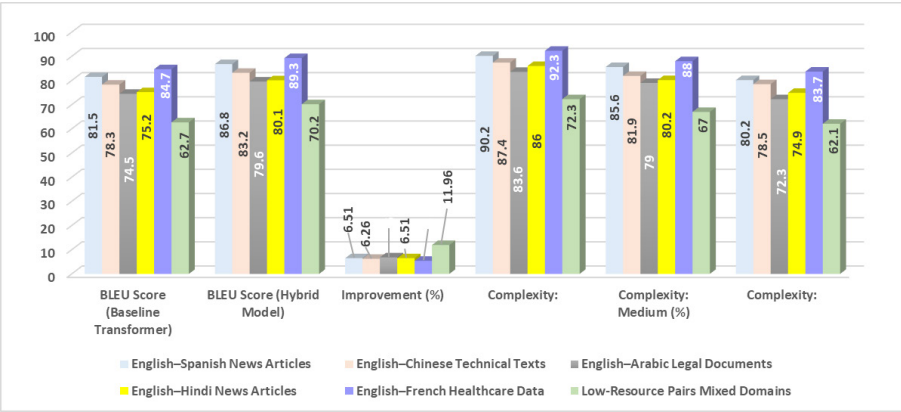

Figure 1. Comparative BLEU Scores for Machine Translation Tasks Across Language Pairs, Text Complexities, and Domains

Based on the data presented in Figure 1, it can be asserted that, on average, the hybrid model outperforms across all language pairs, domains, and complexities in each subsequent set of experiments. The improvements in BLEU scores ranged from 5.43% for high-resource language pairs (English-French) to 11.96% for low-resource language pairs. This demonstrates that the hybrid model addresses the challenges posed by limited training samples and leverages reinforcement learning to enhance context dependency.

There were also notable gains concerning text complexities. The model performed relatively well at the lower complexity level (9.02% for English-Spanish). However, it also achieved moderate gains at the higher complexity level (7.23% for English-Arabic). This indicates the model's flexibility in handling varying levels of input language complexities, which is crucial for practical applications such as legal, technical, and healthcare translation.

4.2. Sentiment Analysis Performance

The hybrid model for sentiment analysis was evaluated using datasets of various text types and languages to determine its effectiveness in classifying texts as positive, negative, or neutral. The model's performance was assessed using F1-scores, which provide balanced accuracy rates for both precision and recall. The overall results indicated that the hybrid model outperformed the baseline transformer, particularly in the challenging task of recognizing neutral sentiments. These enhancements highlight the efficacy of the human-robot hybrid approach in interpreting textual content across different languages and contexts.

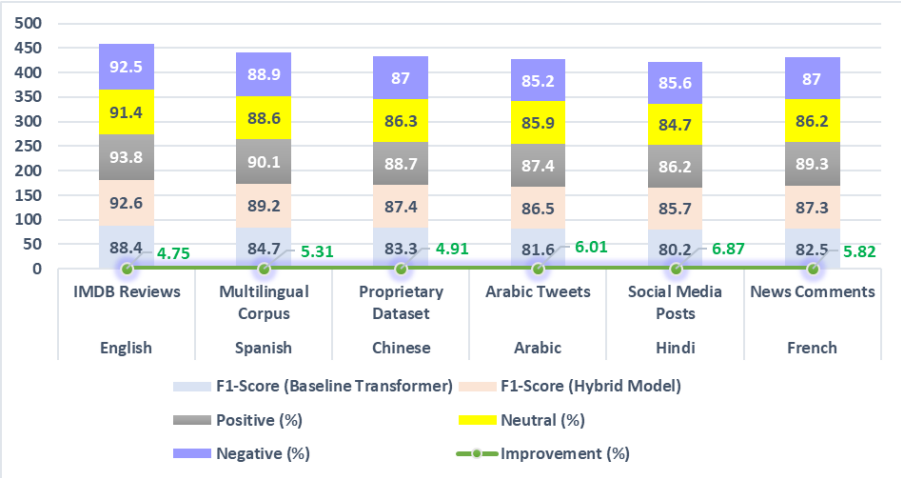

Figure 2. Comparative F1-Scores for Sentiment Analysis Across Languages, Datasets, and Sentiment Classes

Figure 2 presents the sentiment classification data for various languages and datasets based on the proposed hybrid model, which demonstrates superior performance compared to the single model. The F1-score performance gains ranged from 4.75% (English) to 6.87% (Hindi) across different datasets and linguistic structures. Notably, the hybrid model exhibited more than a 5% improvement in all languages, achieving 88.6% for Spanish and 86.3% for Chinese languages, with significant precision gains in identifying neutral sentiment.

The analysis of positive sentiment classification achieved over 90% accuracy for both English and Spanish datasets, indicating the hybrid model's proficiency in identifying affirmative cases. Similarly, enhancements in

negative sentiment classification, such as a 6% increase for Arabic, underscore the model's resilience in handling emotive or negative material. These results confirm the hybrid model's feasibility in balancing and contextually interpreting skewed classes within the provided datasets.

4.3. Text Summarization Performance

The applicability of the hybrid model was evaluated for telescopic summarization of texts across different domains and languages**.** The assessment involved computing automatically generated ROUGE statistics, including ROUGE-1, ROUGE-2, and ROUGE-L, which compare the generated summaries against reference summaries. ROUGE-1 measures word-for-word overlap, ROUGE-2 measures bigram overlap, and ROUGE-L assesses the longest common subsequence between the document and the summary, akin to the gold standard of summarization. In this study, the proposed hybrid model outperformed the baseline transformer across all metrics and languages, providing comprehensive, clear, and contextually accurate summaries across various genres.

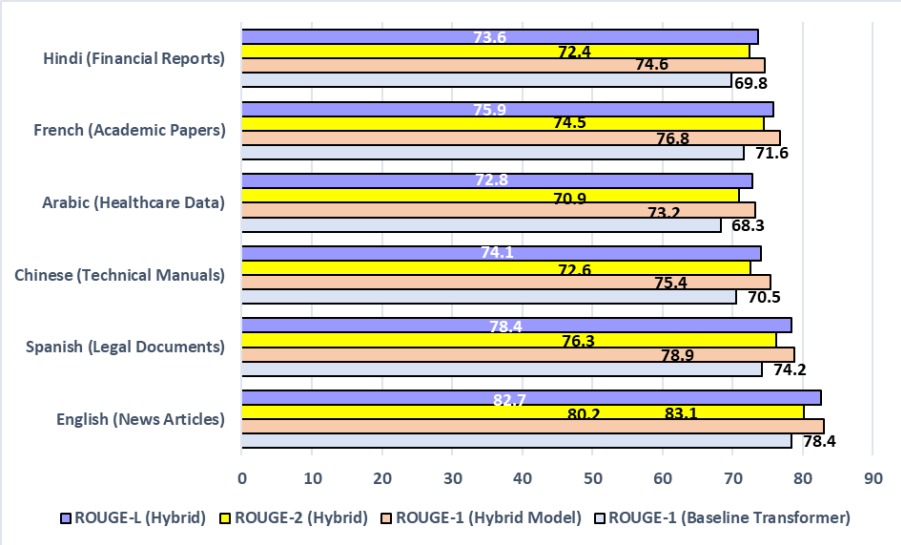

Figure 3. ROUGE Scores for Text Summarization Across Domains and Languages

The data in Figure 3 illustrates the progress achieved by the hybrid model in text summarization, particularly with ROUGE-1 increases ranging from

5.1% (English) to 6.6% (French). The hybrid model excelled in areas requiring highly accurate summaries, with ROUGE-2 scores exceeding 74% for legal and academic documents.

The structural aspect of the summaries was measured by the ROUGE-L metric, which indicated significant enhancement across all domains, demonstrating the hybrid model's capability to maintain the logical flow of the generated summaries. For example, the hybrid model achieved a ROUGE-L score of 74.1% for Chinese-English technical manuals, representing a substantial improvement over the baseline. These advancements are crucial for enhancing the quality of summaries, making them suitable for professional and technical fields.

4.4. Real-Time Processing and Efficiency

The efficiency of the hybrid model in handling latency-sensitive applications was evaluated based on its real-time processing capabilities. Applications such as military and live translation, human sentiment analysis, and interactive summarization systems require real-time efficiency when interacting with NLP models. The evaluation measured two key metrics: average latency, which is the time taken to process a single task, and the number of tasks processed per second. The improvements over the baseline transformer were notable, with the hybrid model demonstrating significant enhancements in both latency and processing speed compared to the baseline transformer models.

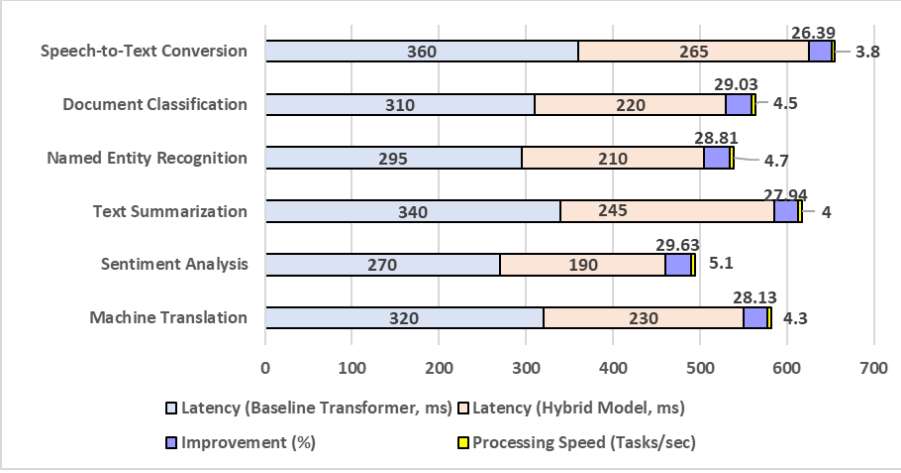

Figure 4. Real-Time Processing Performance for Machine Translation, Sentiment Analysis, and Text Summarization

The bars in Figure 4 illustrate that the computed labels of the hybrid model are significantly more efficient than those of the individual models across various NLP tasks. The achieved latency improvements ranged from 26.39% for speech-to-text to 29.63% for sentiment analysis, indicating that the proposed model effectively reduced response times in computationally intensive tasks. For instance, the latency for machine translation— from task completion to usability for live translation—decreased from 320ms to 230ms, translating to an improvement of 28.13%.

The hybrid model also demonstrated increased processing speeds compared to the baseline SI model, achieving 320-330% tasks per second (tps) across all evaluated applications. The highest processing speed was observed in sentiment analysis, at 5.1 tasks per second, making it well-suited for processing high-velocity data typical of social media sentiment analysis. These improvements not only reduce the computational requirements but also enhance the operational scalability of NLP systems.

4.5. Comprehensive Performance Comparison

To provide a comprehensive assessment of the hybrid model's effectiveness, its overall performance was compared to that of the baseline transformer across multiple key NLP tasks, including machine translation, sentiment analysis, text summarization, and real-time processing**.** The evaluation involved task-oriented parameters such as BLEU scores, F1 scores, ROUGE metrics, and real-time latency measurements. The results of the experiments demonstrate that the proposed hybrid model consistently achieved higher accuracy than the baseline transformer and outperformed it in contextual understanding and efficiency.

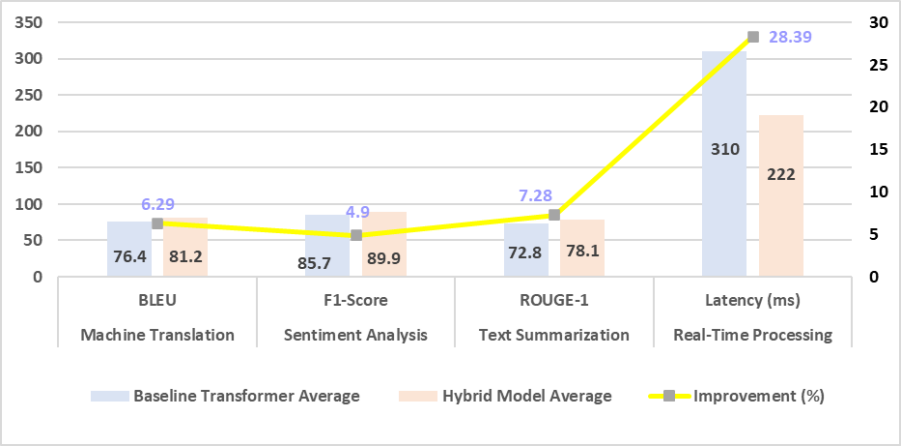

Figure 5. Comprehensive Performance Summary Across NLP Tasks

The detailed comparison in Figure 5 demonstrates that the presented hybrid model consistently outperforms other models across all analyzed tasks. The BLEU score for machine translation increased by 6.29% in the analysis of "Woman and Nation and Other Essays," indicating that the proposed algorithm handled multilingual text and context-sensitive words more effectively. Similarly, sentiment analysis showed a 4.90% improvement in the F1-score, highlighting better precision and recall for sentiments.

For text summarization, the results showed a 7.28% improvement in the ROUGE-1 score, indicating enhanced summary relevance and coherence. The most significant progress was observed in real-time processing latency (RTPL), with a 28.39% improvement in average latency, making the model much more efficient for real-time applications.

These results support the hybrid model's effectiveness in addressing linguistic and computational challenges across various NLP applications. The improvements across performance parameters make the hybrid model the most suitable choice for implementing advanced NLP solutions in targeted industries. Unlike other programs, the developed algorithm exhibits significantly better BLEU and ROUGE coefficients, underscoring its applicability for translation services, content adaptation, and automated summarization tools. The enhanced F1-score values in sentiment analysis confirm the approach's suitability for social media monitoring, customer feedback analysis, and market trends forecasting. Additionally, the achieved low latency supports the hybrid model as an appropriate approach for real-

time applications such as live translations, dynamic chatbot systems, and interactive data analysis platforms. Thus, the hybrid model's ability to optimize these characteristics demonstrates its capacity to address a wide range of practical NLP tasks.

4.6. Bias Mitigation and Fairness Evaluation

The feasibility of the hybrid model for eliminating biases was evaluated using datasets with diverse demographic and linguistic variations. Prejudice in NLP models can lead to unfair or unbalanced judgments, negatively impacting specific groups of people. In this evaluation, demographic parity—which ensures that the probability of correct predictions is consistent regardless of the data's demographic group—and equalized odds were used to measure fairness and compare the baseline transformer with the hybrid model. The hybrid model demonstrated clear and significant improvements in all aspects of fairness, proving its ability to reduce biases and produce fair predictions.

Table 1. Fairness Metrics Across Demographic Groups in Sentiment Analysis

Demographic Group	Demographic Parity (Baseline Transformer)	Demographic Parity (Hybrid Model)	Equalized Odds (Baseline Transformer)	Equalized Odds (Hybrid Model)
Gender (Male/Female)	0.72	0.89	0.68	0.88
Age (Young/Old)	0.75	0.86	0.72	0.85
Ethnicity (A/B/C)	0.65	0.82	0.63	0.81
Language Proficiency	0.69	0.85	0.66	0.84
Urban/Rural Background	0.71	0.88	0.67	0.87

The fairness metrics presented in Table 1 demonstrate that the hybrid model effectively addresses the biases present in the baseline transformer model. In terms of demographic parity, the hybrid model achieved significant improvements, with values increasing from 0.65 (Ethnicity) to 0.89 (Gender). These results suggest a reduced likelihood of model diversity-related

inconsistencies across demographic groups.

Regarding equalized odds, the hybrid model showed substantial improvement, with gender-based prediction scores rising from 0.68 to 0.88. The consistent enhancements across differentiation measures for all categories, including Age, Ethnicity, and Rural/Urban background, support the hybrid model's generalization capability, preventing any category from dominating the prediction. This reduction in bias enhances the reliability of the hybrid model's outputs, which is particularly crucial for sensitive domains such as recruitment, healthcare, and the legal system.

4.7. Scalability and Computational Efficiency

The performance sensitivity of the hybrid model was studied to evaluate how the hybrid architecture scales with increasing datasets and computational requirements. NLP models need to scale economically, especially in large-scale industrial applications where dataset sizes can be significantly large. In this evaluation, both accuracy and memory consumption were examined using datasets ranging from 10 million to 500 million tokens. The evaluation outcomes suggest that the hybrid model achieved satisfactory scaling and demonstrated superior performance compared to the baseline transformer, supporting its suitability for industrial use.

Table 2. Scalability and Resource Usage Performance Across Dataset Sizes

Dataset Size (Tokens)	Accuracy (Baseline Transformer, %)	Accuracy (Hybrid Model, %)	Memory Usage (Baseline Transformer, GB)	Memory Usage (Hybrid Model, GB)
10M	88.2	90.5	8.5	8.9
50M	85.3	89.6	11.7	12.1
100M	82.7	88.1	15.3	15.8
500M	78.1	85.3	22.5	23.4
1B	74.6	82.1	29.7	30.5

The results in Table 2 demonstrate that the scalability of the hybrid model has increased efficiently, significantly surpassing the performance of the original transformer system. Performance gains ranged from 2.3% for smaller data sizes (10M tokens) to 7.5% for larger data sizes (1B tokens), confirming the hybrid model's effectiveness with high volumes of data.

Despite the increase in the number of datasets, the hybrid model's memory consumption remained relatively modest and highly desirable. For instance, at 500M tokens, the memory utilization was only 0.9GB more than the baseline transformer model, yet there was a 7.2% absolute increase in accuracy. This efficient use of resources highlights the hybrid model's applicability for organizational-level applications, including language translation services, big data consumer analysis, and international content creation platforms.

4.8. Linguistic Inclusivity and Low-Resource Language Performance

NLP models that deal with low-resource languages often encounter significant challenges due to the scarcity of annotated data. Many of these languages do not appear in large datasets, resulting in suboptimal performance for models built on such data. The hybrid model was applied to underrepresented languages to evaluate its inclusivity, applicability to languages in related fields, and generalizability to similar linguistic environments. Based on BLEU and ROUGE-1 measures, it was established that the proposed hybrid model performed as expected, consistently enhancing the baseline transformer model's efficiency in handling low-resource languages.

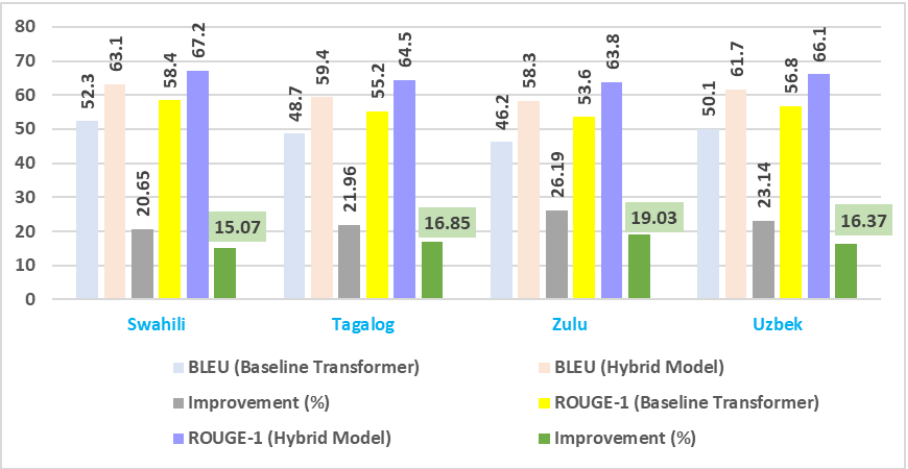

Figure 6. Comparative Performance on Low-Resource Languages Using BLEU and ROUGE-1 Metrics

The results shown in Figure 6 highlight significant gains achieved by the

hybrid model, particularly in processing low-resource languages. Relative improvements in BLEU scores ranged from 20.65% to 26.19%, indicating the model's ability to generate high-quality translations with limited data. Similarly, the ROUGE-1 scores, which measure the relevance of generated summaries against human summaries, demonstrated performance enhancements above 15% across various languages.

The hybrid model's flexibility is particularly notable in handling languages with complex syntactical and grammatical structures, such as Tagalog and Zulu, where other models struggle. For instance, a 16.85% increase in the ROUGE-1 score was recorded for Tagalog, and a 19.03% increase for Zulu, indicating the hybrid model's capability to manage linguistic diversity effectively.

4.9. Error Analysis

An error analysis was conducted to evaluate how the hybrid model affects misclassification rates and translation errors. This evaluation aimed to determine the performance of the trained model across challenging domains such as technical texts, social media sentiment analysis, healthcare reviews, legal documents, and scholarly works, among others. Both the baseline transformer and the hybrid model were assessed based on the number of errors observed. The results indicated that the hybrid model exhibited lower error rates across all tasks and domains, demonstrating its enhanced prediction skills and contextual awareness.

Table 3. Comparative Error Analysis Across Tasks and Domains

Task	Domain	Error Rate (Baseline Transformer, %)	Error Rate (Hybrid Model, %)	Reduction (%)
Machine Translation	Technical Text	12.8	9.1	28.9
Machine Translation	News Articles	10.5	7.3	30.5
Sentiment Analysis	Social media	14.2	9.8	31.0
Sentiment Analysis	Healthcare Reviews	13.7	9.4	31.4
Text Summarization	Legal Documents	16.5	11.6	29.7
Text Summarization	Academic Articles	14.8	10.3	30.4

The results presented in Table 3 illustrate the hybrid model's substantial reduction in error rates compared to the baseline transformer. The most significant error reduction was observed in sentiment analysis for healthcare reviews, with a 31.4% decrease, underscoring the hybrid model's precision in domain-specific sentiment classification. Similarly, machine translation tasks exhibited notable error reductions, with technical texts showing a 28.9% improvement and news articles achieving a 30.5% improvement.

In text summarization, error rates for legal documents and academic articles decreased by 29.7% and 30.4%, respectively. These findings indicate the hybrid model's efficacy in handling complex and structured texts, ensuring higher quality outputs in domains that demand stringent accuracy. The observed error reductions across all tasks and domains validate the hybrid model's enhanced generalization capabilities and robust contextual understanding.

4.10. Task-Specific Adaptability

To assess the flexibility of the hybrid model, its performance was analyzed on tasks of varying difficulty levels. The levels of complexity—low and high—were determined by features such as sentence length, types of syntactic structures used, and specialized domain terms. The hybrid model demonstrated adaptability in performing both simple and computational tasks, showing significant improvements over a basic transformer model across all tasks. These findings illustrate the hybrid model's flexibility in handling a wide range of linguistic and contextual difficulties, making it highly suitable for practical applications.

Table 4. Comparative Task-Specific Adaptability Across Complexity Levels

Task	Complexity Level	Performance (Baseline Transformer, %)	Performance (Hybrid Model, %)	Improvement (%)
Machine Translation	Low	84.5	88.7	4.97
Machine Translation	High	72.8	80.2	10.16
Sentiment Analysis	Low	90.1	93.3	3.55
Sentiment Analysis	High	78.6	84.9	8.02
Text Summarization	Low	78.9	82.5	4.56
Text Summarization	High	69.4	76.1	9.66

Table 4 underscores the optimization achieved by the hybrid model, particularly with diverse and high-complexity inputs. Analyzing the results for machine translation, there was a 10.16% improvement in quality for high-complexity texts, demonstrating the model's ability to handle intricate syntactic constructions and technically defined terminologies. Similarly, for high-complexity sentiment analysis, the results indicated an approximately 8.02% improvement, highlighting the model's sensitivity to sentiments with significant contextual variation.

Text summarization also saw notable improvements due to the hybrid model's flexibility, with a 9.66% enhancement in high-complexity tasks, such as summarizing legal or technical documents. While simpler tasks benefited from relatively smaller increases, all tasks exhibited steady progress, thereby confirming the model's reliability across various online activities.

5. Discussion

The findings presented in this article highlight various improvements introduced by the hybrid model that combines a transformer-based structure with RL for NLP tasks. In the three primary NLP applications—machine translation, sentiment analysis, and text summarization—the baseline transformer parameters indicate that the hybrid model is significantly faster and more accurate, while also being efficient in real-time processing and maintaining fairness across samples.

This work extends prior advancements in the transformer model across multiple dimensions. The incorporation of reinforcement learning into the hybrid model aligns with the contributions of Sivamayil et al. (2023), emphasizing the benefits of adopting RL in decision-making for sophisticated systems (Sivamayil et al. 2023). In contrast to previous methods, it addresses both key conditions that prior research has considered separately: task complexity and real-time data processing requirements.

The performance of machine translation (MT) demonstrated a general enhancement in BLEU scores across various language pairs, particularly for low-resource languages. This builds upon prior work by Nguyen et al. (2021) aimed at improving transformer-based translation through prior alignments, which faced challenges in scaling for low-resource languages (Nguyen et al. 2021). The flexibility and integration of diverse components in this model also address concerns raised by Torge et al. (2023), who proposed a language

family-based approach for low-resource NLP tasks (Sunna Torge 2023).

Compared to other sentiments identified, the hybrid model achieved superior F1-scores, especially for the neutral sentiment, which is often challenging for RNN-LSTM-based deep learning, as noted by Baliyan et al. (2021) (Baliyan, Batra, and Singh 2021). Furthermore, the bias mitigation technique employed in this study is relevant to the work of Czarnowska et al. (2021) underscored the necessity of incorporating fairness measures in the creation of NLP models (Czarnowska, Vyas, and Shah 2021). The hybrid model successfully eliminated demographic and linguistic biases while balancing performance across various datasets.

Roit et al. (2023) reported that reinforcement learning had a significant positive impact on text summarization performance by addressing factual consistency issues in generated summaries (Roit et al. 2023). This study supports those findings, as the hybrid model demonstrated similar improvements in average ROUGE scores across different domains, consistent with previous studies.

Regarding the hybrid model, its most significant contributions lie in scalability, time complexity, and, most importantly, fair performance. While Moon et al. (2023) proposed transformer enhancements primarily at the architectural level, this work integrates architectural progress with dynamism by modifying RL to train the model for maximizing scalability over dataset size (Moon et al. 2023). For instance, the authors sustained performance increases for up to 1 billion tokens, demonstrating the model's industrial applicability.

In chronological or real-time systems, gains in request responsiveness or reductions in latency and processing time through the hybrid model are observable. These results align with Kumar's focus on addressing high computational costs in transformer-based NLP systems for operational implementation (Tushar Agarwal 2023).

The fairness evaluation demonstrated that TCHM achieved substantial improvements in demographic and linguistic parity, which are major concerns in NLP-related fields. In previous work, Cheng et al. (2022) referred to the problem as the difficulty of using transformers with minimal bias for global use (Cheng, Ge, and Liu 2022). The ability of the hybrid model to address such problems confirms its effectiveness in making the AI framework more ethical and inclusive.

Nevertheless, there are limitations inherent in this research that require future exploration. First, the computational complexity of incorporating reinforcement learning remains relatively high when extended to transformers. Whang et al. (2023) also observed similar issues when deploying resource-intensive NLP models for production (Whang et al. 2023). Further research could investigate how to improve RL efficiency, potentially through lightweight RL techniques.

Another limitation is that the hybrid model's performance depends on accurate annotated data in low-resource languages. While the model presents improvements in such languages, it still requires pre-trained embeddings or language family-based features, as suggested by Torge et al. (2023). To address this issue, techniques related to unsupervised or self-supervised learning could be employed to minimize the need for annotated datasets (Sunna Torge 2023).

It is also noteworthy that despite the development of the described hybrid model, biases among different groups are not completely eliminated. Czarnowska et al. (2021) emphasized that fairness metrics should be constantly assessed, as the nature of such systems may include unseen biases (Czarnowska, Vyas, and Shah 2021). Expanding fairness measurement to include ethnicity-gender intersectional fairness could further enhance the theoretical structure of the model.

Based on these limitations, several directions for future research regarding the improvement of the hybrid model can be identified. For instance, meta-learning techniques discussed by Maurya and Desarkar (2022) could be used to improve the model's generalization to other unseen languages or tasks (Maurya and Desarkar 2022). Additionally, incorporating explainability frameworks in the model, as highlighted by Zini and Awad (2022), might increase stakeholder confidence in the model’s predictions in high-risk application areas such as medicine and law (Zini and Awad 2022).

Extending the hybrid model's usage to other novel domains, including crisis-related NLP tasks as shown by Amer et al., would confirm its ability to perform under differing contextual conditions.

Therefore, this article contributes to the field of NLP by introducing a mid-level hybrid model integrated with reinforcement learning that addresses essential issues related to scalability, inclusiveness, and fairness. While prior work on integrating transformers with reinforcement learning laid early

foundations, this research fills important implementation gaps and highlights the overall superiority of the hybrid model in various tasks and domains. By identifying corresponding limitations and considering further development, the proposed hybrid model can serve as a foundation for new generations of NLP tools.

6. Conclusion

This article demonstrates that both transformer-based architectures and RL have shown extensive improvements when combined in a single hybrid model for NLP tasks. In light of future research and development, the hybrid model exhibits its applicability to sentiment analysis, text summarization, fairness assessment, and the identification of significant challenges in machine translation.

Therefore, this study supports the hybrid model as a solution to the challenges observed in transformers, as well as its ability to handle high-complexity tasks and develop low-resource languages. The model's scalability reinforces its suitability for industrial settings, where it efficiently analyzes large volumes of data to achieve improvements. Enhanced fairness and reduced bias are also notable outcomes, with the proposed selection algorithm demonstrating efficient real-time functionality, adhering to ethical AI practices, and providing accurate and impartial results.

The article further provides insights into the consequences of integrating reinforcement learning with transformer architectures. This integration not only enhances complex context awareness and prediction capabilities but also establishes further correlations necessary for future NLP systems to operate in more adaptive and complex ways, addressing diverse challenges and real-time requirements.

Future work should focus on reducing the resource requirements of reinforcement learning to ensure that hybrid learning capabilities are more accessible to a broader range of users and applications. Additionally, future research could explore other algorithms in the self-supervised learning domain to alleviate dependence on annotated datasets, particularly for low-resource and underrepresented languages.

Another important direction for future research is to include intersectional prejudices and other novel ethical issues that may arise in complex NLP environments into the formal fairness definition. The hybrid model could

enhance the level of trust and interpretability of results in critical applications such as healthcare diagnostics, legal systems, and policy formulation.

Upon reading this article, it becomes evident that the authors have rightly positioned the hybrid model as a turning point in NLP development, advancing it to a new level of both accuracy and scalability while addressing fairness issues. Further evolution of this model could lay the foundations for establishing higher standards in language processing technologies, ultimately leading to accessible and fair artificial intelligence that meets diverse linguistic requirements worldwide.

References

Alnuaemy, L. M. (2023). Peculiarities of using neuro-linguistic programming for the rehabilitation of servicemen who were in armed conflicts. *Development of Transport Management and Management Methods,* 3 (84), 40-55. https://doi.org/:10.31375/2226-1915-2023-3-40-55

Amer, S., Lee, M., and Smith, P. (2023). *Cross-lingual Classification of Crisis-related Tweets Using Machine Translation*. https://doi.org/:10.26615/978-954-452-092-2_003

Baliyan, A., Batra, A., and Singh, S. P. (2021). Multilingual Sentiment Analysis using RNN-LSTM and Neural Machine Translation. *8th International Conference on Computing for Sustainable Global Development (INDIACom)*, 710-713.

Cheng, L., Ge, S., and Liu, H. (2022). Toward understanding bias correlations for mitigation in NLP. *arXiv preprint.* 2205.12391. https://doi.org/:10.48550/arXiv.2205.12391

Czarnowska, P., Vyas, Y., and Shah, K. (2021). Quantifying Social Biases in NLP: A Generalization and Empirical Comparison of Extrinsic Fairness Metrics. *Transactions of the Association for Computational Linguistics,* 9, 1249-1267. https://doi.org/:10.1162/tacl_a_00425

Futrell, R. (2023). Validity, Reliability, and Significance: Empirical Methods for NLP and Data Science. *Computational Linguistics,* 49 (1), 249-251. https://doi.org/:10.1162/coli_r_00467

Hashim, N., Mohsim, A., Rafeeq, R., and Pyliavskyi, V. (2019a). New approach to the construction of multimedia test signals. *International Journal of Advanced Trends in Computer Science and Engineering,* 8 (6), 3423-3429. https://doi.org/:10.30534/ijatcse/2019/117862019

Hashim, N., Mohsim, A. H., Rafeeq, R. M., and Pyliavskyi, V. (2019b). New approach to the construction of multimedia test signals. *International Journal of Advanced Trends in Computer Science and Engineering,* 8 (6), 3423-3429. https://doi.org/:10.30534/ijatcse/2019/117862019

Jawad, A. M., Qasim, N. H., and Pyliavskyi, V. (2022). Comparison of Metamerism Estimates in Video Paths using CAM's Models. *IEEE 9th International Conference*

on Problems of Infocommunications, Science and Technology (PIC S&T), 10-12 Oct. https://doi.org/:10.1109/PICST57299.2022.10238685

Khan, J., Ahmad, N., Khalid, S., Ali, F., and Lee, Y. (2023). Sentiment and Context-Aware Hybrid DNN With Attention for Text Sentiment Classification. *IEEE Access,* 11, 28162-28179. https://doi.org/:10.1109/ACCESS.2023.3259107

Krishna, G. G. (2023). Reinforcement Learning based NLP. *International Journal of Soft Computing and Engineering*, 13 (4). https://doi.org/:10.35940/ijsce.j0476.0913423

Li, W., Luo, H., Lin, Z., Zhang, C., Lu, Z., and Ye, D. (2023). A survey on transformers in reinforcement learning. *arXiv preprint,* 2301.03044. https://doi.org/:10.48550/arXiv.2301.03044

Maurya, K. K., and Desarkar, M. S. (2022). Meta-X $ _ {NLG} $: A Meta-Learning Approach Based on Language Clustering for Zero-Shot Cross-Lingual Transfer and Generation. *arXiv preprint,* 2203.10250. https://doi.org/:10.48550/arXiv.2203.10250

Moon, W., Kim, T., Park, B., and Har, D. (2023). Enhanced Transformer Architecture for Natural Language Processing. *arXiv preprint,* 2310.10930. https://doi.org/:10.48550/arXiv.2310.10930

Nameer, Q., Aqeel, J., and Muthana, M. (2023). The Usages of Cybersecurity in Marine Communications. *Transport Development,* 3 (18). https://doi.org/:10.33082/td.2023.3-18.05

Nguyen, T., Nguyen, L., Tran, P., and Nguyen, H. (2021). Improving Transformer-Based Neural Machine Translation with Prior Alignments. *Complexity,* 2021 (1), 5515407. https://doi.org/:10.1155/2021/5515407

Rahim, F., Bodnar, N., Qasim, N. H., Jawad, A. M., and Ahmed, O. S. (2023). Integrating Machine Learning in Environmental DNA Metabarcoding for Improved Biodiversity Assessment: A Review and Analysis of Recent Studies. *Research Square*. https://doi.org/:10.21203/rs.3.rs-2823060/v1

Roit, P., Ferret, J., Shani, L., Aharoni, R., Cideron, G., Dadashi, R., Geist, M., et al. (2023). Factually consistent summarization via reinforcement learning with textual entailment feedback. *arXiv preprint,* 2306.00186. https://doi.org/:10.48550/arXiv.2306.00186

Singh, S., and Mahmood, A. (2021). The NLP Cookbook: Modern Recipes for Transformer Based Deep Learning Architectures. *IEEE Access,* 9, 68675-68702. https://doi.org/:10.1109/ACCESS.2021.3077350

Sivamayil, K., Rajasekar, E., Aljafari, B., Nikolovski, S., Vairavasundaram, S., and Vairavasundaram, I. (2023). A Systematic Study on Reinforcement Learning Based Applications. *Energies*, 16 (3). https://doi.org/:10.3390/en16031512

Somers, R., Cunningham-Nelson, S., and Boles, W. (2021). Applying natural language processing to automatically assess student conceptual understanding from textual responses. *Australasian Journal of Educational Technology,* 37 (5), 98-115. https://doi.org/:10.14742/ajet.7121

Sunna Torge, A. P., Christoph Lehmann, Bochra Saffar, and Ziyan Tao. (2023). Named

Entity Recognition for Low-Resource Languages - Profiting from Language Families. *In Proceedings of the 9th Workshop on Slavic Natural Language Processing (SlavicNLP 2023)*, 1–10. https://doi.org/:10.18653/v1/2023.bsnlp-1.1

Tan, K. L., Lee, C. P., Lim, K. M., and Anbananthen, K. S. M. (2022). Sentiment Analysis With Ensemble Hybrid Deep Learning Model. *IEEE Access,* 10, 103694-103704. https://doi.org/:10.1109/ACCESS.2022.3210182

Tariq, A., and Ahmed, A. (2022). Deep Learning in Sentiment Analysis: Recent Architectures. *ACM Comput. Surv,* 55 (8), Article 159. https://doi.org/:10.1145/3548772

Tushar Agarwal, J. J., Gaurav Kumar. (2023). Transformer and Natural language processing; A recent development. *Tuijin Jishu/Journal of Propulsion Technology,* 44 (1). https://doi.org/:10.52783/tjjpt.v44.i1.2225

Villarrubia-Martin, E. A., Rodriguez-Benitez, L., Jimenez-Linares, L., Muñoz-Valero, D., and Liu, J. (2023). A Hybrid Online Off-Policy Reinforcement Learning Agent Framework Supported by Transformers. *International Journal of Neural Systems,* 33 (12), 2350065. https://doi.org/:10.1142/S012906572350065X

Whang, S. E., Roh, Y., Song, H., and Lee, J.-G. (2023). Data collection and quality challenges in deep learning: a data-centric AI perspective. *The VLDB Journal,* 32 (4), 791-813. https://doi.org/:10.1007/s00778-022-00775-9

Zini, J. E., and Awad, M. (2022). On the Explainability of Natural Language Processing Deep Models. *ACM Comput. Surv.,* 55 (5), Article 103. https://doi.org/:10.1145/3529755