iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** V **Month of publication:** May 2025

DOI: https://doi.org/10.22214/ijraset.2025.71832

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue V May 2025- Available at www.ijraset.com

A Review of Transformer-Based Models for Natural Language Processing (NLP)

Jayasudha J
Assistant Professor, Department of Computer Science, Sri Ramakrishna College of Arts & Science for Women, Coimbatore

Abstract: ***In recent years, Transformer-based architecture has revolutionized the field of Natural Language Processing (NLP), enabling significant advancements across a wide range of tasks such as language modeling, text classification, machine translation, and question answering. This review paper provides a comprehensive overview of the development and evolution of these models, beginning with foundational word embedding techniques and progressing through major transformer architectures such as BERT, RoBERTa, sBERT, MiniLMetc...This paper analyzes core mechanisms such as attention mechanisms, pretraining strategies, and fine-tuning approaches, and highlights how they improve performance compared to traditional NLP models.Additionally, the paper explores recent advancements such as model compression, transfer learning, and multilingual modeling. It also addresses key challenges and future research directions, including model interpretability, computational efficiency, and ethical implications. This review is intended to be a comprehensive resource for researchers and practitioners aiming to understand and apply Transformer-based models in natural language processing.***
Keywords: Transformer Models, Natural Language Processing (NLP), BERT, GPT, RoBERTa, Pretraining, Attention Mechanism, Transfer Learning, Deep Learning, Language Models

I. INTRODUCTION

In recent years, Natural Language Processing (NLP) has been revolutionized by deep learning, especially Transformer-based models. Central to NLP are **text embeddings**dense vector representations that capture semantic and syntactic nuances of words, sentences, or documents. Unlike traditional sparse methods such as one-hot encoding, embeddings provide compact, meaningful representations that enhance tasks like sentiment analysis, question answering, and information retrieval.Approaches like Word2Vec and GloVe produced static embeddings, assigning a single vector per word regardless of context, limiting their ability to capture word meaning variations. This was addressed by contextual embeddings introduced by models such as BERT (Bidirectional Encoder Representations from Transformers), which leverage bidirectional context and large unlabeled corpora to generate dynamic, context-aware word representations. Its successor, RoBERTa, further improved performance by optimizing training strategies. Although BERT excels in many tasks, it is computationally expensive for sentence- or document-level similarity due to repeated processing of sentence pairs. To overcome this, Sentence-BERT (SBERT) employs a Siamese network architecture to generate sentence embedding efficiently, enabling faster semantic similarity and clustering. Addressing efficiency and scalability further, MiniLM offers a distilled, lightweight alternative to BERT and RoBERTa, balancing speed and performance for resource-constrained environments.This paper provides an overview of deep learning embeddingsBERT, RoBERTa, SBERT, and MiniLMhighlighting their architecture, training paradigms, and applications, providing insights for both researchers and practitioners.

II. BACKGROUND

A. Traditional Embeddings Vs Deep Learning and Transformer Models

Prior to the adoption of deep contextual models, NLP largely relied on static word embeddings such as Word2Vec and GloVe, which created fixed vector representations for each word based on global co-occurrence statistics [1][2]. While these methods improved semantic understanding over sparse representations, their inability to adapt embeddings dynamically to context limited their effectiveness for words with multiple meanings.

The introduction of the Transformer architecture revolutionized NLP by replacing recurrent structures with self-attention mechanisms, enabling efficient modeling of long-range dependencies and parallelizable training [3]. This architecture underpins many state-of-the-art models. Building upon Transformers, BERT introduced a novel pretraining method using masked language modeling and next sentence prediction to produce deep bidirectional contextual embeddings, capturing rich semantic and syntactic features sensitive to context [4]. This approach dramatically improved performance on numerous NLP benchmarks.

RoBERTa further advanced BERT by optimizing training strategies such as removing next sentence prediction and training on larger datasets with bigger batches, leading to more robust embeddings [5]. These models, however, were computationally expensive for sentence-level similarity tasks, prompting development of more efficient variants.To overcome these challenges, Sentence-BERT (SBERT) adapted BERT to a Siamese network structure, enabling fast generation of semantically meaningful sentence embeddings suitable for large-scale similarity and clustering tasks . Additionally, MiniLM distilled knowledge from larger models into a smaller architecture, offering a competitive performance-efficiency trade-off ideal for resource-constrained environments .

III. TRANSFORMER-BASED EMBEEDING MODELS

Transformer-based models have revolutionized Natural Language Processing by introducing a novel architecture centered around self-attention mechanisms. Unlike traditional recurrent or convolutional neural networks, Transformers process input sequences in parallel and dynamically weigh the importance of each token relative to others in the context. This enables effective modeling of long-range dependencies and complex language patterns. Since their introduction, Transformer models have become the foundation for many state-of-the-art NLP systems, powering tasks such as language understanding, text generation, and semantic embedding generation with remarkable accuracy and efficiency.

A. BERT

Bidirectional Encoder Representations from Transformers (BERT) [6] introduced a paradigm shift in NLP by leveraging the Transformer encoder architecture to produce deep contextualized word embeddings. Unlike previous models that processed text sequentially, BERT uses a bidirectional self-attention mechanism to capture context from both left and right simultaneously. It is pretrained using two self-supervised tasks: masked language modeling (MLM), where random tokens are masked and predicted, and next sentence prediction (NSP), which models inter-sentence relationships. This pretraining enables BERT to generate embeddings that dynamically reflect word meaning based on surrounding context, significantly improving performance across tasks such as question answering, named entity recognition, and sentiment analysis. However, BERT's computational cost and the requirement for pairwise input processing limit its efficiency for tasks involving large-scale sentence similarity comparisons.

B. RoBERTa

RoBERTa (Robustly optimized BERT approach) [7] builds upon BERT's architecture but focuses on improving training procedures to enhance embedding quality. It removes the NSP task, increases the size of training batches, and trains on larger and more diverse corpora for longer durations. These optimizations enable RoBERTa to learn richer contextual representations and outperform BERT on various benchmarks. RoBERTa maintains the bidirectional Transformer encoder structure but benefits from better generalization and more stable training. Despite improvements, RoBERTa inherits BERT's limitation of computationally expensive inference for tasks requiring pairwise input processing.

C. Sentence-BERT (SBERT)

To overcome the limitations of BERTin generating sentence embeddings for semantic similarity, Sentence-BERT (SBERT) [8] modifies the BERT architecture by adding a Siamese network structure that independently encodes sentences into fixed-size embeddings. SBERT is fine-tuned on sentence pairs using tasks such as natural language inference and semantic textual similarity datasets, enabling the model to produce semantically meaningful sentence vectors in a single forward pass. This drastically reduces inference time and allows efficient similarity computations using simple cosine similarity. SBERT has demonstrated state-of-the-art results in semantic search, clustering, and information retrieval, striking a balance between accuracy and efficiency.

D. MiniLM

MiniLM [9] is a lightweight Transformer-based model designed for faster inference and reduced model size, making it suitable for resource-constrained environments. It employs deep self-attention distillation techniques to transfer knowledge from large teacher models like BERT or RoBERTa into smaller student models while preserving performance. MiniLM maintains competitive accuracy on a range of NLP tasks despite having significantly fewer parameters and faster inference times. This efficiency gain is critical for real-time applications and deployment on devices with limited computational power without sacrificing the quality of contextual embeddings.

IV. TRAINING METHODOLOGIES

Transformer-based embedding models rely on multi-stage training processes to learn effective language representations that can generalize across various NLP tasks. These stages include pretraining on large unlabeled corpora, fine-tuning on specific downstream tasks, and, in some cases, model compression via knowledge distillation.

A. Pretraining Objectives

Pretraining forms the foundation of models like BERT and RoBERTa. Both use self-supervised learning to capture language patterns without labeled data. BERT's pretraining involves two main objectives: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) [10]. MLM randomly masks some tokens in the input sentence and trains the model to predict these masked tokens using the surrounding context, enabling the model to learn bidirectional contextual embeddings. NSP teaches the model to predict if one sentence logically follows another, improving understanding of sentence relationships. RoBERTa [11] enhances this by removing NSP, increasing batch sizes, training on larger datasets, and extending training duration. These improvements lead to better contextual embedding quality and more robust language understanding.

B. Fine-tuning Strategies:

Fine-tuning adapts the pretrained models to specific NLP tasks by training on labeled datasets with supervised objectives. This phase modifies the pretrained weights to optimize performance for tasks like sentiment classification, named entity recognition, or semantic textual similarity. Sentence-BERT (SBERT) [12] fine-tunes BERT using a Siamese or triplet network structure on datasets such as natural language inference and semantic similarity benchmarks. This training enables SBERT to produce fixed-size sentence embeddings that capture semantic meaning efficiently, allowing for fast and accurate sentence-level similarity calculations. Fine-tuning typically involves optimizing loss functions such as cross-entropy or triplet loss depending on the task.

C. Knowledge Distillation (for MiniLM):

MiniLM [13] addresses the challenge of deploying large Transformer models in resource-constrained environments by using knowledge distillation—a model compression technique. In this process, a smaller "student" model is trained to mimic the behavior of a larger "teacher" model by learning from its outputs, intermediate representations, and self-attention patterns. MiniLM introduces deep self-attention distillation, where the student model matches the teacher's attention distributions and hidden states at multiple Transformer layers. This approach enables MiniLM to retain much of the teacher's performance while drastically reducing model size and inference latency, making it suitable for real-time applications and devices with limited computational resources.

V. ARCHITECTURE AND COMPUTATIONAL EFFICIENCY

Transformer-based embedding models differ significantly in their architecture, directly influencing their computational efficiency and usability.

BERT uses a deep bidirectional Transformer encoder with 12 layers (base) or 24 layers (large), each layer containing multiple self-attention heads [14]. This design captures rich contextual information but results in a large model size (~110 million parameters for BERT-base) and high computational costs, limiting its deployment in latency-sensitive or resource-limited environments.

RoBERTa maintains BERT's architecture but improves training by removing the next sentence prediction task, increasing batch sizes, and training on larger datasets [15]. These optimizations enhance performance but do not reduce the model size or computational burden, keeping RoBERTa's resource demands like BERT.

Sentence-BERT (SBERT) modifies BERT into a Siamese network to generate sentence embeddings efficiently [16]. By encoding sentences independently, SBERT reduces the computational load when comparing sentence pairs, making tasks like semantic search and clustering more scalable. Although SBERT shares BERT's parameter count, its inference strategy greatly improves runtime efficiency.

MiniLM applies knowledge distillation to create a compact Transformer model with fewer layers and parameters (~22 million) [17]. Using deep self-attention distillation, MiniLM preserves much of the teacher model's representational power while significantly reducing inference time and memory usage. This balance makes MiniLM ideal for real-time and resource-constrained applications.So BERT and RoBERTa provide powerful but computationally intensive models, SBERT optimizes sentence-level embedding efficiency, and MiniLM offers a lightweight alternative with competitive performance.

Despite their powerful capabilities, Transformer-based embedding models face several inherent challenges that affect their practical deployment. One significant limitation is their large model size and high computational demand.

Models like BERT and RoBERTa contain hundreds of millions of parameters, resulting in substantial memory consumption and slow inference times, which makes them less suitable for real-time applications or deployment on devices with limited resources. Although distilled models such as MiniLM address this issue by reducing model size and improving speed, there remains an inherent trade-off between model compactness and representational accuracy [18,19]. Consequently, achieving efficient yet effective embeddings for diverse use cases continues to be an open challenge.

Anotherlimitation is the nuanced understanding of context and language subtleties. While these models excel at capturing rich semantic and syntactic information, they occasionally struggle with complex linguistic phenomena such as sarcasm, idiomatic expressions, and rare word senses. Additionally, the fixed input length constraint imposed by Transformer architectures can truncate long texts, potentially omitting essential context required for accurate representation. Furthermore, embeddings pretrained on general-domain corpora may not transfer effectively to specialized fields without substantial fine-tuning. Domain adaptation often requires access to annotated domain-specific datasets and considerable computational resources, posing a barrier for many practical applications [20,21].

Addressing these challenges requires ongoing research into more efficient model architectures, improved contextual understanding mechanisms, and better domain adaptation strategies. Innovations in model compression, dynamic context handling, and transfer learning hold promise for overcoming these limitations and enabling wider adoption of Transformer-based embeddings in real-world NLP tasks.

VI. APPLICATIONS AND FUTURE DIRECTIONS

Transformer-based embedding models have been widely adopted across numerous NLP applications due to their ability to capture rich semantic representations. Semantic similarity and clustering tasks benefit significantly from models like SBERT, which generate meaningful sentence embeddings that facilitate efficient comparison and grouping of textual data [22]. In question answering systems, embeddings derived from BERT and RoBERTa enable models to understand the context and intent of questions, improving answer retrieval accuracy and relevance [23, 24]. Information retrieval applications leverage these embeddings to enhance search precision by ranking documents based on semantic relevance rather than simple keyword matching [25]. Beyond these, Transformer embeddings have proven effective in a variety of other NLP tasks such as named entity recognition, sentiment analysis, machine translation, and text summarization, consistently advancing state-of-the-art performance across benchmarks [24].

Future research in Transformer-based embedding models is expected to focus heavily on improving model compression and efficiency. Techniques such as knowledge distillation, pruning, quantization, and architecture search are promising approaches to reduce model size and inference latency without significantly compromising performance [26]. These advancements are crucial for enabling deployment in resource-constrained environments like mobile devices and real-time systems. Additionally, the development of more effective multilingual and cross-domain embeddings remains an important area, aiming to create universal models that can handle diverse languages and specialized domains with minimal fine-tuning [27]. This would facilitate broader accessibility and applicability of NLP technologies worldwide.

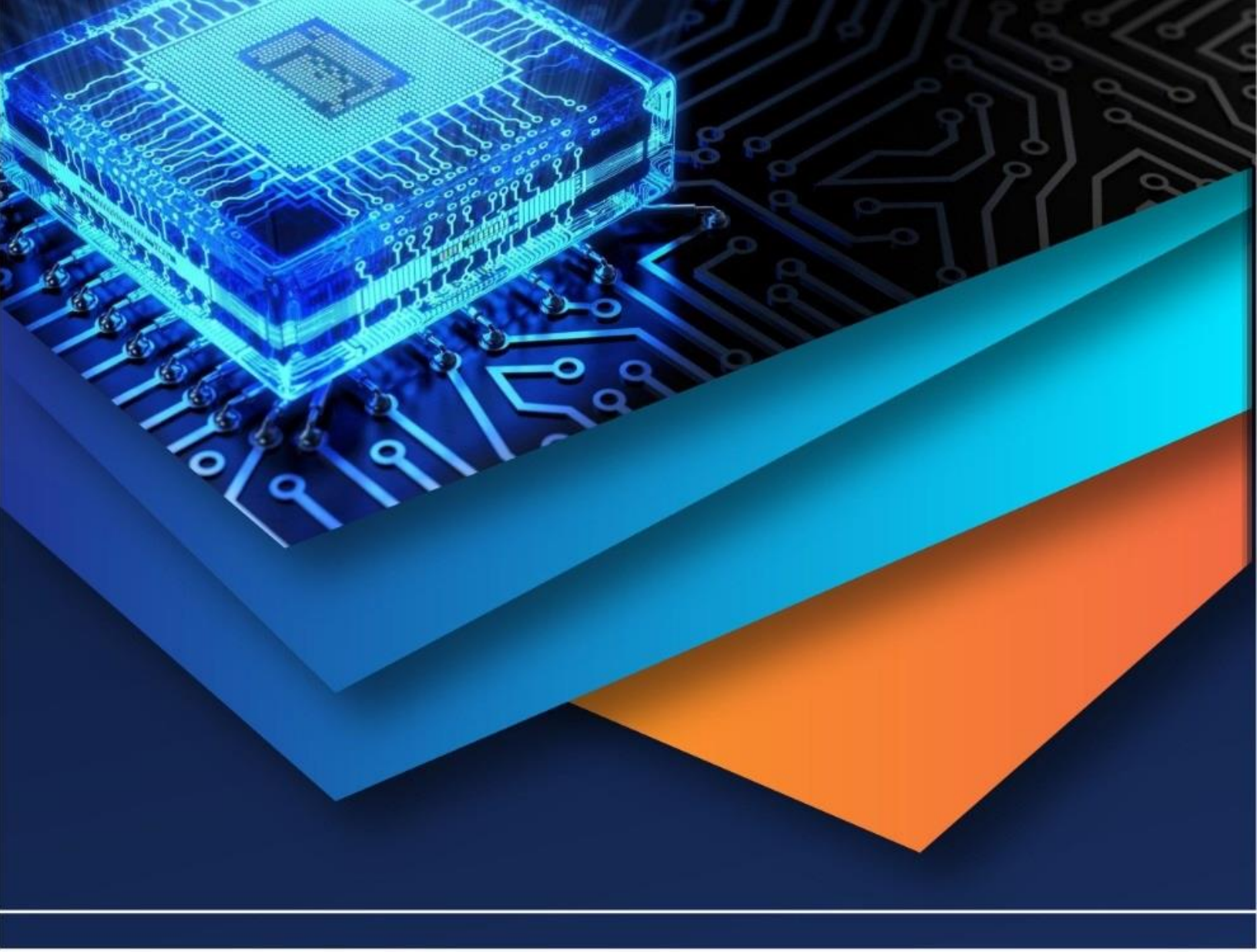
Another exciting frontier involves the integration of Transformer-based embeddings with other modalities, such as vision and audio, to build richer, multi-modal representations. Combining textual embeddings with visual features, for instance, can improve tasks like image captioning, video understanding, and multimodal search, leading to more holistic AI systems that better understand human communication [28]. Research on joint training and alignment of embeddings across modalities is still evolving but holds significant promise for future applications.

VII. CONCLUSION

Transformer-based embedding models have fundamentally transformed the field of natural language processing by enabling rich, context-aware representations that significantly improve performance across a wide range of tasks. Their ability to capture subtle semantic and syntactic nuances has made them indispensable for applications such as semantic similarity, question answering, and information retrieval. However, challenges related to large model sizes, high computational costs, and difficulties in handling domain-specific language and nuanced contexts still limit their widespread deployment, especially in resource-constrained settings.Advancements in model compression, including pruning, quantization, and knowledge distillation, are enabling smaller, faster Transformer models without compromising accuracy. Efforts to develop multilingual and cross-domain embeddings aim to create versatile models adaptable to diverse languages and fields. Integrating embeddings with other modalities like vision and audio promises richer, more comprehensive AI systems. These directions will shape the future of efficient and powerful natural language understanding.

REFERENCES

[1] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.
[2] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL-HLT.
[3] Liu, Y., Ott, M., Goyal, N., et al. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692.
[4] Reimers, N., &Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. EMNLP.
[5] Wang, W., Liu, B., Cho, K., & Gong, Y. (2020). MiniLM: Deep self-attention distillation for task-agnostic compression of pretrained transformers.
[6] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. NeurIPS.
[7] Hutto, C., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. ICWSM.
[8] Lee, J., Yoon, W., Kim, S., et al. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics.
[9] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. Proceedings of ACL, 8440–8451.
[10] Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. Advances in Neural Information Processing Systems, 32.
[11] Nogueira, R., & Cho, K. (2019). Passage Re-ranking with BERT. arXiv preprint arXiv:1901.04085.
[12] Radford, A., Narasimhan, K., Salimans, T., &Sutskever, I. (2018). Improving language understanding by generative pre-training. OpenAI preprint.
[13] Raffel, C., Shazeer, N., Roberts, A., et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. JMLR.
[14] Lewis, M., Liu, Y., Goyal, N., et al. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. ACL.
[15] He, P., Liu, X., Gao, J., & Chen, W. (2021). DeBERTa: Decoding-enhanced BERT with disentangled attention. ICLR.
[16] Zhang, Y., Sun, S., Galley, M., et al. (2020). Dialogpt: Large-scale generative pre-training for conversational response generation. ACL.
[17] Brown, T., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. NeurIPS.
[18] Yang, Z., Dai, Z., Yang, Y., et al. (2019). XLNet: Generalized autoregressive pretraining for language understanding. NeurIPS.
[19] Peters, M. E., Neumann, M., Iyyer, M., et al. (2018). Deep contextualized word representations. NAACL-HLT.
[20] Kiros, R., Zhu, Y., Salakhutdinov, R., et al. (2015). Skip-thought vectors. NeurIPS.
[21] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
[22] Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. EMNLP.
[23] Joulin, A., Grave, E., Bojanowski, P., &Mikolov, T. (2016). Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759.
[24] Bojanowski, P., Grave, E., Joulin, A., &Mikolov, T. (2017). Enriching word vectors with subword information. TACL.
[25] Logeswaran, L., & Lee, H. (2018). An efficient framework for learning sentence representations. ICLR.
[26] Cer, D., Yang, Y., Kong, S.-Y., et al. (2018). Universal Sentence Encoder. arXiv preprint arXiv:1803.11175.
[27] Schuster, T., Ram, O., Barzilay, R., & Jaakkola, T. (2019). Cross-lingual alignment of contextual word embeddings. EMNLP.
[28] Nie, Y., Chen, H., & Bansal, M. (2020). Combining fact extraction and verification with neural semantic matching networks. AAAI.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY