Artificial Intelligence Review (2025) 58:166
https://doi.org/10.1007/s10462-025-11162-5

BERT applications in natural language processing: a review

Nadia Mushtaq Gardazi[1] · Ali Daud[2] · Muhammad Kamran Malik[1] · Amal Bukhari[3] · Tariq Alsahfi[3] · Bader Alshemaimri[4]

Accepted: 19 February 2025 / Published online: 15 March 2025

Abstract

BERT (Bidirectional Encoder Representations from Transformers) has revolutionized Natural Language Processing (NLP) by significantly enhancing the capabilities of language models. This review study examines the complex nature of BERT, including its structure, utilization in different NLP tasks, and the further development of its design via modifications. The study thoroughly analyses the methodological aspects, conducting a comprehensive analysis of the planning process, the implemented procedures, and the criteria used to decide which data to include or exclude in the evaluation framework. In addition, the study thoroughly examines the influence of BERT on several NLP tasks, such as Sentence Boundary Detection, Tokenization, Grammatical Error Detection and Correction, Dependency Parsing, Named Entity Recognition, Part of Speech Tagging, Question Answering Systems, Machine Translation, Sentiment analysis, fake review detection and Cross-lingual transfer learning. The review study adds to the current literature by integrating ideas from multiple sources, explicitly emphasizing the problems and prospects in BERT-based models. The objective is to comprehensively comprehend BERT and its implementations, targeting both experienced researchers and novices in the domain of NLP. Consequently, the present study is expected to inspire more research endeavors, promote innovative adaptations of BERT, and deepen comprehension of its extensive capabilities in various NLP applications. The results presented in this research are anticipated to influence the advancement of future language models and add to the ongoing discourse on enhancing technology for understanding natural language.

Keywords Bidirectional encoder representation for Transformers (BERT) · Natural Language processing (NLP) · Large Language models (LLM) · Deep learning (DL) · BERT applications

Extended author information available on the last page of the article

1 Introduction

Natural Language Processing (NLP) has seen substantial growth, with applications like conversational bots, language translation, voice assistants, and real-time speech translation (Khan et al. 2023). This rapid development is reflected in a broad array of products now available on the market. Recent advancements in NLP have notably improved its ability to achieve near-human language comprehension. Over the years, various approaches have been designed to tackle NLP challenges, including rule-based methods, machine learning techniques, hybrid models, and deep learning (DL), as depicted in Fig. 1 (Pennington et al. 2014).

Deep neural networks (DNN) have significantly advanced various aspects of NLP. In particular, deep learning models like Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM), and Gated Recurrent Units (GRUs) have introduced innovative approaches across NLP domains, as shown in Fig. 2 (Hochreiter and Schmidhuber 1997; Chung et al. 2014). However, RNN and LSTM models often miss essential contextual information in sequences and face challenges such as lengthy training times and the need for extensive datasets.

To address limitations in capturing contextual information, researchers adopted unsupervised neural networks to create distinct vector representations for words through embeddings. Techniques like Word2Vec, GloVe, and FastText (Mikolov et al. 2013; Pennington et al. 2014; Bojanowski et al. 2017) initially generated word vectors but lacked the ability to account for contextual meaning within phrases or documents, representing words independently of surrounding context. To overcome this, NLP frameworks evolved to include contextual embeddings produced by large-scale, pre-trained language models.

The Transformer model (Vaswani et al. 2017) became foundational in NLP, surpassing traditional models like CNNs and RNNs in tasks requiring both comprehension and language generation. This architecture excels at pretraining on large text corpora, significantly improving accuracy in tasks such as text classification, language understanding, coreference resolution, common-sense inference, and machine translation (Yang et al. 2019; Liu et al. 2019a, b; Wang et al. 2019; Lample and Conneau 2019). The Transformer has propelled NLP forward, enabling more robust and nuanced language models capable of deeper lan-

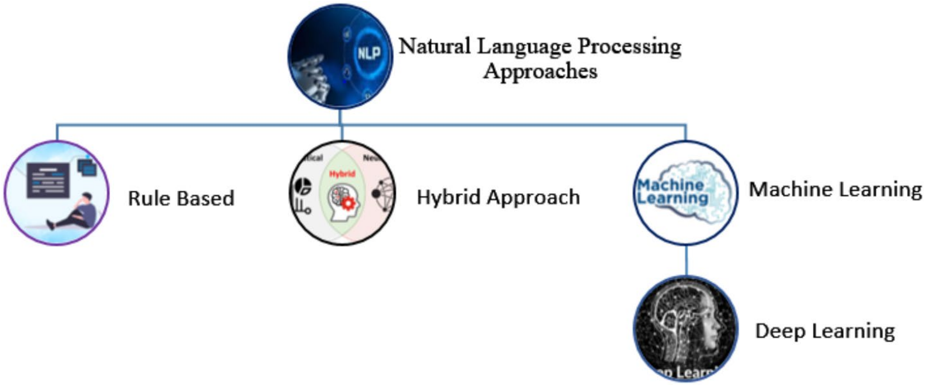


Fig. 1 Different approaches for NLP tasks

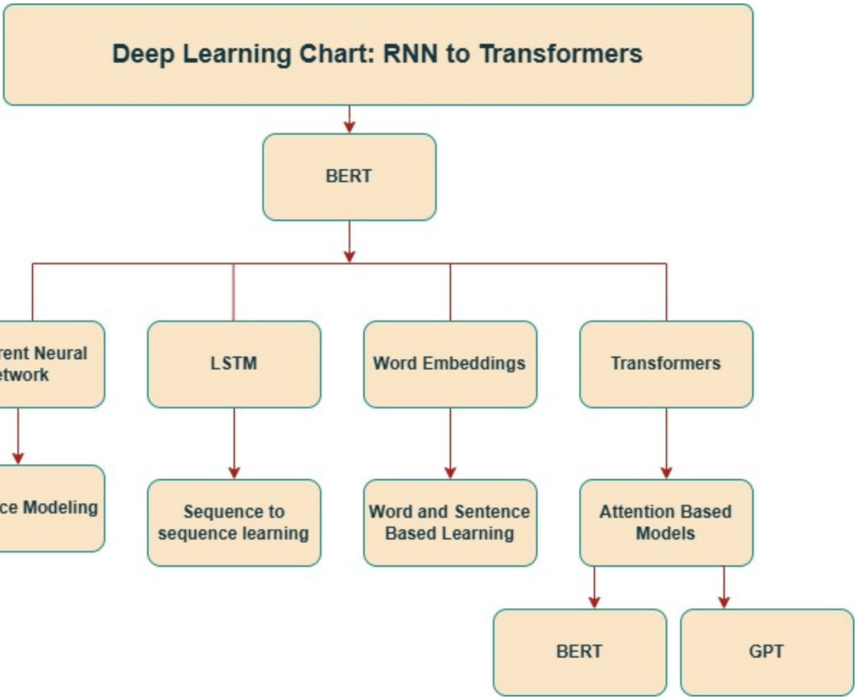


Fig. 2 Deep learning models

guage understanding and analysis. Figure 3 highlights the progression of key models in NLP, including Word2Vec, GloVe, ELMo, GPT, and BERT.

Over time, contextual comprehension in NLP has improved significantly. Early models like Word2Vec and GloVe offered moderate accuracy in capturing word meaning, but more advanced models such as ELMo, GPT, and BERT have made considerable strides in addressing pretraining challenges for contextual understanding. The shift from BERT to GPT-4 marks a pivotal moment in NLP history. Figure 4 illustrates this evolution. BERT's two-way approach revolutionized context understanding, while the creativity and adaptability of GPT models, especially GPT-1, GPT-3, and GPT-4, further propelled the field, making AI increasingly practical and valuable in real-world applications.

The field of pre-trained models in NLP has undergone substantial evolution as researchers have explored various models and their applications (Wang et al. 2023). Earlier studies played a vital role in shaping our understanding of these models' strengths and limitations. Qiu et al. (2020) provide an extensive review of pre-trained models in NLP, covering language models, classification models, and sequence-to-sequence models for tasks like language modeling and text classification (Daud et al. 2017). However, their focus is primarily on English, limiting applicability to other languages and lacking a deeper evaluation of different models.

Koroteev (2021) examined BERT's progress in text analysis, including techniques like annotation and classification. The study highlighted the need to adapt BERT for various NLP tasks, especially in low-resource languages. Aftan and Shah (2023) also analyzed BERT's applications in NLP but fell short of thoroughly examining its effectiveness com-

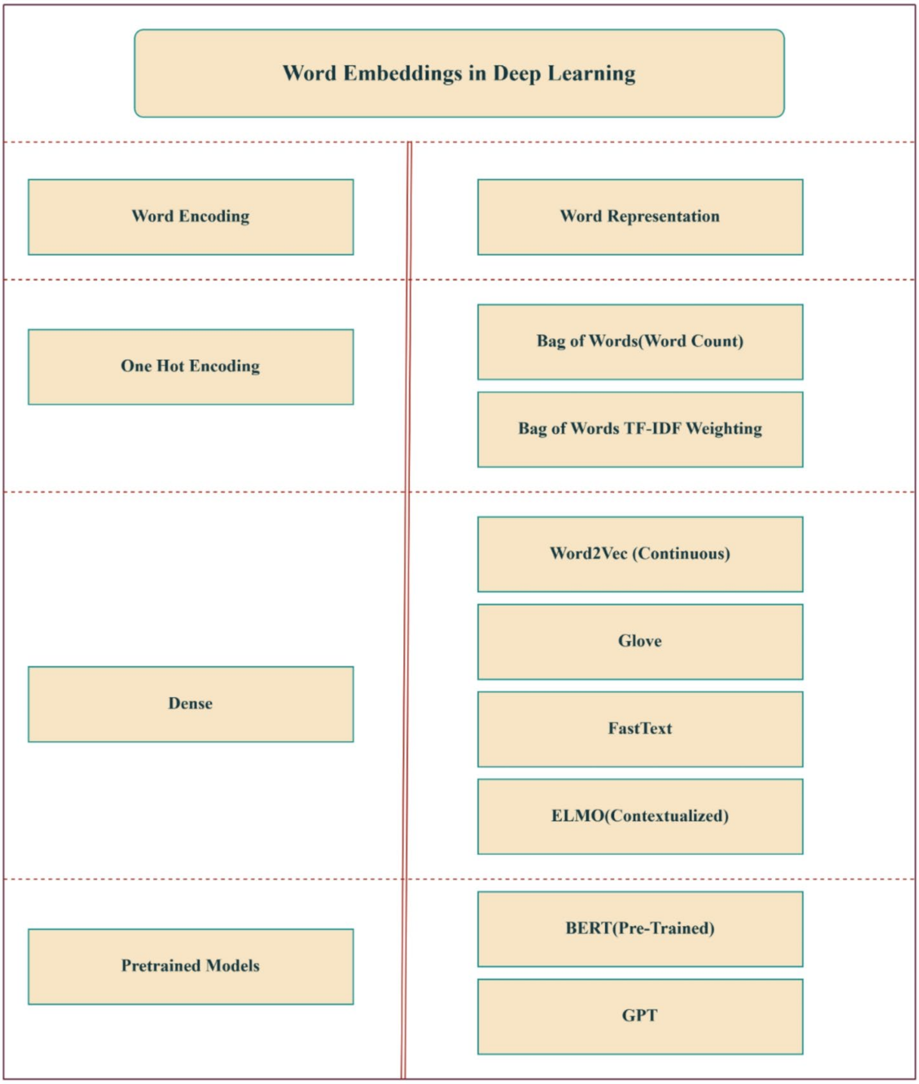


Fig. 3 Word embeddings in deep learning approach

pared to alternative models and did not reference prior studies that extensively explored BERT's potential across tasks.

The previous investigations have offered an overview of BERT and its applications within a specific scope. However, they have yet to compare other language models or techniques conductively. Hence, it is imperative to examine BERT's applications on NLP tasks thoroughly. This review is supposed to include a diverse array of NLP tasks, including named entity recognition (NER), part-of-speech (POS) tagging, question-answering systems (QAS), machine translation (MT), sentiment analysis, and fake review detection. The

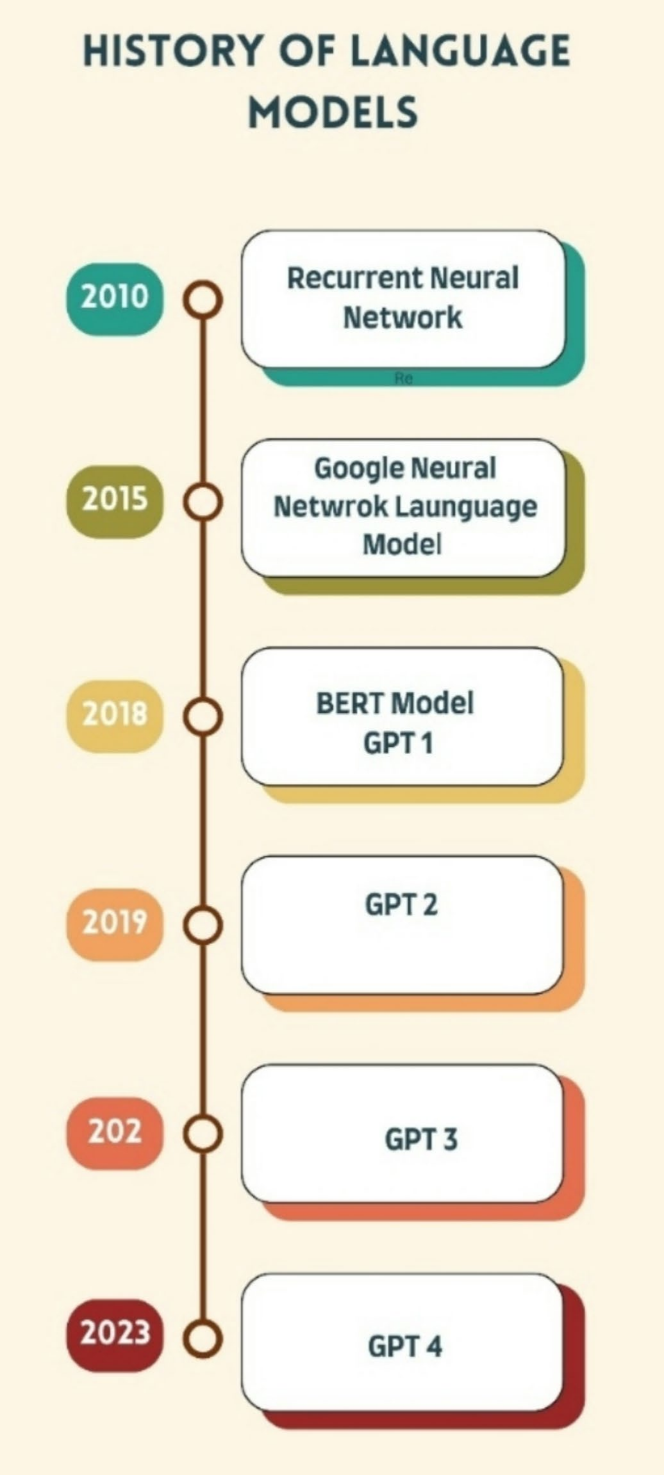


Fig. 4 History of language model

following section highlights the significant contributions made by this paper.Overview of BERT Architecture and its effectiveness for NLP tasks.

- Modification in Bert Architecture.
- Review of previous researches on Application of BERT in various NLP tasks using different languages and its comparison with previous models.
- Insight about Different data sets used in performing NLP tasks using BERT Model.
- Challenges and limitations of BERT performance in NLP tasks.
- Future directions and recommendations.

Overall, this survey paper is providing a comprehensive overview of BERT-based NLP, highlight its strengths and weaknesses, and identify the future research directions in the field. The contributions of the author will be a thorough literature review, detailed explanation of BERT architecture, and critical analysis of its applications and limitations on several NLP tasks.

The paper is precisely organized and presented in the Fig. 5.

2 Application of BERT architecture in NLP tasks

This section provides a thorough explanation of the BERT architecture and a brief overview of its utilization in diverse NLP applications. BERT has demonstrated extraordinary effectiveness as a language model, achieving significant results.

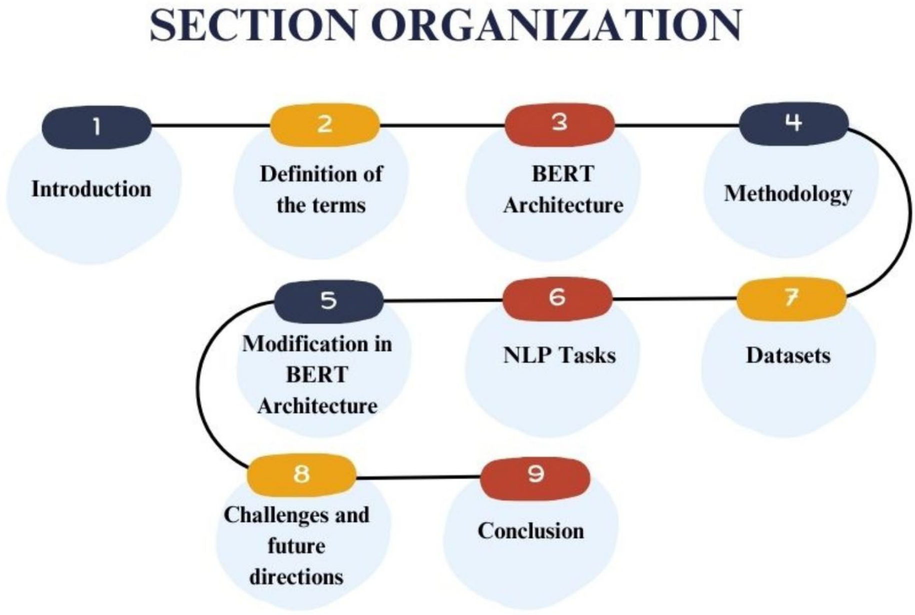


Fig. 5 Section Organization of review paper

2.1 Bert as an embedding

Conventional word embeddings like Word2Vec and GloVe provide static, context-independent word representations. In contrast, BERT generates contextual embeddings that reflect word meanings based on their usage in specific contexts, significantly advancing NLP by storing rich semantic information (Tanaka et al. 2019; Turton et al. 2020). These contextual embeddings outperform static ones in tasks like document classification (Tanaka et al. 2019), idiom recognition (Nedumpozhimana et al. 2022), and question answering (Ma et al., 2019). BERT's bidirectional training enables it to capture context from both directions, improving accuracy compared to earlier unidirectional models (Kaliyar 2020). Contextual embeddings enhance performance in various applications, including disaster tweet prediction (Duraisamy et al. 2023), sequential recommendation (Harte et al. 2023), and conceptual metaphor detection (Li et al. 2023). Fine-tuning BERT for specific tasks has delivered state-of-the-art results, such as in cybersecurity entity recognition (Srivastava et al. 2023) and improving performance in low-resource languages like Amharic (Yeshambel et al. 2023). Overall, BERT and similar contextual models show significant improvements over static embeddings across NLP applications.

2.2 BERT as a transformer

BERT, short for Bidirectional Encoder Representations from Transformers, is a language representation model based on transformer architecture that excels in various NLP tasks (Devlin et al. 2019). Its bidirectional training and masked language modeling enable it to understand context from both directions (Gupta 2024a, b), leading to superior performance in tasks like question-answering and linguistic inference. BERT's success has driven its widespread adoption in both research and industry (Gupta 2024a, b). The model utilizes multiple transformer encoder layers with self-attention mechanisms to process input sequences (Ghojogh and Ghodsi 2020). To handle longer texts, researchers have expanded BERT's capabilities by segmenting inputs and adding layers (Pappagari et al. 2019). Analysis of BERT's hidden states shows it aligns with traditional NLP pipeline stages (van Aken et al. 2019). BERT's proficiency in capturing contextual relationships allows it to set new benchmarks in NLP applications (Kora and Mohammed 2023), from phishing email detection (Otieno et al. 2023) to speech recognition (Djeffal et al. 2023) and Ethereum fraud detection (Hu et al. 2023). Domain-specific variants like Bioformer have been developed for specialized tasks like biomedical text mining (Fang et al. 2023). However, BERT's large size incurs high computational costs, prompting research into more compact models (Fang et al. 2023). Hybrid approaches combining BERT with models like BiGRU further enhance task performance, such as in extractive text summarization (Bano et al. 2023). BERT's influence continues to grow, with numerous studies expanding its application range (Gupta 2024a, b; Aftan and Shah 2023).

2.3 Versions of BERT

The initial versions of BERT are BERTtiny BERT$_{\text{base}}$ and BERT$_{\text{large}}$, with the latter being more complex and computationally expensive. Table 1 representing difference between all variants.

Table 1 Comparison of BERT$_{base}$ and BERT$_{large}$

	Transformer block	Hidden size	Attention head	Parameters
BERT$_{tiny}$	2–4	128–256	4–8	10 M–30 M
BERT$_{base}$	12	786	12	110 M
BERT$_{large}$	24	1024	16	340 M

Table 2 English data sets used to train BERT model

Dataset	Size	References
BOOKCORPUS	16gb	Zhu et al. (2015)
STORIES	31GB	Trinh and Le (2019)
OPENWEBTEXT	38GB	Gokaslan et al. (2019)
CC-NEWS	76GB	Nagel (2016)

The BERT model has made a substantial impact in the advancement of an extensive range of conventional and advanced NLP tasks. Table 2 presents BERT applications on large English corpora, from simple tasks such as tokenization, stop word removal, and stemming to more complex tasks such as NER, POS tagging, MT, and QAS.

3 Methodology

3.1 Steps of writing review paper

The study aimed to comprehensively cover tasks performed by the BERT architecture in NLP. The research methodology involved conducting a literature review, refining research criteria, and synthesizing findings to draw conclusions about the effectiveness of BERT in NLP and identify gaps in the existing literature (see Fig. 9). The study's findings provide a valuable resource for this field's practitioners and scholars.

3.1.1 Step 1: planning of the study

Planning of the study, is further divided into planning and review protocol, which were followed throughout the research. During the planning phase, we decided which tasks would be included in the paper, how various sections would be designed, how different studies would contribute to future research, and how to improve further. In the protocol designing phase, we determined how to implement the research, the procedures to follow, and the protocols to ensure successful completion according to the plan.

The research protocol included the main research questions, search strategies, and inclusion and exclusion criteria. We continuously reviewed and revised the protocol throughout the study to ensure its relevance and effectiveness. By following a rigorous research protocol, we maintained consistency and ensured that the study's findings were reliable and relevant to our research questions.

3.1.2 Step 2: implementation of protocol

The next step was the conducting step, and we tried our best to work according to what was decided in the research planning and protocol. Different search strategies, mainly web research, digital libraries, and books, were applied in the conduct of this research. Then, selection criteria were strictly followed, which are listed below in Fig. 6.

3.1.2.1 Inclusion criteria Papers for the applications of NLP utilizing BERT that are being studied in this review study were gathered by filtering research publications with relevant sub-domain names in their titles using the IEEE and ACM Guide to Computing Literature Collections. We included basic to advanced tasks performed by BERT in NLP. When looking for relevant literature, exact word searching was done inside the titles. Basic keywords and Phrases which were commonly used are BERT architecture,

3.1.2.2 Exclusion criteria After reading the abstract, methodology, results, and conclusion, we decided whether to include this paper. We excluded papers that did not satisfy our selec-

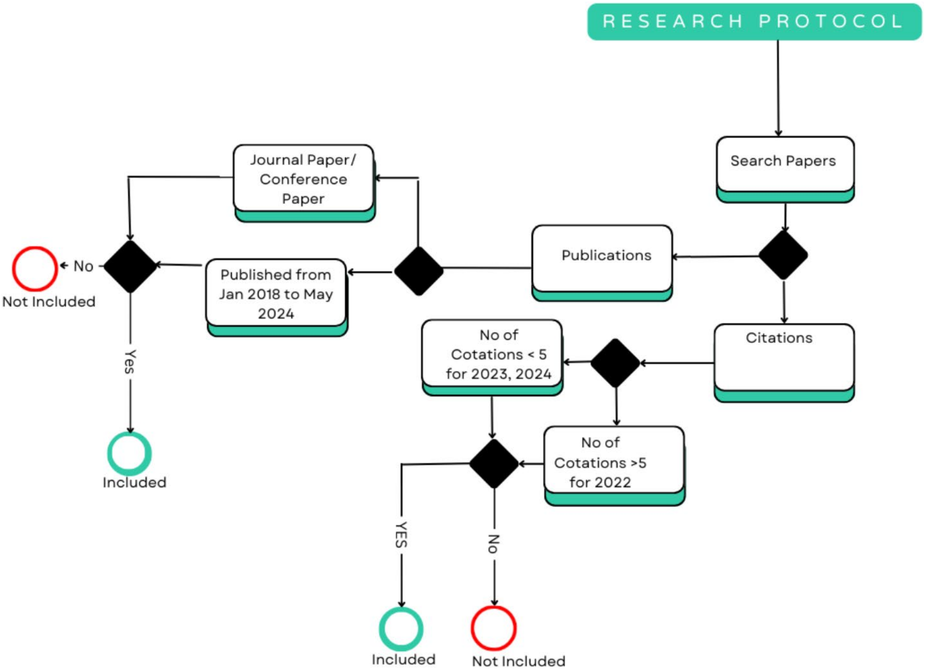

RESEARCH PROTOCOL
Search Papers
Publications
Citations
Journal Paper/ Conference Paper
Published from Jan 2018 to May 2024
No
Not Included
Yes
Included
No of Cotations < 5 for 2023, 2024
No of Cotations >5 for 2022
YES
No
Included
Not Included

Fig. 6 Selection criteria

tion criteria. A total of 150 papers were included in this review article. Papers were considered according to their year of publication and number of citations.

3.1.3 Step 3: results presentation

The last step was the presentation of results. Effective ways were adopted to present results using figures, tables, and other possibilities. In conclusion, the methodology section of this review paper has outlined the three main steps taken to explore the application of BERT in various NLP fields. The first step involved careful planning of the study and research protocol. The second step involved conducting the planning and research protocols, including determining inclusive and exclusive criteria for selecting relevant studies. This step was crucial to ensure the review was comprehensive and focused on relevant research. In the third stage, findings were analyzed and synthesized from appropriate studies to offer a thorough picture of BERT's implementation in NLP. Using a systematic approach, this study gives an accurate and thorough overview of BERT in NLP disciplines. Table 3 depicts tasks and related studies included in this study,

4 Modifications in BERT architecture

4.1 Modification in model

Model size during natural language representation pre training often improves performance on later challenges. However, GPU/TPU memory limits and longer training cycles make model improvement difficult. The BERT model is modified to minimize memory usage and improve training performance. This section details three primary adjustments (see Table 4) and a few additional information.

Table 3 Distribution of papers among all tasks

Tasks	No of studies included in paper
Sentence boundary detection	5
Word segmentations	8
Grammatical error detection and correction	8
Dependency parsing	5
Part of speech tagging	16
NER	19
QA system	18
MT	28
Sentiment Analysis	29
Fake review detection	9
Cross-lingual transfer learning	9

Table 4 Summary of BERT architecture and modifications

Bert architectures	Modifications	References
RoBERTa	Changes in masking strategies	Liu et al. (2019a, b)
ALBERT	Less parameters were used as compared to BERT	Lan et al. (2020)
DistilBERT	Knowledge distillation technique	Sanh et al. (2020)

Table 5 Comparison between BERT and ALBERT

Model	BERT		ALBERT	
	$\text{BERT}_{\text{base}}$	$\text{BERT}_{\text{large}}$	$\text{ALBERT}_{\text{base}}$	$\text{ALBERT}_{\text{Large}}$
Layers	12	24	12	24
Hidden	786	1024	786	1024
Embedding's	786	1024	128	238
Parameters	108 M	334 M	12 M	18 M

4.1.1 ALBERT

A Lite BERT (ALBERT) contains several fewer parameters than BERT (Lan et al. 2020). ALBERT-large has only 18 M parameters, compared to 334 M for BERT-large. Refer to Table 5 for details.

ALBERT reduced model size and improved efficiency by employing a parameter-sharing strategy that reduces the number of parameters compared to BERT.

4.1.2 DistilBERT

DistilBERT, proposed by Sanh et al. (2020), is a compressed version of BERT that employs knowledge distillation for deployment in resource-constrained environments. This reduces training time and computational costs, allowing transformers to be used on minimal computing resources. Knowledge distillation during BERT's pre-training stage resulted in a 40% reduction in model size while maintaining 97% of its understanding skills.

4.1.3 RoBERTa

RoBERTa, or Robustly Optimized BERT Pretraining Approach, is similar to BERT but modifies the masking approach and removes the NSP task (Liu et al. 2019a, b). Enhancements include larger batch sizes, extended training times with more comprehensive datasets, training on longer sequences, and dynamically modifying the masking pattern for training data.

Table 6 demonstrates DistilBERT's lightweight nature, making it suitable for deployment on devices with constrained computational resources. However, for complex tasks requiring comprehensive representation, DistilBERT does not perform as well as the larger BERT model. While RoBERTa is nearly identical to BERT in architecture and model size, it differs significantly in training methodology and datasets. RoBERTa's pre-training modifications enhance its performance on several NLP tasks, often resulting in greater precision and robustness compared to BERT.

Table 6 Comparison between BERT and distilbert

Model		Layers	Hidden	Embeddings	Parameters
BERT	BERT$_{base}$	12	786	786	110 M
	BERT$_{large}$	24	1024	1024	340 M
DistilBERT	DistilBERT$_{base}$	6	786	786	66 M
	DistilBERT$_{Large}$	6	1024	1024	82 M
RoBERTa	RoBERTa$_{base}$	12	786	786	125 M
	RoBERTa$_{Large}$	24	1024	786	355 M

4.1.4 Studies on different variant

Different variants od BERT along with specification and summary are presented in Table 7.

In order to optimize NLP applications, scientists and developers may use these variants, taking into account things like model size, computational effectiveness, and task-specific performance.

5 NLP applications of BERT

Pre-trained language models have transformed NLP applications and systems. An essential paradigm is to train a language model on massive corpora so that it may serve as the foundation upon which an NLP application can be developed and improved. Through applying the BERT framework, we shall examine the improvement of NLP tasks shown in the Fig. 7,

5.1 Sentence boundary detection

BERT significantly improves sentence boundary detection (SBD) by capturing contextual word relationships, making it useful for various NLP tasks in finance and spoken language translation. Models like BERT, RoBERTa, and BiLSTM-CRF have demonstrated effective SBD performance across different datasets, with notable results from studies by Du et al. (2019), Donabauer et al. (2021), and Hayashibe and Mitsuzawa (2020). While BERT-based methods perform well, future research should explore simpler model comparisons and assess limitations, particularly in small datasets. Table 8 presents evaluation scores for different models in SBD tasks across several datasets.

5.2 Tokenization and word segmentation

Tokenization and word segmentation are vital for language understanding, with BERT significantly improving performance in various tasks. BERT has enhanced Chinese word segmentation using Word Piece tokenization (Yang et al. 2019; Wu et al. 2016) and shown superiority in Persian word segmentation (Doostmohammadi et al. 2020). However, challenges remain with multilingual sentences and out-of-vocabulary terms. Hiraoka et al.

Table 7 BERT different variations summary

Model	Specification	Summary	References
BIoBert	Biomedical Text	(Biomedical BERT) Dataset was prepared using biomedical data from PubMed and PMC rather than Wikipedia and books data. NER, relation extraction and Q&A system were developed.	Alsentzer et al. (2019) Lee et al. (2020)
Clinical BERT	Clinical data	Clinical BERT was trained on data contains texts different medical related abbreviations, jargons, professional notes.	Huang et al. (2020)
AraBERT	Arabic text	Arabic Biomedical BERT used a huge Arabic corpus and performed sequence classification, NER and Sentiment Analysis task.	Antoun et al. (2020)
SCIBERT	Science data	It was utilized for performing tasks of text classification and NER on large carpus of scientific research papers.	Beltagy et al. (2019)
DeBERTa	Half training data of RoBERTa	Decoding enhanced Biomedical BERT was an innovative approach designed after combining BERT and RoBERTa. It was used to perform NLP tasks with better results.	He et al. (2021)
AlphaBERT	Character Level Tokens	Without effecting performance character level tokens in English Corpus reduced size. Its architecture was similar to BERT.	Chen et al. (2020).
BERTje	Dutch language text	This was designed on Dutch language and NLP tasks were performed. Pretraining and fine-tuning was performed as in original BERT model.	de Vries et al. (2019)
Bio ALBERT	Lite BERT with Biomedical text	Performed NER task on biomedical data using lite BERT, through this model less memory was consumed with efficient results.	Naseem et al. (2021)
Mobile BERT	Compress and accelerate original BERT.	This model is smaller and faster than BERT$_{base}$. This model can be applied in mobile devices.	Sun et al. (2021)
FlauBERT	Model was trained on large French data	Large French data was used to performed NLP tasks of text classification. It has different versions	Le et al. (2020)
Squeeze BERT	Replaced self-attention with grouped convolution	Grouped convolution is applied (an efficient computer vision technique) on NLP tasks. It's more efficient and robust model than BERT$_{base}$.	Iandola et al. (2020)
Camem BERT	Use transformer model for other languages	BERT is trained mainly on English language, camem BERT train transformer for other languages as well. It helped to perform downstream tasks of NLP	Martin et al. (2020)
Spanbert	Developed by Facebook AI Research (FAIR)	Used in Span based tasks such as questioning answering. It has capacity to train whole span rather than a simple token.	Joshi et al. (2020)
Electra (Efficiently Learning an Encoder that Classifies Token Replacements Accurately):	Developed by Google research	Introduced new pre training objective, "replaced token detection". Works better using minimal computational resources.	Clark et al. (2020)

Table 7 (continued)

Model	Specification	Summary	References
AutoBERT-Zero	Introduced automatic architecture search	Bert architecture was simplified for reduction in computational complexity and memory requirements	Gao (2021)
BERTino	Work on architecture of DistilBERT	It has light weighted, improved performance architecture for Italian language. Better accuracy than BERT$_{BASE}$. Suitable in resource constraint environment.	Muffo and Bertino (2023)
FlexiBERT	Improved flexibility in architecture and resizing can be done dynamically	FlexiBERT is adoptable in different resource constraints	Tuli et al. (2023)
MA-BERT	Matrix Arithmetic- only (Reduced complexity of Original BERT)	As compared to baseline Bert models. It achieved better accuracy and reduced inference time on CPU/GPU.	Ming et al. (2023)
Attention BERT	Improved BERT attention mechanism	It improved code specific representations and syntactic structures.	Sharma et al. (2022)
DrBERT	Refinements in model structure.	Model structure was revised to enhance performance of model.	Liang and Liang (2024)
FinBERT	Model is pretrained for financial text mining	Six Pretraining tasks were performed to demonstrate effectiveness of FinBERT.	Liu et al. (2021)
PatentBERT	Classify Patent document with fine tunning of original BERT architecture	State of the art performance of Patent Documents.	Lee and Hsiang (2019)
ITALIAN-LEGAL-BERT	Model is pre-trained on Italian Civil Law	Domain specific tasks were improved.	Licari and Comandè (2022)
HybridBERT	To encode contextual features, self-attention and poling networks are combined.	The mixed form, which includes both context-aware and label-special word features, is sent to a document encoder so that it can be categorized.	Cai et al. (2020)

(2020) introduced OpTok, an optimization technique for improving tokenization efficiency in BERT-based tasks. Future research should focus on addressing these issues and exploring diverse datasets (Table 9).

.

5.3 Grammatical error detection and correction

Grammatical Error Detection (GED) faces challenges due to limited datasets and inconsistent label distributions. Contextualized word embeddings, especially BERT, have proven effective in overcoming these issues by capturing compositional information from large amounts of unmonitored data. Rei (2017) introduced contextual embeddings as a sophisti-

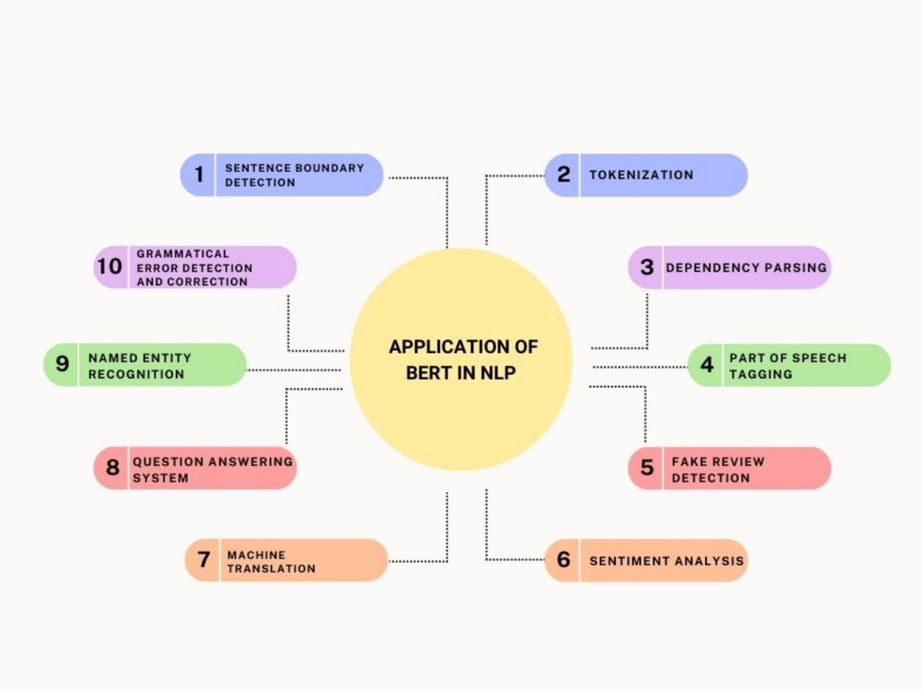

APPLICATION OF BERT IN NLP
1 SENTENCE BOUNDARY DETECTION
2 TOKENIZATION
3 DEPENDENCY PARSING
4 PART OF SPEECH TAGGING
5 FAKE REVIEW DETECTION
6 SENTIMENT ANALYSIS
7 MACHINE TRANSLATION
8 QUESTION ANSWERING SYSTEM
9 NAMED ENTITY RECOGNITION
10 GRAMMATICAL ERROR DETECTION AND CORRECTION

Fig. 7 BERT Implementation in NLP tasks

Table 8 SBD using BERT

Task	Model	Data Set	Year	Evaluation(F1)	References
Sentence boundary Detection	Bert SBD BiLSTM-CRF	FinSBD-2019	2019	Bert SBD: 82.5% BiLSTM-CRF: 89.6%	Du et al. (2019)
	SBD-TT	Stanford Lectures Hybrid Dataset	2020	SBD-TT (Stanford Lectures): 79.83% SBD-TT(Hybrid Dataset): 85.31%	Donabauer et al. (2021)
	BERT	BCCWJ, Jalan-F, Jalan-A	2020	Jalan-F (90.2), Jalan-F+A (95.1), P-BCCWJ(94.8), P-Jalan(92.8)	Hayashibe and Mitsuzawa (2020)

cated solution, and subsequent models like ELMO and Flair were combined with BERT to enhance performance in GED tasks (Akbik et al. 2018; Bell et al. 2019; Peters et al. 2019).

Bell et al. (2019) tested BERT embeddings on multiple datasets, including FCE, CoNLL-2014, and JFLEG, demonstrating that BERTBASE and BERTLARGE outperformed ELMO and Flair. Their findings suggest that the vast training data used for BERT, such as English

Table 9 Word segmentation dataset

Word Segmentation	Datasets	Models	Average F1 score
Chinese Language	MSR, PKU	CRF, Softmax	98.4, 96.5
Persian Language	Bijankhan Test Set	CRF, BERTa, BERTb	0.69, 0.96, 0.98

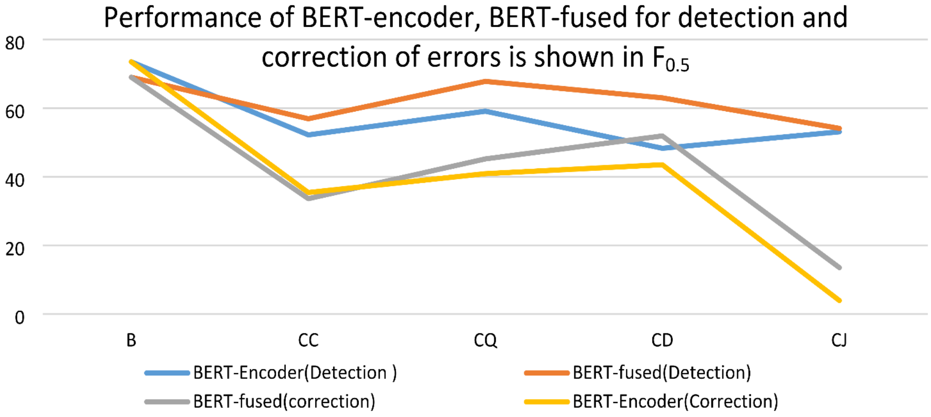


Fig. 8 Error correction "Character level errors are shown by B, CC, CD, CQ and C

Table 10 Grammatical error detection by (Bell et al. 2019)

GED	Model	Data Set	Year	Evaluation				References
1.	ELMO $\text{BERT}_{\text{large}}$ $\text{BERT}_{\text{Base}}$	FCE dataset JFLEG dataset CoNLL-2014	2019	**F1 score (Test)**	FCE dataset	JFLEG dataset	CoNLL-2014	Bell et al. (2019)
				$\text{BERT}_{\text{large}}$	**56.96**	61.52	**45.80**	
				$\text{BERT}_{\text{Base}}$	57.28	**61.98**	46.29	
				ELMO	52.81	58.54	40.15	
				Falre	49.97	54.08	34.35	

Wikipedia and BookCorpus, contributes to its superior performance. In contrast, ELMO and Flair were trained on the One Billion Word Benchmark, leading to performance variations across corpora.

Wang et al. (2020) developed a Chinese Grammatical Error Correction model using BERT-encoder and BERT-fused models, with BERT-encoder excelling at character-level tasks. However, it struggled with sentence-level errors, where BERT-fused showed better performance. Similarly, Li et al. (2020a, b, c) applied BERT to the English FCE dataset for error detection and correction, focusing on sentences with single errors, achieving improved results in their error correction model (Fig. 8).

BERT's superior performance in GED and correction tasks compared to models like ELMO and Flair highlights the importance of selecting appropriate pre-trained embeddings and datasets. However, limitations remain in handling sentence-level errors and multiple errors within a sentence, suggesting areas for further research and improvement (Table 10).

Table 11 Dependency parsing dataset

Languages	Datasets
Slovenian	Slovenian treebank based on ssj500k corpus
Croatian	Treebank by (Agi´c and Ljubeˇsi´c)
Estonian	(Estonian Dependency Treebank)
Finnish	(Finnish Treebank based on Turku Dependency Treebank
English	Gold Standard Dependency Corpus

CroSloEngual BERT demonstrates superior performance compared to mBERT in a monolingual context for all three languages: Croatian, English and Slovenian. The outcomes of dependency parsing task are presented in terms of the unlabeled attachment score (UAS) and labelled attachment score (LAS). When comparing monolingual settings, FinEst BERT demonstrates superior performance compared to mBERT in the Estonian and Finnish languages. The largest difference in performance is shown in the Finnish data. FinEst BERT and mBERT exhibited equivalent performance on English dataset. Results are shown in Table 12 below,

Table 12 Dependency parsing using BERT architecture

	Evaluation (F1 score)						Reference
	mBERT		FinEst		CroSloEngual		
	UAS	LAS	UAS	LAS	UAS	LAS	
Croatian	0.930	0.891	–	–	0.940	0.903	(Ulčar and Robnik-Šikonja 2020)
Slovenian	0.938	0.922	–	–	0.957	0.947	
English	0.917	0.894	0.918	0.895	0.922	0.899	
Finnish	0.898	0.867	0.933	0.915	–	–	
Estonian	0.880	0.848	0.909	0.882	–	–	

5.4 Dependency parsing

The dependency parsing task is to anticipate the hierarchical arrangement that represents the syntactic relationships between words in a given text. (Ulčar and Robnik-Šikonja 2020) used five languages which are shown in Table 11 below along with their datasets used in the task of dependency parsing,

The effectiveness of BERT in dependency parsing varied depending on the language, with certain models, such as CroSloEngual BERT, performing better in some situations.

5.5 Part of speech tagging

The goal of POS tagging is to accurately categorise each token into specific grammatical categories such as verb, adjective, punctuation, adverb, noun, etc. (Khan et al. 2019). (Adesam and Berdičevskis 2021) used Talbanken-SBX, Talbanken-UD, and Eukalyptu datasets, all these resources were annotated using different tag sets. They used a total of five distinct models. KB-Bert, Flair, Stanza, Marmot, and Hunpos. KB-Bert consistently achieves the best accuracy among neural taggers and is also the quickest. Its performance on GPU is equivalent to that of the pre-neural taggers shown in Table 13. The variability in outcomes mostly relies on the choice of corpora and tag sets.

Other factors were embedding's, performance was mainly affected by usage of embeddings n different datasets and the most important thing which effected results was inconsistency in dataset annotation and fine-grained tag sets.

Table 13 Performance of BERT architecture on POS tagging

Task	Model	Data Set	Year	Evaluation						References
POS	BERTje$_{850k}$ BERTje	2.4B tokens	2019	BERTje$_{850k}$, CONLL-2002 = 87.6 BERTje, SoNaR-1 = 88.3						de Vries et al. (2019)
	KBBert Flair Stanza Marmot Hunpos	TB-SBX TB-UD Euk	2021	Accuracy	KB-Bert	Fair	Stanza	Marmot	Hunpos	Adesam and Berdičevskis (2021)
				TB-SBX	72.7	68.9	60.10	55.31	45.47	
				TB-UD	68.8	64.4	57.55	51.11	39.99	
				Euk	59.8	54.1	46.27	40.84	31.86	

Table 14 Performance of BERT architecture on different datasets of POS tagging

Model	Languages	Dataset	Evaluation				References
mBERT, CroSloEngual, FinEst	Finnish, Estonian, English, Croatian Slovenian	Croatian (Agi´c and Ljubeˇsi´c) English, Estonian Finnish (Finnish treebank) Slovenian(Slovenian treebank)		mBERT	CroSloEngual	FinEst	Ulčar and Robnik-Šikonja (2020)
			Croatian	0.98	0.983	–	
			Slovenian	0.987	0.991	–	
			English	0.969	0.972	0.970	
			Finnish	0.970	–	0.981	
			Estonian	0.972		0.970	

Ulčar and Robnik-Šikonja (2020) performed POS task to correctly classify each token using grammatical categories such as, Verb, Noun, adjective, pronoun etc. The languages used were Finnish, Estonian, English, Croatian, and Slovenian. The Table 14 displayed the datasets and BERT models. In the monolingual situation, FinEst and CroSloEngual BERTs outperform mBERT on all languages, except for Croatian, where mBERT and CroSloEngual BERT perform equally.

Malmsten et al. (2020) developed KB-BERT using a Swedish dataset from the National Library of Sweden (KB) to address the lack of data for low-resource languages. This dataset included digitized newspapers, official reports, e-reports, social media text, and Wikipedia text in Swedish. Despite being messy and diverse, this data enabled the model to perform well on downstream tasks of Named Entity Recognition (NER) and Part-of-Speech (POS) tagging.

They used the SUC (Stockholm-Umeå Corpus 3.0) dataset for NER tagging, achieving a notable 92.7% F-measure. For POS tagging with the SUC corpus, they observed a one-percent improvement in accuracy, but performance declined as the dataset size increased, highlighting the need for enhanced pretraining for stable results. Tsai et al. (2019) applied Meta-LSTM (Smith et al. 2018), BERT, and mini B-BERT models to universal POS tagging using CoNLL 2018 Shared Task data. They used 17 labels for POS tagging and performed model distillation, with the distilled model outperforming others (Zeman et al. 2018). In POS tagging, KB-Bert showed consistent accuracy, emphasizing the role of corpora and tag sets. BERT has proven effective for POS tagging in languages like Bangla (Roy et al. 2020), Arabic (Saidi et al. 2021), and Gujarati (Mehta et al. 2024). Some studies proposed enhancing BERT by incorporating POS information into embeddings (Liu et al. 2022) or combining BERT with POS preprocessing (Benamar et al. 2021). Sur (2020) introduced recurrent and singular BERT networks for POS tagging, achieving high accuracy. However, Lim and Park (2020) showed that for morphologically rich languages, character-level representations can sometimes outperform BERT.

5.6 Named entity recognition

NER identifies and extracts nouns or noun phrases from the text, categorizing them into entities like person, place, time, and organization (Malik 2018; Malik and Sarwar 2016). This process transforms unorganized content into organized text (Nadeau and Sekine 2007). The pre-trained BERT language model, built on transformers, has significantly improved NER performance (Peters et al. 2019).

De Vries et al. (2019) used the Dutch CoNLL-2002 dataset and a smaller treebank dataset to perform NER. The datasets were split into training (80%), development (10%), and test (10%) sets. They fine-tuned three BERT models: multilingual BERT, BERTje, and BERTje850k. BERTje outperformed the others on both CoNLL-2002 and SoNar-1 (Fig. 9).

Ulčar and Robnik-Šikonja (2020) trained two multilingual BERT models. FinEst BERT used 3.7 billion tokens in Finnish, Estonian, and English texts, while CroSloEngual BERT used 5.9 billion tokens from Croatian, Slovenian, and English texts. Both models were trained to classify entities into Person (PER), Organization (ORG), Location (LOC), and Other (O). The CroSloEngual and FinEst BERT models did better than the multilingual

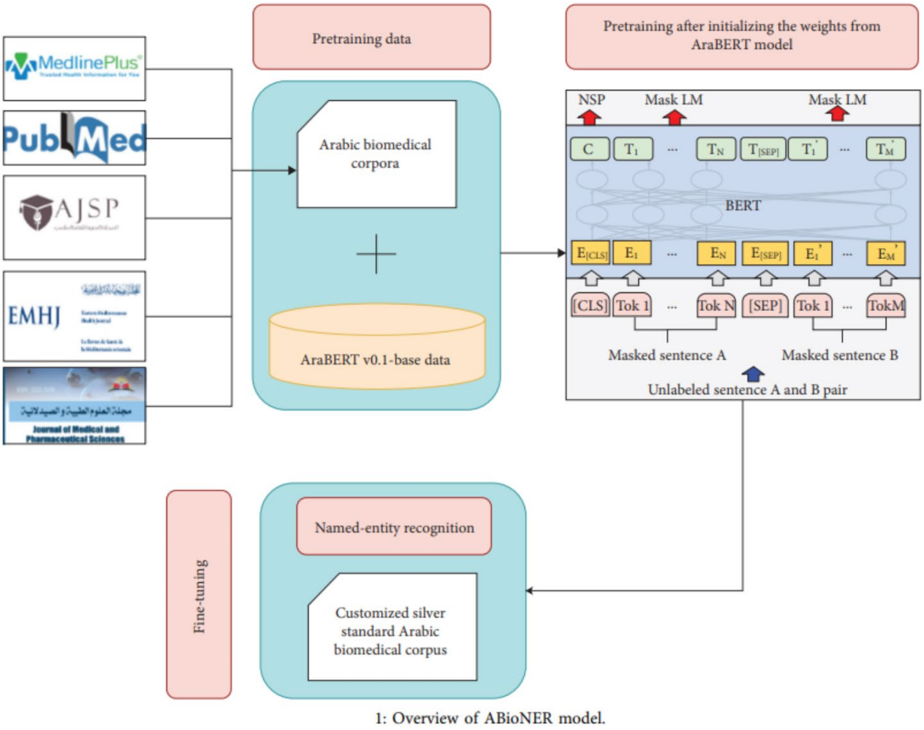

MedlinePlus
PubMed
AJSP
EMHJ
Journal of Medical and Pharmaceutical Sciences
Pretraining data
Arabic biomedical corpora
+
AraBERT v0.1-base data
Pretraining after initializing the weights from AraBERT model
NSP
Mask LM
Mask LM
C
T1
TN
T[SEP]
T1'
TM'
BERT
E[CLS]
E1
EN
E[SEP]
E1'
EM'
[CLS]
Tok 1
Tok N
[SEP]
Tok 1
TokM
Masked sentence A
Masked sentence B
Unlabeled sentence A and B pair
Fine-tuning
Named-entity recognition
Customized silver standard Arabic biomedical corpus
1: Overview of ABioNER model.

Fig. 9 Overview of ABioNER

Table 15 Training on two BERT multilingual models

Task	Model	Data Set		Year	Evaluation				References
NER	Multilingual BERT BERTje$_{850k}$ BERTje	2.4B tokens CONLL-2002, SoNaR-1		2019	**F1 score (Test)**	CONLL-2002	SoNaR-1		de Vries et al. (2019)
					Multilingual BERT	**80.7**	**79.7**		
					BERTje$_{850k}$	**87.6**	**81.1**		
					BERTje	**88.3**	**82.1**		
	mBERT, CroSloEngual, FinEst	Croatian, Slovenian	hr500k, ssj500k	2020		mBERT	CroSloEngual	FinEst	Ulčar and Robnik-Šikonja (2020)
		English	CoNLL 2013		Croatian	0.795	0.894	-	
		Finnish	Finnish news corpus		Slovenian	0.903	0.917	-	
		Estonian	Nime¨uksuste korpus		English	0.940	0.949	0.942	
					Finnish	0.922	-	0.959	
					Estonian	0.906	-	0.930	

Table 16 BERT application on Arabic biometric Corpus

Task	Model	Evaluation		References
NER	AraBERT, BioBERT, ABioNER	Model	F1score	Boudjellal et al. (2021)
		BERT multilingual cased	82.12	
		AraBERT	83.69	
		ABioNER	85.65	

mBERT model in NER, POS labeling, and dependency parsing tasks in a single language or between languages (Table 15).

Regarding the biomedical domain, languages like Arabic do not have as many resources as English. However, Boudjellal et al. (2021) looked at how well AraBERT (Antoun et al. 2020) worked on a small Arabic biomedical corpus compared to the corpus AraBERT used for named entity recognition. Corpora comprised 500k words and two entity types, “disease or syndrome” and “erapeutic or preventive procedure.” The rest of the words were tagged as “O.” The 80% and 20% split was used for training and test datasets, respectively. Performance was compared with AraBERT and BioBERT. ABioNER outperformed (Table 16) AraBERT and BERT multilingual cases (BioBERT). The model’s performance can be improved by using more entity types and a larger corpus.

BERT model along with a layer of CRF was applied on Persian datasets to perform task of NER (Taher et al. 2020). Experiment was applied on two datasets ARAM Poostchi et al. (2018) and PEMA Shahshahani et al. (2018). Results were compiled on both word and phrase level using five named entities. In comparison with previous models’ performance of this model was improved.

Pre train BERT model was applied to perform NER task on contemporary and historical German datasets Labusch et al. (2019) shown in the Table 17 along with No of tokens,

Three named entities (PER, LOC, ORG) were used to perform experiment. They applied BERT basic model and gained state of the art performance without performing extensive optimization and fine tuning. However, BERT large for this much data would be a good

Table 17 NER on contemporary and historical data

Contemporary and Historical data		Size (tokens)
Historical Data Sets	Library of Dr. Friedrich Tessmann (LFT)	70,259
	Austrian National Library (ONB)	28,012
	Digital collection of Berlin State Library (DC-SBB)	47,281
Contemporary Datasets	DE-CoLL-TEST (Sang and De Meulder 2003)	103,387
	DE-CoLL-TRAIN (Sang and De Meulder 2003)	206,931
	GermEval-TEST (Benikova et al. 2014)	96,499
	GermEval-TRAIN (Benikova et al. 2014)	452,853

choice and can be applied in future to gain improved results. In domain of history data more data would be adding to improve performance.

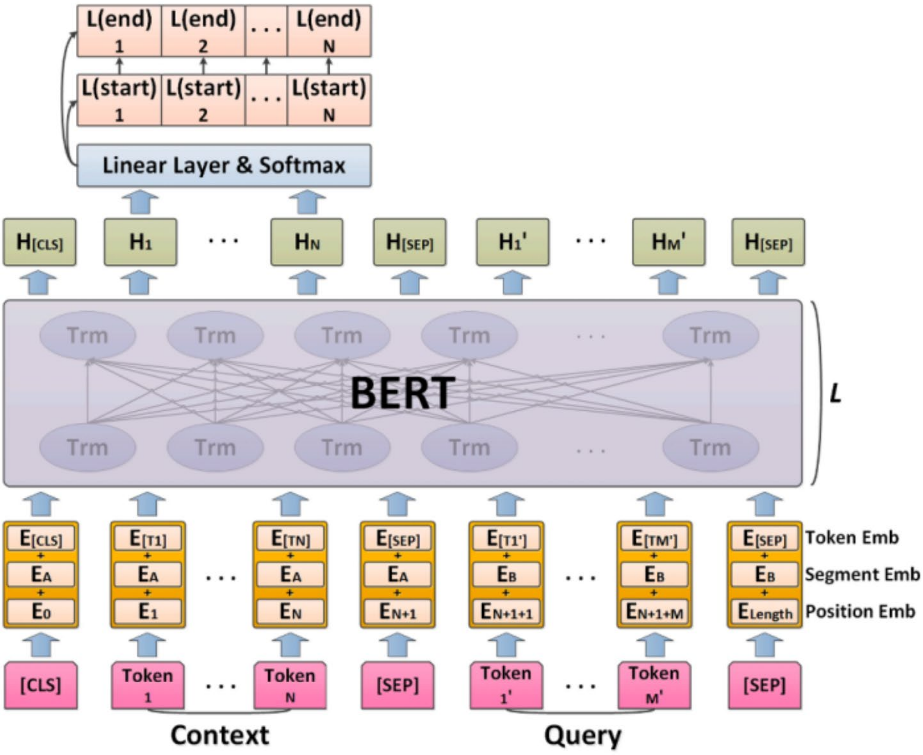
(Sun et al. 2020) used MRC (Machine reading comprehension) method on biomedical datasets to perform tasks of NER using BERT architecture. Previous all studies were sequence labeling problems (Fig. 10).

Chemical, disease and proteins are the three named entities used for extracting information from sentences. To cover biomedical domain, BIOBERT is used. Datasets from chemical entity type are BC4CHEMD and BC5CDR-Chem (Krallinger et al. 2015; Li et al. 2016). C5CDR-Disease Li et al. (2016) and NCBI-Disease Doğan et al. (2014) were from disease entity type while BC2GM Smith et al. (2008), JNLBPA Collier and Kim (2004) were from proteins entity type (Lee et al. 2020). BioBert architecture performed State of the Art results using MRC technique in above mentioned six datasets in biomedical domain. In their paper Sun et al. (2021) compared performance of Bio-BERT(MRC) with previous studies conducted on this domain, results dominated all previous studies (Tables 18 and 19).

Chinese corpora consisted of clinical text collected from Wikipedia and the book Corpus Li et al. (2020b). Targeted named entities were anatomy, disease, symptoms, and exams. Data was collected from different medical fields, such as dermatology and dentistry. Two datasets were used in this research: the CCKS 2017 dataset and the CCKS 2018 dataset. BERT's Pre-train model was used along with three layers: linear, CRF, and BiLSTM-CRF layer. A dictionary and radical features were used to fine-tune. The results obtained were compared with previous research, and it was concluded that this model gave state-of-the-art results compared to others. Results are shown in the Table 20; however, with more refined clinical data and terminology dictionaries, results can be improved further.

The BERT model was used to train Chinese language character embeddings Gong et al. (2019), and a link was established between the BERT model and the Chinese radical-level representation using the BRGU-CRF Model. Radicals are considered the basic unit of Chinese characters. They used the Chinese NER dataset MSRA with three name entities (PER, LOC, ORG) to perform experimentations, and the results are shown in the Table 21,

Results show that adding BERT added more value to results with improved results for Chinese NER. Cho and Lee (2019) used four neural architectures (BiLSTM, BiLSTM-CRF, GRAM-CNN, and BERT model) to evaluate three corpuses NCBI comparatively, GMM, and CDR details in Table 18; they proposed their model named contextual long short-term memory networks with CRF (CLSTM) and compared results in Table 22,



1. Using BERT to perform BioNER in the MRC framework.

Fig. 10 BIoBert Architecture applied on biomedical datasets (Sun et al. 2021)

Table 18 Results of BioBert model on biomedical datasets

Model	Dataset	F1 Measure
BIO-BERT(MRC)	BC4CHEMD	92.92
	BC5CDR- Disease	87.83
	BC5CDR- Chemical	94.19
	NCBI-Disease	90.04
	BC2GM	85.48
	JNLBPA	78.93

Table 19 Named entities along with description

Entity Types		Description
PROTEINAS	Protein	Proteins and genes
NORMLIZABLES	Chemical (+)	Chemicals that are suitable to externalization
NO- NORMLIZABLES	Chemical (-)	Chemicals that are not suitable to externalization
UNCLEAR	Other	Miscellaneous Entities

Table 20 BERT performance on NER datasets

	F1 measure	
	CCKS 2017	CCKS2018
FT-BERT+BiLSTM+CRF	91.31	88.80
FT-BERT+BiLSTM+CRF+Dictionary(ensemble)	91.60	89.56

Table 21 Experimentation results on MSRA datasets

Models	F1 Score
Word2Vec+BGRU-CRF+radical	90.45
BERT+BGRU-CRF+radical	95.42

Table 22 Medical NER datasets

Corpus Name	Entity Type	References
National center for Biotechnology Information	Disease names	Smith et al. (2008)
BioCreative II Gene Mention	Gene Names	Li et al. (2016)
BioCreative V Chemicals Disease Relationship (CDR)	Disease names and chemical names	Moen and Ananiadou (2013)

Table 23 Performance of BERT on different NER datasets

Model	F1 Measure		
	CBI	GM	CDR
BiLSTM	80.71	72.33	81.88
BiLSTM-CRF	83.37	80.30	85.50
GRAM-CNN	84.18	79.53	85.79
BERT	80.90	81.65	85.72
CLSTM (Word+Character Level)	85.31	81.14	86.33

Akhtyamova (2020) scraped the subset of SciELO documents based on some heuristics in Spanish biomedical literature. He retrieved 1,368,080 sentences containing 86,851,275 tokens. SPARC corpus used to perform experiments contained 16,504 sentences and 396,988 tokens and included four entity types: Normalizables, No_Normalizables Proteinas, and Unclear. They experimented with three models' standard embedding, General-domain BERT, and In-domain BERT. Their comparative results are shown in the table below (Table 24),

Table 24 Results of models applied on SPACCC Corpus

Model	F-score
Standard embeddings	87
General-domain BERT	84
In-domain BERT	89

Three training options were integrated into CLSTM (Cho and Lee 2019): a character-level model, a word-level model, and a word+character-level model. Comparative outcomes demonstrated that their model outperformed alternative neural architectures on the CBI and CDR datasets. In contrast, the performance of the BERT model was marginally superior to that of the other models (Table 23).

Table 25 Performance of BERT $_{large}$-BiLSTM-CRF and BERT$_{wwm}$-BiLSTM-CRF

Model	F1 Score
Word2Vec-BiLSTM-CRF	83.37
BERT $_{large}$-BiLSTM-CRF	94.32
BERT$_{wwm}$-BiLSTM-CRF	96.87

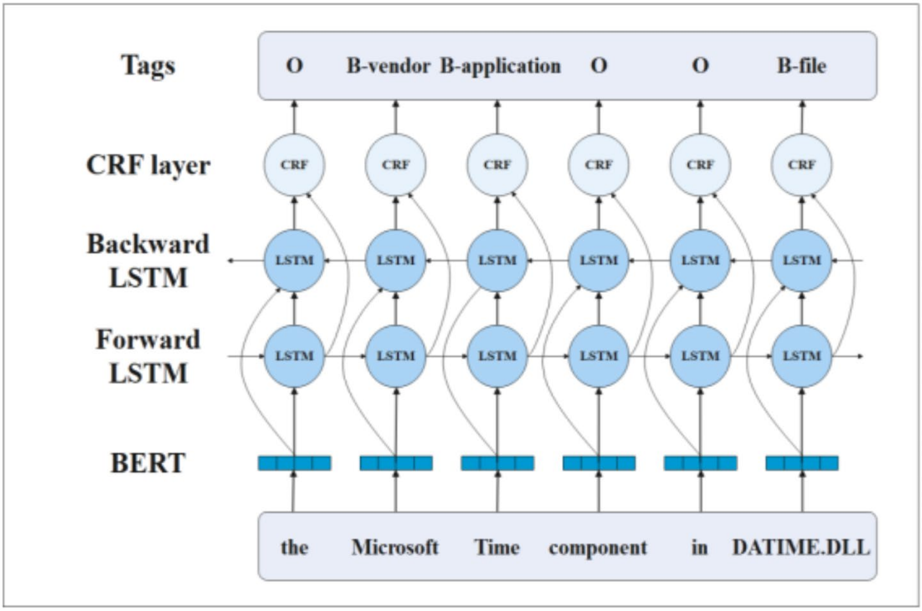


Fig. 11 Demonstration of BERT + LSTM + CRF architectures

During experimentation, In-domain BERT embeddings performed well on Spanish biomedical dataset. General- domain BERT showed pathetic results as compared to other two. They also conducted error analysis in order to find out reasons behind performance of their models. Zhou et al. (2021) performed NER on cyber security data, they used open-source corpus provided by (Bridges et al. 2014).

They used seven entity types (Vendor, application, version, edition, OS, File, Hardware). They used three models Word2Vec-BiLSTM-CRF, BERT $_{large}$-BiLSTM-CRF and BERT-$_{wwm}$-BiLSTM-CRF (Table 25).

Architecture of BERT-BiLSTM and CRF is shown in the figure below, Both BERT $_{large}$-BiLSTM-CRF and BERT$_{wwm}$-BiLSTM-CRF showed significant gain (Fig. 11).

Pre-train BERT model maps words in a vector space. This word embedding will be transferred to the next layer as an input, BiLSTM will combine input features into output vectors, and the CRF layer generates word tags after decoding the word sequence (Zhou et al. 2021). Souza et al. (2020) combined BERT with CRF to perform a named entity recognition task for the Portuguese language corpus. They used two corpora, the first HAREM and the second MiniHAREM corpora (Santos et al. 2006; Freitas et al. 2010). They used three models: multilingual BERT-Base, Portuguese BERT-Base, and Portuguese BERT-Large (Table 26).

Although the Portuguese BERTLARGE models exhibit the maximum performance levels in both scenarios, their performance degrades when implemented in the feature-based

Table 26 BERT application on datasets provided by (Freitas et al. 2010)

Models Architecture	F1 measure	
	Total	Selective
ML-BERTBASE-LSTM	69.59	76.35
ML-BERTBASE-LSTM-CRF	72.14	77.76
ML-BERTBASE	73.78	78.25
ML-BERTBASE-CRF	74.15	79.44
PT-BERTBASE-LSTM	74.30	80.09
PT-BERTBASE-LSTM-CRF	75.69	81.66
PT-BERTBASE	77.98	83.03
PT-BERTBASE-	77.73	82.68
PT-BERTLARGE-LSTM	72.50	78.53
PT-BERTLARGE-LSTM-CRF	74.86	80.37
PT-BERTLARGE	77.92	83.30
PT-BERTLARGE-CRF	78.67	83.24

approach. Although they perform less well than their more minor variants, they remain superior to the Multilingual BERT. Furthermore, BERTLARGE models do not significantly enhance the selective scenario compared to BERTBASE models. The reason for this was the limited size of the NER dataset.

Studies have applied BERT-based models to NER tasks in Kannada (Hebbar et al. 2023), German legal texts (Darji et al. 2023), Chinese medical records (Ma and Chen 2023), and aerospace requirements (Ray et al., 2023). Researchers have proposed enhancements to BERT-based NER models, including the integration of BiLSTM and span pointer decoding for low-resource scenarios (Weng and Zhang 2023), dynamic fusion of BERT layers with whitening (Liang and Shi 2023), and the incorporation of Biaffine layers for improved entity boundary detection (Wang and Gu 2023). In NER tasks, BERT has shown proficiency in several languages, including Arabic, Persian, German, Chinese, and Portuguese. The research showed how vital BERT is for outperforming earlier models, particularly when tailored to particular domains, such as historical or biomedical data.

5.7 Question answering

Question Answering (QA) systems autonomously respond to inquiries using natural language, supported by knowledge information systems (Aithal et al. 2021). BERT-based QA systems have advanced significantly through neural architectures and embeddings (Devlin et al. 2019; Catelli et al., 2021). Yang et al. (2020) used a large corpus of Wikipedia articles with BERTserini, an end-to-end open-domain QA system integrating BERT and the Anserini IR toolkit. BERTserini demonstrated a notable improvement with an exact match score of 38.6, F1 of 46.1, and recall of 85.8. Despite the large corpus, it retrieved many irrelevant sentences, a challenge for further refinement and potential multilingual upgrades.

Yu et al. (2020) introduced TransTQA, using ALBERT for automated responses by extracting answers from analogous questions. TransTQA applied transfer learning with technical knowledge, outperforming BERT-Rerank and BERTserini. It ranks responses based on similarity scores, returning the top three ranked answers. Alzubi et al. (2023) proposed COBERT, a dual algorithmic system combining a retriever and a reader for complex queries using the CORD-19 dataset. COBERT achieved an Exact Match (EM) score of 80.6

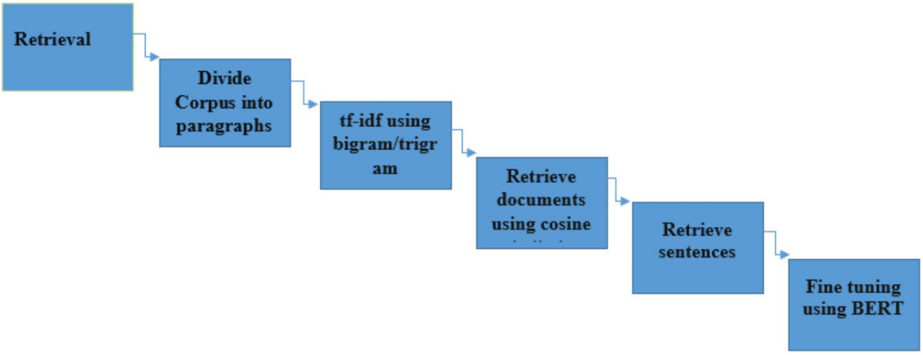

Retrieval
Divide Corpus into paragraphs
tf-idf using bigram/trigr am
Retrieve documents using cosine
Retrieve sentences
Fine tuning using BERT

Fig. 12 Steps to perform Q/A system by (Aithal et al. 2021) (Table 27)

Table 27 Performance of BERT on QA datasets proposed by (Yu et al. 2020; Aithal et al. 2021)

Task	Model	Data Set	Year	Evaluation			References
Question Answering System	SASE LSTM Albert	TechQA Stack Unix	2020		TechQA	Stack Unix	Yu et al. (2020)
				SASE LSTM	56.05	55.20	
				Albert	70.63	71.32	
	BERT	SQuAD 2.0 SQuAD 1.1	2021	(Efficiency) Unanswerable Questions SQuAD 2.0: 48% SQuAD 1.1: 91% Irrelevant Questions SQuAD 2.0: - SQuAD 1.1:100%			Aithal et al. (2021)

and an F1 score of 87.3, surpassing previous models. It processes user requests, providing a one-line answer, title, and detailed paragraph from the scientific literature.

Aithal et al. (2021) conducted experiments on SQuAD 2.0 and SQuAD 1.1 datasets using BERT, addressing irrelevant and unanswerable questions through a similarity mechanism, reducing computational resources (Fig. 12).

Yang et al. (2020) used TVQA datasets for video QA, employing BERT for video and subtitle semantics. Their model showed superior performance with BERT but had constraints with full-length subtitles. Qu et al. (2019) introduced a history answer embedding method in ConvQA, incorporating dialogue history into a BERT-based model. Tests on the QuAC dataset showed the effectiveness of their techniques, with HAE delivering similar performance to FlowQA but with greater training efficiency. Mozannar et al. (2019) tackled open domain QA using Wikipedia, introducing the Arabic Reading Comprehension Dataset (ARCD) and employing Machine Translation for Arabic-SQuAD (Fig. 13).

BERT has dramatically improved response accuracy and has been a crucial part of QA systems. DistilBERT outperformed preceding pre-trained models, and TransTQA, in particular, showed exceptional performance. Recent research on question-answering systems using BERT has focused on improving performance for long text sequences (Ramaraj et al. 2024), combining BERT with GPT for healthcare applications (Jeong et al. 2023), and fine-tuning BERT models for biomedical research papers (Pudasaini and Shakya 2023).

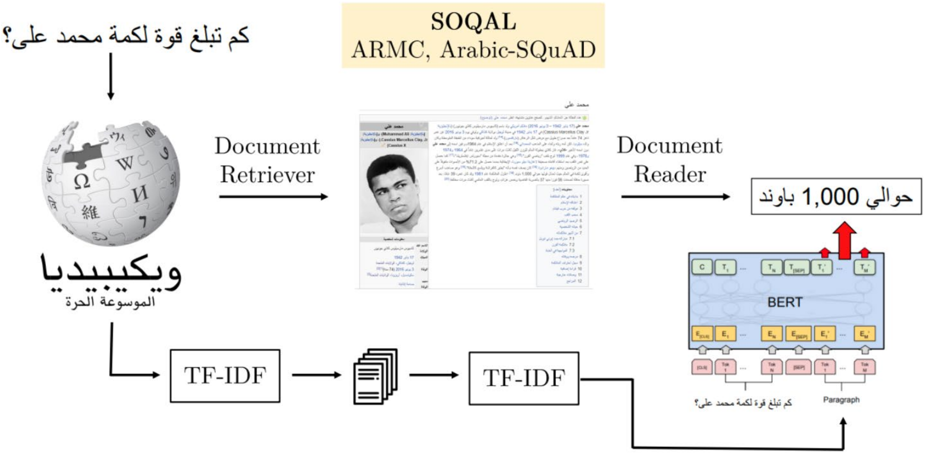


Fig. 13 Model of Machine translation applied on ARCD and Arabic-SQuAD proposed by (Mozannar et al. 2019)

Researchers have also explored integrating knowledge graph embeddings with BERT for domain-specific question answering in nuclear power (Ma et al. 2023), chemistry (Zhou et al. 2023), and bridge inspection (Yang et al. 2023). Other approaches included BERT for question summarization and sentence similarity in key information extraction (Sharma et al. 2023). These studies demonstrate the versatility of BERT-based models in various question-answering tasks and highlight the potential for improved performance through domain-specific adaptations and integration with other techniques like knowledge graph embeddings.

5.8 Machine translation

The Neural Machine Translation (NMT) aims to transform a source language sentence into the target language's associated alternative. An attention module, an encoder, and a decoder comprise an NMT model. The input sequence is encoded into hidden representations, which the decoder then maps to the target sequence. In machine translation, it is crucial to perform bilingual tasks as compared to monolingual tasks while performing pre-training and fine-tuning. However, significant results are achieved in other tasks like NER and POS. To address this issue (Clinchant et al. 2019) presented research where BERT models can be exploited for supervised NMT. They explored the impact of BERT on translation quality by integrating the pre-trained BERT model with previous neural machine translation models. For this purpose, three monolingual corpora are used: NMT-src, Wiki, and News, containing 4.5 M, 72 M, and 210 m, respectively.

In their study, Imamura and Sumita (2019) undertook NMT tasks on English-to-Vietnamese (En-Vi) and English-to-German (En-DE) datasets. The Dist. dataset, sourced from the Workshop on Statistical Machine Translation (WMT) (Bojar et al. 2014), was used for these experiments. Additionally, they conducted experiments on a low-resource setup, using the Eng. to Vi corpus provided by the International Workshop on Spoken Language Translation(Bentivogli et al., 2016). This diverse range of datasets allowed them to test the robustness of their NMT model across different language pairs and resource levels.

Table 28 APT to acquire knowledge from Pre-train model

Machine language datasets		Size
Workshop on Statistical Machine Translation WMT-17	Chinese to English (ZH→EN)	7.5 million Sentence Pair
WMT-14	English to German (EN→DE) German to English (DE→EN)	4.5 million sentence pair

Table 29 MT for different datasets

BLUE Score using BERT-fused Model	
English to German	30.45
German to English	36.11
English to Spanish	41.4
English to French	38.7
English to Chinese	28.2
English to German	30.78
English to French	43.78

Weng et al. (2020) conducted research using two bilingual datasets and monolingual data in English, German, and Chinese. Their findings revealed that fine-tuning does not always provide sufficient information for generating machine translation tasks. This insight is crucial for understanding the limitations of the NMT model and can guide future research in this area (Table 28).

They proposed a framework (APT) that leverages pre-trained models for NMT tasks, achieving significant improvements. Zhu et al. (2020) introduced the BERT-fused model for machine translation, using IWSLT’14 datasets for English to German, Spanish, French, and Chinese translations (training sizes: 160k, 183k, 236k, and 235k respectively), and WMT’14 datasets for English to German and French translations (training sizes: 4.5 million and 36 million). Due to high computational costs, they used BERTbase for WMT tasks and BERTLarge for IWSLT tasks. The BERT-fused model utilized layers instead of embeddings to interact with representations and employed attention modules like BERT-encoder and BERT-decoder. The BLEU scores are shown below (Table 29):

MODEL	Encoder	Decoder	EN→DE	DE→EN	ZH→EN
APT framework	GPT	BERT	28.89	34.32	25.98
	BERT	GPT	29.23	34.84	26.21
	GPT	GPT	28.97	34.26	26.01
	BERT	BERT	29.02	34.67	26.46

Previous studies slightly improved these scores, but their proposed work had certain limitations. Their model required additional storage costs and inference time, which will be addressed to get more robust results.

Shimanaka et al. (2019) conducted research on a Sentence pair encoder that simultaneously encodes Machine translation hypotheses and reference translation. The pre-trained encoder is refined by employing machine language modeling and next-sentence prediction.

Instead of encoding each sentence separately, they used a sentence pair encoding technique in which sentences are encoded in pairs.

Results improved after using sentence pair encoding, pre-training, and fine-tuning. Since machine translation was applied to translate all other languages into English, performance can be improved through the proper implementation of BERT. However, this technique can also be applied to the translation of languages other than English (Table 30).

Zhang et al. (2021) proposed a BERT-fused Joint Attention (BERT – JAM) model, which is designed on BERT multilayer representations and includes a joint attention module that incorporates BERT representation with encoder/decoder representation (Shimanaka et al. 2019) (Table 31).

A three-phase optimization strategy was used to train BERT-JAM, refining BERT and reducing catastrophic confusion. BERT-JAM was tested on five IWSLT low-resource translation tasks: IWSLT'14 English-German, German-English, English-Spanish, IWSLT'17 English-French, and English-Chinese, as well as in a high-resource setting with the WMT'14 English-German dataset (4.5 million sentence pairs). Results in indicated that BERT-JAM surpasses current models, achieving state-of-the-art BLEU scores across translation tasks, highlighting the impact of integrating pre-trained BERT with neural machine translation (NMT) models.

Chen et al. (2023) introduced a BERT-enhanced NMT model that improved performance across tasks, while Dai et al. (2023) integrated syntactic knowledge with BERT via graph attention networks to enrich source representations. Zhang et al. (2023) demonstrated that multilingual knowledge distillation could align cross-lingual embeddings, and Jha et al. (2023) developed a multilingual NMT system for Indian languages using the mT5 transformer. Mohtashami et al. (2023) showed that large language model knowledge can improve translation quality assessment for low-resource languages. Sharma et al. (2023) reviewed machine translation systems, analyzing classical, statistical, and deep learning models.

Recent research highlights BERT's promise in machine translation across languages and applications, with effective applications in translation scoring (Cui and Liang 2024), grammar error detection (Qing 2024), and cross-language translation via prompt engineering (Pourkamali and Sharifi 2024). Techniques like QE-fusion have enhanced translation quality in large language models (Vernikos and Popescu-Belis 2024). BERT's influence has driven pretraining and transfer learning adoption (Gupta 2024a, b), and in some document-level translations, models rival GPT-4's performance (Wu et al. 2024). Simultaneous translation using large language models now achieves results on par with top baselines (Koshkin et al. 2024), while BERT with linguistic features has improved machine reading comprehension (Li and Zhang 2024).

Sentiment Analysis.

Many studies have explored using BERT in aspect-based sentiment analysis (ABSA). Hoang et al. (2019) and Ansar et al. (2021) demonstrate BERT's potential in this task, with Ansar proposing a refined aspect extraction methodology to improve efficiency and accuracy. Sun et al. (2019) further enhance BERT's performance by introducing deep context

Table 30 Performance of BERT $_{\text{BASE}}$ on different MT datasets

BERT $_{\text{BASE}}$	Czech - English	German - English	Finnish - English	Latvian - English	Russian - English	Turkish - English	Chinese - English
	CS→EN	DE→EN	FI→EN	Lv→En	RU→EN	TR→EN	ZH→EN
	0.720	0.761	0.857	0.828	0.788	0.798	0.763

Table 31 Summary of data sets

POS Tagging	No of Tokens	POS tags	Sources	References
Talbanken-SBX(TB-SBX)	96,346	25		Nivre et al. (2016)
Talbanken-UD(TB-UD)	96,858	16		Nivre (2014)
Eukalyptus treebank(Euk)	99,909	13		Adesam et al. (2015)
NER	**No of Tokens**		**Sources**	**References**
Historical Datasets				
Library of Dr. Friedrich Tessmann (LFT)	70,259		Newspaper	Sang and De Meulder (2003)
Austrian National Library (ONB)	28,012		Newspaper	
Digital collection of Berlin State Library (DC-SBB)	47,281		Newspaper	
Contemporary Datasets				
DE-CoLL-TEST	103,387		German Newspaper	Sang and De Meulder (2003)
DE-CoLL-TRAIN	206,931			
GermEval-TEST	96,499		German Wikipedia, Online newspapers	Benikova et al. (2014)
GermEval-TRAIN	452,853			
Medical Datasets	Entity Types			
BC4CHEMD	Chemical/Drug		89,679	Collier and Kim (2004)
BC5CDR- Disease	Disease		14,228	Doğan et al. (2014)
BC5CDR- Chemical	Chemical / Drug		14, 228	Krallinger et al. (2015)
NCBI-Disease	Disease		7639	Li et al. (2016)
BC2GM (Smith et al. 2008)	Protein/ Gene		20,510	Habibi et al. (2017)
JNLBPA	Protein/ Gene		22,562	Lee et al. (2020)
Chinese Biomedical Corpora				
National center for Biotechnology Information (NCBI)	Disease names		792 (abstracts)	Cho and Lee (2019)
BioCreative II Gene Mention (GM)	Gene Names		20,000 sentences	
BioCreative V Chemicals Disease Relationship (CDR)	Disease names and chemical names		(1500 abstracts)	
Spanish Biomedical NER corpus				
SPACCC Corpus (Akhtyamova 2020)	Sentences 16,504	Words 396,988	Normalizables, No Normalizables, Proteinas, Unclea	
Portuguese NER Dataset				
First HAREM	95,585		Location, Person, Organization, Value, Things Date, Event, Title, Abstraction and Other.	Souza et al. (2020)
MiniHAREM	64,853			

Table 31 (continued)

POS Tagging	No of Tokens	POS tags	Sources	References
Data sets for NER	Languages		Entity types	Moon, wasthy, Ni & Florian, (2019).
CoNLL NER '02/'03	Spanish, Dutch, English German		PER, LOC, ORG, MISC	
Onto Notes 5.0	Arabic, Chinese, English		18 NER types	
Datasets for Machine Translation Tasks				
Datasets	**Training Size**			**Reference**
IWSLT'14	Eng to Ger		160k	Imamura and Sumita (2019)
	Eng to Spa		183k	
IWSLT'17	Eng to Fre		236k	
	Eng to Chi		235k	
WMT'14	Eng to Ger		4.5 M	
	Eng to Fre		36 M	
WMT (Workshop on Statistical Machine Translation) 17	Chinese to English (ZH→EN)		7.5 million Sentence Pair	Weng et al. (2020)
WMT 14	English to German (EN→DE) English to German (DE→EN)		4.5 million sentence pair	

features and constructing an auxiliary sentence, respectively. Karimi (2020Li et al. (2020a, b, c) propose modules and a gating mechanism to improve BERT's performance in aspect extraction and sentiment classification. Lakshmidevi (2023) combines BERT for aspect extraction with diverse ML classifiers for sentiment analysis, achieving high accuracy on benchmark datasets.

Chen et al. (2020) and Chinnalagu and Durairaj (2022) further support this, with the former proposing a fine-grained sentiment analysis model based on BERT and the latter finding BERT to outperform other deep learning models. Joshy and Sunda (2022) also confirm BERT's superior performance compared to other transformer-based models. However, Shih (2021) and Deng et al. (2023) identify challenges in BERT's performance, particularly in handling implicit evaluative texts and needing further model supervision. Despite these challenges, Li and Chen (2023) underscore BERT's significant advantages over traditional models in public sentiment analysis during the COVID-19 pandemic.

Sahoo et al. (2023) found that RoBERTuito outperformed other BERT models in sentiment analysis on Twitter data. Bikku et al. (2023) reported that BERT surpassed traditional machine learning algorithms in large-scale social media data sentiment analysis. Deepa et al. (2021) provided a comprehensive review of BERT's architecture and performance in sentiment analysis, highlighting its ability to capture contextual information. Errami et al. (2023) emphasized the effectiveness of BERT in sentiment analysis, with Errami specifically noting its superior performance in Arabic sentiment analysis. Chinedu et al. (2023) compared the performance of VADER and RoBERTa, an extension of BERT, in sentiment analysis, with RoBERTa demonstrating higher accuracy. Batra et al. (2021) explored different strategies for analyzing BERT-based models in sentiment analysis, with the ensemble approach and compressed BERT model showing significant improvements over existing tools. Yin et al. (2024) and Tripty et al. (2024) found that BERT-based models outperformed other methods in e-commerce and code-mixed language sentiment analysis. Yin et al. (2024) make a valuable contribution to the field of sentiment analysis by demonstrating the effectiveness of BERT on e-commerce platforms. It is well-structured and presents a coherent narrative from problem statement to solution and evaluation. Zyout and Zyout(2024) and Branco et al. (2024) demonstrated the effectiveness of BERT in student feedback and Portuguese restaurant review sentiment analysis, with Zyout also highlighting the importance of an attention layer. He et al. (2024) and Verma et al. (2024) proposed hybrid models incorporating BERT, with He achieving improved accuracy in film review sentiment analysis and Verma using BERT in combination with XGBOOST for ChatGPT reviews.

Fake Review Detection.

Recent research on fake review detection using BERT architecture has shown promising results. Multiple studies have demonstrated the effectiveness of BERT-based models in accurately identifying fake reviews across various domains, including hotels, restaurants, and doctors (Refaeli and Hájek 2021; Zabeen et al. 2023; Lu et al. 2023a, b). A study by Refaeli and Hájek in 2021 showed that using a distributed representation from the context-aware BERT model led to significant improvements in accuracy. However, it is still not clear what role different content representations play. They got 91% accuracy on a set of reviews from the public and 73% accuracy on a set of reviews from a third party, Yelp. By combining BERT with LSTM and Monte Carlo Dropout, adding sentiment features (Mewada et al. 2023), and fusing multimodal features (Li and Chen 2023), researchers have looked into different ways to make the models more robust. Zabeen et al. (2023) represented a study with an accuracy of 91.02% and an F1-score of 90.02%. They concluded that BERT beats LSTM in simulated review identification. BERT improves its ensemble accuracy to 91.75% by including Monte Carlo Dropout approaches, displaying excellent performance and resilience. Although BERT's more enormous parameter count increases its computing needs, it seems more efficient in identifying bogus reviews than LSTM (Zabeen et al. 2023). The research conducted by Lu et al. (2023a, b) proposed a novel fake review detection model combining Text CNN, BERT, and SKEP. Over three benchmark datasets, BSTC showed better performance than all baseline models. Particularly on the Hotel dataset, BSTC obtained an F1-score of 93.36%, well above the baseline models, and an accuracy of 93.44%; by systematically collecting contextual, semantic, and sentiment information, these findings highlight how well BSTC detects bogus reviews.

Transfer learning techniques using BERT variants like RoBERTa, ALBERT, and DistilBERT have also been investigated (Gupta et al. 2021). Some studies have proposed novel architectures, such as SentiBERT (Mewada et al. 2023) and BSTC (Lu et al. 2023a, b), which combine BERT with other neural network components. Overall, BERT-based models have consistently outperformed traditional machine learning approaches, achieving accuracies ranging from 87 to 94% across different datasets (Mir et al. 2023; Li and Chen 2023). Future research should investigate integrating many pre-trained language models to enhance performance in spotting bogus reviews.

Cross-lingual transfer learning.

Recent studies into cross-lingual methodologies for low-resource languages have examined diverse techniques to improve transfer learning and alleviate the constraints of current models. Devlin et al. (2019) presented multilingual BERT (mBERT), showcasing its ability for zero-shot transfer between languages in tasks like named entity recognition and part-of-speech tagging. Lu et al. (2023a, b) refined mBERT for the extinct Tangut language, whereas Pfeiffer et al. (2020) introduced MAD-X, an adapter-based framework that surpasses state-of-the-art techniques in named entity recognition by facilitating fast cross-lingual transfer. The emergence of BERT-based models has markedly enhanced Natural Language Processing (NLP) in multiple languages, including Italian. Tamburini (2020) illustrates how these contextualized word embeddings have enhanced performance in Italian NLP tasks. Wu and Dredze (2020) emphasized that mBERT's efficacy markedly diminishes for low-resource languages in contrast to high-resource languages, with monolingual BERT models exhibiting even less performance. Xia et al. (2021) proposed MetaXL, a meta-learning system designed to convert representations from auxiliary languages to target low-resource languages, enhancing tasks such as sentiment analysis and named entity recognition. Snaebjarnarson et al. (2023) illustrated the advantages of utilizing closely related languages for Faroese, whereas Ansell et al. (2023) suggested a cohesive strategy that integrates parameter-efficient adaptation, machine translation, and multi-target training in diverse resource-constrained contexts. Deode et al. (2023) expanded the research by developing multilingual sentence BERT models utilizing synthetic corpora, demonstrating efficacy for Indian languages. Guarasci et al. (2022) evaluated mBERT's syntactic transfer capabilities across Italian, French, and English, concluding that although the model manages certain syntactic links, it has difficulties with language-specific phenomena such as the pro-drop characteristic in Italian. Artetxe et al. (2020) highlighted that even advanced models encounter challenges due to structural discrepancies, such as syntax and morphology, between linguistically distant languages. These researches highlighted the potential of utilizing evolutionary information, combining various adaptation strategies, and capitalizing on the intrinsic cross-lingual characteristics of multilingual models to improve performance in low-resource languages. However, ongoing issues like linguistic diversity and syntactic complexity still require the advancement of more effective pre-training methods and novel representation learning strategies.

6 Data sets

In this section, different datasets are compiled which are used in the section of NLP tasks. Different data sets are presented here in this section,

7 Challenges and future directions

7.1 Challenges and limitations of Bert

The increasing prevalence of larger models gives rise to many issues. Pre-existing language models belonging to the BERT family have established themselves as the leading models in many.

NLP tasks. Despite the potential applications and remarkable achievements, the BERT architecture is subject to some constraints that have been extensively examined by many researchers. The limitations of the BERT model have been extensively examined by researchers.

Following Fig. 14 represents all challenges and their details are mentioned below,

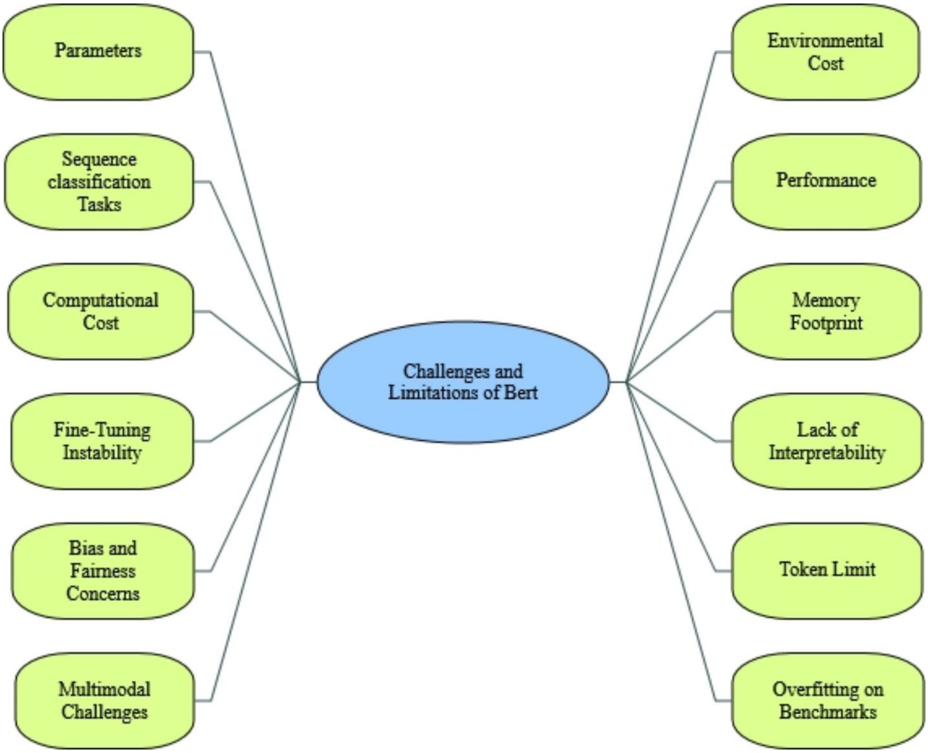


Fig. 14 Challenges and limitations of BERT

7.1.1 Parameters

Implementing BERT-based models in resource-limited settings is complicated, with many parameters determining their performance. Considering the prevailing models in the field, which often include several parameters in the range of hundreds of millions or even billions, encountering these constraints when attempting to expand the size of our models is a foreseeable occurrence. The training pace might be noticeably impeded in distributed training due to the direct correlation between the communication overhead and the number of parameters included in the model. In response to this challenge, current research endeavors have focused on compressing BERT into a more compact model.

7.1.2 Environmental cost

When these models are executed in real-time on the device, there is the possibility of developing innovative and exciting language processing applications (Schwartz et al. 2019; Strubell et al., 2019). However, the exponential increase in computational requirements for scaling these models poses an environmental cost and might prevent broad acceptance.

7.1.3 Sequence classification tasks

According to Kovaleva et al. (2019), BERT is capable of encoding syntactic, semantic, and linguistic aspects. However, it may not use these qualities in downstream tasks. Negation is disregarded by the model proposed by (Ettinger 2020), and the integration of Conditional Random Fields (CRF) may be necessary to enhance its efficacy in some tasks and languages, particularly in the context of sequence classification tasks (Souza et al. 2020).

7.1.4 Performance

According to Rehbein et al. (2020) findings of the research indicate that the use of pretrained BERT embedding's in transfer learning does not consistently provide superior performance compared to other neural architectures. It is emphasized that the selection of an appropriate model and data representation plays a critical role in achieving optimal results.

7.1.5 Computational cost

The use of BERT, particularly its bigger iterations, requires significant computer resources.

The process of training BERT from the beginning might provide challenges for individual researchers or small institutions, as highlighted by (Strubell et al., 2019).

7.1.6 Memory footprint

The deployment of BERT for on-device applications, such as mobile phones, presents challenges primarily attributed to the substantial size of its model parameters. Proposed solutions, such as DistilBERT or TinyBERT, have been suggested to tackle this issue, although with some compromises (Sanh et al. 2020).

7.1.7 Fine-Tuning instability

Fine-tuning of BERT Dodge et al. (2020), might result in instability and necessitates careful selection of hyper parameters.

7.1.8 Lack of interpretability

Neural networks, including BERT, are often criticized as 'black boxes'. Efforts are underway to understand their decision-making processes better, but it is an ongoing challenge (Jacovi and Goldberg 2020).

7.1.9 Bias and fairness concerns

BERT, being trained on large-scale internet text, may inherit societal biases present in those texts. Such biases can influence downstream tasks in unintended ways (Bender et al. 2021).

7.1.10 Token limit

BERT has a maximum token limit (typically 512 tokens), which can be limiting for certain tasks that require processing longer contexts.

7.1.11 Multimodal challenges

While BERT excels at textual tasks, integrating it with other data types (e.g., visual or auditory data) requires additional architectures and complexities.

7.1.12 Overfitting on benchmarks

The community is concerned that constant fine-tuning and testing on benchmark datasets may lead to overfitting on those specific benchmarks, thereby reducing models' generalizability.

BERT has undeniably pushed the boundaries of NLP capabilities, but like all models, it has challenges. While solutions and workarounds exist for some of these issues, they usually come with trade-offs. The challenges present exciting avenues for future research, particularly in making BERT more efficient, interpretable, and unbiased.

7.2 Future recommendations

This study provides researchers with a framework for gaining novel insights into applying the Bert architecture to various NLP tasks. It contributes to the exploration of additional performance enhancement techniques.

The pre-training and fine-tuning paradigm in BERT is a solid way to use large amounts of textual data and make the generalized model work better for specific tasks using smaller labeled datasets. This methodology has produced cutting-edge outcomes and catalyzed an onslaught of subsequent models, including Roberta, DistilBERT, and ALBERT, that endeavor to enhance and broaden the initial framework. However, like all technologies, BERT is not without its challenges. Its resource-intensive nature, the potential for biases, and the intricacies of fine-tuning mean there is still room for improvement in its application and optimization. The recommendations are presented in the Fig. 15 below,

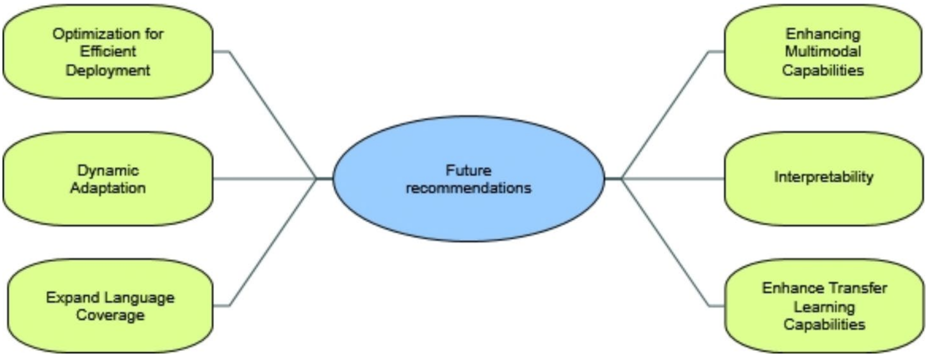

Optimization for Efficient Deployment
Dynamic Adaptation
Expand Language Coverage
Future recommendations
Enhancing Multimodal Capabilities
Interpretability
Enhance Transfer Learning Capabilities

Fig. 15 Future recommendations

7.2.1 Optimization for efficient deployment

Future work should focus on making BERT and its variants more lightweight for deployment in resource-constrained environments like mobile devices. Alotaibi and Dossari (2024) outline promising future explorations to improve model accuracy and give insight into future potential architectures. Methods to prune, quantize, or distill the model are promising avenues. Given the increasing awareness of societal biases embedded in LLM, future applications of BERT should incorporate strategies for bias detection and mitigation. This might include curated training data, post-hoc analysis, or techniques that neutralize biases during fine-tuning (Peng and Zhao 2023) (Fig. 16).

7.2.2 Enhancing multimodal capabilities

Koroteev (2021) suggested recommended digital world is not limited to text, future applications could focus on integrating BERT with other modalities, such as visual or auditory data, to build truly multimodal models.

Fine-Tuning with Domain-Specific Data: BERT's general nature means it might not capture the nuances of specific domains (e.g., medical or legal). Future applications should consider fine-tuning with domain-specific corpora to ensure accuracy and relevance in specialized tasks.

7.2.3 Dynamic adaptation

Instead of static models, future applications might focus on models that can dynamically adapt over time, Sarkar et al. (2023) learning from new data and evolving based on the specific requirements of tasks or users.

7.2.4 Interpretability

As the application of NLP models becomes more widespread, understanding their decision-making processes becomes crucial. Efforts should be made to make BERT, and models like it, more interpretable to non-experts (Aftan and Shah 2023).

7.2.5 Expand language coverage

Though BERT has multilingual versions, there's still a significant skew towards English. Efforts should be directed towards creating BERT-like models for underrepresented languages and dialects (Nozza et al. 2020).

7.2.6 Enhance transfer learning capabilities

Further research can be conducted to understand the layers and neurons of BERT better, allowing more effective transfer learning and scaling to billions of parameters where knowledge from one task can benefit another (Gupta 2024a, b).

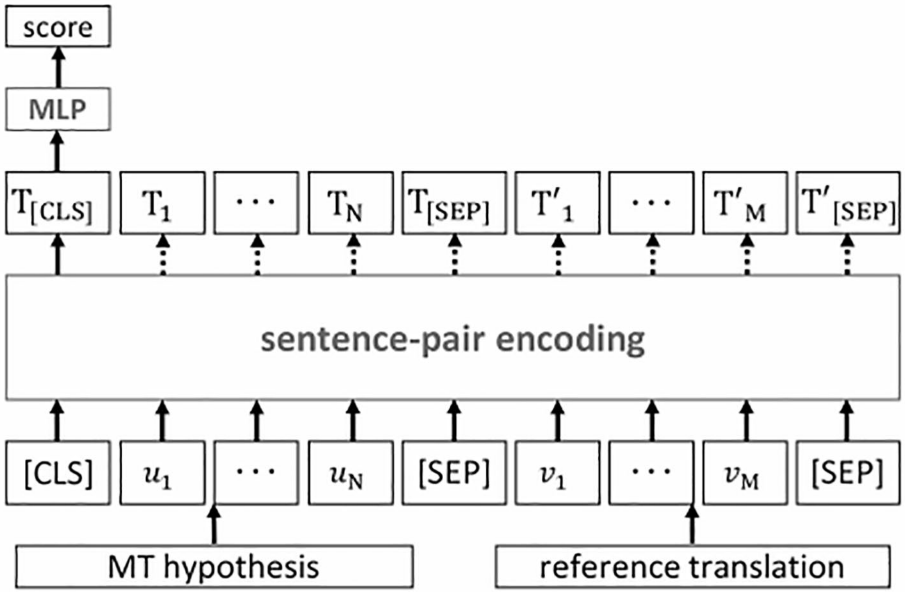


Figure 2: BERT sentence-pair encoding.

Fig. 16 Sentence pair encoding using BERT

8 Conclusions

In conclusion, the BERT has emerged as a versatile and powerful tool across a spectrum of NLP tasks. Its ability to capture contextual relationships and learn from large datasets has led to remarkable improvements in various benchmarks. However, challenges such as model selection, domain-specific fine-tuning, and handling diverse linguistic contexts remain for further exploration and refinement. The foundation that BERT laid for NLP research will most likely continue to grow, resulting in even more complex language models and better performance across various applications. Challenges and future directions reveal the dynamic nature of the field, highlighting the need for continuous innovation and adaptation. As the journey through BERT's landscape continues, this review paper is a valuable resource for researchers, guiding them toward a nuanced understanding of BERT's applications and inspiring further advancements in NLP.

Author contributions Nadia and Ali have written a major part of the paper under the supervision of Ali and Kamran. Amal and Tariq have helped design and improve the methodology and improved the paper initial draft with Nadia and Ali. Kamran, Tariq and Bader have helped in improving the paper sections, such as, review methodology, datasets, performance evaluation and challenges and future directions. All authors are involved in revising the manuscript critically and have approved the final version of the manuscript.

Data availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.

References

Adesam Y, Berdičevskis A (2021) Part-of-speech tagging of Swedish texts in the neural era. *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, 200–209. https://aclanthology.org/2021.nodalida-main.20/

Adesam Y, Bouma G, Johansson R (2015) Defining the Eukalyptus forest–the Koala treebank of Swedish. *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, 1–9. https://aclanthology.org/W15-1804.pdf

Aftan S, Shah H (2023) A survey on BERT and its applications. 2023 20th Learn Technol Conf (L&T) 161:166. https://ieeexplore.ieee.org/abstract/document/10092289/

Aithal SG, Rao AB, Singh S (2021) Automatic question-answer pairs generation and question similarity mechanism in question answering system. Appl Intell 51(11):8484–8497. https://doi.org/10.1007/s10489-021-02348-9

Akbik A, Blythe D, Vollgraf R (2018) Contextual string embeddings for sequence labeling. *Proceedings of the 27th International Conference on Computational Linguistics*, 1638–1649. https://aclanthology.org/C18-1139/?utm_campaign=piqcy&utm_medium=email&utm_source=Revuenewsletter

Akhtyamova L (2020) Named entity recognition in Spanish biomedical literature: Short review and BERT model. *2020 26th Conference of Open Innovations Association (FRUCT)*, 1–7. https://ieeexplore.ieee.org/abstract/document/9087359/

Alotaibi T, Al-Dossari H (2024) A review of fake news detection techniques for arabic language. Int J Adv Comput Sci Appl 15:1

Alsentzer E, Murphy JR, Boag W, Weng WH, Jin D, Naumann T, McDermott M (2019) Publicly available clinical BERT embeddings. *arXiv preprint arXiv:1904.03323*

Alzubi JA, Jain R, Singh A, Parwekar P, Gupta M (2023) COBERT: COVID-19 question answering system using BERT. Arab J Sci Eng 48(8):11003–11013

Ansar W, Goswami S, Chakrabarti A, Chakraborty B (2021) An efficient methodology for aspect-based sentiment analysis using BERT through refined aspect extraction. J Intell Fuzzy Syst 40(5):9627–9644

Ansell J, Smith A, Kumar R (2023) A unified approach combining parameter-efficient adaptation, machine translation, and multi-target training for low-resource languages. J Comput Linguistics 49(2):345–367

Antoun W, Baly F, Hajj H (2020) Arabert: Transformer-based model for arabic language understanding. *arXiv Preprint arXiv:2003.00104*. https://arxiv.org/abs/2003.00104

Artetxe M, Labaka G, Agirre E (2020) *On the structural disparities between linguistically distant languages in cross-lingual models*. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1234–1245

Bano S, Khalid S, Tairan NM, Shah H, Khattak H (2023) Summarization of scholarly articles using BERT and BiGRU: deep learning-based extractive approach. J King Saud Univ Comput Inf Sci 35:101739

Batra H, Punn NS, Sonbhadra SK, Agarwal S (2021) Bert-based sentiment analysis: A software engineering perspective. In *Database and Expert Systems Applications: 32nd International Conference, DEXA 2021, Virtual Event, September 27–30, 2021, Proceedings, Part I 32* (pp. 138–148). Springer International Publishing

Bell S, Yannakoudakis H, Rei M (2019) Context is Key: Grammatical Error Detection with Contextual Word Representations. Proc Fourteenth Workshop Innovative Use NLP Building Educational Appl. https://doi.org/10.18653/v1/W19-4410

Beltagy I, Lo K, Cohan A (2019) SciBERT: A pretrained Language model for scientific text. ArXiv:1903.10676. http://arxiv.org/abs/1903.10676

Benamar A, Bothua M, Grouin C, Vilnat A (2021) Easy-to-use Combination of POS and BERT Model for Domain-Specific and Misspelled Terms. *NL4AI@AI*IA*

Bender EM, Gebru T, McMillan-Major A, Shmitchell S (2021) On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Benikova D, Biemann C, Kisselew M, Pado S (2014) Germeval 2014 named entity recognition shared task: Companion paper. *Workshop Proceedings of the 12th Edition of the KONVENS Conference*, 104–112. https://hildok.bsz-bw.de/files/283/03_00.pdf

Bentivogli L, Bisazza A, Cettolo M, Federico M (2016) Neural versus Phrase-Based machine translation quality: A case study. ArXiv:1608.04631. http://arxiv.org/abs/1608.04631

Bikku T, Jarugula J, Kongala L, Tummala ND, Donthiboina NV (2023), June Exploring the effectiveness of BERT for sentiment analysis on large-scale social media data. In *2023 3rd International Conference on Intelligent Technologies (CONIT)* (pp. 1–4). IEEE

Bojanowski P, Grave E, Joulin A, Mikolov T (2017) Enriching word vectors with subword information. Trans Assoc Comput Linguistics 5:135–146

Bojar O, Buck C, Federmann C, Haddow B, Koehn P, Leveling J, Monz C, Pecina P, Post M, Saint-Amand H (2014) Findings of the 2014 workshop on statistical machine translation. *Proceedings of the Ninth Workshop on Statistical Machine Translation*, 12–58. https://aclanthology.org/W14-3302.pdf

Boudjellal N, Zhang H, Khan A, Ahmad A, Naseem R, Shang J, Dai L (2021) ABioNER: A BERT-based model for Arabic biomedical named-entity recognition. Complexity 2021:1–6

Branco A, Parada D, Silva M, Mendonça F, Mostafa SS, Morgado-Dias F (2024) Sentiment analysis in Portuguese restaurant reviews: application of transformer models in edge computing. Electronics 13(3):589

Bridges RA, Jones CL, Iannacone MD, Testa KM, Goodall JR (2014) *Automatic Labeling for Entity Extraction in Cyber Security* (arXiv:1308.4941). http://arxiv.org/abs/1308.4941

Cai D, Zhao H (2016) *Neural Word Segmentation Learning for Chinese* (arXiv:1606.04300). arXiv. http://arxiv.org/abs/1606.04300

Cai L et al (2020) A hybrid BERT model that incorporates label semantics via adjustive attention for multi-label text classification. Ieee Access 8:152183–152192

Catelli R, Fujita H, De Pietro G, Esposito M (2022) Deceptive reviews and sentiment Polarity: effective link by exploiting BERT. Expert Syst Appl 209:118290

Chen Y-P, Chen Y-Y, Lin J-J, Huang C-H, Lai F (2020) Modified bidirectional encoder representations from transformers extractive summarization model for hospital information systems based on character-level tokens (AlphaBERT): Development and performance evaluation. JMIR Med Inf 8(4):e17787

Chen X, He B, Hui K, Sun L, Sun Y (2023) Dealing with textual noise for robust and effective BERT reranking. Inf Process Manag 60(1):103135

Chinedu EQ, Asogwa EC, Sunday BT, Onyeizu NM, Obulezi JO (2023) Unraveling emotions: contemporary approaches in sentiment analysis. J Sen Net Data Comm 3(1):223–230

Chinnalagu A, Durairaj AK (2022), December Comparative analysis of BERT-base transformers and deep learning sentiment prediction models. In 2022 11th International Conference on System Modeling & Advancement in Research Trends (SMART) (pp. 874–879). IEEE

Cho H, Lee H (2019) Biomedical named entity recognition using deep neural networks with contextual information. BMC Bioinform 20(1):735. https://doi.org/10.1186/s12859-019-3321-4

Chu Y, Xu J, Zhou X, Yang Q, Zhang S, Yan Z, Zhou J (2023) Qwen-audio: advancing universal audio Understanding via unified large-scale audio-language models. Preprint arXiv:2311.07919.

Chung J, Gulcehre C, Cho K, Bengio Y (2014) Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*

Clark K, Luong MT, Le QV, Manning CD (2020) Electra: Pre-training text encoders as discriminators rather than generators. *arXiv:2003.10555*

Clinchant S, Jung KW, Nikoulina V (2019) *On the use of BERT for Neural Machine Translation* (arXiv:1909.12744). arXiv. http://arxiv.org/abs/1909.12744

Collier N, Kim J-D (2004) Introduction to the bio-entity recognition task at JNLPBA. *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications (NLPBA/BioNLP)*, 73–78. https://aclanthology.org/W04-1213.pdf

Cui Y, Liang M (2024) Automated Scoring of Translations with BERT Models: Chinese and English Language Case Study. *Applied Sciences*

Dai T, Zhao J, Li D, Tian S, Zhao X, Pan S (2023) Heterogeneous deep graph convolutional network with citation relational BERT for COVID-19 inline citation recommendation. Expert Syst Appl 213:118841

Darji H, Mitrović J, Granitzer M (2023) German BERT Model for Legal Named Entity Recognition. *International Conference on Agents and Artificial Intelligence*

Daud A, Khan W, Che D (2017) Urdu language processing: a survey. Artif Intell Rev 47(3):279–311. https://doi.org/10.1007/s10462-016-9482-x

de Andrade C, Belém FM, Cunha W, França C, Viegas F, Rocha LC, Gonçalves MA (2023) On the class separability of contextual embeddings representations - or the classifier does not matter when the (text) representation is so good! Inf Process Manag 60:103336

de Paiva BBM, Pereira PD, de Andrade CMV, Gomes VMR, Souza-Silva MVR, Martins KPMP, Marcolino MS (2023) Potential and limitations of machine meta-learning (ensemble) methods for predicting COVID-19 mortality in a large inhospital Brazilian dataset. Sci Rep 13(1):3463

de Vries W, van Cranenburgh A, Bisazza A, Caselli T, van Noord G, Nissim M (2019) BERTje: A Dutch BERT model. ArXiv. ArXiv:1912.09582. http://arxiv.org/abs/1912.09582

Deepa MD (2021) Bidirectional encoder representations from Transformers (BERT) Language model for sentiment analysis task. Turkish J Comput Math Educ (TURCOMAT) 12(7):1708–1721

Deng L, Yin T, Li Z, Ge Q (2023) Sentiment analysis of comment data based on BERT-ETextCNN-ELSTM. Electronics 12(13):2910

Deode M, Patel S, Singh T (2023) Creating multilingual sentence BERT models using synthetic corpora for Indian languages. IEEE Trans Nat Lang Process 10(1):89–101

Devlin J, Chang M-W, Lee K, Toutanova K (2019) BERT: Pre-training of deep bidirectional Transformers for Language Understanding. ArXiv. ArXiv:1810.04805. http://arxiv.org/abs/1810.04805

Djeffal N, Kheddar H, Addou D, Mazari AC, Himeur Y (2023) Automatic Speech Recognition with BERT and CTC Transformers: A Review. *2023 2nd International Conference on Electronics, Energy and Measurement (IC2EM), 1*, 1–8

Dodge S, Gao S, Tomko M, Weibel R (2020) Progress in computational movement analysis – towards movement data science. Int J Geogr Inf Sci 34(12):2395–2400. https://doi.org/10.1080/13658816.2020.1784425

Doğan RI, Leaman R, Lu Z (2014) NCBI disease corpus: A resource for disease name recognition and concept normalization. J Biomed Inform 47:1–10

Donabauer G, Kruschwitz U, Corney D (2021) Making sense of subtitles: sentence boundary detection and speaker change detection in unpunctuated texts. Companion Proc Web Conf 2021 357–362. https://doi.org/10.1145/3442442.3451894

Doostmohammadi E, Nassajian M, Rahimi A (2020) Persian Ezafe recognition using Transformers and its role in Part-Of-Speech tagging. ArXiv. ArXiv:2009.09474. http://arxiv.org/abs/2009.09474

Du J, Huang Y, Moilanen K (2019) AIG Investments. AI at the FinSBD task: Sentence boundary detection through sequence labelling and BERT fine-tuning. *Proceedings of the First Workshop on Financial Technology and Natural Language Processing*, 81–87. https://aclanthology.org/W19-5513.pdf

Duraisamy P, Duraisamy M, Periyanayaki M, Natarajan Y (2023) Predicting Disaster Tweets using Enhanced BERT Model. *2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS)*, 1745–1749

Errami M, Ouassil MA, Rachidi R, Cherradi B, Hamida S, Raihani A (2023), May Investigating the Performance of BERT Model for Sentiment Analysis on Moroccan News Comments. In *2023 3rd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)* (pp. 1–8). IEEE

Ettinger A (2020) What BERT is not: lessons from a new suite of psycholinguistic diagnostics for Language models. Trans Assoc Comput Linguistics 8:34–48

Fang L, Chen Q, Wei C, Lu Z, Wang K (2023) Bioformer: an efficient transformer language model for biomedical text mining

Foo S, Li H (2004) Chinese word segmentation and its effect on information retrieval. Inf Process Manag 40(1):161–190

Freitas C, Carvalho P, Gonçalo Oliveira H, Mota C, Santos D (2010) Second HAREM: Advancing the state of the art of named entity recognition in Portuguese. *Quot; In Nicoletta Calzolari; Khalid Choukri; Bente Maegaard; Joseph Mariani; Jan Odijk; Stelios Piperidis; Mike Rosner; Daniel Tapias (Ed) Proceedings of the International Conference on Language Resources and Evaluation (LREC 2010)(Valletta 17–23 May de 2010) European Language Resources Association*. https://comum.rcaap.pt/bitstream/10400.26/20499/2/FreitasetalLREC2010.pdf

Gao J, Xu H, Shi H, Ren X, Philip LH, Liang X, Li Z (2022), June Autobert-zero: Evolving bert backbone from scratch. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, No. 10, pp. 10663–10671)

Ghojogh B, Ghodsi A, Karray F, Crowley M (2020) Locally linear embedding and its variants: Tutorial and survey. *arXiv preprint arXiv:2011.10925*

Gokaslan A, Cohen V, Pavlick E, Tellex S (2019) *Openwebtext corpus*

Gong C, Tang J, Zhou S, Hao Z, Wang J (2019) Chinese named entity recognition with bert. DEStech Trans Comput Sci Eng 12

Guarasci R, Silvestri S, De Pietro G, Fujita H, Esposito M (2022) BERT syntactic transfer: A computational experiment on Italian, French, and english languages. Comput Speech Lang 71:101261

Guo Z, Nguyen ML (2020) Document-Level Neural Machine Translation Using BERT as Context Encoder. *AACL*

Gupta R (2024a) Bidirectional encoders to state-of-the-art: a review of BERT and its transformative impact on natural language processing. *Информатика. Экономика. Управление - Informatics. Economics. Management*

Gupta S (2024b) The impact of BERT on natural Language processing: A review and future directions. Artif Intell Rev 47(3):309–325. https://doi.org/10.1007/s10462-023-10123-4

Gupta P, Gandhi S, Chakravarthi B (2021) Leveraging Transfer learning techniques- BERT, RoBERTa, ALBERT and DistilBERT for Fake Review Detection. *Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation*

Habibi M, Weber L, Neves M, Wiegandt DL, Leser U (2017) Deep learning with word embeddings improves biomedical named entity recognition. Bioinformatics 33(14):i37–i48

Harte J, Zorgdrager W, Louridas P, Katsifodimos A, Jannach D, Fragkoulis M (2023), September Leveraging large language models for sequential recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems* (pp. 1096–1102)

Hayashibe Y, Mitsuzawa K (2020) Sentence Boundary Detection on Line Breaks in Japanese. *Proceedings of the Sixth Workshop on Noisy User-Generated Text (W-NUT 2020)*, 71–75. https://aclanthology.org/2020.wnut-1.10/

He P, Liu X, Gao J, Chen W (2021) *DeBERTa: Decoding-enhanced BERT with Disentangled Attention* (arXiv:2006.03654). arXiv. http://arxiv.org/abs/2006.03654

He C, Zhu X, Le Y, Liu Y, Yin J (2024) SEBERTNets: Sequence Enhanced BERT Networks for Event Entity Extraction Tasks Oriented to the Finance Field. *arXiv preprint arXiv:2401.11408*

Hebbar S, N BAR, Supriya MS (2023) M., G, N.V., & L, S. Named Entity Recognition Using BERT Model for Kannada Language. *2023 International Conference on Recent Advances in Information Technology for Sustainable Development (ICRAIS)*, 212–216

Hiraoka T, Okazaki N (2024) Knowledge of Pretrained Language Models on Surface Information of Tokens. *arXiv preprint arXiv:2402.09808*

Hoang M, Bihorac OA, Rouces J (2019) Aspect-based sentiment analysis using bert. In *Proceedings of the 22nd nordic conference on computational linguistics* (pp. 187–196)

Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780. http://arxiv.org/abs/1909.00100

Hu J, Hu R, Wang Z, Li D, Wu J, Ren L, Wang M (2023), October Collaborative Fraud Detection: How Collaboration Impacts Fraud Detection. In *Proceedings of the 31st ACM International Conference on Multimedia* (pp. 8891–8899)

Huang K, Altosaar J, Ranganath R (2020) *ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission* (arXiv:1904.05342). arXiv. http://arxiv.org/abs/1904.05342

Iandola FN, Shaw AE, Krishna R, Keutzer KW (2020) *SqueezeBERT: What can computer vision teach NLP about efficient neural networks?* (arXiv:2006.11316). arXiv. http://arxiv.org/abs/2006.11316

Imamura K, Sumita E (2019) Recycling a pre-trained BERT encoder for neural machine translation. *Proceedings of the 3rd Workshop on Neural Generation and Translation*, 23–31. https://aclanthology.org/D19-5603/

Jacovi A, Goldberg Y (2020) *Towards Faithfully Interpretable NLP Systems: How should we define and evaluate faithfulness?* (arXiv:2004.03685). arXiv. http://arxiv.org/abs/2004.03685

Jeong SW, Kim CG, Whangbo TK (2023) Question Answering System for Healthcare Information based on BERT and GPT. *2023 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON)*, 348–352

Jha K, Karmakar S, Saha S (2023) Graph-BERT and Language model-based framework for protein–protein interaction identification. Sci Rep 13(1):5663

Joshi M, Chen D, Liu Y, Weld DS, Zettlemoyer L, Levy O (2020) Spanbert: improving pre-training by representing and predicting spans. Trans Association Comput Linguistics 8:64–77

Joshy A, Sundar S (2022), December Analyzing the performance of sentiment analysis using Bert, Distilbert, and Roberta. In *2022 IEEE international power and renewable energy conference (IPRECON)* (pp. 1–6). IEEE

Kaliyar RK, Goswami A, Narang P, Sinha S (2020) FNDNet–a deep convolutional neural network for fake news detection. Cogn Syst Res 61:32–44

Karimi A, Rossi L, Prati A (2020) Improving bert performance for aspect-based sentiment analysis. *arXiv preprint arXiv:2010.11731*

Khan W, Daud A, Nasir JA, Amjad T, Arafat S, Aljohani N, Alotaibi FS (2019) Urdu part of speech tagging using conditional random fields. Lang Resour Evaluation 53(3):331–362. https://doi.org/10.1007/s10579-018-9439-6

Khan W, Daud A, Khan K, Muhammad S, Haq R (2023) Exploring the frontiers of deep learning and natural Language processing: A comprehensive overview of key challenges and emerging trends. Nat Lang Process J 4:100026. https://doi.org/10.1016/j.nlp.2023.100026

Kim K-M, Heo M-O, Choi S-H, Zhang B-T (2017) *DeepStory: Video Story QA by Deep Embedded Memory Networks* (arXiv:1707.00836). arXiv. http://arxiv.org/abs/1707.00836

Kora R, Mohammed A (2023) A Comprehensive Review on Transformers Models For Text Classification. *2023 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC)*, 1–7

Koroteev MV (2021) *BERT: A Review of Applications in Natural Language Processing and Understanding* (arXiv:2103.11943). arXiv. http://arxiv.org/abs/2103.11943

Koshkin R, Sudoh K, Nakamura S (2024) TransLLaMa: LLM-based Simultaneous Translation System. *ArXiv, abs/2402.04636*

Kovaleva O, Romanov A, Rogers A, Rumshisky A (2019) *Revealing the Dark Secrets of BERT* (arXiv:1908.08593). arXiv. http://arxiv.org/abs/1908.08593

Krallinger M, Rabal O, Leitner F, Vazquez M, Salgado D, Lu Z, Leaman R, Lu Y, Ji D, Lowe DM, Sayle RA, Batista-Navarro RT, Rak R, Huber T, Rocktäschel T, Matos S, Campos D, Tang B, Xu H, Valencia A (2015) The CHEMDNER corpus of chemicals and drugs and its annotation principles. J Cheminform 7:1. https://doi.org/10.1186/1758-2946-7-S1-S2

Labusch K, Zu S, Kulturbesitz B, Neudecker C, Zellhöfer D (2019) *BERT for Named Entity Recognition in Contemporary and Historical German*

Lakshmidevi N, Swain SK, Vamsikrishna M (2023) September A Hybrid Enhancing Aspect-Based Sentiment Analysis with BERT for Aspect Extraction and Diverse ML Classifiers. In *2023 International Conference on Network, Multimedia and Information Technology (NMITCON)* (pp. 01–08). IEEE

Lample G, Conneau A (2019) Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*

Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R (2020) *ALBERT: A Lite BERT for Self-supervised Learning of Language Representations* (arXiv:1909.11942). arXiv. http://arxiv.org/abs/1909.11942

Le H, Vial L, Frej J, Segonne V, Coavoux M, Lecouteux B, Allauzen A, Crabbé B, Besacier L, Schwab D (2020) *FlauBERT: Unsupervised Language Model Pre-training for French* (arXiv:1912.05372). arXiv. http://arxiv.org/abs/1912.05372

Lee JS, Hsiang J (2019) Patentbert: Patent classification with fine-tuning a pre-trained bert model. *arXiv preprint arXiv:1906.02124*

Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J (2020) BioBERT: A pre-trained biomedical Language representation model for biomedical text mining. Bioinformatics 36(4):1234–1240

Lei J, Yu L, Bansal M, Berg TL (2019) *TVQA: Localized, Compositional Video Question Answering* (arXiv:1809.01696). arXiv. http://arxiv.org/abs/1809.01696

Li X, Chen L (2023) Fake Review Detection Using Deep Neural Networks with Multimodal Feature Fusion Method. *2023 IEEE 29th International Conference on Parallel and Distributed Systems (ICPADS)*, 2869–2872

Li J, Zhang Y (2024) The Death of Feature Engineering? BERT with Linguistic Features on SQuAD 2.0. *ArXiv, abs/2404.03184*

Li J, Sun Y, Johnson RJ, Sciaky D, Wei C-H, Leaman R, Davis AP, Mattingly CJ, Wiegers TC, Lu Z (2016) BioCreative V CDR task corpus: A resource for chemical disease relation extraction. *Database*, *2016*. https://academic.oup.com/database/article/doi/10.1093/database/baw068/2630414?ref=https%3A%2F%2Fgithubhelp.com&login=true

Li Y, Anastasopoulos A, Black AW (2020a) *Towards Minimal Supervision BERT-based Grammar Error Correction* (arXiv:2001.03521). arXiv. http://arxiv.org/abs/2001.03521

Li Y, Anastasopoulos A, Black AW (2020b) *Towards Minimal Supervision BERT-based Grammar Error Correction* (arXiv:2001.03521). arXiv. http://arxiv.org/abs/2001.03521

Li L, Ma R, Guo Q, Xue X, Qiu X (2020c) Bert-attack: Adversarial attack against bert using bert. *arXiv preprint arXiv:2004.09984*

Li Y, Wang S, Lin C, Guerin F, Barrault L (2023) FrameBERT: Conceptual Metaphor Detection with Frame Embedding Learning. *ArXiv, abs/2302.04834*

Liang W, Liang Y (2024) DrBERT: Unveiling the Potential of Masked Language Modeling Decoder in BERT pretraining. arXiv preprint arXiv:2401.15861

Liang M, Shi Y (2023) Named Entity Recognition Method Based on BERT-whitening and Dynamic Fusion Model. *2023 5th International Conference on Natural Language Processing (ICNLP)*, 191–197

Licari V (2022) ITALIAN-LEGAL-BERT: Pre-training on Italian civil law corpora. J Comput Law 35(2):211–225. https://doi.org/10.1093/jcl/ztac024

Lim K, Park J (2020) Part-of-speech tagging using multiview learning. IEEE Access 8:195184–195196

Liu K, Wang X, Wei N, Song Z, Li D (2019a) Accurate quantification and transport Estimation of suspended atmospheric microplastics in megacities: implications for human health. Environ Int 132:105127

Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019b) *RoBERTa: A Robustly Optimized BERT Pretraining Approach* (arXiv:1907.11692). arXiv. http://arxiv.org/abs/1907.11692

Liu Z, Huang D, Huang K, Li Z, Zhao J (2021) Finbert: A pre-trained financial language representation model for financial text mining. In *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence* (pp. 4513–4519)

Liu W, Lin S, Gao B, Huang K, Liu W, Huang Z, Feng J, Chen X, Huang F (2022) BERT-POS: Sentiment Analysis of MOOC Reviews Based on BERT with Part-of-Speech Information. *International Conference on Artificial Intelligence in Education*

Lu J, Zhan X, Liu G, Zhan X, Deng X (2023a) BSTC: A Fake Review Detection Model Based on a Pre-Trained Language Model and Convolutional Neural Network. *Electronics*

Lu X, Liu W, Jiang S, Liu C (2023b), March Multilingual BERT cross-lingual transferability with pre-trained representations on Tangut: A survey. In *2023 5th International Conference on Natural Language Processing (ICNLP)* (pp. 229–234). IEEE

Ma B, Chen L (2023) Named entity recognition in medical field based on BERT model. *Other Conferences*

Ma J, Liu J, Lin Q, Wu B, Wang Y, You Y (2021) Multitask learning for visual question answering. IEEE Trans Neural Networks Learn Syst 34(3):1380–1394

Ma Z, Yan K, Wang H (2023) BERT-based Question Answering using Knowledge Graph Embeddings in Nuclear Power Domain. *2023 26th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, 267–272

Malik MK (2018) Urdu named entity recognition and classification system using artificial neural network. ACM Trans Asian Low-Resource Lang Inform Process 17(1):1–13. https://doi.org/10.1145/3129290

Malik MK, Sarwar SM (2016) Named entity recognition system for postpositional languages: Urdu as a case study. Int J Adv Comput Sci Appl 7:10

Malmsten M, Börjeson L, Haffenden C (2020) *Playing with Words at the National Library of Sweden—Making a Swedish BERT* (arXiv:2007.01658). arXiv. http://arxiv.org/abs/2007.01658

Martin L, Muller B, Suárez PJO, Dupont Y, Romary L, de la Clergerie ÉV, Seddah D, Sagot B (2020) CamemBERT: A Tasty French Language Model. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7203–7219. https://doi.org/10.18653/v1/2020.acl-main.645

Mehta H, Kumar Bharti S, Doshi N (2024) Comparative Analysis of Part of Speech(POS) Tagger for Gujarati Language using Deep Learning and Pre-Trained LLM. *2024 3rd International Conference for Innovation in Technology (INOCON)*, 1–3

Mewada A, Dewang RK, Goldar P, Maurya SK (2023) SentiBERT: A Novel Approach for Fake Review Detection Incorporating Sentiment Features with Contextual Features. *Proceedings of the 2023 Fifteenth International Conference on Contemporary Computing*

Mikolov T, Chen K, Corrado G, Dean J (2013) *Efficient Estimation of Word Representations in Vector Space* (arXiv:1301.3781). arXiv. http://arxiv.org/abs/1301.3781

Ming NW, Wang Z, Liu C, Goh RSM, Luo T (2022) Ma-bert: Towards matrix arithmetic-only bert inference by eliminating complex non-linear functions. In *The Eleventh International Conference on Learning Representations*

Mir AQ, Khan FY, Chishti MA (2023) Online Fake Review Detection Using Supervised Machine Learning And BERT Model. *ArXiv, abs/2301.03225*

Mohtashami A, Jaggi M (2024) Random-access infinite context length for transformers. Adv Neural Inf Process Syst 36:52

Moon T, Awasthy P, Ni J, Florian R (2019) Towards lingua franca named entity recognition with bert. *arXiv preprint arXiv:1912.01389*

Mozannar H, Hajal KE, Maamary E, Hajj H (2019) Neural Arabic question answering. *arXiv preprint arXiv:1906.05394*

Muffo M, Bertino E (2023) Bertino: An italian distilbert model. *arXiv preprint arXiv:2303.18121*

Nadeau D, Sekine S (2007) A survey of named entity recognition and classification. Lingvisticae Investigationes 30(1):3–26. https://doi.org/10.1075/li.30.1.03nad

Nagel S (2016) Cc-news. *URL: Http://Web.Archive.Org/Save/CommoncrawlOrg/2016/10/Newsdatasetavailable*.

Napoles C, Sakaguchi K, Tetreault J (2017) *JFLEG: A Fluency Corpus and Benchmark for Grammatical Error Correction* (arXiv:1702.04066). arXiv. http://arxiv.org/abs/1702.04066

Naseem U, Khushi M, Reddy V, Rajendran S, Razzak I, Kim J (2021) Bioalbert: A simple and effective pre-trained language model for biomedical named entity recognition. *2021 International Joint Conference on Neural Networks (IJCNN)*, 1–7. https://ieeexplore.ieee.org/abstract/document/9533884/

Nedumpozhimana V, Klubicka F, Kelleher JD (2022) Shapley Idioms: Analysing BERT Sentence Embeddings for General Idiom Token Identification. Front Artif Intell 5:e205

Ng M, Fleming T, Robinson M, Thomson B, Graetz N, Margono C, Mullany EC, Biryukov S, Abbafati C, Abera SF (2014) Global, regional, and National prevalence of overweight and obesity in children and adults during 1980–2013: A systematic analysis for the global burden of disease study 2013. Lancet 384(9945):766–781

Nguyen T, Li Z, Spiegler V, Ieromonachou P, Lin Y (2018) Big data analytics in supply chain management: A state-of-the-art literature review. Comput Oper Res 98:254–264

Nivre J (2014) Universal dependencies for swedish. *Proceedings of the Swedish Language Technology Conference (SLTC)*, 5

Nivre J, De Marneffe M-C, Ginter F, Goldberg Y, Hajic J, Manning CD, McDonald R, Petrov S, Pyysalo S, Silveira N (2016) Universal dependencies v1: A multilingual treebank collection. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 1659–1666. https://aclanthology.org/L16-1262/

Nogueira R, Cho K (2020) *Passage Re-ranking with BERT* (arXiv:1901.04085). arXiv. http://arxiv.org/abs/1901.04085

Nozza D, Bianchi F, Hovy D (2020) What the [MASK]? Making Sense of Language-Specific BERT Models. *ArXiv, abs/2003.02912*

Otieno DO, Namin AS, Jones KS (2023) The Application of the BERT Transformer Model for Phishing Email Classification. *2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*, 1303–1310

Pappagari R, Zelasko P, Villalba J, Carmiel Y, Dehak N (2019) December). Hierarchical Transformers for long document classification. 2019 IEEE automatic speech recognition and Understanding workshop (ASRU). IEEE, pp 838–844

Peng N, Dredze M (2017) *Improving Named Entity Recognition for Chinese Social Media with Word Segmentation Representation Learning* (arXiv:1603.00786). arXiv. http://arxiv.org/abs/1603.00786

Peng Z, Zhao Y (2023) Triple-Compressed BERT for Efficient Implementation on NLP Tasks. *2023 3rd International Conference on Electronic Information Engineering and Computer Science (EIECS)*, 1162–1165.T

Pennington J, Socher R, Manning C (2014) Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://doi.org/10.3115/v1/D14-1162

Peters ME, Ruder S, Smith NA (2019) *To Tune or Not to Tune? Adapting Pretrained Representations to Diverse Tasks* (arXiv:1903.05987). arXiv. http://arxiv.org/abs/1903.05987

Pfeiffer J, Houlsby N, Gurevych I (2020) *MAD-X: An adapter-based framework for efficient cross-lingual transfer. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 6543–6551

Poostchi H, Borzeshi EZ, Piccardi M (2018) BiLSTM-CRF for Persian named-entity recognition Arman-PersoNERCorpus: The first entity-annotated Persian dataset. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. https://aclanthology.org/L18-1701.pdf

Pourkamali N, Sharifi SE (2024) Machine Translation with Large Language Models: Prompt Engineering for Persian, English, and Russian Directions. *ArXiv, abs/2401.08429*

Pudasaini S, Shakya S (2023) Question Answering on Biomedical Research Papers using Transfer Learning on BERT-Base Models. *2023 7th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, 496–501

Qing Y (2024) Design and application of automatic english translation grammar error detection system based on BERT machine vision. Scalable Comput Pract Exp 25:2088–2102

Qiu X, Sun T, Xu Y, Shao Y, Dai N, Huang X (2020) Pre-trained models for natural Language processing: A survey. Sci China Technological Sci 63(10):1872–1897. https://doi.org/10.1007/s11431-020-1647-3

Qu C, Yang L, Qiu M, Croft WB, Zhang Y, Iyyer M (2019) BERT with History Answer Embedding for Conversational Question Answering. *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1133–1136. https://doi.org/10.1145/3331184.3331341

Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Liu PJ (2020) Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res 21(140):1–67

Ramaraj V, Appa Swamy MV, Prince EE, Kumar C (2024) Improving the BERT model for long text sequences in question answering domain. Int J Adv Appl Sci 8:96

Refaeli D, Hájek P (2021) Detecting Fake Online Reviews using Fine-tuned BERT. *Proceedings of the 2021 5th International Conference on E-Business and Internet*

Rehbein I, Ruppenhofer J, Schmidt T (2020) Improving sentence boundary detection for spoken language transcripts. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 7102–7111. https://aclanthology.org/2020.lrec-1.878/

Rei M (2017) *Semi-supervised Multitask Learning for Sequence Labeling* (arXiv:1704.07156). arXiv. http://arxiv.org/abs/1704.07156

Roy K, Hasan, Fuhad KM, Mohammed N, Hasan RABBYA, Nahar N, J., Rahman F (2020) Bangla Part of Speech Tagging Using Contextual Embeddings and Oversampling Techniques

Sahoo A, Chanda R, Das N, Sadhukhan B (2023), August Comparative Analysis of BERT Models for Sentiment Analysis on Twitter Data. In *2023 9th International Conference on Smart Computing and Communications (ICSCC)* (pp. 658–663). IEEE

Saidi R, Jarray F, Mansour M (2021) A BERT Based Approach for Arabic POS Tagging. *International Work-Conference on Artificial and Natural Neural Networks*

Sang EFTK, De Meulder F (2003) *Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition* (arXiv:cs/0306050). arXiv. http://arxiv.org/abs/cs/0306050

Sanh V, Debut L, Chaumond J, Wolf T (2020) *DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter* (arXiv:1910.01108). arXiv. http://arxiv.org/abs/1910.01108

Santos FC, Pacheco JM, Lenaerts T (2006) Cooperation prevails when individuals adjust their social ties. PLoS Comput Biol 2(10):e140

Sarkar S, Babar MF, Hassan MM, Hasan M, Karmaker S (2023) Exploring Challenges of Deploying BERT-based NLP Models in Resource-Constrained Embedded Devices. *ArXiv, abs/2304.11520*

Schwartz D, Toneva M, Wehbe L (2019) Inducing brain-relevant bias in natural language processing models. *Advances in Neural Information Processing Systems*, *32*. https://proceedings.neurips.cc/paper_files/paper/2019/hash/2b8501af7b64d1aaae7dd832805f0709-Abstract.html

Shahshahani MS, Mohseni M, Shakery A, Faili H (2018) *PEYMA: A Tagged Corpus for Persian Named Entities* (arXiv:1801.09936). arXiv. http://arxiv.org/abs/1801.09936

Shaik Vadla MK, Suresh MA, Viswanathan VK (2024) Enhancing product design through AI-driven sentiment analysis of Amazon reviews using BERT. Algorithms 17(2):59

Sharma R, Chen F, Fard F, Lo D (2022) May An exploratory study on code attention in BERT. In *Proceedings of the 30th IEEE/ACM International Conference on Program Comprehension* (pp. 437–448)

Sharma KV, Singh K, Sharma K, Gupta J (2023) Question summation and sentence similarity using BERT for key information extraction. Int J Res Appl Sci Eng Technol

Shavarani HS, Sarkar A (2021) Better Neural Machine Translation by Extracting Linguistic Information from BERT. *Conference of the European Chapter of the Association for Computational Linguistics*

Shih CF, Tseng YH, Yang CW, Chen PE, Chou HY, Tan LH, Hsieh SK (2021), October What confuses BERT? Linguistic Evaluation of Sentiment Analysis on Telecom Customer Opinion. In *Proceedings of the 33rd Conference on Computational Linguistics and Speech Processing (ROCLING 2021)* (pp. 271–279)

Shimanaka H, Kajiwara T, Komachi M (2019) *Machine Translation Evaluation with BERT Regressor* (arXiv:1907.12679). arXiv. http://arxiv.org/abs/1907.12679

Smith L, Tanabe LK, Ando RJN, Kuo C-J, Chung I-F, Hsu C-N, Lin Y-S, Klinger R, Friedrich CM, Ganchev K, Torii M, Liu H, Haddow B, Struble CA, Povinelli RJ, Vlachos A, Baumgartner WA, Hunter L, Carpenter B, Wilbur WJ (2008) Overview of BioCreative II gene mention recognition. *Genome Biology*, *9*(S2), S2. https://doi.org/10.1186/gb-2008-9-s2-s2

Smith A, Bohnet B, de Lhoneux M, Nivre J, Shao Y, Stymne S (2018) *82 Treebanks, 34 Models: Universal Dependency Parsing with Multi-Treebank Models* (arXiv:1809.02237). arXiv. http://arxiv.org/abs/1809.02237

Snaebjarnarson G, Jónsson H, Björnsson H (2023) Leveraging closely related languages for improving NLP tasks in Faroese. J Artif Intell Res 73:987–1003

Souza F, Nogueira R, Lotufo R (2020) *Portuguese Named Entity Recognition using BERT-CRF* (arXiv:1909.10649). arXiv. http://arxiv.org/abs/1909.10649

Srivastava S, Paul B, Gupta D (2023) Study of word embeddings for enhanced cyber security named entity recognition. Procedia Computer Science

Strubell E, Ganesh A, McCallum A (2019) *Energy and Policy Considerations for Deep Learning in NLP* (arXiv:1906.02243). arXiv. http://arxiv.org/abs/1906.02243

Sun S, Cheng Y, Gan Z, Liu J (2019) Patient knowledge distillation for bert model compression. *arXiv preprint arXiv:1908.09355*

Sun Z, Yu H, Song X, Liu R, Yang Y, Zhou D (2020) *MobileBERT: A Compact Task-Agnostic BERT for Resource-Limited Devices* (arXiv:2004.02984). arXiv. http://arxiv.org/abs/2004.02984

Sun C, Yang Z, Wang L, Zhang Y, Lin H, Wang J (2021) Biomedical named entity recognition using BERT in the machine reading comprehension framework. J Biomed Inform 118:103799

Sur C (2020) RBN: enhancement in Language attribute prediction using global representation of natural Language transfer learning technology like Google BERT. SN Appl Sci 2(1):22

Taher E, Hoseini SA, Shamsfard M (2020) *Beheshti-NER: Persian Named Entity Recognition Using BERT* (arXiv:2003.08875). arXiv. http://arxiv.org/abs/2003.08875

Tamburini F (2020) How BERTology Changed the State-of-the-Art also for Italian NLP. Proceedings of the Seventh Italian Conference on Computational Linguistics CLiC-it 2020

Tanaka H, Shinnou H, Cao R, Bai J, Ma W (2019) Document Classification by Word Embeddings of BERT. *International Conference of the Pacific Association for Computaitonal Linguistics*

Tikayat Ray A, Pinon-Fischer OJ, Mavris DN, White RT, Cole BF (2023) aeroBERT-NER: Named-Entity Recognition for Aerospace Requirements Engineering using BERT. *AIAA SCITECH 2023 Forum*

Trinh TH, Le QV (2019) A Simple Method for Commonsense Reasoning (arXiv:1806.02847). arXiv. http://arxiv.org/abs/1806.02847

Tripty Z, Nafis M, Chowdhury A, Hossain J, Ahsan S, Das A, Hoque MM (2024), March CUETSentimentSillies@ DravidianLangTech-EACL2024: Transformer-based Approach for Sentiment Analysis in Tamil and Tulu Code-Mixed Texts. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages* (pp. 234–239)

Tsai H, Riesa J, Johnson M, Arivazhagan N, Li X, Archer A (2019) *Small and Practical BERT Models for Sequence Labeling* (arXiv:1909.00100).

Tuli S, Dedhia B, Tuli S, Jha NK (2023) FlexiBERT: are current transformer architectures too homogeneous and rigid? J Artif Intell Res 77:39–70. http://www2.lingfil.uu.se/SLTC2014/abstracts/sltc2014_submission_7.pdf

Turton J, Vinson DP, Smith R (2020) Deriving Contextualised Semantic Features from BERT (and Other Transformer Model) Embeddings. *ArXiv, abs/2012.15353*

Ulčar M, Robnik-Šikonja M (2020) FinEst BERT and CroSloEngual BERT: Less Is More in Multilingual Models. In P. Sojka, I. Kopeček, K. Pala, & A. Horák (Eds.), *Text, Speech, and Dialogue* (Vol. 12284, pp. 104–111). Springer International Publishing. https://doi.org/10.1007/978-3-030-58323-1_11

Van Aken B, Winter B, Löser A, Gers FA (2019), November How does bert answer questions? a layer-wise analysis of transformer representations. In *Proceedings of the 28th ACM international conference on information and knowledge management* (pp. 1823–1832)

Van Noord G, Bouma G, Van Eynde F, De Kok D, Van der Linde J, Schuurman I, Sang ETK, Vandeghinste V (2013) Large scale syntactic annotation of written Dutch: Lassy. *Essential Speech and Language Technology for Dutch: Results by the STEVIN Programme*, 147–164

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser \Lukasz, Polosukhin I (2017) Attention is all you need. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper/7181-attention-is-all

Verma N, Elbayad M (2024) Merging text transformer models from different initializations. *arXiv preprint arXiv:2403.00986*

Vernikos G, Popescu-Belis A (2024) Don't Rank, Combine! Combining Machine Translation Hypotheses Using Quality Estimation. *ArXiv, abs/2401.06688*

Wang Y (2020) Extending multilingual BERT to low-resource languages. J Multiling Multicultural Dev 41(5):537–551. https://doi.org/10.1080/01434632.2020.1731144

Wang P, Gu J (2023) Named entity recognition of electronic medical records based on BERT-BiLSTM-Biaffine Model. J Phys: Conf Ser 2560

Wang A, Pruksachatkun Y, Nangia N, Singh A, Michael J, Hill F, Levy O, Bowman S (2019) Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in Neural Information Processing Systems*, *32*. https://proceedings.neurips.cc/paper_files/paper/2019/hash/4496bf24afe7fab6f046bf4923da8de6-Abstract.html

Wang H, Kurosawa M, Katsumata S, Komachi M (2020) *Chinese Grammatical Correction Using BERT-based Pre-trained Model* (arXiv:2011.02093). arXiv. http://arxiv.org/abs/2011.02093

Wang Y, Cui L, Zhang Y (2021) Improving Skip-Gram embeddings using BERT. IEEE/ACM Trans Audio Speech Lang Process 29:1318–1328

Wang H, Li J, Wu H, Hovy E, Sun Y (2023) Pre-Trained Language models and their applications. Engineering 25:51–65. https://doi.org/10.1016/j.eng.2022.04.024

Weng M, Zhang W (2023) Named Entity Recognition Based on BERT-BiLSTM-SPAN in Low Resource Scenarios. *2023 15th International Conference on Computer Research and Development (ICCRD)*, 32–37

Weng R, Yu H, Huang S, Cheng S, Luo W (2020) Acquiring knowledge from pre-trained model to neural machine translation. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(05), 9266–9273. https://ojs.aaai.org/index.php/AAAI/article/view/6465

Wu Y, Dredze M (2020) *Analyzing the performance decline of multilingual BERT on low-resource languages*. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1120–1130

Wu Y, Schuster M, Chen Z, Le QV, Norouzi M, Macherey W, Krikun M, Cao Y, Gao Q, Macherey K, Klingner J, Shah A, Johnson M, Liu X, Kaiser Ł, Gouws S, Kato Y, Kudo T, Kazawa H, Dean J (2016) *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation* (arXiv:1609.08144). arXiv. http://arxiv.org/abs/1609.08144

Wu Z, Chen Y, Kao B, Liu Q (2020) Perturbed masking: Parameter-free probing for analyzing and interpreting BERT. *arXiv preprint arXiv:2004.14786*

Wu M, Vu T, Qu L, Foster G, Haffari G (2024) Adapting Large Language Models for Document-Level Machine Translation. *ArXiv, abs/2401.06468*

Xia Y, Li X, Zhou Z (2021) MetaXL: A meta-learning framework for transforming representations in low-resource languages. IEEE Trans Neural Networks Learn Syst 32(5):2045–2058

Xu J, Deng Y, Guo Y, Ney H (2007) Domain dependent statistical machine translation. *Proceedings of Machine Translation Summit XI: Papers*. https://aclanthology.org/2007.mtsummit-papers.68.pdf

Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov RR, Le QV (2019) Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems*, *32*. https://proceedings.neurips.cc/paper/2019/hash/dc6a7e655d7e5840e66733e9ee67cc69-Abstract.html

Yang Z, Garcia N, Chu C, Otani M, Nakashima Y, Takemura H (2020) Bert representations for video question answering. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1556–1565. http://openaccess.thecvf.com/content_WACV_2020/html/Yang_BERT_representations_for_Video_Question_Answering_WACV_2020_paper.html

Yang B, Luo X, Sun K, Luo MY (2023) August Recent progress on text summarisation based on bert and gpt. In *International Conference on Knowledge Science, Engineering and Management* (pp. 225–241). Cham: Springer Nature Switzerland

Yannakoudakis H, Briscoe T, Medlock B (2011) A new dataset and method for automatically grading ESOL texts. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 180–189. https://aclanthology.org/P11-1019.pdf

Yeshambel T, Mothe J, Assabie Y (2023) Learned text representation for amharic information retrieval and natural Language processing. Inf 14:195

Yin H, Liu X, Wu Y, Arini HM, Mohawesh R (2023), October A BERT-Based Semantic Enhanced Model for COVID-19 Fake News Detection. In *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data* (pp. 1–15). Singapore: Springer Nature Singapore

Yu W, Wu L, Deng Y, Mahindru R, Zeng Q, Guven S, Jiang M (2020) A technical question answering system with transfer learning. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 92–99. https://aclanthology.org/2020.emnlp-demos.13/

Zabeen S, Hasan A, Islam MF, Hossain MS, Rasel AA (2023) Robust Fake Review Detection Using Uncertainty-Aware LSTM and BERT. *2023 IEEE 15th International Conference on Computational Intelligence and Communication Networks (CICN)*, 786–791

Zeman D, Hajic J, Popel M, Potthast M, Straka M, Ginter F, Nivre J, Petrov S (2018) CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies. *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 1–21. https://aclanthology.org/K18-2001/

Zhang Z, Wu S, Jiang D, Chen G (2021) BERT-JAM: maximizing the utilization of BERT for neural machine translation. Neurocomputing 460:84–94

Zhang X, Malkov Y, Florez O, Park S, McWilliams B, Han J, El-Kishky A (2023), August Twhin-bert: A socially-enriched pre-trained language model for multilingual tweet representations at twitter. In *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining* (pp. 5597–5607)

Zhou S, Liu J, Zhong X, Zhao W (2021) Named entity recognition using BERT with whole world masking in cybersecurity domain. 2021 IEEE 6th International Conference on Big Data Analytics (ICBDA), 316–320. https://ieeexplore.ieee.org/abstract/document/9403180/

Zhou X, Zhang S, Agarwal M, Akroyd J, Mosbach S, Kraft M (2023) Marie and BERT—A knowledge graph embedding based question answering system for chemistry. ACS Omega 8:33039–33057

Zhu Y, Kiros R, Zemel R, Salakhutdinov R, Urtasun R, Torralba A, Fidler S (2015) Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *Proceedings of the IEEE International Conference on Computer Vision*, 19–27. https://www.cv-foundation.org/openaccess/content_iccv_2015/html/Zhu_Aligning_Books_and_ICCV_2015_paper.html

Zhu J, Xia Y, Wu L, He D, Qin T, Zhou W, Li H, Liu T-Y (2020) Incorporating BERT into Neural Machine Translation (arXiv:2002.06823). arXiv. http://arxiv.org/abs/2002.06823

Zyout I, Zyout MA (2024) Sentiment analysis of student feedback using attention-based RNN and transformer embedding. Int J Artif Intell 13(2):2173–2184

Authors and Affiliations

Nadia Mushtaq Gardazi[1] · Ali Daud[2] · Muhammad Kamran Malik[1] · Amal Bukhari[3] · Tariq Alsahfi[3] · Bader Alshemaimri[4]

✉ Ali Daud
alimsdb@gmail.com

[1] Department of Data Science, Faculty of Computing and Information Technology, University of the Punjab, Lahore, Pakistan

[2] Faculty of Resilience, Rabdan Academy, Abu Dhabi, United Arab Emirates

[3] Department of Information Systems and Technology, Collage of Computer Science and Engineering, University of Jeddah, Jeddah, Saudi Arabia

[4] Software Engineering Department, College of Computing and Information Sciences, King Saud University, Riyadh, Saudi Arabia