A Survey on BERT and Its Applications

Sulaiman Aftan
Department of Computer Science,
Texas Tech University,
Lubbock, TX 79709, USA
https://orcid.org/0000-0003-2093-9894

Habib Shah
Department and College of Computer Science,
King Khalid University,
Abha 62529, Saudi Arabia
habibshah.uthm@gmail.com
hurrahman@kku.edu.sa

2023 20th Learning and Technology Conference (L&T) | 979-8-3503-0030-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/LT58159.2023.10092289

***Abstract*— A recently developed language representation model named Bidirectional Encoder Representation from Transformers (BERT) is based on an advanced trained deep learning approach that has achieved excellent results in many complex tasks, the same as classification, Natural Language Processing (NLP), prediction, etc. This survey paper mainly adopts the summary of BERT, its multiple types, and its latest developments and applications in various computer science and engineering fields. Furthermore, it puts forward BERT's problems and attractive future research trends in a different area with multiple datasets. From the findings, overall, the BERT and their recent types have achieved more accurate, fast, and optimal results in solving most complex problems than typical Machine and Deep Learning methods.**

Keywords—BERT, Machine Learning, Natural Language Processing model, bidirectional encoder.

I. INTRODUCTION

The outstanding development of emerging tools and techniques in science and technology changed the individual need and various fields of industry, medicine, science, and technology. The best way to use automated technologies for many populations and industrial requirements is to save time and money and obtain more accurate results than typical systems[1]. These tools have different foundations and backgrounds, including computer science and engineering. Every day, researchers are coming up with new, improved automated tools to prove the benefits of these systems for solving real and complex problems. These tools include pattern, face recognition, and other emerging areas[2], [3]. Different methods and improvements have been used to increase the accuracy of these methods in complex situations. From exact methods, various learning algorithms such as heuristic and non-heuristic, direct and indirect methods, local and global search algorithms based on computational or natural intelligence behaviors, and online automated methods[4]. The most famous machine learning techniques can easily be simulated, understood, and configured to solve real-world science, engineering, and complex social problems with different parameters and tasks such as classification, prediction, etc.[5]. Several classifications of algorithms, such as Supervised, Unsupervised, Reinforcement, hybrid, Improved, emerging, stochastics and pre-trained learning algorithms, were implemented to categorize machine learning-based techniques[6].

Among the methods mentioned above, metaheuristics and ML are more famous. At the same time, the advanced version of deep learning has increased the motivation of computer science and mathematical researchers. It has increased the effectiveness of the methods with highly accurate and quick results in various problems domain[7], [8]. In addition, machine learning and deep learning are approached with a significant history of achievement in solving problems in nonlinear, complex, difficult optimization, science, engineering, medicine, social, E-Systems, entrepreneurship, customer feedback, and other fields[2], [6]. The efficiency of these methodologies has increased, as has the motivation of researchers from all backgrounds to the automated, semi-automated, dynamic and parallel human processing methods used in DL.

The survey research is structured as follows: The BERT model, its kinds, and its history of performance success in many applications are explained in Parts II and III. The overall analysis of BERT models in the various issue domains is presented in Section IV, along with recommendations for further research.

II. BERT MODEL

BERT is a bidirectional DL-based model that analyzes various kinds of text accepting from the left and right sides instead of just from a single direction. The most recent BERT model adopts the Transformer structure, which contains multiple encoded layers [9]. It has demonstrated its benefits on various computational tasks, including inference and semantic understanding, NLP and text segmentation, classification, etc. Feed-Forward Neural Network, the encoder's capacity to give

979-8-3503-0030-7/23/$31.00 ©2023 IEEE

attention to itself, and how the decoder transfers data affect BERT performance.

Using the Transformer, BERT became famous and attracted various researchers to apply it to a couple of problems for obtaining the state of art results of other computational-based methods. It is a feature-based pre-trained model utilized for various tasks; fine-tuning involves many trainer acquisitions and can be used to retrain an automated mathematical model that has already been trained using the same or relevant big data. Pre-training and fine-tuning are the two phases of the BERT model[10]. The BERT model uses unlabeled data from various pre-training tasks on the problem. Notably, the model has been pre-trained and is fine-tuned using labeled data from the downstream functions even though it has the first initialization parameters. Because pre-trained representations are utilized as features from the two sides (left to right and right to left), it is also known as a feature-based methodology[10]. The fundamental BERT model for input representation is shown in Figure 1.

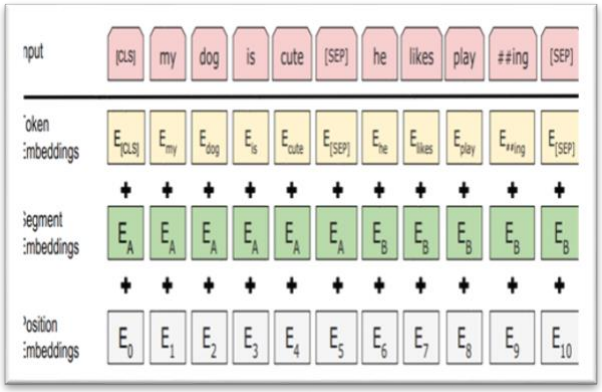

Fig. 1. Using typical BERT for input representation.

Each phrase has a BERT classification token at the beginning, separated by a BERT classification token [SEP]. Furthermore, a learned embedding is added, which identifies sentences (A or B) to each token to aid the model in differentiating between the various sentences. In this procedure, E with various samples has been designated as input sets at the first hidden vector of [CLS] symbol as $C \in R^H$. The i^{th} input token at the end of the last hidden layers that can be $T_i \in R^H$ [10].

A. *Types of BERT*

The BERT model performs effectively in various natural and artificial applications, particularly those involving NLP [10]. There are multiple types of BERT, such as DistilBERT[11], BioBERT[12], Clinical BERT[13], AraBERT[14] RoBERTa[15], DeBERTa [15], and so on. Various BERT models have been applied successfully to some problems, such as classification, prediction, NLP analysis, etc. Some of the BERT types are given in the following figure 2.

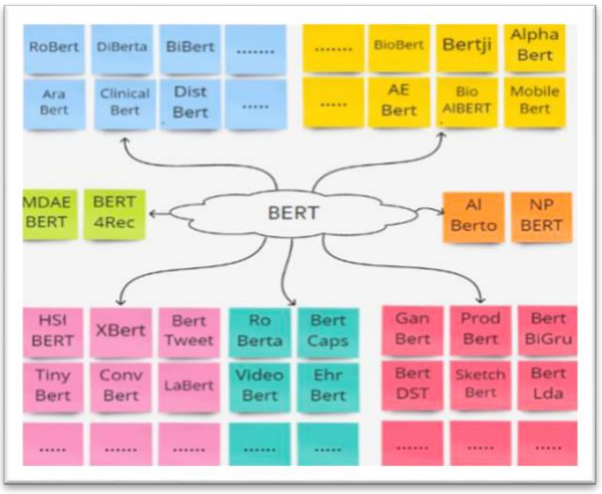

Fig. 2: Various types of BERT.

III. BERT APPLICATIONS

Initially, the basic BERT (Bert Base and Bert Large) was used for eleven NLP tasks, which produced more than 4.6% and 7.7 % accuracy and absolute improvement, respectively[10]. To classify sentiment and extract coverage-based sub-sentences from the text for entire sentences without the Prediction Next Sentence, the simulated trained and tested Robustly Optimized BERT Approach (RoBERTa) NLP model is the suitable model[9]. The simulation results obtained through the 80% training and 20% for testing from RoBERT were outstanding than other methods, including typical BERT[16]. For food named-entity recognition, the three types of BERT Models, such as BioBERT etc., results reached close to 93 and 94 % for macro F1 scores successfully based on 75 % training and 25 % testing in distinguishing food versus non-food entities. One of the famous types of BERT is called A Lite Biomedical Bidirectional Encoding Representation Transformer (ALBERT). Both were used for Feature Extraction in Hyperspectral Images. The simulation results obtained from simulation results BERT were more outstanding than other methods in most of the tests [17].

Following the findings, the proposed BIBERT outperformed BERT in various famous NLP tasks with multiple datasets, such as Semantic Similarity, Natural Language Inference, Sentiment Analysis, etc. The WikiText-103 benchmark dataset was successfully used and simulated along with typical BERT and the new model DIBERT [18]. Detecting fake news is a challenging task; however, using the BERT model, the

accuracy reached 9.23%, F-Score of 99.15%, a precision of 98.86% and recall came 99.46%, which are much more accurate than typical ML and DL methods[19]. In addition, using two performance assessment matrices, the standard BERT outperformed both the primary and sophisticated Recurrent Neural Network (RNN) models for categorizing news text: generic model classification accuracy and loss value.

The X-BERT (BERT for eXtreme Multi-label Text Classification) was proposed and simulated for the Eurlex-4K, Wiki10-28K, AmazonCat-13K, and Wiki-500K datasets along with PD-Sparse, fastText, FastXML, and Parabel. The proposed X-BERT produced better results quicker than the other listed methods[20]. In 2021, BERT + BiLSTM w/ Att was presented and used as a data pre-processor and embedding generator for a Bi-LSTM for Detecting subjectively biased statements; the results obtained by the BERT model achieved state-of-the-art accuracy than BiLSTM + Glove, BERT + BiLSTM w/o Atta and typical BERT[21].

The newly developed fine-tuned BERT Large (Cased) model outperforms four datasets in MAP and MRR, except WikiQA, for predicting and including predicted answer types. Along with various types of deep and typical machine and metaheuristics learning algorithms, such as k-Nearest Neighbors, Bio-Inspired methods, eXtreme Gradient Boosting (XGB), exact usual methods and other emerging tools are successfully used for NLP tasks along with NPBERT[22]. To address the public health problem of foodborne infections, the BERTweet model was presented in 2021. It aimed to extract entities from Twitter relevant to foodborne illnesses and discover unreported cases. Once more, the proposed BERTweet is used successfully in tweets containing foodborne illnesses, such as those having lettuce, tea, soup, fries, sushi, coffee, pizza, chicken, spices, cream, meat, taco, burger, chocolate, milk, cheese, ice, salad chip, beer, fish, and beef[23].

The ERT4Rec model has been proposed and successfully used on four benchmark datasets (Beauty, Steam, ML-1m and ML-20m)[24]. The new proposed AlBERTos is based on typical BERT used to model Italian Social Media writing style Language. Regarding statistical performance matrices such as F1 Score and Precision, the AlBERTo model's performance is exceptional[25]. The typical BERT model and other methods used for Dialog State Tracking with WoZ 2.0 dataset, where the BERT was with the final compressed model, achieved seven times faster and eight times smaller. Another famous type, Multi-Domain Aspect Extraction (MDAE), utilizes BERT to address the context-based semantic distance [26]. The MDAE-BERT and aspect extraction-based BERT have provided higher accuracies and F1-Macro than the LSTM-based approach in Restaurants and Laptops datasets from the 2014 SemEval ABSA feature extraction context. The BERT model is also used successfully with high precision in solving, detecting, predicting, and measuring natural hazards based on Twitter datasets [27]. This paper successfully simulated seven approaches for the seven well-known online business datasets, including "CNNs, DistilBERT, RoBERTv RNNs, LSTMs, BERT, and MBERT," to predict user loyalty in mobile applications. Compared to other models, the DBERT model achieved the maximum accuracy with 50%, 80%, and 30% training for the Hitman dataset according to predicted results with various training ratios, such as 80, 50, and 30%, [28].

Other BERT models, such as BERT4Bitter, a model for improving the prediction of bitter peptides, text classification, and BERT-DST Scalable end-to-end dialogue state tracking, have also produced exciting and successful results in the application of document classification , pandemic disease prediction [29], hands-on question answering systems [30], food information extraction [31], sentiment classification [32], fake news detection [33], and text classification. DNABERT language in the genome, Bert4sessrec: Content-based video relevance prediction, Learning Video Representations, BERT4Rec: Sequential Recommendation, Sketch-BERT [34], BERTTimbau Brazilian Portuguese , X-BERT: extreme multi-label text classification. MDAE-BERT: Multi-Domain Aspect Extraction [35], KG-BERT: for knowledge graph completion [36].

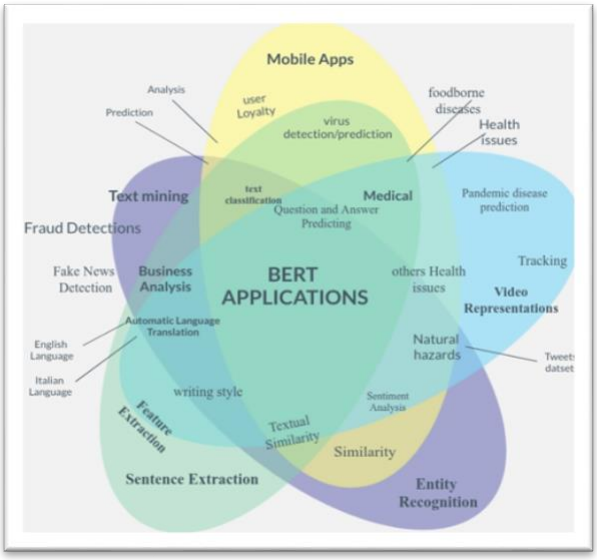

Fig. 3. Applications of BERT.

The results show that most BERT models have been successfully applied in a few CS fields, including NLP English, Italian, classification, and detection. The applications utilized by BERT and their various categories are shown in Figure 3.

We didn't discover much research generally except for a few publications in the commercial and medical domains. Also, there are limitations regarding the widely utilized scientific development procedures of hybridization and merging. Most BERT types have been enhanced using a single variable or set of parameters and repeated datasets for practical purposes. Nonetheless, the various BERT types have produced the best results compared to more advanced approaches. Hybridization approaches can be pretty valuable in improving the BERT model based on several algorithms' past successes. For instance, metaheuristics and other computational techniques can quickly improve these numerous BERT kinds. Emerging significant data problems can be resolved using conventional ML, DL, heuristic, and non-heuristic strategies, enhancing the effectiveness of the BERT's traditional and contemporary kinds.

IV. CONCLUSION

In the past five years, traditional and new BERT types have been extensively used in computer science, engineering, medicine, and other fields. According to the analysis of the results using various matrices and factors, it is one of the models developing, adapting, and appeals the fastest. In addition to general complex challenges, BERT successfully achieves state-of-the-art text mining, text processing, text analysis, and text detection. Previous research has demonstrated that modern BERTs perform more effectively than conventional ones. Although many challenging problems in natural science and engineering still need to be solved, and the existing BERT model and its new variations are not yet fully developed, the BERT model has shown an outstanding ability to learn.

REFERENCES

[1] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," SN Comput Sci, vol. 2, no. 3, p. 160, 2021, doi: 10.1007/s42979-021-00592-x.

[2] K. H. Teoh, R. C. Ismail, S. Z. M. Naziri, R. Hussin, M. N. M. Isa, and M. Basir, "Face Recognition and Identification using Deep Learning Approach," J Phys Conf Ser, vol. 1755, no. 1, p. 12006, 2021, doi: 10.1088/1742-6596/1755/1/012006.

[3] H. Shah, "Using new artificial bee colony as probabilistic neural network for breast cancer data classification," Frontiers in Engineering and Built Environment, vol. ahead-of-p, no. ahead-of-print, Jan. 2021, doi: 10.1108/FEBE-03-2021-0015.

[4] D. Sattar and R. Salim, "A smart metaheuristic algorithm for solving engineering problems," Eng Comput, vol. 37, no. 3, pp. 2389–2417, 2021, doi: 10.1007/s00366-020-00951-x.

[5] S. Kumar and M. Zymbler, "A machine learning approach to analyze customer satisfaction from airline tweets," J Big Data, vol. 6, no. 1, pp. 1–16, 2019.

[6] J. Qiu, Q. Wu, G. Ding, Y. Xu, and S. Feng, "A survey of machine learning for big data processing," EURASIP J Adv Signal Process, vol. 2016, no. 1, p. 67, 2016, doi: 10.1186/s13634-016-0355-x.

[7] W. G. Hatcher and W. Yu, "A Survey of Deep Learning: Platforms, Applications and Emerging Research Trends," IEEE Access, vol. 6, pp. 24411–24432, 2018, doi: 10.1109/ACCESS.2018.2830661.

[8] J. Shen and M. O. Shafiq, "Short-term stock market price trend prediction using a comprehensive deep learning system," J Big Data, vol. 7, no. 1, p. 66, 2020, doi: 10.1186/s40537-020-00333-6.

[9] U. Özdil, B. Arslan, D. E. Taşar, G. Polat, and Ş. Ozan, "Ad Text Classification with Bidirectional Encoder Representations," in 2021 6th International Conference on Computer Science and Engineering (UBMK), 2021, pp. 169–173. doi: 10.1109/UBMK52708.2021.9558966.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[11] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," arXiv preprint arXiv:1910.01108, 2019, [Online]. Available: http://arxiv.org/abs/1910.01108

[12] J. Lee et al., "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," Bioinformatics, vol. 36, no. 4, pp. 1234–1240, 2020, doi: 10.1093/bioinformatics/btz682.

[13] E. Alsentzer et al., "Publicly Available Clinical," in arXiv preprint arXiv:1904.03323, 2019, pp.

72–78. doi: 10.18653/v1/w19-1909.
[14] W. Antoun, F. Baly, and H. Hajj, "Arabert: Transformer-based model for arabic language understanding," arXiv preprint arXiv:2003.00104, 2020.
[15] Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv preprint arXiv:1907.11692, 2019, [Online]. Available: http://arxiv.org/abs/1907.11692
[16] J. Lim, I. Sa, H. S. Ahn, N. Gasteiger, S. J. Lee, and B. MacDonald, "Subsentence Extraction from Text Using Coverage-Based Deep Learning Language Models," Sensors , vol. 21, no. 8. 2021. doi: 10.3390/s21082712.
[17] I. O. Sigirci, H. Ozgur, and G. Bilgin, "Feature Extraction with Bidirectional Encoder Representations from Transformers in Hyperspectral Images," in 2020 28th Signal Processing and Communications Applications Conference (SIU), 2020, pp. 1–4. doi: 10.1109/SIU49456.2020.9302204.
[18] A. Wahab and R. Sifa, "DIBERT: Dependency Injected Bidirectional Encoder Representations from Transformers." Oct. 18, 2021. doi: 10.36227/techrxiv.16444611.v2.
[19] H. Jwa, D. Oh, K. Park, J. M. Kang, and H. Lim, "exBAKE: Automatic fake news detection model based on Bidirectional Encoder Representations from Transformers (BERT)," Applied Sciences (Switzerland), vol. 9, no. 19, p. 4062, 2019, doi: 10.3390/app9194062.
[20] W.-C. Chang, H.-F. Yu, K. Zhong, Y. Yang, and I. Dhillon, "X-bert: extreme multi-label text classification with using bidirectional encoder representations from transformers," arXiv preprint arXiv:1905.02331, 2019.
[21] E. V. Tunyan, T. A. Cao, and C. Y. Ock, "Improving Subjective Bias Detection Using Bidirectional Encoder Representations from Transformers and Bidirectional Long Short-Term Memory," International Journal of Cognitive and Language Sciences, vol. 15, no. 5, pp. 329–333, 2021.
[22] T.-H. Nguyen-Vo, Q. H. Trinh, L. Nguyen, T. T. T. Do, M. C. H. Chua, and B. P. Nguyen, "Predicting Antimalarial Activity in Natural Products Using Pretrained Bidirectional Encoder Representations from Transformers," J Chem Inf Model, Aug. 2021, doi: 10.1021/acs.jcim.1c00584.
[23] D. Tao, D. Zhang, R. Hu, E. Rundensteiner, and H. Feng, "Crowdsourcing and machine learning approaches for extracting entities indicating potential foodborne outbreaks from social media," Sci Rep, vol. 11, no. 1, p. 21678, 2021, doi: 10.1038/s41598-021-00766-w.
[24] F. Sun et al., "BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer," in Proceedings of the 28th ACM international conference on information and knowledge management, 2019, pp. 1441–1450. doi: 10.1145/3357384.3357895.
[25] M. Polignano, V. Basile, P. Basile, M. de Gemmis, and G. Semeraro, "AlBERTo: Modeling Italian Social Media Language with BERT," IJCoL. Italian Journal of Computational Linguistics, vol. 5, no. 5–2, pp. 11–31, 2019.
[26] B. N. Dos Santos, R. M. Marcacini, and S. O. Rezende, "Multi-Domain Aspect Extraction Using Bidirectional Encoder Representations From Transformers," IEEE Access, vol. 9, pp. 91604–91613, 2021, doi: 10.1109/ACCESS.2021.3089099.
[27] Z. Wang, T. Zhu, and S. Mai, "Disaster Detector on Twitter Using Bidirectional Encoder Representation from Transformers with Keyword Position Information," in 2020 IEEE 2nd International Conference on Civil Aviation Safety and Information Technology (ICCASIT, 2020, pp. 474–477. doi: 10.1109/ICCASIT50869.2020.9368610.
[28] Z. H. Kilimci, "Prediction of user loyalty in mobile applications using deep contextualized word representations," Journal of Information and Telecommunication, pp. 1–20, 2021, doi: 10.1080/24751839.2021.1981684.
[29] L. Rasmy, Y. Xiang, Z. Xie, C. Tao, and D. Zhi, "Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction," NPJ Digit Med, vol. 4, no. 1, p. 86, 2021, doi: 10.1038/s41746-021-00455-y.
[30] N. Sabharwal and A. Agrawal, "BERT Model Applications: Question Answering System," in Hands-on Question Answering Systems with BERT, Springer, 2021, pp. 97–137.
[31] E. Atagün, B. Hartoka, and A. Albayrak, "Topic Modeling Using LDA and BERT Techniques: Teknofest Example," in 2021 6th International Conference on Computer Science and Engineering (UBMK), 2021, pp. 660–664. doi: 10.1109/UBMK52708.2021.9558988.
[32] Z. WANG, H. WU, H. LIU, and Q.-H. CAI, "Bert-Pair-Networks for Sentiment Classification," in 2020 International Conference on Machine

Learning and Cybernetics (ICMLC), 2020, pp. 273–278. doi: 10.1109/ICMLC51923.2020.9469534.

[33] H. Jwa, D. Oh, K. Park, J. M. Kang, and H. Lim, "exBAKE: Automatic fake news detection model based on Bidirectional Encoder Representations from Transformers (BERT)," Applied Sciences (Switzerland), vol. 9, no. 19, p. 4062, 2019, doi: 10.3390/app9194062.

[34] F. Sun et al., "BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer," in Proceedings of the 28th ACM international conference on information and knowledge management, 2019, pp. 1441–1450. doi: 10.1145/3357384.3357895.

[35] B. N. Dos Santos, R. M. Marcacini, and S. O. Rezende, "Multi-Domain Aspect Extraction Using Bidirectional Encoder Representations From Transformers," IEEE Access, vol. 9, pp. 91604–91613, 2021, doi: 10.1109/ACCESS.2021.3089099.

[36] L. Yao, C. Mao, and Y. Luo, "KG-BERT: BERT for Knowledge Graph Completion," arXiv preprint arXiv:1909.03193, 2019, [Online]. Available: http://arxiv.org/abs/1909.03193