Performance analysis of different BERT implementation for event burst detection from social media text

Dharmendra Mangal[1], Hemant Makwana[2]
[1]Department of CSE, Faculty of Engineering, Medi-Caps University, Indore, India
[2]Institute of Engineering and Technology, DAVV, Indore, India

Article Info

Article history:

Received Jun 6, 2024
Revised Oct 17, 2024
Accepted Oct 30, 2024

Keywords:

BERT
Event detection
Natural language processing
Performance analysis
Social media
Transformers

ABSTRACT

The language models play very important role in natural language processing (NLP) tasks. To understand natural languages, the learning models are required to be trained on large corpus. This requires a lot of time and computing resources. The detection of information like events, and locations from text is an important NLP task. As events detection is to be done in real-time so that immediate actions can be taken, hence we need efficient decision-making models. The pertained models like bi-directional encoders representation from transformers (BERT) gaining popularity to solve NLP problems. As BERT based models are pre-trained on large language corpus it requires very less time to adapt for domain specific NLP task. Different implementations of BERT have been proposed to enhance efficiency and applicability of the base model. The selection of right implementation is essential for overall performance of NLP based system. This work presents the comparative insights of five widely used BERT implementations named as BERT-base, BERT-large, Distill BERT, Robust BERT approach (RoBERTa-base) and RoBERT-large for event detection from the text extracted from social media streams. The results show that Distill-BERT model outperforms on basis of performance metric like precision, recall, and F1-score while the fastest to train also.

Corresponding Author:

Dharmendra Mangal
Department of CSE, Faculty of Engineering, Medi-Caps University
Indore, India
Email: managldharmendra83@gmail.com

1. INTRODUCTION

The social media becomes the information sharing platform. The detection of events is very crucial task especially for that needs attention and response by administrative officers. The events like flood, earthquakes, riots, and accidents need immediate response [1]. The timely response requires timely detection. This helps the government officials to react on correct time. The messages shared on social media can be sensed to detect the events [2]. The text messages are used as input to detect the events hence event detection is basically a natural language processing (NLP) task. Traditionally NLP-toolkit (NLTK) has been used for event detection as it is very simple to use. But NLTK relies on manual features extraction as compared to deep learning approaches which provides automatic feature extraction mechanism [3]. Consequently, recent works are based on deep learning models. Although NLTK is used for data preprocessing. Most of the deep learning models-based approaches are rely on text representation features like word embeddings, part-of-speech tags, named entities, and word position. The word embedding is the most prominent one as it maps input token (word) into numeric value. This is mandatory for input processing by any deep learning model as they are fundamentally mathematical in nature [4]. Most of the deep learning methods proposed are convolution neural network (CNN) and recurrent neural network (RNN) based [5].

The field of event detection attracts significant research interest, leading to the publication of numerous research papers. Various research groups employ different anomaly detection techniques, NLP tools, modalities, and social networks, focusing on diverse applications [6]. Several surveys have been conducted to analyze this wealth of information. For example, Zhou *et al*. [7] conducted an analysis of social event detection approaches from a modality perspective. Additionally, the survey includes performance comparisons of various methods using multiple public datasets. While the survey does not prioritize the evaluation processes used in reviewing papers, it is crucial to understand how authors gather and label data, as well as the metrics used to assess algorithm quality from a comparative standpoint.

The issue of comparing event detection approaches is emphasized by [8], who also present their proposal for achieving reproducible research in event detection techniques based on Twitter data. Their main concept involves creating a simulated Twitter stream with specific parameters to assess various methods. This comparison enables the implementation of different approaches within the same environment. Nevertheless, this approach may not be appropriate for solutions that rely on diverse data modalities due to the absence of metadata for evaluation of algorithm quality.

Weiler *et al.* [9], [10] have authored multiple papers focusing on an in-depth analysis of metrics suitable for evaluation. They recommended the utilization of the duplicate events rate (DERate) to achieve a more precise evaluation of methods. Furthermore, in consideration of challenges associated with data markup, the authors suggested the use of metrics that can be automatically obtained. For instance, quantitative metrics like memory usage or execution time. They also proposed the use of a precision metric based on search engine results for queries related to the identified events as a qualitative measure. The incorporation of such a metric enables a more equitable comparison of different methods.

Applying transformer like BERT for event detection is very new idea. Very few researchers have been worked on it. There are various BERT implementations have been proposed. The Figure 1 shows the basic architecture of BERT based models. The BERT base model is the origin of all the different BERT implementations [11]. There are multiple encoders (let N) present in BERT based model [12].

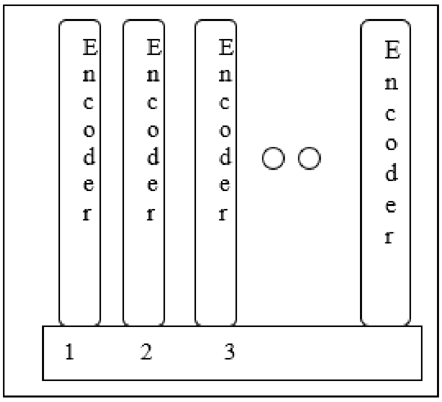

Figure 1. BERT based model general architecture

In the year 2017 new deep learning network architecture was proposed by Google research named as transformers. This network architecture is based on the idea of ‘attention’ based learning. The ‘attention’ in learning means understanding the word context in both the directions in a text. The BER, GPT, and embeddings from language models (ELMo) are primarily transformer models [13], [14].

Word embedding is the central feature in most of the NLP task [15]. The attention mechanism enhances the word embedding performance. Most of the previous works have used word embedding models like Word2Vec and GloVe. But pre- trained language model like BERT has proved its ability to provide better results in most of the NLP task such as event detection form text message [16]. The BERT is a transformer-based model which follows self-supervised and transfer learning [17], [18]. BERT was created to jointly train the left and right contexts in order to pre-train deep bidirectional representations from the unlabeled text. Its trained model serves as the mind, which can subsequently control are the increasingly vast collections discoverable information and queries that are tailored to the individual's needs. This procedure referred as transfer learning [19]. The BERT model pre-trained on a vast collection of unlabeled text encompassing Wikipedia (2,500 million words) and literary works (English). Due to the extensive training on a large text corpus, the model gradually develops a comprehensive and thorough understanding of English language [20].

2. METHOD

The notion of transformer in NLP is revolutionary one. The transformers are basically deep learning model-based systems with attention [21]. All BERT implementations i.e. BERT-base, BERT-large, Distill-BERT, RoBERTa-base, and RoBERTa-large are compared in this work. The brief description of these models is given in following paragraphs.

The BERT model process 512 tokens input and outputs the vector representation of the sequence i.e. word embedding [22]. This output can have one or two segments, with the first token always being [CLS], containing the specific classification embedding, and another special token, [SEP], used to separate the segments. BERT organizes the final hidden state h of the first token [CLS] in order to process the complete sequence for text classification tasks. To obtain the predicted probabilities from the trained model a SoftMax classifier is included at the top of the BERT model [23]. The data set needs to be converted into vectors before being input into the classifier because it is initially in text form. BERT learns contextual embeddings instead of context-free embeddings, unlike Word2Vec. Even though there are different models for text vectorization, BERT carries out tokenization using the WordPiece approach. The foundation of BERT is a stack of encoder layers. The number of encoder layers is where BERT base and BERT-large diverge. In the BERT-large model, there are 24 layers of encoders layered on top of one another, compared to 12 layers in the BERT base model [24].

The DistilBERT decreases the size of BERT by 40% while keeping 97% of BERT's performance [25]. DistilBERT removes pooler and token-type embeddings to resemble the BERT model. Distillation is the technique of approximating a larger network's full output distributions using a smaller network after the larger network has been trained. This is analogous to posterior approximation in certain respects, and Kulback Leiber divergence is one of the key optimization procedures applied [26].

RoBERTa, short for robustly optimized BERT approach, was introduced by Facebook [27]. It involves retraining BERT using an enhanced training methodology, a larger dataset, and increased computing power. The RoBERTa model is implemented similar to the BERT model, with a minor change to the embeddings and a setup for pretraining RoBERTa models. Although it shares the same architecture as BERT, RoBERTa utilizes a byte-level pair encoding (BPE) tokenizer similar to GPT-2 and employs a different pretraining scheme. Large mini-batches, a higher byte-level BPE, dynamic masking, and entire phrases without NSP loss are all used in the training of RoBERTa. By eliminating the next sentence prediction (NSP) task from BERT's pre-training and implementing dynamic masking to alter the masked token during training epochs, RoBERTa enhances the training process [28]. The experiment demonstrated that larger batch training sizes are more beneficial in the training process. Notably, in addition to BERT training 16 GB of books Corpus and English Wikipedia data, RoBERTa utilizes 160 GB of text for pre-training. In the RoBERTa-large model, there are 24 layers of encoders layered on top of one another, compared to 12 layers in the RoBERTa base model [29].

The Figure 2 shows the methodology adapted in this work. After preprocessing of the dataset, all the five models are applied sequentially. Finally, their performance is compared. The detailed architectural differences among all five models are shown in Table 1. It shows comparison among the considered BERT variants based on characteristics like number of encoders, hidden layers, self-attention heads and decision parameters.

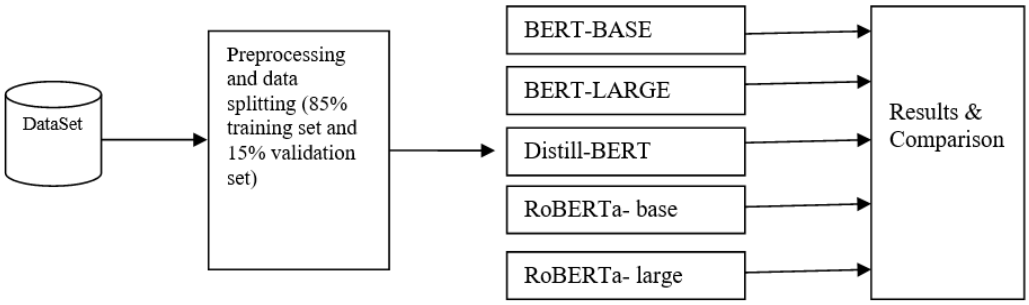

Figure 2. Methodology

Table 1. Architectural differences among five different BERT-implementations

Model	# Encoders (N)	Hidden layer size	# Self attention heads	# Parameters
BERT-base	12	768	12	110M
BERT-large	24	1024	16	340M
Distill-BERT	6	768	22	66M
RoBERTa-base	12	768	12	125M
ROBERTa-large	24	1024	16	355M

3. RESULTS AND DISCUSSION

The experiment dataset contains more than 10K text messages taken from X (Twitter) platform. This data set is divided into two parts: training set (85%) and testing set (15%). The training of all the candidate models has been performed in 10 epochs. Also, ADAM optimizer is used for hyper-parameter tuning. The hyper-parameters include accuracy and loss value. Factually, accuracy is a metric that describe percentage of the test or validation data correctly labeled whereas loss value is the average distance between the true values and the values predicted by the model. The datasets used comprises more than 10K text messages (tweets) available from Kaggle platform. The Figure 3 to Figure 7 shows the learning curves for different BERT models for accuracy and loss value respectively.

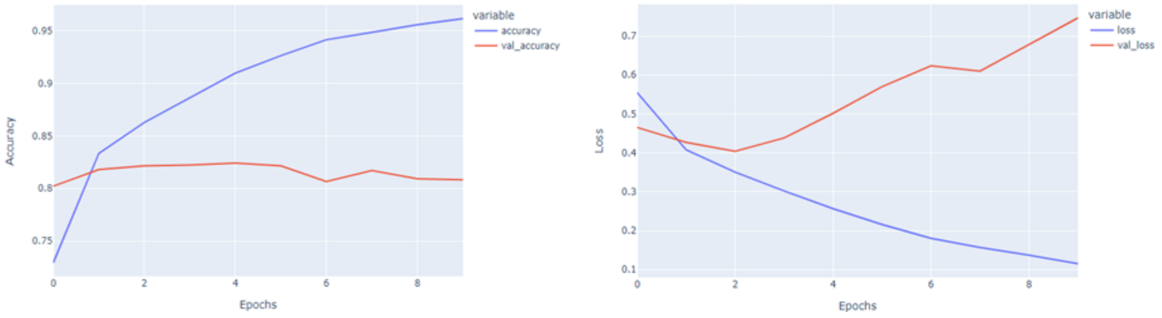

Figure 3. Hyper-parameters (accuracy and loss) tuning for BERT-base model in 10 epochs

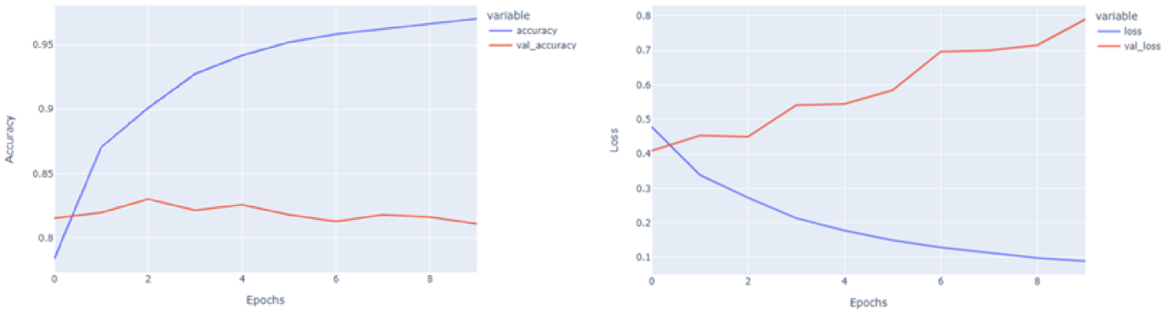

Figure 4. Hyper-parameters (accuracy and loss) tuning for BERT-large model in 10 epochs

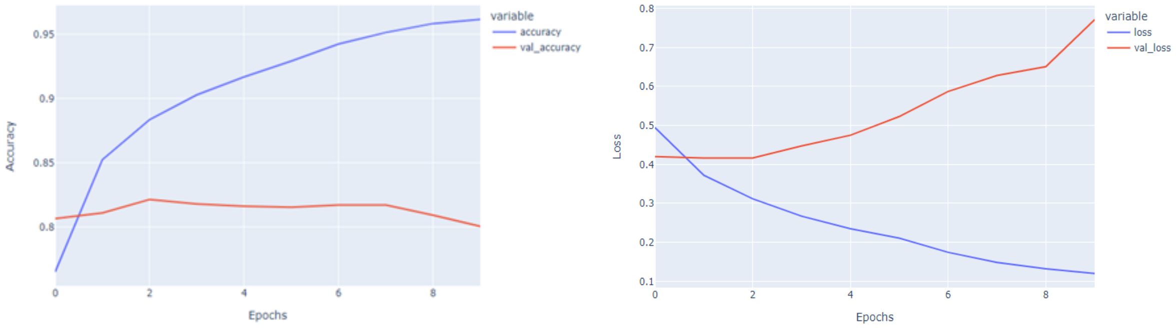

Figure 5. Hyper-parameters (accuracy and loss) tuning for distill-BERT model in 10 epochs

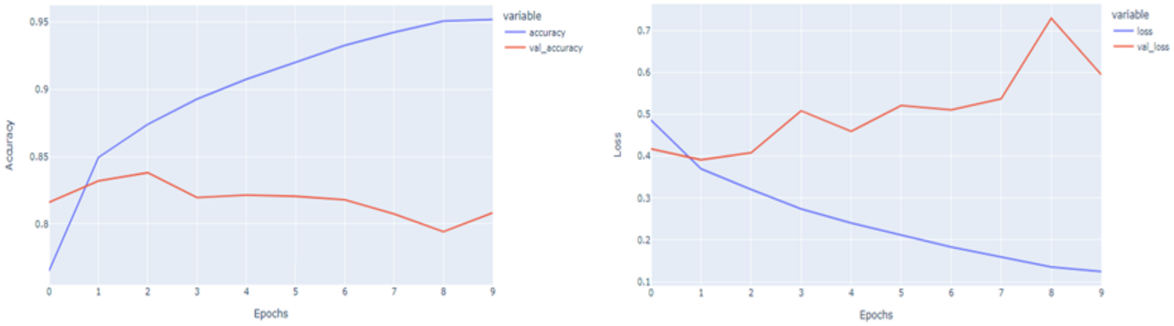

Figure 6. Hyper-parameters (accuracy and loss) tuning for RoBERTa-base model in 10 epochs

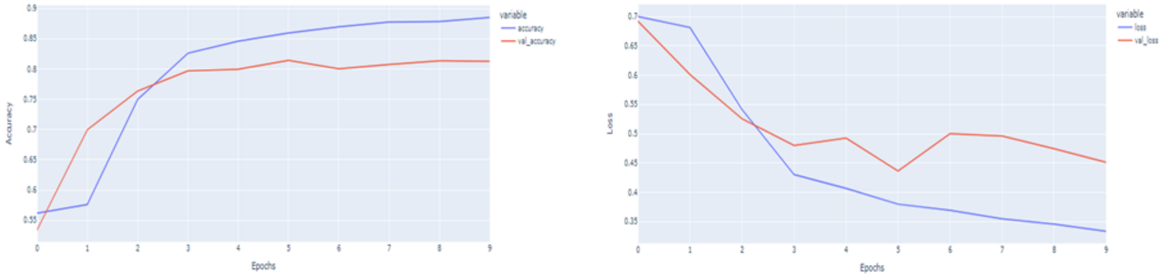

Figure 7. Hyper-parameters (accuracy and loss) tuning for RoBERTa-base model in 10 epochs

In summary, the Figure 8 and 9 shows the performance comparison among all the five model during fine tuning in 10 iteration (epochs) of learning. There are 3 performance metrics are used for comparison i.e. precision, recall, and F1-score. Figure 10 shows the relation among them.

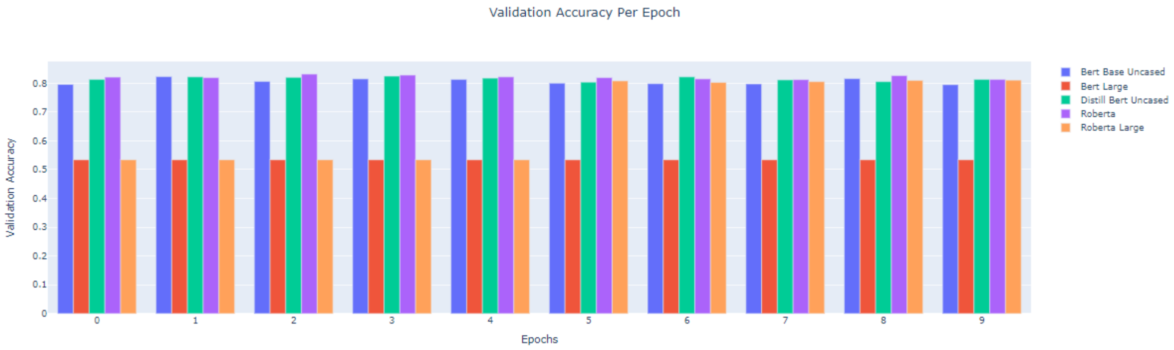

Figure 8. Comparison of accuracy during fine tuning in each epoch

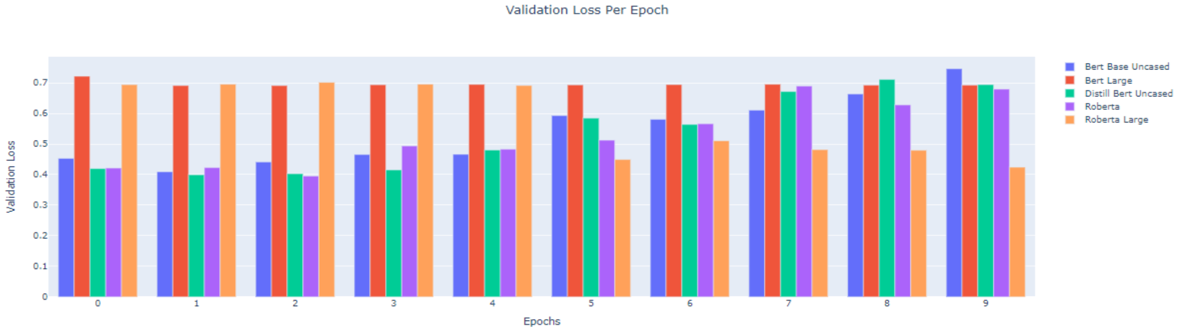

Figure 9. Comparison of loss-value during fine tuning in each epoch

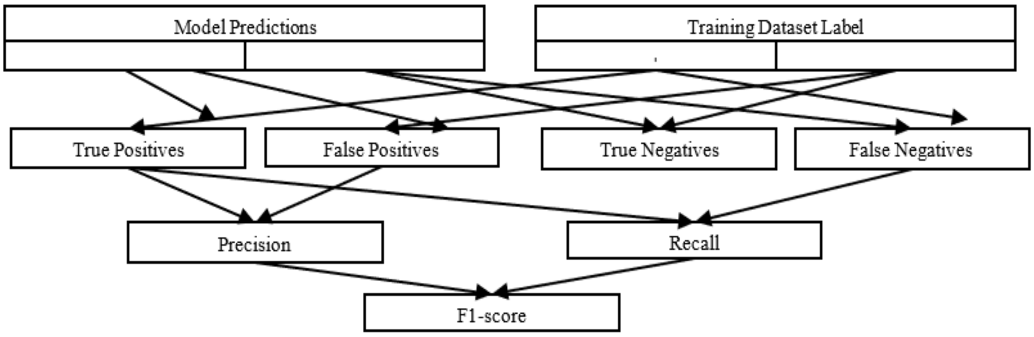

Figure 10. Deriving precision, recall and F1-score from raw measurements

As we can consider the event burst identification as binary classification problem i.e. whether the text message indicating an event burst (positive +) or not (negative -). Thus, the model predictions are based on these binary classes. The training data is labeled with binary labels (1 for positive and 0 for negative). True positives (true+) is count of correctly identified positive labels whereas false positive (false+) indicates count of positive label classified incorrectly. Similarly true negatives (true-) are the count of correctly identified negative labels whereas false negatives(false-) are count of incorrectly classified negative labels by the model. We can derive the values of precision and recall using the following formulas:

$$Precision = Count(true+)/Count(true+) + Count\ (false+)$$

$$Recall = Count\ (true\ +)/Count\ (true\ +) + Count\ (false\ -)$$

The F1-score is the harmonic mean of precision and recall value. It is formulated as:

$$F1 - score = (2 * Precision * Recall)/Precision + Recall)$$

These metrics has been computed for different BERT variants as summarized in Table 2. It shows precision, recall, and F1 value of the compared BERT variants for both positive labeled and negative labeled data in the dataset. This study is unique and comprehensive one. Table 3 shows the differences with some previous work.

Table 2. Comparison of different BERT variants based on performance metrics

Model	Precision (label 0)	Precision (label 1)	Recall (label 0)	Recall (label 0)	F1-score (label 0)	F1-score (label 1)
BERT-base	0.88	0.93	0.96	0.82	0.91	0.87
BERT-large	0.88	0.94	0.96	0.83	0.92	0.88
Distill-BERT	0.9	0.95	0.96	0.87	0.93	0.91
RoBERTa-base	0.89	0.95	0.97	0.85	0.93	0.9
ROBERTa-large	0.89	0.85	0.88	0.85	0.89	0.85

Table 3. Comparison with previous work

Author(s)	Models compared	Corpus used	Findings
Turchin *et al.* [20]	BERT base, clinical BERT, and BioBERT	Medical text documents	Clinical BERT outperforms.
Zeinali *et al.* [23]	BERT base, span BERT, clinical BERT, and BioBERT	Electronic health records	Clinical BERT outperforms.
Cortiz [24]	BERT base, Distil- BERT, RoBERTa, XLNet, and ELECTRA	Emotion's datasets	DistilBERT outperforms
Proposed work	BERT base/large, Distil-BERT, and RoBERTa base/large	Disaster event tweets	DistilBERT and RoBERTa base outperforms.

4. CONCLUSION

Use of pre-trained language models in NLP applications led to significant performance gain. As various pre-trained models are available careful comparison between them is a challenging task. This helps the NLP solution architect to choose the most appropriate among the all available. In this work we carefully observe the performance of different BERT implementations on event detection task from social media text. The five popular BERT implementations namely BERT-base, BERT-large, Distill-BERT, RoBERTa–base, and RoBERTa-large are compared based on the performance metrics precision, recall, and F1-score. In conclusion, we have found that Distill-BERT implementation trained on event detection dataset outperforms among all other while RoBERTa-base model performed impressive almost equal to Distill-BERT model. This study further extended for areas like medical document analysis, travel blog analysis for disease prediction and spatial information extraction respectively.

REFERENCES

[1] M. Ding, C. Zhou, H. Yang, and J. Tang, "CogLTX: applying BERT to long texts," *Advances in Neural Information Processing Systems*, vol. 2020-December, 2020.

[2] M. Zaheer *et al.*, "Big bird: transformers for longer sequences," *Advances in Neural Information Processing Systems*, vol. 2020-December, 2020.

[3] C. Casula and S. Tonelli, "Hate speech detection with machine-translated data: the role of annotation scheme, class imbalance and undersampling," *CEUR Workshop Proceedings*, vol. 2769, 2020, doi: 10.4000/books.aaccademia.8345.

[4] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, vol. 106, pp. 249–259, 2018, doi: 10.1016/j.neunet.2018.07.011.

[5] A. Vaswani *et al.*, "Attention is all you need in advances in neural information processing systems," *Search PubMed*, pp. 5998–6008, 2017.

[6] Y. Zhu *et al.*, "Aligning books and movies: towards story-like visual explanations by watching movies and reading books," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2015 International Conference on Computer Vision, ICCV 2015, pp. 19–27, 2015, doi: 10.1109/ICCV.2015.11.

[7] H. Zhou, H. Yin, H. Zheng, and Y. Li, "A survey on multi-modal social event detection," *Knowledge-Based Systems*, vol. 195, p. 105695, May 2020, doi: 10.1016/j.knosys.2020.105695.

[8] A. Weiler, H. Schilling, L. Kircher, and M. Grossniklaus, "Towards reproducible research of event detection techniques for Twitter," in *Proceedings - 6th Swiss Conference on Data Science, SDS 2019*, IEEE, Jun. 2019, pp. 69–74. doi: 10.1109/SDS.2019.000-5.

[9] A. Weiler, M. Grossniklaus, and M. H. Scholl, "Evaluation measures for event detection techniques on twitter data streams," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9147, pp. 108–119, 2015, doi: 10.1007/978-3-319-20424-6_11.

[10] A. Weiler, M. Grossniklaus, and M. H. Scholl, "An evaluation of the run-time and task-based performance of event detection techniques for Twitter," *Information Systems*, vol. 62, pp. 207–219, Dec. 2016, doi: 10.1016/j.is.2016.01.003.

[11] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in NLP," *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pp. 3645–3650, 2020.

[12] J. Pennington, R. Socher, and C. D. Manning, "GloVe: global vectors for word representation," *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pp. 1532–1543, 2014, doi: 10.3115/v1/d14-1162.

[13] T. Wolf *et al.*, "HuggingFace's transformers: state-of-the-art natural language processing," *arXiv preprint 1910.03771*, pp. 1–8, 2019.

[14] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint 1503.02531*, Mar. 2015, [Online]. Available: http://arxiv.org/abs/1503.02531.

[15] A. F. Adoma, N. M. Henry, and W. Chen, "Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition," *2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing, ICCWAMTIP 2020*, pp. 117–121, 2020, doi: 10.1109/ICCWAMTIP51612.2020.9317379.

[16] D. Chatterjee, "Making neural machine reading comprehension faster," *arXiv preprint 1904.00796*, Mar. 2019, [Online]. Available: http://arxiv.org/abs/1904.00796.

[17] D. Mangal and H. Makwana, "Extracting geo-references from social media text using bi-long short term memory networks," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 35, no. 2, pp. 1263–1270, Aug. 2024, doi: 10.11591/ijeecs.v35.i2.pp1263-1270.

[18] C. Bucilă, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA: ACM, Aug. 2006, pp. 535–541. doi: 10.1145/1150402.1150464.

[19] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," *ACL-HLT 2011 - Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, pp. 142–150, 2011.

[20] A. Turchin, S. Masharsky, and M. Zitnik, "Comparison of BERT implementations for natural language processing of narrative medical documents," *Informatics in Medicine Unlocked*, vol. 36, p. 101139, 2023, doi: 10.1016/j.imu.2022.101139.

[21] R. Tang, Y. Lu, L. Liu, L. Mou, O. Vechtomova, and J. Lin, "Distilling task-specific knowledge from BERT into simple neural networks," Mar. 2019, [Online]. Available: http://arxiv.org/abs/1903.12136

[22] S. Subramanian, A. Rahimi, T. Baldwin, T. Cohn, and L. Frermann, "Fairness-aware class imbalanced learning," *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*, pp. 2045–2051, 2021, doi: 10.18653/v1/2021.emnlp-main.155.

[23] N. Zeinali, A. AlBashayreh, W. Fan, and S. G. White, "Comparison of BERT Implementations for enhanced cancer symptoms extraction from electronic health records," in *Proceedings - 2024 IEEE 1st International Conference on Artificial Intelligence for Medicine, Health and Care, AIMHC 2024*, IEEE, Feb. 2024, pp. 18–19. doi: 10.1109/AIMHC59811.2024.00011.

[24] D. Cortiz, "Exploring transformers models for emotion recognition: a comparision of BERT, DistilBERT, RoBERTa, XLNET and ELECTRA," in *ACM International Conference Proceeding Series*, New York, NY, USA: ACM, Aug. 2022, pp. 230–234. doi: 10.1145/3562007.3562051.

[25] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: a pretrained language model for scientific text," *arXiv preprint 1903.10676*, Mar. 2019, [Online]. Available: http://arxiv.org/abs/1903.10676.

[26] J.-S. Lee and J. Hsiang, “PatentBERT: patent classification with fine-tuning a pre-trained BERT model,” *arXiv preprint 1906.02124*, May 2019, [Online]. Available: http://arxiv.org/abs/1906.02124

[27] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, “LEGAL-BERT: The muppets straight out of law school,” *Findings of the Association for Computational Linguistics Findings of ACL: EMNLP 2020*, pp. 2898–2904, Oct. 2020, doi: 10.18653/v1/2020.findings-emnlp.261.

[28] M. B. A. McDermott, B. Yap, P. Szolovits, and M. Zitnik, “Structure inducing pre-training,” *arXiv preprint 2103.10334*, Mar. 2021, [Online]. Available: http://arxiv.org/abs/2103.10334

[29] X. Du and C. Cardie, “Event extraction by answering (almost) natural questions,” in *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 671–683. doi: 10.18653/v1/2020.emnlp-main.49.

BIOGRAPHIES OF AUTHORS

Dharmendra Mangal is assistant professor in Medi-Caps University, Indore also Ph.D. scholar at IET, DAVV, Indore. The research interest includes NLP and machine learning. He can be contacted at email: mangaldharmendra83@gmail.com.

Dr. Hemant Makwana is associate professor at IET, DAVV, Indore. The research interest includes computer graphics and computer architecture. He is Ph.D. in computer engineering. He can be contacted at email: hmakwana@ietdavv.edu.in.