2023 Third International Conference on Ubiquitous Computing and Intelligent Information Systems (ICUIS) | 979-8-3503-0698-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICUIS60567.2023.00066

Leveraging DistilBERT and BERT for the Detection of Online Sexism: A Comparative Analysis

Sowmen Mitra
Department of Computer Science and Technology
Department Of Artifical Intelligence
Hebei University of Technology
Beijing,China
Sowmenmitra7@gmail.com

Proma Kanungoe
Department Of Information Technology
School Of Information Technology And Engineering
Vellore Institute Of Technology
Vellore, India
proma.kanungoe2018@vitalum.ac.in

Ali Newaz Chowdhury
Department of Computer Science and Engineering,
Ahsanullah University of Science and Technology
Dhaka, Bangladesh
alinewaz33@gmail.com

***Abstract*—Online sexism perpetuates harmful gender stereotypes and biases, leading to an environment rife with prejudice and injustice. This not only erodes human well-being by inducing feelings of emotional distress and worthlessness, particularly among women and marginalized genders, but also stifles the free exchange of ideas by creating hostile digital spaces. These spaces suppress voices, limit participation, and hinder meaningful interactions. In our study, we utilized the Bidirectional Encoder Representations from Transformers (BERT) and DistilBERT to swiftly identify sexist comments. Experimental results indicate that BERT significantly outperforms both DistilBERT and other leading architectures. With an emphasis on sustainable AI practices, we've optimized both models for efficiency. In evaluating their efficacy, we've considered both their F1 score and their environmental impact.**

Keywords—Sexism, Transformer, Bidirectional Encoder Representations from Transformers (BERT), Distillated Bidirectional Encoder Representations from Transformers (DistilBERT), Text Classification.

I. Introduction

Expressing feelings on online social media platforms is facile as a consequence of its democratic nature [1]. According to [2], approximately 4.49 billion people interact through these platforms on a regular basis. These frequent interactions allow users to express their feelings regarding versatile topics where harassment is a fatal topic that can create hatred between communities. According to a statement from the United Nations Development Program (UNDP), threats or violence in online media have a real life impact [3]. Sexism can be listed in the subcategory of hate speech, a declining factor in society [4]. These types of incidents are most frequent in Asian continents, where it has been handled as a sensitive topic [5]. Platforms may support diversity and representation by fostering a more welcoming online atmosphere. This can make it possible for everyone to participate in online debates and activities without worrying about becoming the target of discrimination or harassment.

Online sexism includes threats of harm, casual use of gender slurs, descriptive and emotive attacks, and sexual objectification through social media [6]. There should be a lenient way of detecting online sexism because of the frequent usage of such occurrences. This system can be beneficial in identifying online sexism within a shorter period of time, where public concern is the topmost priority. Online sexism has a direct influence on cyberattacks that significantly affect social tranquillity [7]-[9] Detection of online sexism is an arduous task as there can be many subcategories under this topic. Due to the complexity and variety of hate speech categories, it is difficult for machines to understand the patterns significantly [10]. The detection of online sexism is a point of discussion in Artificial Intelligence (AI). In this research, the authors focus on detecting online sexism from numerous platforms. The dataset is gathered from different social media sites. Online sexism must be detected within a shorter period so that no other effect can occur. Taking this thing into account, authors have taken a transformer-based approach where DistilBERT has been trained in order to identify online sexism at an earlier stage.

Our main contributions can be summarized in three points:

- **Specialized dataset:** This paper has curated a dataset of comments related to online sexism, enabling the training and evaluation of models for detecting such content.
- **Deep learning-based approach:** By utilizing transformer models like BERT and DistilBERT, this paper has developed a powerful deep learning approach to identify and classify instances of online sexism.
- **Performance evaluation and future prospects:** This paper has rigorously evaluated author's approach using various metrics and discussed future directions, including the integration of additional models and the development of a comprehensive framework to combat online sexism.

II. Literature Review

The diverse user base of social media has led to a broad spectrum of applications and challenges. Among the emergent issues, there's been a discernible surge in sexist comments [11]. Post the Covid-19 outbreak, there was an observable increase in sexism, especially towards Chinese individuals. Jiang et al. dedicated efforts to compile a Chinese dataset encapsulating online

979-8-3503-0698-9/23/$31.00 ©2023 IEEE
DOI 10.1109/ICUIS60567.2023.00066

sexism comments, though the study did not address data quality measurements [11]. Parallelly, Hewitt delved into classifying misogynist tweets, employing a multiclass classification strategy [12]. Anzivino et al. conducted an exhaustive survey on misogynous tweets [13], paving the way for subsequent studies that explored online sexism in various languages and aimed to develop instantaneous classification models [14]-[15].

Traditional machine learning algorithms, while potent, often falter in accurately identifying online sexism. Their inability to aptly preserve sequence information is a notable constraint [16]-[17]. Recurrent Neural Networks (RNN), for instance, grapple with sentences characterized by extended sequences. They frequently encounter the vanishing gradient problem, complicating the training phase [19].

In contrast to their predecessors, transformer models have demonstrated a remarkable capability to identify online sexism with heightened precision and in reduced time frames [18]. Their resilience against the issues faced by RNNs, such as the vanishing gradient problem, is mainly attributable to the self-attention mechanism. This mechanism facilitates efficient capturing of dependencies between words spaced far apart in sentences. A cornerstone in the realm of transformer models, the Bidirectional Encoder Representations from Transformers (BERT), was pioneered by Google AI [20]. Its extensive parameter count does introduce computational complexities but also underpins its superior performance. For applications that prioritize lighter models, DistilBERT emerges as a viable alternative. Introduced in [21], DistilBERT is a distilled version of BERT, retaining only 66 million of the original's 344 million parameters. Despite its compactness, DistilBERT mirrors the functionality of BERT closely, owing to the distillation technique where a smaller model is trained to emulate a larger one. Consequently, DistilBERT demands substantially lower memory compared to BERT. However, in the trade-off, BERT remains unparalleled in accuracy, with reduced fine-tuning and pre-training durations compared to its counterparts.

The crux of this research revolves around juxtaposing the DistilBERT and traditional BERT models, especially in contexts characterized by constrained training durations and computational resources. A rigorous evaluation, incorporating key performance metrics, revealed promising outcomes.

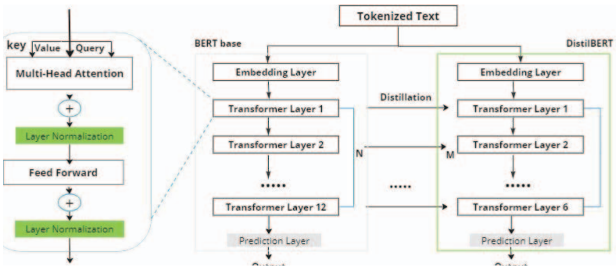

Fig. 1. DistilBERT Architecture

Figure 1 depicts the total workflow of this research, where it can be observed various techniques have been integrated for training the DistilBERT model. To observe the result precisely K-cross validation has been applied along with data loss at every phase.

III. Data Analysis

A. Dataset Description

In this research, the authors are focused on using a specialized dataset where the dataset consists of comments from different subcategories of online sexism. The whole dataset is split into three segments, namely training, validation and test dataset. The test dataset is reserved away from the model so that the model cannot be trained on that specific set. In the training dataset, there are several columns available. Table I shows the detailed description of the attributes.

TABLE I. Attributes Names With Description

Attribute Name	Description
Rewire-id	Year of gathering and language of a comment.
Text	Comments gathered from numerous sources
Label-sexiest	Whether a comment is sexist of not
Label-category	What type of sexiest comment is that
Label-vector	In which sub category the sexiest comment resides.

The training dataset consists of a total of 14,000 comments, which have been divided into training and validation sets for the research purposes. The authors have conducted a detailed analysis of the dataset, particularly focusing on the parametric description. One specific attribute of interest is the "Label-sexiest" category, which indicates whether a comment is classified as sexist or not. The distribution of data in this category is presented in Table II, providing insights into the prevalence of sexist comments within the dataset.

TABLE II. Value Distribution for Label-Sexiest

Label Sub category	Value counts
Not sexiest	10602
SexiestTABLE II	3398

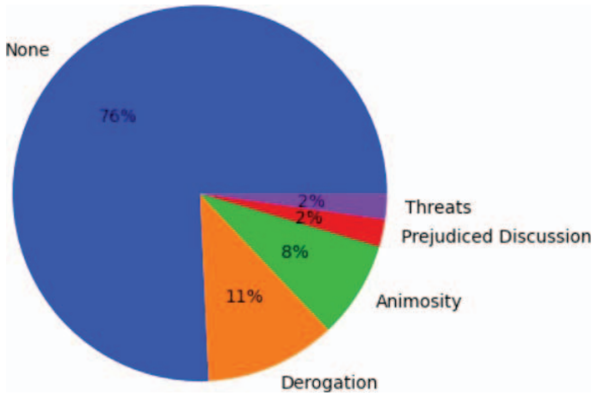

Fig. 2. Distribution of Label-category attribute

Figure 2 represents the pictorial format for the label-category attributes. Among 14000 comments, it has been observed 76% of the comments do not fall into the online sexism category. Among 3398 comments 11% comments are derogatory where the next dominant subcategory is Animosity comments. Prejudice discussions and threats share the same amount of percentage of 2%.

Next, the authors have paid attention to analyzing the Label- vector field. Figure 3 shows the data distribution for the Label-vector section. As described earlier, 10602 comments do not belong to the online sexism section. The Label-vector is subcategorized into 11 sections, where most comments are in descriptive attacks. Table III shows the data distribution of these subcategories.

TABLE III. Description of the Subcategories of Label Vectors

Sub category of Label-vector	Number of Comments
Descriptive attacks	717
Aggressive and emotive attacks	673
Casual use of gender slurs	637
Immutable gender differences	417
Supporting systemic discrimination	258
Incitement of harm and emotive attacks	254
Dehumanizing-attacks and overt sexual objectification	200
Supporting mistreatment of individual women	75
backhanded gendered compliments	64
Threats of harm	56
Condescending explanations or unwelcome advice	47
Total	3398

Finally, the attribute text has been observed, it contains 3398 comments regarding different online sexism categories of numerous lengths. There are stopwords, punctuations and other remarks available. Before feeding data into the model preprocessing is required to achieve greater results.

B. Data Preprocessing

- **Tokenization**: To provide the sentences into the BERT model, unnecessary columns have been dropped at first. After that, the sentences are tokenized to be provided in both DistilBERT and BERT models. The tokenization involves splitting the text into individual words into subword units.
- **Padding and truncating:** Both BERT and DistilBERT require all input sentences to be in the same length. In order to match the maximum length, shorter sequences must be padded with special tokens, while longer sequences must be trimmed.
- **Segment Embeddings:** BERT and DistilBERT uses segment embeddings for processing the data. So, authors have performed segment embeddings here for data preprocessing purposes.

Later, all the sentences have been converted into numerical IDs that have been directly fed to applied models.

C. Pytorch

Pytorch is a popular deep learning library that has been widely utilized for training deep learning models. Both CPU and GPU computations are supported by the Pytorch library. In this research, Pytorch is utilized for training both BERT and DistilBERT models.

IV. Research Methodologies

A. BERT Model Description

- **Origins of BERT:** BERT, an acronym for Bidirectional Encoder Representations from Transformers, is a transformative model introduced by Google AI [20]. It was groundbreaking because, unlike previous models that processed words in a sentence in a unidirectional manner (either from left to right or vice versa), BERT considers the entire context of a word by looking at it from both directions.
- **Architecture**: BERT employs the transformer architecture, an attention mechanism that enables the model to focus on specific parts of the input text. Instead of analyzing individual or adjacent words in isolation, the transformer looks at the entire text, allowing for a deeper understanding of context.
- **Pre-training and Fine-tuning:** BERT's design includes two main steps: pre-training and fine-tuning. In the pre-training phase, BERT is trained on a massive text corpus (like BooksCorpus and Wikipedia) without specific labels. It learns to predict missing words in a sentence, allowing it to understand context. In the fine-tuning phase, BERT is adapted to specific tasks (like question-answering or sentiment analysis) using labeled data, ensuring it can be tailored to various NLP applications.
- **Embeddings and Contextual Understanding:** What sets BERT apart is its ability to generate contextual embeddings. Traditional models generate the same embedding (numerical representation) for a word, regardless of its context. BERT, in contrast, considers the surrounding words, leading to different embeddings for the same word based on its usage. This nuance allows for a richer understanding of language.
- **Aplications:** Due to its deep contextual understanding, BERT has found applications in a range of NLP tasks, including but not limited to sentiment analysis, question-answering, named entity recognition, and more. Its bidirectional approach ensures that it captures nuances often missed by other models, making it a top choice for many researchers and industry professionals.
- **BERT vs. Other Models:** Compared to its predecessors and some contemporaries, BERT stands out for its accuracy. Its bidirectional context analysis combined with the vast number of parameters (344 million) ensures it achieves state-of-the-art results on many NLP benchmarks. However, this complexity also means that BERT requires significant computational resources, leading to innovations like DistilBERT, which seek to maintain performance while reducing resource demands.

The model class is inherited from torch.nn.Module.The first layer of our model is a Bert model layer that outputs sequence output and pooler output with shape (1, 768). The second layer is the drop out layer with a dropout of 0.3. And the last layer is a Linear layer with an input shape of 768 and an output of shape 12 as this paper has a total of 12 classes.

Then comes our forward function that takes input id's , attention masks, and token-type ids then feed them to the Bert Model first the output of the Bert model is fed to the dropout layer then the linear layer takes the output of the dropout layer, and outputs the predicted output which is the output of the forward function. The forward function is used by the torch.nn.Module to train, test and predict. This is the basic architecture of our model. The bidirectional training approach allows one to have a proper understanding of the data context. In this research, the hyperparameters of the BERT models are fine-tuned.

The fine-tuned hyperparameters of the BERT model aredescribed below in Table IV:

TABLE IV. Hyperparameter Tunning For DistilBERT

Name of the hyperparameters	Fine-tuned value
Maximum length	256
Train batch size	64
Validation batch size	16
Training and Validation ratio	70 to 30 Percent
Learning rate	0.01
Epochs	20
Random state	12
Loss function	BCEWithLogitsLoss

B. DistilBERT Model Description

BERT (Bidirectional Encoder Representations from Transformers) is a language model that was developed by Hugging Face. DistilBERT is a compressed and distilled version of BERT. While being smaller and easier to learn and use, it still retains the majority of the BERT's key characteristics. Here authors are focused on applying the DistilBERT architecture, as it takes much less time. Here DistilBERT model creates a student network for imitating the training procedure of a larger model. The student network is trained to use fewer parameters while minimizing the disparity between its predictions and those of the teacher network during the training phase. To do this, a distillation loss term is added to the overall loss function that was employed during training. The difference between the soft objectives is measured by the distillation loss term.

TABLE V. HYPERPARAMETER TUNING FOR BERT ARCHITECTURE

Name of the hyperparameters	Fine-tuned value
Maximum length	256
Train batch size	16
Validation batch size	16
Training and Validation ratio	70 to 30 Percent
Learning rate	0.00001
Epochs	10
K cross validation	12
Random state	42
Loss function	BCEWithLogitsLoss

C. Evaluation Metrics

To understand the performance of the architectures, authors have focused on observing certain criteria, which include F1- score, Precision, Recall and Training time for the architectures.

$$F1 - Score = \frac{TP+FP}{TP+FP+TN+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$\text{Recall}=\frac{TP}{TP+FN} \quad (3)$$

V. RESULT ANALYSIS

At first the result of the BERT model is observed, the 12 cross validation result has been taken into account. training and validation phase has been described in Table VI, The best 9 epochs have been considered in the below mentioned table.

TABLE VI. Result Analysis Of BERT Architecture

Epoch	Precision	Recall	F1-score	Validation loss	Validation F1-score
1	95.93%	73.31%	82.67%	0.000123	77.05%
2	97.23%	76.03%	84.99%	0.000416	78.24%
3	97.39%	78.58%	86.64%	0.000401	78.77%
4	97.28%	81.68%	88.58%	0.000406	78.73%
5	97.42%	85.62%	90.88%	0.000495	79.41%
6	97.70%	89.70%	93.39%	0.000476	78.95%
7	97.77%	92.29%	94.83%	0.000372	81.64%
8	98.40%	95.34%	96.79%	0.000435	80.54%
9	98.95%	96.60%	97.72%	0.000385	82.72%

A Graphical Processing Unit, or GPU, was used in conjunction with the PyTorch framework so that the BERT model could be trained. Approximately 12.76 hours were spent on the training procedure in its entirety. On the other hand, the DistilBERT model was trained using a GPU as well, although the process took a lot less time—specifically 5.36 hours—than the other two. The performance metrics of the DistilBERT model on the provided dataset are presented in Table VII. These metrics demonstrate the efficacy and efficiency of the DistilBERT model in comparison to the BERT model.

TABLE VII. DISTILBERT Model Evaluation Metrics Result

F1-Score	Final Data Loss
63.63	1.5146

So, in terms of training data F1-score and Validation data F1-score. Both cases BERT model has performed significantly well than DistilBERT model.

Finally in Table VIII, this has compared the elapsed time and electricity cost of these models. From there, this paper can see that training a BERT model is more expensive than a DistilBERT model.

Table VIII. Elapsed Time And Electricity Cost Of Bert And Distilbert

Model Name	Elapsed Time	Electricity Cost
BERT	12.76 hrs	USD 7.96
DistilBERT	5.36 hrs	USD 4.01

Comparisons with the state-of-the-art architectures have been shown in Table IX:

TABLE IX. F1 SCORE COMPARISON WITH STATE-OF-THE-ART MODELS

Model Name	F1-score in percentage
BERT	**96.25**
LSTM	57.54
GRU	61.28

In the ensuing visual representation, we compare the F1-scores of three state-of-the-art models on our dataset. The chart succinctly depicts BERT's significant performance lead over the LSTM and GRU models. The F1-scores, expressed in percentages, allow for an immediate grasp of each model's efficacy for our specific task.

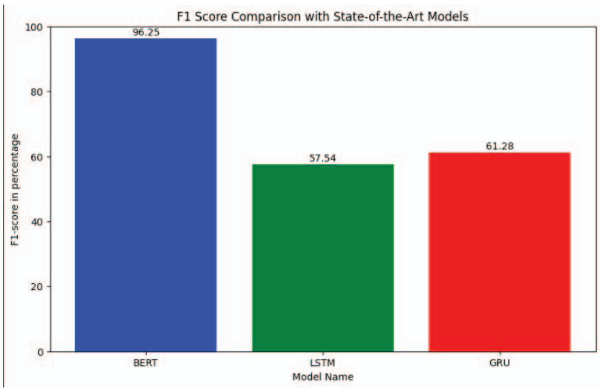

Fig. 3. Comparison of F1-scores for BERT, LSTM, and GRU models on the chosen dataset.

VI. CONCLUSION

With the aim of addressing the rampant rise in online sexism, particularly heightened in the aftermath of the Covid-19 pandemic, this research emerges as a crucial beacon in the technological and social arenas. Beyond merely highlighting the imperative to track and counteract toxic online discourse, the study adeptly leverages the capabilities of contemporary transformer models, namely BERT and DistilBERT, situating the work at the forefront of textual classification. Through an expansive multilingual analysis, it accentuates the universal issue of digital sexism and calls for globally relevant countermeasures. The nuanced examination of the balance between computational speed, as demonstrated by DistilBERT, and precision, as exemplified by BERT, provides profound implications for real-world scenarios where operational limits intersect with the demand for immediate digital surveillance. This work is a testament to the quest for cultivating a more respectful online culture, ingeniously intertwining advanced AI prowess with the overarching aspiration for a harmonious digital society.

In the study, the researchers primarily zone in on the detection of online sexism via the BERT and DistilBERT frameworks. A meticulous comparative analysis revealed the superior efficacy of the BERT model in pinpointing online sexist content, showcasing a higher F1 score. Conversely, DistilBERT marked its distinction with a swifter training process and a more economical footprint. Looking ahead, the authors aspire to incorporate an array of transformer models to gauge their respective efficacies. Furthermore, there's a keen interest in devising a dedicated deep learning framework optimized for the robust detection of online sexism.

REFERENCES

[1] Acheampong, F. A., Nunoo-Mensah, H., Chen, W. (2021). Transformer models for text-based emotion detection: a review of BERT-based approaches. Artificial Intelligence Review, 1-41.

[2] Van Aken, B., Winter, B., Lo¨ser, A., Gers, F. A. (2019, November). How does bert answer questions? a layer-wise analysis of transformer representations. In Proceedings of the 28th ACM international confer- ence on information and knowledge management (pp. 1823-1832).

[3] Jiang, Y., Sharma, B., Madhavi, M., Li, H. (2021). Knowledge distilla- tion from bert transformer to speech transformer for intent classification. arXiv preprint arXiv:2108.02598.

[4] Zhou, C., Li, Q., Li, C., Yu, J., Liu, Y., Wang, G., ... Sun, L. (2023). A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. arXiv preprint arXiv:2302.09419.

[5] Sanh, V., Debut, L., Chaumond, J., Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

[6] Bell, B. T., Cassarly, J. A., Dunbar, L. (2018). Selfie-objectification: Self-objectification and positive feedback ("likes") are associated with frequency of posting sexually objectifying self-images on social media. Body image, 26, 83-89.

[7] Fox, J., Cruz, C., Lee, J. Y. (2015). Perpetuating online sexism offline: Anonymity, interactivity, and the effects of sexist hashtags on social media. Computers in human behavior, 52, 436-442.

[8] Samghabadi, N. S., Patwa, P., Pykl, S., Mukherjee, P., Das, A., Solorio, T. (2020, May). Aggression and misogyny detection using BERT: A multi-task approach. In Proceedings of the second workshop on trolling, aggression and cyberbullying (pp. 126-131).

[9] Rallabandi, S., Singhal, S., Seth, P. (2023). SSS at SemEval-2023 Task 10: Explainable Detection of Online Sexism using Majority Voted Fine-Tuned Transformers. arXiv preprint arXiv:2304.03518

[10] Saleh, H., Alhothali, A., Moria, K. (2023). Detection of hate speech

using bert and hate speech word embedding with deep model. Applied Artificial Intelligence, 37(1), 2166719.

[11] Govers, J., Feldman, P., Dant, A., Patros, P. (2023). Down the Rabbit Hole: Detecting Online Extremism, Radicalisation, and Politicised Hate Speech. ACM Computing Surveys

[12] Bashar, M. A., Nayak, R., Suzor, N. (2020). Regularising LSTM classifier by transfer learning for detecting misogynistic tweets with small training set. Knowledge and Information Systems, 62, 4029-4054

[13] Anzivino, E., Fioriti, D., Mischitelli, M., Bellizzi, A., Barucca, V., Chiarini, F., Pietropaolo, V. (2009). Herpes simplex virus infection in pregnancy and in neonate: status of art of epidemiology, diagnosis, therapy and prevention. Virology journal, 6(1), 1-11

[14] Chiu, K. L., Collins, A., Alexander, R. (2021). Detecting hate speech with gpt-3. arXiv preprint arXiv:2103.12407

[15] Kirk, H. R., Yin, W., Vidgen, B., Ro¨ttger, P. (2023). SemEval- 2023 Task 10: Explainable Detection of Online Sexism. arXiv preprint arXiv:2303.04222.

[16] Hartvigsen, T., Gabriel, S., Palangi, H., Sap, M., Ray, D., Kamar, E. (2022). Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. arXiv preprint arXiv:2203.09509

[17] Dietterich, T. G. (2002). Machine learning for sequential data: A review. In Structural, Syntactic, and Statistical Pattern Recognition: Joint IAPR International Workshops SSPR 2002 and SPR 2002 Windsor, Ontario, Canada, August 6–9, 2002 Proceedings (pp. 15-30). Springer Berlin Heidelberg

[18] Yang, A., Zhang, W., Wang, J., Yang, K., Han, Y., Zhang, L. (2020). Re- view on the application of machine learning algorithms in the sequence data mining of DNA. Frontiers in Bioengineering and Biotechnology, 8, 1032

[19] Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., Zhang, W. (2021, May). Informer: Beyond efficient transformer for long sequence time-series forecasting. In Proceedings of the AAAI conference on artificial intelligence (Vol. 35, No. 12, pp. 11106-11115).

[20] Noor, A. K., Burton, W. S., Bert, C. W. (1996). Computational models for sandwich panels and shells

[21] Adoma, A. F., Henry, N. M., Chen, W. (2020, December). Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition. In 2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP) (pp. 117-121). IEEE