2019 ACM/IEEE Symposium on Architectures for Networking and Communications Systems (ANCS)

Online Sexism Detection Using Transformer

I. Authors Details

Team Members Name:
Sovon Chakraborty, Id 22366023; Muhammed Yaseen Morshed Adib, Id 22366020; Md. Sajeebul Islam SK, 22366027

II. 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 Programme (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.

The major contribution of this research can be summarized as Providing a deep learning-based approach for identifying online sexism instantly so that social media can be used by all types of users. Upon detection of such actions, abusers can be warned or legal actions can be taken by the respective authority. The remaining reports are organized as follows where section II shows the recent research in the respective field along with the usage of transformer models in recent times. The necessary methodologies of this research are stated in section III. The analysis of experimental results is discussed in section IV. Section V represents the conclusion and future perspective of this research in the near future.

III. Literature Review

Diverse communities use social media; as a result, the applications become wider. The amount of sexiest comments has increased significantly [11]. Jiang et al. worked on building a Chinese dataset containing online sexism comments as after Covid 19 outbreak, sexism towards Chinese people has significantly increased. No data quality measurements have not been shown in this research. Hewitt has classified misogynist tweets [12], where authors have worked on multiclass classifications. Nevertheless, a survey on misogynous tweets has been surveyed by Anzivino et al [13]. Later, sexism in other languages has been studied to propose a model that can classify online sexism instantly [14]- [15]. Traditional Machine learning algorithms face difficulties in identifying online sexism properly as the sequence information can not be preserved properly [16]- [17]. Transformer models have significant improvement in identifying online sexism precisely within a shorter period of time [18]. The main boundary here is the time needed in the training phase. Recurrent neural networks (RNN) face complexities while handling sentences with long sequences, where the vanishing gradient problem is a big issue [19]. Self-attention mechanism allows us to overcome such problems in the case of transformer models. Capturing dependencies between distant words using self-attention is regarded as an efficient way. Another significant advantage of transformer models is the pre-training phase, where models are trained on large corpora. Bidirectional Encoder Representations from Transformers (BERT) was first introduced by Google AI [20], where the amount of parameters is enormous. These more significant parameters allow the models to have computational complexities. An efficient transformer model DistilBERT has been proposed in [21], where the model is lighter than BERT with 66 million parameters. The parameters in BERT are 344 million, almost six times that of DistilBERT. The model is trained on a specific technique named distillation. Distillation involves training a model that is small, which is later utilized to mimic the behaviour of a large model. The memory requirements for DitilBERT are much lesser than the BERT model. On the contrary, BERT architecture performs significantly well in terms of accuracy. Fine-tuning and pre-training phases are smaller than other versions of BERT. In

978-1-7281-4387-3/19/$31.00 © 2019 IEEE

this research, authors are focused on applying DistilBERT and traditional BERT models and observing their performance in the particular dataset, which includes shorter training times and lesser computational complexities. The experimental result shows an excellent achievement of results where some major evaluation metrics have been used.

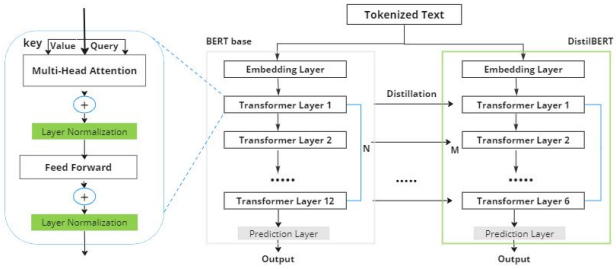

Fig. 1. DistilBERT Architecture

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.

IV. 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 can not be trained on that specific set. In the training dataset, there are several columns available. Table I shows the detailed description of the attributes.

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.

TABLE I
Attributes Names with Description

In total there are 14000 comments in the training dataset which is split for training and validation purposes. Authors have decided to look closely at the detailed parametric description. Talking about the Label-sexiest attributes, Table II shows the distribution of data on that specific category.

Label Sub category	Value counts
Not sexiest	10602
Sexiest	3398

TABLE II
Value distribution for Label-sexiest

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%.

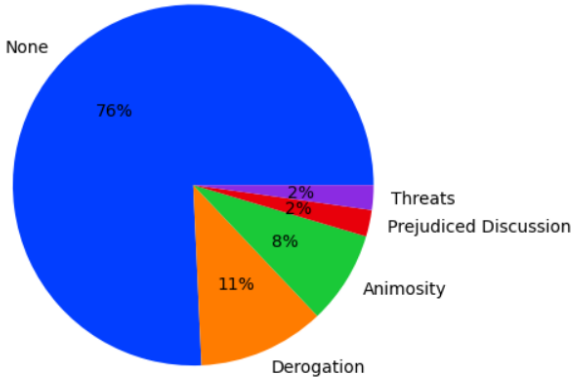

Fig. 2. Distribution of Label-category attribute

Next, the authors have paid attention to analyzing the Lable-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.

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

TABLE III
Description of the Subcategories of Label Vectors

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

1) 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.

2) 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.

3) 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.

V. Research Methodologies

A. BERT Model Description

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 dropout 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 we have 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 are described below in Table IV:

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

TABLE IV
HYPERPARAMETER TUNING FOR BERT ARCHITECTURE

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.

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

TABLE V
HYPERPARAMETER TUNING FOR DISTILBERT

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.

$$F1 - Score = \frac{TP + FP}{TP + FP + TN + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

VI. Result Analysis

At first the result of the BERT model is observed, the 12 cross validation result has been taken into account. The average F1 score, average Precision, average recall for training and validation phase has been described in the Table VI,. The best 9 epochs have been considered in the below mentioned table.

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%

TABLE VI
RESULT ANALYSIS OF BERT ARCHITECTURE

For training this BERT model, we have used Graphical processing unit (GPU) that has been run on Pytorch. On the other hand , if we observe the training time for the whole procedure was approximately 12.76 hrs.

On the other hand, DistilBERT model has also been run on the GPU but it takes much fewer time which is 5.36 hrs. Table VII shows the performance of DistilBERT model on that particular dataset.

F1-Score	Final Data Loss
63.63	1.5146

TABLE VII
DISTILBER MODEL EVALUATION METRICS RESULT

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, we have compared the elapsed time and electricity cost of these models. From there, we can see that training a BERT model is more expensive than a DistilBERT model.

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

TABLE VIII
ELAPSED TIME AND ELECTRICITY COST OF BERT AND DISTILBERT

VII. CONCLUSION

In this research, authors have focused on detecting online sexism using BERT and DistilBERT model. A comparative study has also been performed. From there it can be observed BERT model is more efficient in detecting online sexism with more F1-score on the other hand training process is much faster in the case of DistilBERT model. The DistilBERT model is also more cost efficient. Authors are planning to apply more transformer models in near future to evaluate the performances of them. Authors are also focused on proposing a deep learning architecture that will detet online sexism efficiently.

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., Löser, 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 conference on information and knowledge management (pp. 1823-1832).

[3] Jiang, Y., Sharma, B., Madhavi, M., Li, H. (2021). Knowledge distillation 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., Röttger, 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). Review 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