A GAN-BERT Based Approach for Bengali Text Classification with a Few Labeled Examples

Raihan Tanvir, MD Tanvir Rouf Shawon, Md Humaion Kabir Mehedi,

Md Motahar Mahtab, and Annajiat Alim Rasel

BRAC University, 66 Mohakhali, Dhaka - 1212
{raihan.tanvir, tanvir.rouf.shawon, humaion.kabir.mehedi,
md.motahar.mahtab}@g.bracu.ac.bd
annajiat@gmail.com
https://www.bracu.ac.bd/

Abstract. Basic machine learning algorithms or transfer learning models work well for language categorization, but these models require a vast volume of annotated data. We need a better model to tackle the problem because labeled data is scarce. This problem may have a solution in GAN-BERT. To classify Bengali text, we developed a GAN-BERT based model, which is an adapted version of BERT. We used two different datasets for this purpose. One is a hate speech dataset, while the other is a fake news dataset. To understand how the GAN-BERT and traditional BERT models behave with Bangla datasets, we experimented with both. With a small quantity of data, we were able to get a satisfactory result using GAN-BERT. We also demonstrated how the accuracy increases as the number of training samples increases. A comparison of performance between traditional BERT based Bangla-BERT and our GAN-Bangla-BERT model is also shown here, where we can see how these models react to a small number of labeled data.

Keywords: BNLP · GAN · BERT · SS-GAN · GAN-BERT

1 Introduction

Natural Language Processing is a branch of artificial intelligence that enables machines to read, comprehend, and derive meaning from sentences. Text and speech both provide a wealth of information. It's because, as intelligent beings, we communicate primarily through writing and speaking. Sentiment analysis, cognitive assistant, spam filtering, spotting bogus news, and real-time language translation are all tasks that NLP can perform for us. There are roughly 6,500 languages spoken along the different borders of the world. English is basically used as a universal language all over the globe. That is why all the extra ordinary NLP research works cover English as their base language to use in their architectures. This architecture, which is based on the English language, should be replicated in other local languages as well, because local languages are widely

spoken throughout the world. Even most of the famous social networking sites are giving options for native languages. Aside from that, news periodicals produced in native tongues are now fairly prevalent. Bangla is one of the most prominently spoken native languages. It is used by a large number of people around the world. It is spoken as first language in Bangladesh, as well as second most spoken language in India . A large number of individuals currently use social media and other networking platforms. These social media networks promote freedom of speech, but some people abuse it by expressing their venomous thoughts toward others. There has been a lot of effort put into recognizing hate speech in the Bengali language domain [10][4].

Besides, a large number of printed and online newspapers are published everyday in Bangladesh. Many rumors are circulated via news articles published in low-rated and/or unlicensed newspapers, which can lead to riots across the country. People are sometimes misled by parody news also. There is a desperate need for an automated technique to recognize this type of news. For this, a variety of algorithms and transformer-based models, such as BERT or CNN, [1] is used.

As these models rely so heavily on data, they require a huge number of tagged data. However, tagged data is difficult to come by. The GAN-Bert [3] architecture has been proposed as a solution to this problem. It performs admirably with tiny amounts of tagged data. We attempted to evaluate this model in our work using the Bengali language to see how it reacts. In our paper, we mentioned some previous papers related to our work in section II. Section III demonstrates the background architecture of the models we are working on. In section IV the structure of our model has been described. Sections V and VI will show the datasets we are using and the experiments on the datasets, respectively.

2 Previous Works

The traditional BERT architecture [5] was introduced by Jacob Devlin et al. of the Google AI Language platform. BERT is used to train a deep two way explanation from unannotated textual data by conditioning both the context in every steps. Without making significant task-specific architecture changes, specific models can be created to perform a variety of tasks by adjusting only the last layer of a pre-executed BERT network, like language reasoning or language classification. Using the MLM pre-training target, this model overcomes the unidirectionality limitation. The MLM masks some tokens arbitrarily using the fed values, aiming to forecast native lexicon id maintaining the context. Unlike pre-training for left-to-right language models, the MLM goal permits the representation to integrate the contexts of both sides. In this paper, they worked with 11 different datasets to test their model. The accuracy they found for various datasets is quite satisfactory. These findings show that deep unidirectional designs can benefit even low-resource tasks.

The architecture we are working on is the GAN-BERT [3] architecture proposed by Danilo Croce et al. introduced us with a semi supervised learning

which is based on Generative Adversarial Networks [7]. The authors presented the architecture that extends the adjusting of models like BERT with non tagged samples. Experiments revealed that the number of annotated instances needed can be considerably decreased (down to 50-100 annotated examples) while still achieving good results in a variety of sentence categorization tasks. The authors here used Bidirectional Encoder (BERT), which is a variation of transfer learning. They extended the fine-tuning of architectures similar to BERT with unannotated text samples in a generative adversarial notion. In this paper, the authors used four different datasets to compare the results of their model with other BERT base models. The results of this paper's evaluations suggest that a form of BERT can increase the robustness of such systems without adding to the inference costs. In fact, the generating network is only employed during training, and only the discriminator is used during inference.

Claudia Breazzano et al. proposed MT-GAN-BERT [2] where they followed the basic structure of BERT and built a semi-supervised model. To solve numerous tasks at once, their model employs a multi-task learning technique. The input examples are encoded using a single BERT-based model, whereas the classification steps are implemented using several linear layers, resulting in a significant reduction in processing costs.

3 Background Architectures

3.1 Semi-supervised GANs

The Generative Adversarial Network (GAN) [7] is an architecture that uses large, unlabeled datasets to train a generator network via a discriminator network. In the context of GAN [7] environment, SS-GANs [11] provide semi-supervised learning. The discriminator network is trained for $k+1$ class goal. In $\mathcal{D}$, actual samples are categorized into one of the intended k classes, while synthetically created examples are categorized into the $k+1$ class.

For discriminator $\mathcal{D}$ and generator $\mathcal{G}$, let p_D and p_G indicate the true data distribution and generated data distribution, respectively. The aim of $\mathcal{D}$ is expanded as follows for training a semi-supervised k-class classifier. Let, $p_m(\hat{y} = y \mid x, y = k+1)$ be the likelihood that a generic example x is associated with the fake class, and $p_m(\hat{y} = y \mid x, y \in (1, \ldots, k))$ express the chance that x is regarded as real, thus falling into one of the target classes. The loss function of the discriminator is defined as: $L_{\mathcal{D}} = L_{\mathcal{D}_{\text{sup.}}} + L_{\mathcal{D}_{\text{unsup.}}}$, each term is expressed as defined (1) and (2) respectively.

$$L_{\mathcal{D}_{\text{sup.}}} = -\mathbb{E}_{x,y\sim p_d} \log\left[p_{\text{m}}(\hat{y} = y \mid x, y \in (1, \ldots, k))\right] \tag{1}$$

$$\begin{aligned} L_{\mathcal{D}_{\text{unsup.}}} = &-\mathbb{E}_{x\sim p_d} \log\left[1 - p_{\text{m}}(\hat{y} = y \mid x, y = k+1)\right] \\ &- \mathbb{E}_{x\sim\mathcal{G}} \log\left[p_{\text{m}}(\hat{y} = y \mid x, y = k+1)\right] \end{aligned} \tag{2}$$

$\mathcal{D}_{\text{sup.}}$ quantifies the error in identifying the incorrect class to a real sample from the initial k categories. $\mathcal{D}_{\text{unsup.}}$ evaluates the mistake in erroneously recognizing

a real, unannotated example as fake and failing to recognize a fake example. Simultaneously, $\mathcal{G}$ is supposed to provide instances that are comparable to those drawn from the actual distribution. According to [11], $\mathcal{G}$ should generate data approximating the distribution of true data as closely as possible. Let f(x) represent the activation of a $\mathcal{D}$ intermediate layer.The $\mathcal{G}$'s feature matching loss is thus given by (3). According (3), the $\mathcal{G}$ should synthesize instances whose intermediate representations supplied as input to $\mathcal{D}$ are closely identical to the real ones. The mistakes caused by fake examples correctly identified by $\mathcal{D}$ are also taken into account in $\mathcal{G}$ loss (4). So, the overall $\mathcal{G}$ loss is defined as (5).

$$L_{G_{\text{feature matching}}} = \|\mathbb{E}_{x \sim p_d} f(x) - \mathbb{E}_{x \sim \mathcal{G}} f(x)\|_2^2 \tag{3}$$

$$L_{\mathcal{G}_{\text{unsup.}}} = -\mathbb{E}_{x \sim \mathcal{G}} \log\left[1 - p_m(\hat{y} = y \mid x, y = k+1)\right] \tag{4}$$

$$L_{\mathcal{G}} = L_{\mathcal{G}_{\text{feature matching}}} + L_{\mathcal{G}_{\text{unsup}}} \tag{5}$$

While SS-GANs are often utilized with image inputs, in [3], it is shown that they can be exploited in combination with BERT [5] over inputs encoding linguistic information.

3.2 GAN-BERT: Semi-supervised GAN with BERT

BERT [5] is a transfer learning strategy that involves pretraining a model on generic tasks and then fine-tuning it on the specified tasks. Transfer learning has been proven to be useful in a variety of computer vision tasks [6]. BERT is an immensely deep model that is pretrained on enormous corpora of raw texts before being fine-tuned with target-labeled data. The transformer is the core of BERT, which is an attention-based [12] system that adapts contextual connections between words or sub-words.

BERT offers context-dependent embeddings of the words that construct phrases, as well as a sentence embedding that encodes semantics at the sentence level. BERT's pre-training is designed to collect such data by utilizing incredibly large corpora. Following pre-training, BERT enables the encoding of (i) individual words in a sentence, (ii) the complete sentence, and (iii) sentence pairings in specialized embeddings. These may be utilized as input to other layers to perform tasks such as sequence labeling, sentence classification, and relational learning. This is achieved by utilizing labeled examples to build task-specific layers and fine-tuning the overall architecture.

The capacity of BERT is extended in GAN-BERT [3] by using SS-GANs [11] for the fine-tuning step. Two components are added to an already trained BERT model to fine-tune it, one is the layers for goal tasks, ii) SS-GAN layers, which facilitate semi-supervised learning.

The SS-GAN architecture was developed on top of BERT by including i) a discriminator $\mathcal{D}$ for categorizing the samples and ii) an adversarial generator $\mathcal{G}$. $\mathcal{G}$ is a MLP that receives a 100-dimensional noise vector drawn from $N\left(\mu, \sigma^2\right)$ as input and produces a vector h_{fake} in real data distribution. The discriminator is also a MLP that accepts a vector h_* in real data distribution as input, where

h_* can either h_{fake} generated by the generator or h_{CLS} for unlabeled or labeled instances from the real distribution. As explained in SS-GAN [11], the last layer of $\mathcal{D}$ is activated by a soft-max function that produces a $k+1$ vector of logits.

$\mathcal{D}$ should assign the actual instances into one of the k classes during the forward step. Each samples should be predicted into $k+1$ class when they are fake. As discussed in section 2, the training process attempts to optimize two opposing losses, namely L_D and L_G.

Unlabeled instances only contribute to (2). This means, they are considered in the loss computation only if they are erroneously categorized into the $k+1$ category. Their contribution to the loss is omitted for the rest of the cases. As a result, the annotated samples only contribute to (1). Finally, the examples generated by $\mathcal{G}$ contribute to both (2) and (4). So, $\mathcal{D}$ is punished for not detecting samples generated by $\mathcal{G}$ and otherwise. Modifications to the BERT weights are made when updating $\mathcal{D}$, in order to fine-tune its inner representations, taking into consideration both labeled and unlabeled data. After training is completed, $\mathcal{G}$ is excluded, but the rest of the original BERT model is retained for inference. This indicates that, no additional cost is required at inference time.

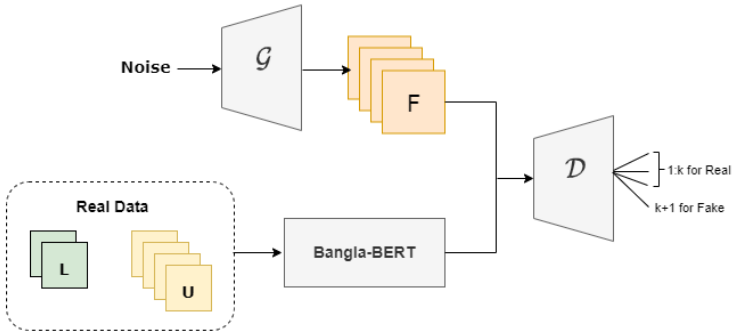

Fig. 1: SS-GAN on top of a pretrained Bangla-BERT model; $\mathcal{G}$ produces fake samples (F), $\mathcal{D}$ detects whether a sample is fake on not from labeled (L) and unlabeled (U) real data.

4 GAN-Bangla-BERT

In this paper, we have proposed a technique for implementing semi-supervised GAN [11] for different downstream tasks on top of a pretrained BERT model following the techniques demonstrated in [11]. For implementing our idea, we have used a pretrained BERT based model for the Bengali language. The Bangla-Bert-BaseBangla-BERT-Base[1] is used as the base model. It's a pretrained language model of Bengali language based on mask language modeling described

[1] `https://huggingface.co/sagorsarker/bangla-bert-base`

in BERT [5]. Currently, the Bangla-BERT-Base model follows the bert-base-uncased model architecture.We employed semi-supervised learning using a GAN framework on top of Bangla-BERT as demonstrated in [3]. The architecture of our proposed model is illustrated in Fig. 1

5 Dataset

Two different dataset were used to evaluate the efficiency of our model on two different downstream task. Initially, we explored the capability of SS-GANs [11] for bangla fake news detection task. For this purpose, we used the *BanFakeNews* [8] dataset, to train the models. The dataset contain more than 50000 news in Bengali, among which only 7000 are labeled authentic and 1299 are labeled as 3 types of spurious samples. We took 1300 random samples from the authentic part and all the data from the false part for our models. We also experimented

Table 1: Different class labels with their corresponding frequency in *BankFake-News* Dataset (**left**) and *Bengali-Hate-Speech-Dataset* (**right**) respectively

Class label	**Frequency**	**Class label**	**Frequency**
Satire	1136	Personal hate	2189
Click-baits	82	Political hate	1738
Fake	81	Religious hate	957
Authentic	7202	Geopolitical hate	814

with Bangla hate speech detection tasks to explore the capability of GAN-BERT. To evaluate our technique, we used the Bengali-Hate-Speech-Dataset [9]. This dataset has 4 different categories. It has two versions. We have used the version 2.0. It contains a total of 175 distinct abusive terms. The frequency of each category in the both dataset is shown in table 1.

6 Experiments and Results

In this section, we will discuss about various approaches conducted for analyzing performances of our proposed techniques on two downstream tasks with varying training settings like number of training examples and others.

6.1 Bangla Fake News Detection Task

With the aim of fine tuning the model described earlier, we trained the networks using *BanFakeNews* dataset. Among the samples, only the labeled instances were used for feeding them into generative networks as labeled samples. Training was performed using four different amount of annotated data, which were $64, 128, 256$ and 512. In Fig. 2a, the accuracy of the model trained with different numbers

of labeled samples are shown. The obtained performance metrics shows that, the achieved result is comparable with state-of-the-art models with many fewer percentages of labeled data. Some sample predictions are provided below:
"নির্মাতা গিয়াস উদ্দিন সেলিম কথা বলেছেন চ্যানেল আই অনলাইনের সাথ 'মনপুরা' থেকে 'স্বপ্নজাল', চলছে নতুন চলচ্চিত্র 'অপারেশন জ্যাকপট'-এর কাজ এবং বাংলাদেশের সিনেমা নিয়ে বলেছেন তার দর্শনের কথা " - which is authentic news and our model also predicts it as authentic.
"চীনা ভাষা মান্দারিনকে সরকারি ভাষার স্বীকৃতি দেয়ার প্রস্তাব দিয়েছে পাকিস্তান পার্লামেন্টের উচ্চকক্ষ " - this misleading news is correctly predicted as fake by our model.
"হ্যাঁ..অবিশ্বাস্য শুনতে হলেও এটাই সত্যি!! এরকম ঘটনা ঘটেছে পাকিস্থানে ১৯ বছরের সেই যুবক হস্তমৈথুন করে নিজের নাম গিনেস বুকে লিখিয়েছে "- now this news is actually a clickbait, but the prediction is satire.

So our model predicts whether the news is fake or authentic very competently, but in some cases it makes mistakes in detecting the type of fake news it is, like in the previous example, deciding whether it's just misleading fake news, satire or clickbait.

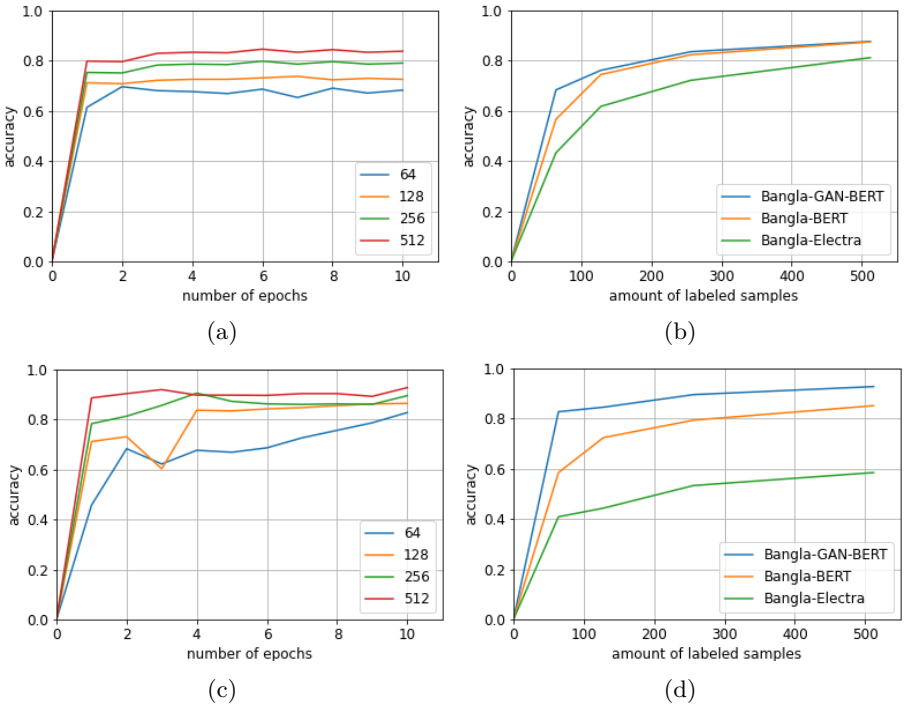

Fig. 2: Accuracy of Bangla-GAN-BERT on BanFakeNews (**top**) and BanglaHate-Speech (**bottom**) dataset for different number of epochs (**left**) and comparison with Bangla-BERT-Base and Bangla-Electra (**right**) respectively for different amount of labeled data:64,128,256,512

We compared the results with Bangla-BERT-Base[1] and Bangla-Electra[2]. For this purpose, we trained both the models with the same dataset, keeping the same settings as our proposed techniques. A comparison of accuracy among the models are shown in Fig. 2b. When only a very small amount of labeled data, i.e: 64 samples, are provided as annotated examples, our model outperforms both the Bangla-BERT-Base and Bangla-Electra by a margin of 17.13% and 36.7% respectively. Then when the samples are doubled the performance gap reduced to 2.19% and 18.8% respectively. And, when the number of labeled samples are increased to 512 the Bangla-BERT-Base models catches up with our proposed model, while Bangla-Electra still lag behind with 7.36% less accuracy.

Table 2: Performance metrics for Bangla-GAN-BERT on BanFakeNews *top* and Bengali-Hate-Speech *bottom* dataset respectively. and for different number of training example

Size	**Accuracy(%)**	**Precision**	**Recall**	**F1**
64	68.2	0.632	0.682	0.656
128	73.2	0.679	0.732	0.705
256	73.8	0.722	0.738	0.719
512	75.4	0.739	0.754	0.734
64	85.4	0.958	0.854	0.882
128	89.1	0.941	0.891	0.907
256	91.4	0.867	0.914	0.886
512	92.6	0.857	0.926	0.890

6.2 Bangla Hate Speech Detection Task

We have exploited the efficiency of GAN-BERT for Bangla hate speech detection. For this purpose, we fine tuned the Bangla-BERT-Base [1] model with the Bengali-Hate-Speech-Dataset [9] using generative adversarial networks. Like the previous approach, we also followed the same mechanism here. From the entire dataset, we sampled four different amount of examples and trained the model. Since the dataset is more compact, the generative models capture the real data distribution more accurately compared to the BanFakeNews [8] dataset, hence the model prediction capability has improved and become more efficient in detecting hate speech in Bangla. Some samples of predictions made by the trained model with the Bengali-Hate-Speech-Dataset are shown below:

"কর্পোরেট নারীরা কী পতিতা-দের থেকে কম?" - this quote is a personal hate speech and our model correctly predicted so.

"মালাউনের বাচ্চারা আর কোনোদিন ভালো হবেন - it's predicted as personal hate, but the true label is religious hate speech. Due to the contextual information model, it didn't produce the same output as the ground truth.

[2] `https://huggingface.co/monsoon-nlp/bangla-electra`

"বোরকা পরা নারীদের দেখে বুঝা যায় না, তারা মানুষ, ভূত না জানোয়ার, বলেছেন লিটন" - this is an expression of religious hate and predicted accurately by the model. In Fig. 2c, the accuracy of the model trained with different amount of labeled samples are demonstrated. From Fig. 2d, we could see that there is a large gap between the accuracy of the models. Our GAN-BERT based model surpasses both the models by a large margin. While it outperformed Bangla-BERT-Base by a difference of 29% to 18.07% for training the models with 64 to 512 labeled samples, the Bangla-Electra [2] has failed to obtain comparable results by a consistent gap in accuracy of around 45.5%. The confusion matrix for 512 training example is shown in Fig. 3 and accuracy, precision, recall and F1 score value of *BanFakeNews* and Bengali-Hate-Speech-Datase for different size of the training samples is given in table 2. So, clearly our model has proven the capability of semi-supervised GAN with resource constrained environments.

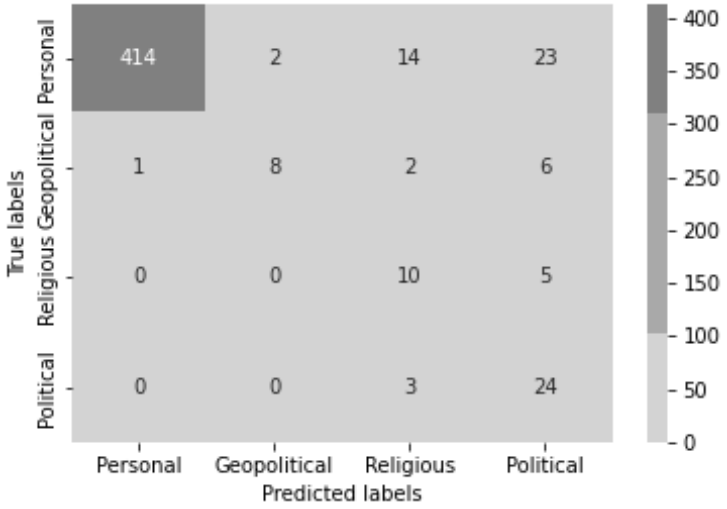

Fig. 3: Confusion matrix of Bangla-GAN-BERT on Bengali-Hate-Speech dataset for 512 training example

7 Conclusion

In this work, we have exploited the capability of GAN-BERT for BNLP with low resource settings. From the experiments, it is evident that, fine tuning a BERT architecture with small amount of annotated data yields poor performance. In such resource-constrained conditions, a generative adversarial network-based model may come in handy. From the experimental results, it is found that our fine tuned model on top of a Bangla-BERT-Base achieved comparable results

with state-of-the-art models using very limited labeled data and outperforms the baseline model[1] and Bangla-Electra[2] by a considerably wide margin. We wish to employ this methodology for many other downstream tasks in BNLP which are deprived due to a scarcity of resources.

References

1. Adib, Q.A.R., Mehedi, M.H.K., Sakib, M.S., Patwary, K.K., Hossain, M.S., Rasel, A.A.: A deep hybrid learning approach to detect bangla fake news. In: 2021 5th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT). pp. 442–447 (2021). https://doi.org/10.1109/ISMSIT52890.2021.9604712
2. Breazzano, C., Croce, D., Basili, R.: Mt-gan-bert: Multi-task and generative adversarial learning for sustainable language processing (2021)
3. Croce, D., Castellucci, G., Basili, R.: GAN-BERT: Generative adversarial learning for robust text classification with a bunch of labeled examples. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 2114–2119. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.191, `https://aclanthology.org/2020.acl-main.191`
4. Das, A.K., Al Asif, A., Paul, A., Hossain, M.N.: Bangla hate speech detection on social media using attention-based recurrent neural network. Journal of Intelligent Systems **30**(1), 578–591 (2021)
5. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. CoRR **abs/1810.04805** (2018), `http://arxiv.org/abs/1810.04805`
6. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation (2014)
7. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks (2014)
8. Hossain, M.Z., Rahman, M.A., Islam, M.S., Kar, S.: BanFakeNews: A dataset for detecting fake news in Bangla. In: Proceedings of the 12th Language Resources and Evaluation Conference. pp. 2862–2871. European Language Resources Association, Marseille, France (May 2020), `https://aclanthology.org/2020.lrec-1.349`
9. Karim, M.R., Chakravarti, B.R., P. McCrae, J., Cochez, M.: Classification benchmarks for under-resourced bengali language based on multichannel convolutional-lstm network. In: 7th IEEE International Conference on Data Science and Advanced Analytics (IEEE DSAA,2020). IEEE (2020)
10. Karim, M.R., Dey, S.K., Islam, T., Sarker, S., Menon, M.H., Hossain, K., Hossain, M.A., Decker, S.: Deephateexplainer: Explainable hate speech detection in under-resourced bengali language. In: 2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA). pp. 1–10 (2021). https://doi.org/10.1109/DSAA53316.2021.9564230
11. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. p. 2234–2242. NIPS'16, Curran Associates Inc., Red Hook, NY, USA (2016)
12. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need (2017)