Multilingual hate speech detection using semi-supervised Generative Adversarial Network

Khouloud Mnassri, Reza Farahbakhsh, Noel Crespi

Multilingual Hate Speech Detection using Semi-Supervised Generative Adversarial Network

Khouloud Mnassri, Reza Farahbakhsh, Noel Crespi

Samovar, Telecom SudParis, Institut Polytechnique de Paris, 91120 Palaiseau, France.
{khouloud.mnassri,reza.farahbakhsh,noel.crespi}@telecom-sudparis.eu

Abstract. Online communication has overcome linguistic and cultural barriers, enabling global connection through social media platforms. However, linguistic variety introduced more challenges in tasks such as the detection of hate speech content. Although multiple NLP solutions were proposed using advanced machine learning techniques, data annotation scarcity is still a serious problem urging the need for employing semi-supervised approaches. This paper proposes an innovative solution— a multilingual Semi-Supervised model based on Generative Adversarial Networks (GAN) and mBERT models, namely SS-GAN-mBERT. We managed to detect hate speech in Indo-European languages (in English, German, and Hindi) using only 20% labeled data from the HASOC2019 dataset. Our approach excelled in multilingual, zero-shot cross-lingual, and monolingual paradigms, achieving, on average, a 9.23% F1 score boost and 5.75% accuracy increase over baseline mBERT model.

Keywords: Hate Speech, offensive language, semi-supervised, GAN, mBERT, multilingual, social media.

1 Introduction

Social media platforms like Twitter and Facebook have been growing in popularity in recent years as means of communication and connection. Unfortunately, an increasing concern has been illustrated, along with this expansion, that many people have reported encountering hate speech and offensive content on these platforms [1]. In fact, due to the anonymity provided by these tools, users are becoming more free to express themselves, and sometimes engaging in hateful actions [2]. In addition, offensive content is no longer restricted to human scripting, but it's crucial to acknowledge that Generative AI and Large Language Models (LLMs) can also generate it, which emphasizes further the need for robust content moderation. Moreover, due to the enormous volume of multilingual content spread online, it has become more difficult to manually regulate it. However, there have been several initiatives to automate the detection of hateful and offensive content in multilingual settings, which remains a challenging task [3]. Indeed, most of the existing machine learning solutions (monolingual and

multilingual) have used supervised learning approaches [3], where transfer learning techniques, based on pre-trained Large Language Models LLMs, have proven to give outstanding results. In fact, Transformer-based architectures, such as BERT (Devlin et al., 2019 [4]), have been demonstrated to achieve state-of-the-art performance in a variety of hate speech detection tasks. As a result, a large number of BERT-based approaches were presented in this field [5–8] etc. Moreover, multilingual transformers, particularly mBERT (multilingual BERT), have been implemented in the multilingual domain. This model has provided cutting-edge performance in cross-lingual and multilingual settings, where several studies demonstrate its usefulness in many languages especially in low-resource ones [9] etc. While these approaches have made remarkable advances, they still have difficulties obtaining enough annotated data, which is further complicated in multilingual hate speech detection tasks. More specifically, acquiring such high-quality labeled corpora is expensive and time-consuming [10]. Adding to that, multilingual robust models often depend on enormous linguistic resources, which are mostly available in English (as a rich-resource language). As a result, these models encounter generalization issues that yield decreased performance when used with low-resource languages [11].

As a solution for these deficiencies, Semi-Supervised SS-Learning was introduced in order to decrease the need for labeled data. It enables building generalizable efficient models with unlabeled corpora using only small-sized of annotated samples. Thus, SS-Learning was largely used in NLP for hate speech detection tasks [12, 13]. One of these SS techniques is Generative Adversarial Network (GAN) [14], which is based on an adversarial process, where a "discriminator" learns to distinguish between real and generated instances, produced by a "generator" that mimics data based on a distribution. An extension of GANs is Semi-Supervised SS-GANs, where the "discriminator" also allocates a class to each data sample [15]. It becomes a remarkable solution in semi-supervised learning in hate speech detection, widely used combined with pre-trained Language models like SS-GAN-BERT [16] (non-English language).

In this paper we propose a semi-supervised generative adversarial framework, in which we incorporate mBERT for multilingual hate speech and offensive language detection, and we hereby refer to the introduced model as SS-GAN-mBERT. This procedure leverages mBERT's ability to generate high-quality text representations and to adapt to unlabeled data, contributing to enhancing the GAN's generalization for hate speech detection in multiple languages. Even though GAN-BERT has been utilized for different non-English languages in NLP, the semi-supervised GAN-mBERT approach remains underexplored specially in multilingual hate speech detection. Therefore, this study aim to fill this gap by proposing the SS-GAN-mBERT model for hate speech and offensive language detection across English, German, and Hindi. The key contributions are as follows:

- We proposed an SS-GAN-mBERT model, in multilingual and cross-lingual settings, and we compared with baseline semi-supervised mBERT, evaluating the impact of adopting GAN on improving pre-trained models' performance.

- Training across three scenarios: multilingual, cross-lingual (zero-shot learning), and monolingual, in order to examine linguistic feature sharing within Indo-European languages and prove their crucial role in enhancing text classification tasks.
- Exploration of SS-GAN's progressive influence in improving performance through iterative labeled data increase in a multilingual scenario.

2 Literature Survey

2.1 GAN for Hate Speech detection

In order to address the challenge of imbalance labeling with hateful tweets, Cao et al. [17] (2020) presented HateGAN, a deep generative reinforcement learning network. Inspired by Yu et al. (2017) [18] (SeqGAN), their reinforcement learning-based component encourages the generator to produce more hateful samples in English by introducing a reward policy gradient to direct its generation function. Their results indicate that HateGAN enhances hate speech identification accuracy. Although their contribution in implementing reinforcement learning, there wasn't a detailed explanation of its influence on the model's performance, nor a significant improvement in the results. Therefore, we won't consider this method in our approach for the moment.

2.2 GAN-BERT

GAN-BERT was first introduced by Croce et al. [19] (2020) as a viable solution to deal with the lack of annotated data. They've seen that using semi-supervised learning could be beneficial in this case in order to improve the generalization performance within the availability of little amount of labeled data. As a result, they proposed GAN-BERT, an extension of BERT model combined with generative adversarial network and fine-tuned on labeled and unlabeled data. They implemented their model on several classification datasets, and they found that the performance of their semi-supervised model gets better every time increasing the size of labeled dataset. Moreover, Jiang et al. [20] used CamemBERT, and ChouBERT in order to build GAN-BERT models. They also worked on examining varied losses over changing the number of labeled and unlabeled samples in the training French datasets in order to provide greater understanding into when and how to train GAN-BERT models for domain-specific document categorization. Adding to that, Jain et al. [21] worked on consumer sentiment analysis using GAN-BERT within aspect fusion. They extracted several service features from consumer evaluations and merged them with word sequences, before feeding them into the model.

2.3 GAN-BERT for Hate Speech detection

Ta et al. [22] handled the Detection of Aggressive and Violent INCIdents from Social Media in Spanish (DAVINCIS@IberLEF2022). In order to increase the dataset size, they used back translation for data augmentation, implementing the models of Helsinki-NLP. By translating the original tweets in Spanish to English, French, German, and Italian, then translating them back to English to be used in

the BERT-based model, they managed to balance the dataset and fill the violent label deficiency. Moreover, working on Bengali both hate speech and fake news detection, Tanvir et al. [16] used Bangla-BERT based GAN-BERT model. They compared its performance with Bangla-BERT baseline, to interpret the benefit of implementing GAN, especially on a small amount of data samples. In addition, Santos et al. [23] proposed an ensemble of two semi-supervised models in order to automatically generate a hate speech dataset in Portuguese with reduced bias. The first model incorporates GAN-BERT network, where they used Multilingual BERT and BERTimbau, while the second model is based on label propagation to propagate labels from existing annotated corpora to unlabeled dataset.
Overall, the existing hate speech detection methods based on GAN-BERT have shown effectiveness, especially in languages apart from English. These approaches have focused on languages such as Spanish, Portuguese, and Bengali, and have used personalized BERT variants that were pre-trained specifically for these languages, working on monolingual approaches. The goal of our paper is to build a multilingual BERT-based semi-supervised generative adversarial model. This method involves simultaneously training in many languages, including English, German, and Hindi within labeled and unlabeled data, in order to share linguistic features. The primary goal of this research is to determine the influence of GAN-based algorithms in the context of multilingual text classification, with a particular emphasis on their performance on unlabeled datasets.

3 Methodology

3.1 Semi-Supervised Generative Adversarial Network: SS-GAN

Starting with understanding the general concept of Generative Adversarial Networks, GAN was first introduced by Goodfellow et al., 2014 [14], composed basically from two components: a "generator" (G) and a "discriminator" (D). During training, the generator generates synthetic data while the discriminator determines whether the data is real or fake. In this context, G aims to generate data samples that increase the difficulty for D to recognize them from real data, whereas the latter aims to enhance its capacity to distinguish between these data samples. As a result, G generates progressively more realistic data. After that, Salimans et al. [15] introduced, in 2016, Semi-Supervised SS-GANs, a variant of GANs that enables semi-supervised learning in GAN network, which means that D allocates also a label to the data samples. Overall, Table 1 sums up a simple illustration of the roles and related loss functions in mathematical formulas of both GAN's D and G. First of all, let p_{real} and p_g denote the real data and generated data distribution respectively, $p(\hat{y} = y|x, y = k+1)$ the probability that a sample data x is associated with the fake class, and $p(\hat{y} = y|x, y \in (1...k))$ the probability that x is considered real.

3.2 SS-GAN-mBERT

Starting with a pre-trained mBERT model[1], we fine-tuned it by adding GAN layers for semi-supervised learning. More specifically, assuming we are working

[1] https://github.com/google-research/bert/blob/master/multilingual.md

Table 1. Roles and Loss Functions for the Discriminator D and Generator G in SS-GAN frameworks

	D	G
Role	Training within $(k+1)$ labels, D assigns "real" samples to one of the designated $(1, ..., k)$ labels, whereas allocating the generated samples to an additional class labeled as $k+1$.	Generating samples that are similar to the real distribution p_{real} as much as possible.
Loss function	$L = L_{\text{sup}} + L_{\text{unsup}}$ where: $L_{\text{sup}} = -E_{x,y\sim p_{\text{real}}} \log[p(\hat{y} = y\|x, y \in (1, \ldots, k))]$ and $L_{\text{unsup}} = -E_{x,y\sim p_{\text{real}}} \log[1 - p(\hat{y} = y\|x, y = k+1)]$ $-E_{x\sim G} \log[p(\hat{y} = y\|x, y = k+1)]$	L is the error of correctly identifying fake samples by D $L = L_{\text{matching}} + L_{\text{unsup}}$ where: $L_{\text{matching}} = \|E_{x\sim p_{\text{real}}} f(x) - E_{x\sim G} f(x)\|_2^2$ and $L_{\text{unsup}} = -E_{x\sim G} \log[1 - p(\hat{y} = y\|x = k+1)]$

L_{sup} is the error in wrongly assigning a label to a real data sample.
L_{unsup} is the error in wrongly assigning a fake label to a real (unlabeled) data sample.
$f(x)$ represents the activation or feature representation on an intermediate layer of D.
$L_{matching}$ is the distance between the feature representations of real and generated data.

on classifying a sentence $s = (s1, ..., sn)$ over k classes, mBERT outputs an $n+2$ vector representations in R_d: (h_{CLS}, $h_{s1}...h_{sn}$, h_{SEP}). As a result, h_{CLS} representation will be used as a sentence embedding for our classification task. As illustrated in Figure 1, we combined the GAN architecture on top of mBERT by including an adversarial generator G and a discriminator D for final classification.

We took both G and D as a multi-layer perception MLP. First of all, G takes a 50-dimensional noise vector and generates a vector $h_{fake} \in R_d$. Then, this vector can be received by the discriminator D along with the representation vector of real data (labeled and unlabeled) produced by mBERT: h_{CLS}. After that, the last layer of the discriminator D, which is a softmax activation layer, will output 3 vectors of logits (for the 3 classes for our task: 'hateful and offensive', 'normal', and 'is real or fake?' classes). More specifically, during traning, if real data are sampled ($h = h_{CLS}$), D will classify them into the 2 classes of the hateful data ('hateful and offensive' or 'normal'), otherwise, if $h = h_{fake}$, D will classify them into all of the 3 classes.

No cost at inference time: During the inference phase, the generator G is no longer utilized after training, but the remainder of the original mBERT model and the discriminator D are maintained for classification (inference phase). This means that utilizing the model for final classification doesn't require any additional computational resources overhead [19].

4 Experiments and Results

4.1 Dataset

In the HASOC track at FIRE 2019, Mandl et al. [24] created an Indo-European Language corpora for Hate Speech and Offensive Content identification, extracted from Twitter and Facebook. They provided three publicly available datasets[2] in English, German, and Hindi, which presents respectively 40.82%, 26.63% and 32.54% of the total training dataset. For each language, they provide the train and test datasets labeled in three subtasks. In the first subtask, the data is

[2] https://hasocfire.github.io/hasoc/2019/

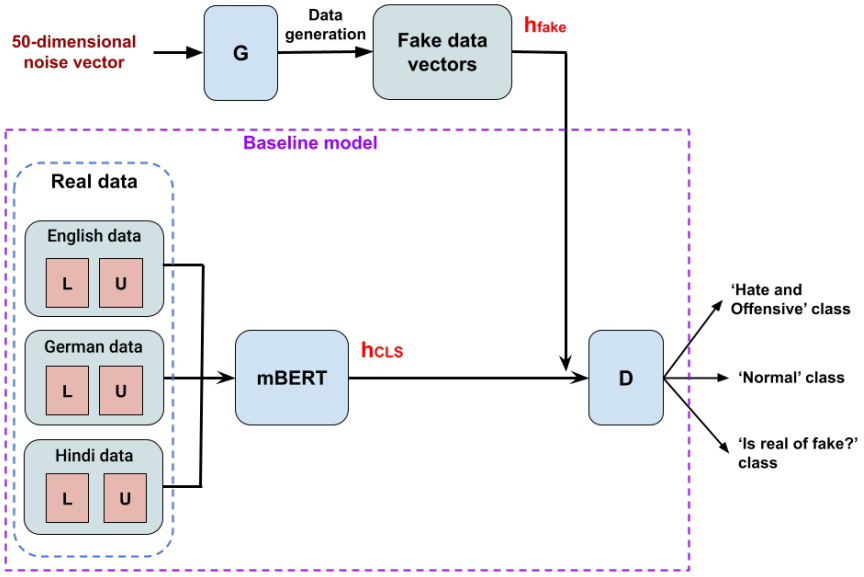

Fig. 1. Structure of SS-GAN-mBERT model for multilingual hate speech detection. "L" refers to labeled data subset & "U" refers to unlabeled data subset. Given a random noise vector, The GAN generator G generates fake data samples and outputs vectors $h_{fake} \in Rd$, which are used as input to the discriminator D, along with the representations of L and U data executed by mBERT as $h_{CLS} \in Rd$ vectors for each of the given languages.

binary labeled into (HOF) Hate and Offensive, and (NOT) Non Hate-Offensive. Figure 2 displays the class distribution of each language in this training dataset. As for the test set, English contain 34,71%, German 25.59%, and Hindi 39.68%. In our work, we considered the first subtask. Taking the training set, we divided it into 80% (∼11.5k) for the Unlabeled set (U), and 20% (∼3k) for the Labeled set (L), keeping the same class distribution. We selected this division because we aim to prove the efficiency of using GAN to train on small-size labeled datasets. We also present the evolution of our SS-GAN-mBERT model's performance (F1 macro score) using progressive percentages of labeled dataset L. We analyze the influence of increasing this amount of data in order to prove the importance of implementing GAN within a pre-trained language model to be efficient enough with the least amount of labeled data. This means that even with few annotated samples, SS-GAN-mBERT can give pretty good classification results, unlike using pre-trained language models alone, which require a lot of annotated datasets to be able to give similar performance.

4.2 Experiments and Analysis

Training scenarios We are focusing on training two models, SS-GAN-mBERT and baseline semi-supervised mBERT. First of all, as part of our multilingual approach, our training process will consider all three languages of our dataset

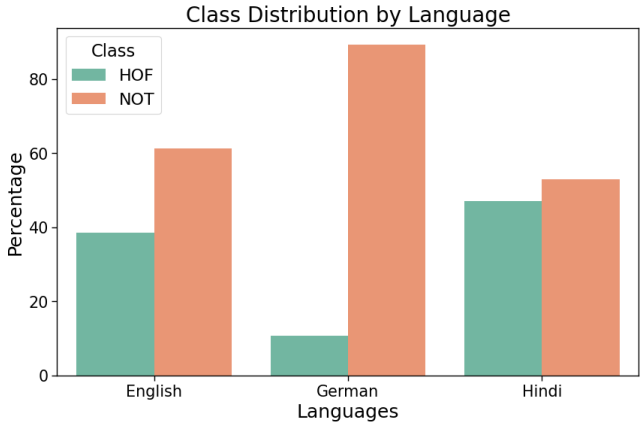

Fig. 2. Class distribution over languages in HASOC2019 training dataset. Note: In this corpora, English presents 40.82%, German 26.63%, and Hindi 32.54%.

(English, German, and Hindi). Utilizing linguistic features and patterns that are shared across these languages, we aim to analyze the influence of this method on our model performance. As a result, we will evaluate model results for each language separately using separate test sets provided by HASOC2019 for our evaluation process. Adding to that, we will consider a cross-lingual scenario, we will train our models on the English dataset because of its rich linguistic resources and its size compared to the other two languages. Then, using a zero-shot learning paradigm for the other two languages, we will evaluate these models. Lastly, by training models separately on each language, we are investigating the monolingual scenario. This method contributes to a richer understanding of model behavior across many linguistic contexts by providing insights into the complexities and difficulties unique to each language.

Models implementation Based on the computational resources used in the training process, we made the architecture of GAN as simple as possible. In fact, the Generator is implemented as a Multi-Layer Perceptron MLP with one hidden layer, it is used to generate fake data vectors. More specifically, it transforms noise vectors, which are extracted from a standard normal distribution $N(0, 1)$ (Where its values are sampled from the standard normal probability distribution with a mean (μ) of 0 and a standard deviation (σ) of 1). The generator consists of a linear layer that transforms the input noise vector of size 50 to a hidden size vector of 512, followed by a 0.2 LeakyReLU activation layer and a dropout layer with a rate of 0.1.
Similar to the generator, the discriminator is another MLP with one hidden layer, it is composed of a linear transformation layer with a 0.2 leakyReLU activation, followed by dropout layer (of rate 0.1). The linear layer outputs class logits with 3 outputs including a separate class for fake/real data. These logits are then directed to a softmax activation layer in order to derive class probabilities. This architecture is used for our final classification task.

To build our SS-GAN-mBERT model, we used "BERT-Base Multilingual Cased"[3]: trained on 104 languages, this transformer is composed of 12 layers, 768 hidden size and 12 attention heads, and it has 110M parameters. We selected the model state 'Cased' as it's mainly suggested for languages with non-Latin alphabets (e.g. Hindi). Moreover, our models have been implemented using Pytorch[4] and trained using batch size of 32 on Google Colab Pro[5] (V100 GPU environment with 32 GB of RAM). We set the maximum length variable to 200, and we train our models on 5 epochs, with a learning rate of $1e-5$ and AdamW optimizers for both the discriminator and the generator. We used Accuracy and F1 macro score as evaluation metrics to measure our models results displayed in Table 2.

Table 2. Results of SS-GAN-mBERT in monolingual, cross-lingual and multilingual training on HASOC2019 dataset.

		English		**German**		**Hindi**	
		Acc.	**F1**	**Acc.**	**F1**	**Acc.**	**F1**
Monolingual training	**Baseline mBERT**	0.638	0.601	**0.842**	0.485	0.696	0.693
	SS-GAN-mBERT	0.731	0.673	0.811	0.538	0.754	0.754
Cross-lingual training	**Baseline mBERT**			0.657	0.502	0.567	0.557
	SS-GAN-mBERT			0.704	0.561	0.636	0.63
Multilingual training	**Baseline mBERT**	0.736	0.699	0.820	0.583	0.737	0.736
	SS-GAN-mBERT	**0.753**	**0.708**	0.771	**0.609**	**0.783**	**0.783**

In cross-lingual training, we implement zero-shot learning: training on English and testing on German and Hindi.

Results & Analysis Considering the three training paradigms: Monolingual, zero-shot Cross-lingual, and Multilingual, the results in Table 2 illustrate that SS-GAN-mBERT consistently outperforms the baseline mBERT. In the context of multilingual training scenario, SS-GAN-mBERT proved to be an effective option for improving performance, achieving the highest overall results, compared to monolingual and cross-lingual training. The model shows 6.5% increase in accuracy and 6.4% rise in F1 score in Hindi, compared to the baseline model. These results highlight the model's efficiency in employing multilingual data to improve its linguistic representation and hence increase its classification capability. The same improvement is highlighted in zero-shot cross-lingual training, where SS-GAN-mBERT demonstrated the highest results getting to ∼12% increase in both the accuracy and in F1 macro score for Hindi. This doesn't hide both of the models' remarkable results in the monolingual paradigm getting the most increased accuracy of ∼84% in German. Overall, with SS-GAN-mBERT continually surpassing the baseline in all training situations, this underlines the effectiveness of adversarial training in improving the model's capacity to recognize fine-grained linguistic features, which proved to be enhanced further with the increase in the number of languages. Adding to that, since we're dealing

[3] https://github.com/google-research/bert/blob/master/multilingual.md
[4] https://pytorch.org/
[5] https://colab.research.google.com/signup

with an imbalanced dataset, we're considering F1 macro scores to analyze the performance of our models, thus, comparing the languages output, we can say that our models gain the highest performance in Hindi. This can be due to the size of the corresponding dataset, which is bigger than German, so it's reasonable why getting the lower performance for the latter language.

5 Discussions and Future directions

5.1 Discussions

Improving performance through iterative Labeled data increase: Based on the results we obtained, as illustrated in Table 2, we took the best training paradigm, which is multilingual training, tested on Hindi, and we reiterated the training of both of the models while progressively increasing the annotated dataset L. Maintaining the same size of unlabeled material U, we start by sampling only 1% of L (which presents very few samples, 29 samples), then raising the labeled set size with 5%, 10%, 20% etc. As we already explained in previous subsection 4.2, we will consider F1 macro score metric.

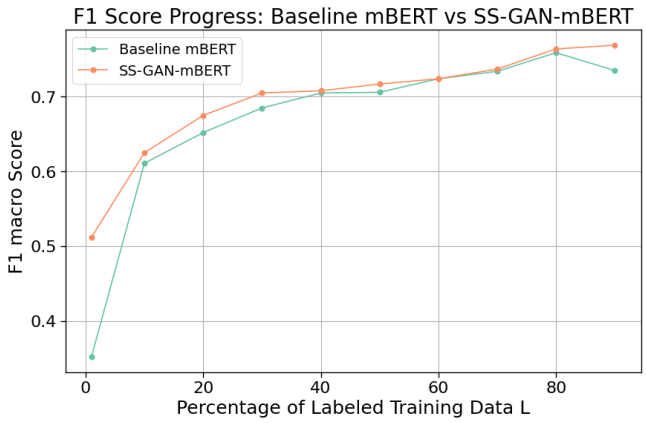

Fig. 3. F1 Score Progress on Hindi: Baseline mBERT vs SS-GAN-mBERT in multilingual training.

Based on Figure 3, we can clearly observe the difference between the baseline and SS-GAN-mBERT models, especially when using the smallest percentage of L data, and even with the use of almost the total amount of labeled data (80% - 90%), the baseline couldn't reach the performance of SS-GAN-mBERT. Moreover, it was also evident that SS-GAN-mBERT managed to reach the same performance as the baseline model, with a very less amount of labeled data (e.g. we can see the same F1 macro score attained by SS-GAN-mBERT with 1% of L while the baseline needed more than 6% to reach it). Another aspect to consider is the requirement for labeled data. In fact, in this semi-supervised

framework (whether within GAN-mBERT or mBERT alone), we can see that with the training unlabeled sets provided U, both of the models didn't need a big volume of annotated data. More specifically, as presented in Figure 3, baseline mBERT started giving F1 macro score of more than 0.7 with ∼40% of L while SS-GAN-mBERT needed only ∼30% to reach this performance, this indicates the benefits of implementing SS-learning as it helps to reduce the necessity to data labeling. Overall, we managed to show, through these experiments, that the need for annotated instances is reduced when the GAN structure is applied over SS-mBERT, it can be reduced more when further improving the structure of GAN, which could be our next step in future work to implement more complex GAN structures with more hidden layers in both the generator and the discriminator.
Computational cost at inference time: Considering the cost at inference time as already mentioned in subsection 3.2, we measured the time both of the models took in each of the training paradigms, and we didn't observe a huge difference (the maximum time gap was 16 minutes in one training scenario), which proves that the training time of SS-GAN-mBERT remains quite similar to that of the baseline model. This suggests that the SS-GAN-mBERT is an effective choice for situations where both training efficiency and robustness are important because its usefulness in inference time doesn't require significantly more extended training duration. However, this is still related to the simple structure of our GAN's generator (MLP), which could increase the time gap when implementing a more complex structure. Overall, this opens new directions we aim to examine for future research.

5.2 Future directions

We have chosen a constant noise vector of size 50 as input to our GAN's generator. We selected this value based on the results of the first experiments we made and on the computational efficiency provided. In the future, we aim to develop strategies that automatically optimize the generator to set the best noise vector size for any dataset. For instance, Wasserstein GAN could help provide the diversity of the data produced by the generator, thus, improving training stability [25]. Moreover, in dealing with the problem of class imbalance, we aim to reduce the effect of this issue by implementing new data augmentation solutions such as back translation [22], or working on GAN's data augmentation. Although this task still needs more work to improve GAN's accuracy, there have been many good attempts we aim to explore such as Conditional GAN [26]. Furthermore, we aim to generalize better and employ more advanced multilingual Large Language Models (LLMs) like BLOOM, GPT-3. Although this procedure requires more computational resources, we aim to start with smaller architectures like GPT-2, and Distil-GPT [27] and we seek to explore their performance within the SS-GAN model for future research.

6 Conclusion

In this paper, we introduced a Semi-Supervised Generative Adversarial SS-GAN-mBERT model, which achieved remarkable performance in both multilingual and

zero-shot cross-lingual hate speech detection for English, German, and Hindi. Our method emphasizes the usefulness of using semi-supervised learning to address the challenge of data labeling scarcity, yielding impressive results, which were further improved via Generative Adversarial Network (GAN).

References

1. *Social Media and Democracy: The State of the Field, Prospects for Reform*. SSRC Anxieties of Democracy. Cambridge University Press, 2020.
2. Paula Fortuna and Sérgio Nunes. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4), jul 2018.
3. Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. Towards multidomain and multilingual abusive language detection: a survey. *Personal and Ubiquitous Computing*, 27(1):17–43, 2023.
4. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol 1*, pages 4171–4186, Minneapolis, Minnesota, 2019.
5. Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. A bert-based transfer learning approach for hate speech detection in online social media. In *Complex Networks and Their Applications VIII*, pages 928–940, Cham, 2020. Springer International Publishing.
6. Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. Hate speech detection and racial bias mitigation in social media based on bert model. *PloS one*, 15(8):e0237861, 2020.
7. Khouloud Mnassri, Praboda Rajapaksha, Reza Farahbakhsh, and Noel Crespi. Bert-based ensemble approaches for hate speech detection. In *IEEE GLOBECOM*, pages 4649–4654, 2022.
8. Khouloud Mnassri, Praboda Rajapaksha, Reza Farahbakhsh, and Noel Crespi. Hate speech and offensive language detection using an emotion-aware shared encoder. *arXiv preprint arXiv:2302.08777*, 2023.
9. Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. Cross-lingual few-shot hate speech and offensive language detection using meta learning. *IEEE Access*, 10:14880–14896, 2022.
10. György Kovács, Pedro Alonso, and Rajkumar Saini. Challenges of hate speech detection in social media: Data scarcity, and leveraging external resources. *SN Computer Science*, 2:1–15, 2021.
11. Wenjie Yin and Arkaitz Zubiaga. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science*, 7:e598, 2021.
12. Ashwin Geet D'Sa, Irina Illina, Dominique Fohr, Dietrich Klakow, and Dana Ruiter. Label propagation-based semi-supervised learning for hate speech classification. In *Proceedings of the First Workshop on Insights from Negative Results in NLP*, pages 54–59, Online, November 2020. Association for Computational Linguistics.
13. Safa Alsafari and Samira Sadaoui. Semi-supervised self-learning for arabic hate speech detection. In *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 863–868, 2021.

14. Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Commun. ACM*, 63(11):139–144, oct 2020.
15. Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. Improved techniques for training gans. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.
16. Raihan Tanvir, Md Tanvir Rouf Shawon, Md Humaion Kabir Mehedi, Md Motahar Mahtab, and Annajiat Alim Rasel. A gan-bert based approach for bengali text classification with a few labeled examples. In *Distributed Computing and Artificial Intelligence, 19th International Conference*, pages 20–30, 2023.
17. Rui Cao and Roy Ka-Wei Lee. HateGAN: Adversarial generative-based data augmentation for hate speech detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6327–6338, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.
18. Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. Seqgan: Sequence generative adversarial nets with policy gradient. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, page 2852–2858, 2017.
19. Danilo Croce, Giuseppe Castellucci, and Roberto Basili. 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*, pages 2114–2119, Online, July 2020.
20. Shufan Jiang, Stéphane Cormier, Rafael Angarita, and Francis Rousseaux. Improving text mining in plant health domain with gan and/or pre-trained language model. *Frontiers in Artificial Intelligence*, 6, 2023.
21. Praphula Kumar Jain, Waris Quamer, and Rajendra Pamula. Consumer sentiment analysis with aspect fusion and gan-bert aided adversarial learning. *Expert Systems*, 40(4):e13247, 2023.
22. Hoang Thang Ta, Abu Bakar Siddiqur Rahman, Lotfollah Najjar, and Alexander Gelbukh. Gan-bert: Adversarial learning for detection of aggressive and violent incidents from social media. In *Proceedings of IberLEF, CEUR-WS*, 2022.
23. Raquel Bento Santos, Bernardo Cunha Matos, Paula Carvalho, Fernando Batista, and Ricardo Ribeiro. Semi-Supervised Annotation of Portuguese Hate Speech Across Social Media Domains. In João Cordeiro, Maria João Pereira, Nuno F. Rodrigues, and Sebastião Pais, editors, *11th SLATE Conference*, volume 104, pages 11:1–11:14, 2022.
24. Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th Annual Meeting of the Forum for Information Retrieval Evaluation*, page 14–17. Association for Computing Machinery, 2019.
25. Gustavo H. de Rosa and João P. Papa. A survey on text generation using generative adversarial networks. *Pattern Recogn.*, 119(C), nov 2021.
26. Kanishka Silva, Burcu Can, Raheem Sarwar, Frederic Blain, and Ruslan Mitkov. Text data augmentation using generative adversarial networks – a systematic review. *Journal of Computational and Applied Linguistics*, 1:6–38, Jul. 2023.
27. Zhang Ze Yu, Lau Jia Jaw, Wong Qin Jiang, and Zhang Hui. Fine-tuning language models with generative adversarial feedback, 2023.