Natural Language Engineering (2023), 1–20
doi:10.1017/S1351324923000396

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

SSL-GAN-RoBERTa: A robust semi-supervised model for detecting Anti-Asian COVID-19 hate speech on social media

Xuanyu Su, Yansong Li, Paula Branco and Diana Inkpen

School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, ON, Canada
Corresponding author: Diana Inkpen; Email: diana.inkpen@uottawa.ca

(Received 10 November 2021; revised 27 June 2023; accepted 8 July 2023)

Abstract

Anti-Asian speech during the COVID-19 pandemic has been a serious problem with severe consequences. A hate speech wave swept social media platforms. The timely detection of Anti-Asian COVID-19-related hate speech is of utmost importance, not only to allow the application of preventive mechanisms but also to anticipate and possibly prevent other similar discriminatory situations. In this paper, we address the problem of detecting Anti-Asian COVID-19-related hate speech from social media data. Previous approaches that tackled this problem used a transformer-based model, BERT/RoBERTa, trained on the homologous annotated dataset and achieved good performance on this task. However, this requires extensive and annotated datasets with a strong connection to the topic. Both goals are difficult to meet without employing reliable, vast, and costly resources. In this paper, we propose a robust semi-supervised model, SSL-GAN-RoBERTa, that learns from a limited heterogeneous dataset and whose performance is further enhanced by using vast amounts of unlabeled data from another related domain. Compared with the RoBERTa baseline model, the experimental results show that the model has substantial performance gains in terms of Accuracy and Macro-F1 score in different scenarios that use data from different domains. Our proposed model achieves state-of-the-art performance results while efficiently using unlabeled data, showing promising applicability to other complex classification tasks where large amounts of labeled examples are difficult to obtain.

Keywords: Hate speech detection; Deep learning; Semi-supervised learning

1. Introduction

When the COVID-19 pandemic struck at the end of 2019, it caused irreversible harm to public health and extensive damage to the world economy (Lora, Palumbo, and Brown 2021), along with an explosive wave of discriminatory speech against Asians on social media. According to CNN reports (Johnson and John 2021), a series of violent acts against Asians broke out in the United States (in Atlanta) in early 2021. Moreover, the increasing number of Anti-Asian speech on social media has caused widespread mental health issues in those targeted by it (Vidgen and Derczynski 2020). Therefore, it is crucial to build a timely Anti-Asian hate speech detection algorithm to prevent hate speech from unfettered spreading and evolving.

However, building a robust Anti-Asian speech detection algorithm in the context of the sudden COVID-19 outbreak is challenging (Vidgen and Derczynski 2020). First, it is not trivial to massively acquire and annotate emerging speech on social media towards a specific event. The lack of annotation could result in COVID-related Anti-Asian speech as a low-resource corpus, thus leading to data scarcity in terms of the model parametrization process. Existing works

(de Gibert *et al.* 2018; Yuan *et al.* 2019; Mozafari, Farahbakhsh, and Crespi 2020; Wei *et al.* 2021; Ali *et al.* 2022) on resolving data scarcity in the hate speech domain focus on transfer learning techniques, which rely on leveraging well-annotated heterogeneous datasets (i.e., toxicity, reactionary or racist speech datasets) to parameterize the model and use the learned weights to initialize models which are trained on low-resource target hate speech data. Such settings may not be satisfactory in the absence of a clear definition of "hate speech." While many people have an intuitive understanding of what hate speech is, this does not easily translate to a clear set of characteristics that can be assumed to share the same feature with other heterogeneous datasets (Bigoulaeva, Hangya, and Fraser 2021). In addition, our empirical experiments using transfer learning techniques show that the model's performance is far from satisfactory when trained with heterogeneous datasets. Second, one may adopt the recent prevalent massively pre-training paradigm to pre-train the whole model with social media-specific datasets such as BERTweet (Nguyen, Vu, and Nguyen 2020) and COVID-Twitter-BERT (CT-BERT) (Müller, Salathé, and Kummervold 2023). Although these in-domain pre-trained models improve the performance of low-resource hate speech detection tasks, imposing the massive pre-training paradigm inevitably introduces overwhelming computation costs. In effect, it takes 672 hours to train BERTweet and 120 hours to train COVID-Twitter-BERT on multiple TPU servers, which is inaccessible for most individual developers. Therefore, we focus on the following research question: **Can we build a robust Anti-Asian speech detection algorithm in a more accessible and efficient way?**

To address the above challenges, we propose Semi-Supervised Learning-based Generative Adversarial Network with RoBERTa (SSL-GAN-RoBERTa), a robust transformer-based Anti-Asian speech detector that uses unannotated Anti-Asian related data to implicitly learn the features of Anti-Asian Speech and improve model performance given arbitrary annotated heterogeneous datasets. Specifically, SSL-GAN-RoBERTa introduces a "generator" to produce samples resembling real data distribution and a "discriminator" that is trained jointly with the generator to detect the Anti-Asian speech from a given sample while discriminating whether it was generated from the "generator." Our framework is based on the key finding that semi-supervised generative adversarial networks can improve discriminator's inner representations based on the unlabeled data distribution given arbitrary downstream tasks (Salimans *et al.* 2016). Besides that, SSL-GAN-RoBERTa is grounded on the complement generator theory (Dai *et al.* 2017) and introduces more learnable parameters in the discriminator than the generator, thus leading to separate distribution of generated representation and real data representation. This setting guarantees that the discriminator can obtain correct decision boundaries in high-density areas within labeled, unlabeled, and generated data representations and build a robust discriminator in a semi-supervised learning fashion.

We carried out experiments with multiple heterogeneous datasets to study the effectiveness of our proposed framework. Results show that SSL-GAN-RoBERTa can consistently improve model performance with unannotated Anti-Asian Speech datasets and achieves comparable performance with pre-training-based methods with lower computation costs. Moreover, we show that the complement generator theory in the GAN structure is helpful for the model to capture data features, therefore enhancing the model's performance score.[a] Our results confirm previous theoretical findings of Dai *et al.* (2017) and the experiments in Salimans *et al.* (2016) and Ulyanov *et al.* (2018).

Our main contributions: Our experiments demonstrate that SSL-GAN-RoBERTa can effectively make use of unlabeled data under the adversarial training paradigm and that it can consistently outperform the transformer encoder model for hate speech classification. Below we highlight the main contributions of this paper.

[a]The performance metrics used will be explained in Section 5.

https://doi.org/10.1017/S1351324923000396 Published online by Cambridge University Press

- We propose SSL-GAN-RoBERTa, a robust alternative to tackle hate speech detection in the case of insufficient data annotation, which could utilize unlabeled in-domain data to improve downstream task performance in a semi-supervised learning fashion given an arbitrary annotated cross-domain heterogeneous dataset.
- We carry out a comprehensive set of experiments to demonstrate that our method outperforms cross-domain transfer learning and achieves comparable performance with in-domain pre-training, with a lower computational cost.

This paper is organized as follows. In Section 2, we discuss the related work. In Section 3, we describe our proposed semi-supervised SSL-GAN-RoBERTa algorithm. Section 4 provides the details of the tested knowledge transfer scenarios and the used datasets. In Section 5, we experimentally evaluate our proposed method for the Anti-Asian COVID-19 hate speech detection task. Moreover, we also carried out an experimental evaluation of our algorithm in the sentiment classification domain. Finally, Section 6 concludes the paper and discusses future research directions.

2. Related work

Anti-Asian COVID-19 hate speech: The COVID-19 crisis has triggered hate speech on multiple social media platforms. In particular, during the COVID-19 pandemic, we witnessed a growing hate speech trend toward Asian communities. Existing works (Fan, Yu, and Yin 2020; He *et al.* 2021; Costello *et al.* 2021; Tekumalla *et al.* 2022; Li and Ning 2022) propose various analyses on decomposing the hate speech attributes from demographics and lexicon-based emotions perspectives based on social media information (i.e., user profile, tweets hashtag, sentiment distribution), which provide critical insights in raising public awareness and mitigating future crises. However, as the content of hate speech rapidly evolves through time, it imposes a critical challenge in classifying emerging hate speech with the unseen format. Only a few studies (Fan *et al.* 2020; Agarwal and Chowdary 2021) address this issue through an ensemble learning-based adaptive model. However, such settings may not be satisfactory in the particular case of hate speech towards Asian communities during the COVID-19 pandemic, where the well-annotated dataset is unavailable. Obtaining high-quality annotations in such a particular domain is demanding and costly. Still, it is necessary to address this problem and to develop robust solutions that can better leverage limited data available (even if it is in the form of unlabeled data). The algorithm we propose here addresses this gap providing an effective solution for the Anti-Asian COVID-19 hate speech detection problem. Moreover, we show that our solution can extend to similar tasks suffering from analogous challenges.

Anti-Asian COVID-19 related datasets: To resolve the hate speech detection, several hate speech dataset (Founta *et al.* 2018; Davidson, Bhattacharya, and Weber 2019; Abroshan *et al.* 2021) in multiple languages have been made available in recent years by the community towards the development of automatic hate speech detection. However, the conflated concepts of hate speech, abusive/offensive language, and inconsistent cultural understanding of hate speech have made hate speech detection towards a specific community non-trivial (Davidson *et al.* 2017; Founta *et al.* 2018; Mathew *et al.* 2021). To resolve this challenge, Mathew *et al.* (2021) propose HateXplain by collecting social media posts from Twitter and Gab platforms and manually annotating these posts into the following categories: {Hate Speech, Normal, Offensive} towards several vulnerable communities (i.e., African American, Muslim, and Asian groups) in North America. Although HateXplain provides a valuable annotated dataset, it is yet unexplored whether we can build a generalizable hate speech detection framework without a time-consuming manual annotation process. To tackle this challenge, we collect general hate speech (i.e., hate speech that was collected before Covid pandemic) from Twitter [East Asian Prejudice dataset (EA)

https://doi.org/10.1017/S1351324923000396 Published online by Cambridge University Press

(Vidgen *et al.* 2020)], Gab [Gab Hate Corpus (GHC) (Kennedy *et al.* 2021)] and Wikipedia [Toxicity (Thain *et al.* 2017)] platforms and propose a semi-supervised learning framework SSL-GAN-RoBERTa to investigate how to effectively learn hate speech representation with unannotated Covid-related data. We list detailed information about the datasets in Section 4.

Previous architectures for NLP classification problems: Most of the natural language processing (NLP) classification tasks are based on supervised learning, which requires a comprehensive model structure and massive annotated datasets. BERT/RoBERTa with transformer-based structure and massive pre-trained models achieve state-of-the-art performance on Natural Language Understanding tasks through fine-tuning (e.g., Liu *et al.* 2019; Devlin *et al.* 2019; Wang *et al.* 2018, 2019). However, such a setting may not be satisfactory in the case of the low-resource corpus. For instance, while doing cross-lingual Natural Language Inference, the model performance degrades with the low-resource language. To tackle this challenge, Conneau *et al.* (2020) propose pre-training the whole RoBERTa model with the multilingual datasets and substantially increasing the model parameter scale. Another alternative to tackle the low-resource corpus is through a semi-supervised learning model that was recently proposed for embedding BERT and GAN, the GAN-BERT model (Croce, Castellucci, and Basili 2020). This strategy proposes a solid methodology to combine the BERT encoder with the generative adversarial model, which could improve the overall model performance with a limited annotated dataset on specific language tasks. Our solution is different as it uses the RoBERTa model, explores a more complex structure with complement generator theory, and investigates the use of different amounts of unlabeled data.

Previous semi-supervised GAN applications: Generative Adversarial Network (GAN) models are frequently used approaches. Most of the GAN applications are designed to aim at the computer vision area. However, they do provide a lot of feasible semi-supervised learning approaches with GAN structure. In this context, a GAN-based model that caught our attention is context-conditional GAN [CC-GAN (Denton, Gross, and Fergus 2017)], which utilize unlabeled image combined with generated noise to obtain representation learning among the dataset; the discriminator would benefit from real/fake classification task and supervised classification task. CatGAN (Springenberg 2016) with a similar context information extraction architecture in semi-supervised learning fashion overscore the traditional supervised learning-based model. Another approach was introduced by Miyato *et al.* (2018), which uses virtual adversarial training to improve the robustness of SGANs against small perturbations in the input data. In a different study, Berthelot *et al.* (2019) proposed a method that uses adversarial training to generate realistic unlabeled data, which can be used to improve the performance of a classifier trained on a limited amount of labeled data. Other works have explored the use of SGANs in text domains, such as sentiment analysis (Li and Ye 2018; Lee *et al.* 2019). Overall, SGANs have shown promising results in improving classification performance in text datasets through generating authentic datasets, and there is ongoing research in this area to explore their potential further.

Complement Generator: Based on the theoretical analysis presented in Dai *et al.* (2017) and empirical experiments presented by Salimans *et al.* (2016) and Ulyanov *et al.* (2018), there exists a contradiction between generated data quality and feature matching discrimination accuracy, which indicates that the GAN model cannot propagate the optimal parameters for both the generator and the discriminator in the same optimization process. To obtain the optimal discriminator to make the correct decision boundary, constructing a complement generator is essential in order to improve classification performance. Referring to the model parameter setting from the source code provided in Dai *et al.* (2017), we include additional linear layers to the RoBERTa encoder, with the goal of increasing the learnability of the discriminator's optimization phase.

https://doi.org/10.1017/S1351324923000396 Published online by Cambridge University Press

BERT VS RoBERTa: BERT (Bidirectional Encoder Representations from Transformers) (Devlin *et al.* 2019) outperforms the latest technology of NLP on several challenging tasks (Wang *et al.* 2019). Its performance improvement can be attributed to the bidirectional detection ability, the two pre-trained tasks of the Masked Language Model and the Next Sentence Prediction (NSP), along with a large amount of data from Google's computing store (Devlin *et al.* 2019).

RoBERTa (Liu *et al.* 2019) was based on the BERT model, and it has a model structure for pre-training on more extensive data and computing resources. RoBERTa removed the NSP task from BERT's pre-training and introduced Dynamic Masking so that the masked token changes during the training epochs. The model uses 160 GB of text for pre-training (Liu *et al.* 2019), including the 16 GB of the Books Corpus and the English Wikipedia used in BERT. The additional data are the CommonCrawl News dataset with 63 million articles (76 GB), a web text corpus (38 GB), and stories from Common Crawl (31 GB).

RoBERTa adjusted the pre-training data volume and model pre-training tasks based on BERT, according to the above analysis and comparison. For these reasons, we propose to use the RoBERTa model to be integrated with GAN architecture in our work.

3. SSL-GAN-RoBERTa

This section describes SSL-GAN-RoBERTa, our proposed solution for COVID-19-related Anti-Asian hate speech detection. The central idea of the proposed SSL-GAN-RoBERTa is to integrate RoBERTa and GAN models while simultaneously leveraging unlabeled data in a semi-supervised fashion. Our proposed method consists of the following components: (i) the combined usage of RoBERTa and GANs; (ii) the inclusion of a mechanism for improving the model's performance by introducing complement generator theory to extend the features provided by RoBERTa; and (iii) the use of both labeled and unlabeled data in the RoBERTa Encoder module. To achieve our goals, SSL-GAN-RoBERTa is structured into three main components: a generator, a discriminator, and a RoBERTa encoder. Figure 1 provides an overview of the main components of SSL-GAN-RoBERTa. The overall idea of our algorithm is to incorporate both labeled and unlabeled data in a trainable RoBERTa encoder while stacking a number of extra linear layers after the RoBERTa encoder to fill the complement generator theory. Finally, a GAN is included in the process through a trainable generator and discriminator. We initialize a Gaussian noise as the input of the generator, while the discriminator's inputs are the data from the generator and RoBERTa encoder. Below, we describe the details of SSL-GAN-RoBERTa.

Let B, S, and M represent the batch size, the maximum sentence length, and word embedding size, respectively. For the RoBERTa Encoder component E, we implement a pre-trained RoBERTa R with extra linear layers stacked as a feature encoder to extract the features from both the labeled and unlabeled data. We consider a generator component G which takes noise $N(B, S, M)$ to generate features coordinated with the input features and a discriminator component D whose goal is the classification of the actual labeled data L into K classes, the generated data F into a $K+1$th class (representing fake data), and the actual unlabeled data U into K possible classes. Let $< x, y >$ be a data sample, and the corresponding label, P_D and P_G denote the distribution from the discriminator and the generator, respectively. The objective functions for the discriminator are displayed in Equations (1), (2), (3), and (4).

$$\exp_D = \max_D \left(\exp_L + \exp_U + \exp_F \right) \tag{1}$$

$$\exp_L = \mathbb{E}_{x,y \in L} \log_{P_D} (y|x, y \leq K) \tag{2}$$

$$\exp_U = \begin{cases} \mathbb{E}_{x \in U} \log_{P_D} (y|x, y \leq K) & \text{if } D(x) = K+1 \\ 0 \quad \text{otherwise} \end{cases} \tag{3}$$

$$\exp_F = \mathbb{E}_{x \in P_G} \log_{P_D} (y|x, y = K+1) \tag{4}$$

https://doi.org/10.1017/S1351324923000396 Published online by Cambridge University Press

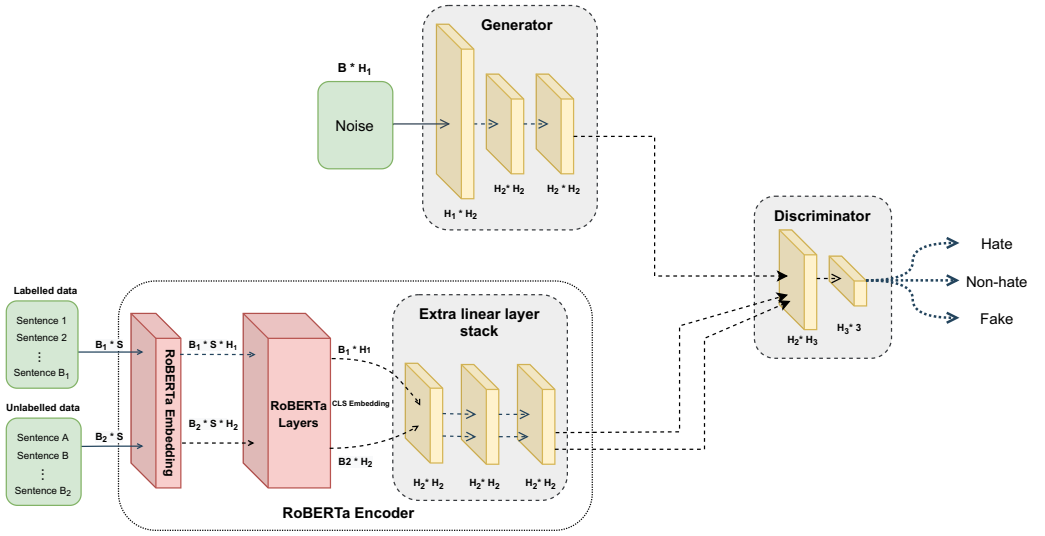

Figure 1. SSL-GAN-RoBERTa consists of 3 main components: (1) A RoBERTa encoder that extracts sentence-level features from both labeled $B_1 * S$ and unlabeled data $B_2 * S$, along with an extra linear layers stack that compresses the sentence-level features into $B * H_2$. (2) A generator that transforms the random noise $B * H_1$ into $B * H_2$. (3) A discriminator that classifies the sentence-level embedding into three categories (Hate, Non-hate, Fake). Each linear layer is concatenated with the Leaky-ReLU activation layer and dropout.

The discriminator is a classifier model that will classify the actual data, calibrate on unlabeled data only when the unlabeled data are classified as $K + 1$th class, and identify the generated data.

The objective function for the generator is defined by Equation (5).

$$\exp_G = \min_G \left(\mathbb{E}_{x \in L} D(x) - \mathbb{E}_{z \in N} D(G(z))\right) \tag{5}$$

The role of the generator is to generate more authentic data corresponding to real data. More specifically, the goal of the generator expectation function is to minimize the cross-entropy between discriminator's output probabilistic distribution and the mini-batch annotated data label. This way, the training method of the generator is basically the same as the optimization process of the autoencoder. The essence of the generator here is to generate a latent feature space consistent with the real data. The above methodology is aligned with our related research on representation learning by the semi-supervised GAN models and thus enhances the performance of both the generator and the discriminator.

4. Knowledge transfer scenarios and datasets

4.1 Knowledge transfer scenarios tested

Our main application domain is Anti-Asian COVID-19 hate speech detection. This predictive task is very specific, and thus, the available annotated datasets are scarce. For this reason, we considered different knowledge transfer scenarios where we use data, either labeled or unlabeled, from related domains in addition to the target domain data.

As shown in Figure 3, in contrast to the conventional semi-supervised learning knowledge transfer scenario from Figure 2, our target is to leverage the knowledge from the unannotated domain to the unannotated target, where knowledge transfer scenarios are considered in terms of the different labeled and unlabeled datasets used in our experiments.

https://doi.org/10.1017/S1351324923000396 Published online by Cambridge University Press

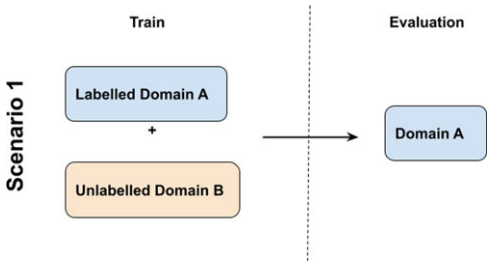

Figure 2. In-domain transfer: Scenario using data from the same domain A to train and test and including unlabeled data from a different related domain B.

We evaluate our proposed algorithm in these two scenarios (in-domain and cross-domain transfer) and observe the impact on the performance under these situations. In the first scenario (in-domain transfer), we assume there is enough labeled data from the specific domain to train and test and we use unlabeled data from a different domain to improve the performance in our specific application domain. In scenario two (cross-domain transfer), we consider that we only have a small amount of labeled data from our target domain which we use for testing. Regarding the training stage, we use unlabeled data from the same application domain as well as labeled data from another related domain.

To proceed with the experiments, we need to consider adequate datasets related to our main predictive task involving Anti-Asian COVID-19 hate speech detection, which we present in the next section. We also extend the experiments to a different domain to show that our solution also presents advantages in other tasks beyond the Anti-Asian COVID-19 hate speech detection.

4.2 Datasets for Anti-Asian COVID-19 hate speech

To address our problem, we use five different datasets for model training and testing (Table 1). We begin by describing the three datasets we considered for training our model. Then, we describe the dataset that was used for testing, and finally, we present the details of the dataset we used as a source of unlabeled data.

Toxicity: This dataset includes over 100,000 annotated comments from the English Wikipedia (Thain *et al.* 2017). Multiple annotators labeled each comment via CrowdFlower on whether it is a toxic (positive) or a healthy (negative) contribution (binary). We kept the three columns: "comment," "split," and "toxicity" for model training (the "split" column is used for train, valid, test sets partitions), and removed all the other columns not related to our paper.

Gab Hate Corpus (GHC): This dataset is the largest annotated hate speech corpus to date, containing 27,665 annotated posts from a social networking service call gab.ai (Kennedy *et al.* 2021). Each post is annotated by a minimum of three trained annotators (based on a coding typology synthesized from the definition of hate speech from research in law, computing, psychology, and sociology) into two categories (Hate, note hate).

East Asian Prejudice (EA): This database provides 20,000 tweets annotated by experienced analysts. The tweets were collected in the year 2020, between January 1st and March 17th using a total of 14 selected hashtags that are related to both East Asia. The dataset contains four classes: (a) Hostility against East Asia; (b) Criticism of East Asia; (c) Meta-discussions of East Asian prejudice; (d) Neutral (Vidgen *et al.* 2020). We produce the binary classification on the EA dataset by setting "hostility" as hate speech, "criticism," "meta-discussion" and "neutral" as non-hate speech. The EA training set includes 16,000 examples in the training set, 2000 in the validation set, and 2000 in the test set (Note: We are using EA's training set as our labeled data and the EA validation set in our ablation study).

https://doi.org/10.1017/S1351324923000396 Published online by Cambridge University Press

Table 1. Main characteristics of the datasets used for the main task on Anti-Asian COVID-19 hate speech detection (positive class: class related to hate speech/toxic comments; negative class: non-hate speech/non-toxic comments).

Dataset	Reference	Lab./unlab.	No. cases	No. positive class cases	No. negative class cases
Toxicity	Thain et al. (2017)	Labeled	1,598,289	232,055	1,366,234
GHC	Kennedy et al. (2021)	Labeled	27,665	11,249	16,416
EA	Vidgen et al. (2020)	Labeled	20,000	5331	14,669
CHT	He et al. (2021)	Labeled	2319	678	1641
UCH	He et al. (2021)	Unlabeled	27,000	Not applicable	Not applicable

4.2.1 Test set

COVID-HATE tweet (CHT): CHT is a manually annotated dataset of 2319 COVID-19-related racial hate tweets categorized into four classes: *hate, counterhate, neutral,* and *non-Asian aggression*. He et al. team adopted a keyword-based approach to target relevant tweets on Twitter based on *Covid-19* and *hate keywords* database. We created a binary classification task by categorizing the four classes into *hate* (positive class) and *non-hate* (negative class). Our hate class matches the existing hate class while the remaining classes (counterhate, neutral, non-Asian aggression) are aggregated into the non-hate (negative) class. The CHT dataset is a part of the COVID-HATE dataset (He *et al.* 2021), the largest dataset of Anti-Asian hate and counterhate spanning 3 months, containing over 30 million tweets. From this large database, the CHT is the only (small) subset that has been labeled.

4.2.2 Unlabeled set

Unlabeled COVID-HATE (UCH): As mentioned above, the COVID-HATE dataset contains a large number of Anti-Asian tweets (over 30 million tweets). However, the majority of these tweets are unlabeled. Only 2319 examples among them are annotated and constitute the CHT test set previously described. We selected 27k data, converted the IDs into tweets using the Twitter API. We then used different amounts of these 27,000 extracted unlabeled tweets in our experiments.

4.3 Datasets on a different domain: sentiment classification

In order to evaluate the robustness of the performance of our proposed algorithm in a different task, we included another set of experiments on a different domain: sentiment classification. Our predictive task goal is the prediction of positive or negative sentiments in text. In this case, we considered the Stanford Sentiment Treebank (SST-2) (Socher *et al.* 2013) dataset as the small labeled data, and we included the IMDB movie review dataset as the unlabeled data. The goal of using these datasets is to simulate a situation similar to the one observed for Anti-Asian COVID-19 hate speech where a small set of annotated cases exist and a large volume of unlabeled data is available. By also observing the performance of our proposed algorithm in this domain, we aim at showing its robustness and usefulness on similar tasks. Below we provide more details on the datasets considered.

4.3.1 Training and test sets

The Stanford Sentiment Treebank (SST) (Socher *et al.* 2013) dataset in the GLUE benchmark (Wang *et al.* 2018) is used as our labeled dataset. The SST-2 is a corpus including fully labeled parse trees that enables a thorough examination of sentiment's effects in language. The corpus consists of 215,154 unique phrases that were annotated by three human judges. SST-2 includes

https://doi.org/10.1017/S1351324923000396 Published online by Cambridge University Press

Table 2. Main characteristics of the datasets used for the second task on sentiment prediction.

Dataset	Reference	Lab./unlab.	No. cases	No. positive cases	No. negative cases
SST-2	Socher et al. (2013)	Labeled	11,286	5644	5642
IMDB	Maas et al. (2011)	Unlabeled	19,811	Not applicable	Not applicable

short sentences with an average sequence length of 19, and spanning from a minimum length of 2 to a maximum length of 52. In the multi-class-associated task, each phrase has the following labels: negative, somewhat negative, neutral, somewhat positive, or positive. For binary classification, the labels are aggregated in two classes as follows: negative or somewhat negative vs somewhat positive or positive. Neutral sentences are discarded. We have used this binary classification version (SST-2) of the SST dataset. The SST-2 dataset is available with a training and testing partition that we use in the experiments. The SST-2 training set includes 8117 training set, 2125 validation set, and the 1044 test set.

4.3.2 Unlabeled dataset

We selected the IMDB movie review data (Maas *et al.* 2011), which has some similarities to SST-2 data to be used as the source of unlabeled data. IMDB dataset consists of long sentence movie reviews with an average sequence length of 231, and spanning from a minimum length of 4 to a maximum length of 2486. Since SST-2 includes short sentences and IMDB consists of long sentences, we first filter out the IMDB movie reviews with a sequence length smaller than 52. We obtained 19811 unlabeled examples satisfying this condition on IMDB dataset. The use of this unlabeled data will help to validate the results of our proposed SSL-GAN-RoBERTa model in an additional task (Table 2).

5. Experiments

In this section, we provide a set of experiments to demonstrate the usefulness of the proposed SSL-GAN-RoBERTa algorithm for the Anti-Asian COVID-19 hate speech detection task. We start with our main experiment to compare with our baseline methods. Then, we provide a set of experiments to evaluate the performance when using different amounts of unlabeled sequences[b] in SSL-GAN-RoBERTa. Then, we conduct experiments on the sensitivity of the hyperparameters of the model, which show that new architecture modifications proposed in SSL-GAN-RoBERTa provide important performance gains. We also provide an error analysis displaying examples of the different types of errors we observed in our model to understand the model's strengths and weaknesses. Finally, we assess the effectiveness of the proposed algorithm in a different domain, such as sentiment classification, to determine if the solution is generalizable to other contexts.

5.1 Baselines

We compare our SSL-GAN-RoBERTa method with existing pre-training-based methods BERTweet (Nguyen *et al.* 2020) and COVID-Twitter-BERT (CT-BERT) (Müller *et al.* 2023) and conventional transfer learning techniques with multiple heterogeneous datasets. As for the in-domain pre-trained CT-BERT[c] and BERTweet,[d] we directly downloaded the corresponding

[b] Sequence represents sentence of words.
[c] https://huggingface.co/digitalepidemiologylab/covid-twitter-bert
[d] https://huggingface.co/vinai/bertweet-base

https://doi.org/10.1017/S1351324923000396 Published online by Cambridge University Press

Table 3. Results of transfer learning techniques, pre-training techniques, and our SSL-GAN-RoBERTa fine-tuned by toxicity, GHC, and EA datasets on Anti-Asian detection performance (CHT test set).

Methods	Accuracy	Macro-F1 score	Model adaptation/training time
Cross-domain transfer learning			
$\text{RoBERTa}_{\text{Toxicity}}$	$0.6245_{\pm 0.0153}$	$0.5362_{\pm 0.0131}$	2.5 h
$\text{RoBERTa}_{\text{GHC}}$	$0.7176_{\pm 0.0141}$	$0.2013_{\pm 0.0136}$	1.1 h
$\text{RoBERTa}_{\text{EA}}$	$0.7993_{\pm 0.0127}$	$0.6814_{\pm 0.0124}$	1.2 h
In-domain pre-training			
$\text{BERTweet}_{\text{EA}}$	$0.7818_{\pm 0.0174}$	$0.5734_{\pm 0.0165}$	673.2 h
- *Zero-Shot* (Nguyen *et al.* 2020)	$0.7023_{\pm 0.0021}$	$0.4981_{\pm 0.0034}$	672 h
$\text{CT-BERT}_{\text{EA}}$	$0.8249_{\pm 0.0137}$	$0.7892_{\pm 0.0155}$	121.2 h
- *Zero-Shot* (Müller *et al.* 2023)	$0.7598_{\pm 0.0014}$	$0.6156_{\pm 0.0009}$	120 h
SSL-GAN-RoBERTa			
$\text{SSL-GAN-RoBERTa}_{\text{Toxicity}}$	$0.7593_{\pm 0.0191}$	$0.6115_{\pm 0.0183}$	2.8 h
$\text{SSL-GAN-RoBERTa}_{\text{GHC}}$	$0.7582_{\pm 0.0163}$	$0.5316_{\pm 0.0156}$	1.4 h
$\text{SSL-GAN-RoBERTa}_{\text{EA}}$	$\mathbf{0.8533}_{\pm 0.0128}$	$\mathbf{0.8142}_{\pm 0.0136}$	1.5 h

checkpoints from the HuggingFace repository and fine-tuned with EA dataset (as EA is the only source coming from Twitter among three heterogeneous datasets) to maximize the potential of tweets-related pre-trained models. As for transfer learning techniques, we independently fine-tune the $\text{RoBERTa}_{\text{Large}}$ model for Anti-Asian speech detection with each of the three datasets GHC, Toxicity, and EA.

5.2 Implementation details

To demonstrate the effectiveness of SSL-GAN-RoBERTa, we use $\text{RoBERTa}_{\text{Large}}$ as the backbone model. Experiments are performed on an RTX 2080Ti GPU. We adopt Adam as the optimizer. For the training, we set the learning rate to 2e-6 for all methods. The batch size is set to 16, and we train all methods for a total of 3 epochs. We set the maximum sequence length to 128 tokens. For SSL-GAN-RoBERTa, we set the number of unlabeled sequences to 9000 and implement the same Adam optimizer with a learning rate of 1e-6 for both the generator and discriminator. We add three 1024×1024 extra linear layers at the end of the $\text{RoBERTa}_{\text{Large}}$ model to impose the complement generator theory. We provide the results of both the Accuracy and the F1 score for the positive class (hate speech class).

5.3 Main results

Table 3 shows the results of cross-domain transfer learning, in-domain pre-training, and SSL-GAN-RoBERTa on the CHT Anti-Asian speech detection dataset. As the EA dataset is essentially East Asian Prejudice speech from Twitter, which share a similar topic and identical platform with the CHT dataset, we can observe that model fine-tuned by the EA dataset outperforms the models fine-tuned on the GHC and Toxicity datasets. The overall performance of in-domain pre-training techniques is higher than cross-domain transfer learning, which, however, introduces overwhelming computation costs in the pre-training process. Our SSL-GAN-RoBERTa

https://doi.org/10.1017/S1351324923000396 Published online by Cambridge University Press

Table 4. Results of the different models trained on EA training set and tested on CHT test set and EA test sequences using different amounts of unlabeled sequences from UCH dataset.

	CHT test set		EA test set	
# of unlabeled sequences	Accuracy	Macro-F1	Accuracy	Macro-F1
0k	$0.8271_{\pm 0.0182}$	$0.7951_{\pm 0.0155}$	$0.9038_{\pm 0.0131}$	$0.8778_{\pm 0.0118}$
2k	$\mathbf{0.8547}_{\pm 0.0127}$	$0.8191_{\pm 0.0146}$	$0.9145_{\pm 0.0125}$	$0.8822_{\pm 0.0124}$
6k	$0.8503_{\pm 0.0141}$	$0.8118_{\pm 0.0126}$	$0.9047_{\pm 0.0116}$	$0.8902_{\pm 0.0132}$
9k	$0.8533_{\pm 0.0128}$	$0.8142_{\pm 0.0136}$	$\mathbf{0.9215}_{\pm 0.0141}$	$\mathbf{0.8991}_{\pm 0.0129}$
12k	$0.8526_{\pm 0.0119}$	$\mathbf{0.8164}_{\pm 0.0144}$	$0.9085_{\pm 0.0137}$	$0.8813_{\pm 0.0135}$
18k	$0.8471_{\pm 0.0124}$	$0.8129_{\pm 0.0133}$	$0.9109_{\pm 0.0141}$	$0.8842_{\pm 0.0127}$
27k	$0.8516_{\pm 0.0127}$	$0.8134_{\pm 0.0142}$	$0.9011_{\pm 0.0111}$	$0.8823_{\pm 0.0124}$

consistently outperforms cross-domain transfer learning techniques with respect to each dataset resource and achieves comparable performance with in-domain pre-training techniques with much shorter training time. The results suggest that semi-supervised learning in an adversarial training paradigm can effectively utilize the in-domain unannotated sequences and improve model's performance in the absence of annotated sequences.

5.4 Effect of the amount of unlabeled sequences

We investigate the effect of using different amounts of unannotated sequences and verify the generalizability of our proposed methods given arbitrary heterogeneous datasets. Therefore, we independently use EA, Toxicity, and GHC as training sequences and gradually increase the number of UCH unlabeled sequences in SSL-GAN-RoBERTa. We also test the in-domain performance of these heterogeneous datasets, respectively. We set the different amounts of unlabeled sequences to $\{0k, 2k, 6k, 9k, 12k, 18k, 27k\}$. Table 4 summarizes the performance results obtained with different amounts of unlabeled sequences when using the CHT and EA test sets. Figure 4 displays the overall results for combinations of training/test datasets.

Result analysis on overall performance: Figure 4 shows that the overall model performance on CHT consistently increases across the three datasets when adding more UCH unannotated sequences. Similar phenomena can be observed in the in-domain test where adding CHT unannotated sequences can also improve the model performance tested on the EA, Toxicity, and GHC test sequences, respectively. The results confirm our intuition that applying adversarial training with unannotated sequences can help the model learn better representations, thus resulting in strong model performance.

Result analysis on CHT dataset: In the experiments with the CHT test dataset (Figure 4 and Table 4), we can observe two critical points of noticeable improvement. First, when we test our SSL-GAN-RoBERTa model without the use of any unlabeled sequences, the Macro-F1 and Accuracy of the model have been greatly improved when compared against the baseline RoBERTa model. Secondly, when we gradually added unlabeled sequences into the learning process, we observe that the performance of the model is further improved. The accuracy reaches a peak after adding 2k unlabeled sentences, and the Macro-F1 score achieves its peak after adding 12k of unlabeled sentences, which shows that SSL-GAN-RoBERTa can effectively utilize unlabeled sequences

https://doi.org/10.1017/S1351324923000396 Published online by Cambridge University Press

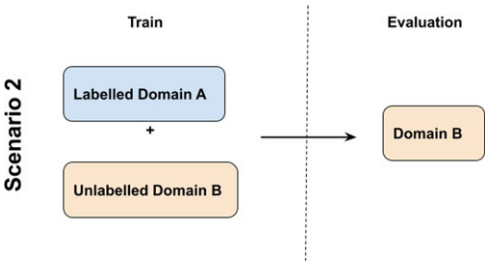

Figure 3. Cross-domain transfer: Scenario using data from a related domain B to train and unlabeled data from the target domain A, and testing on domain A.

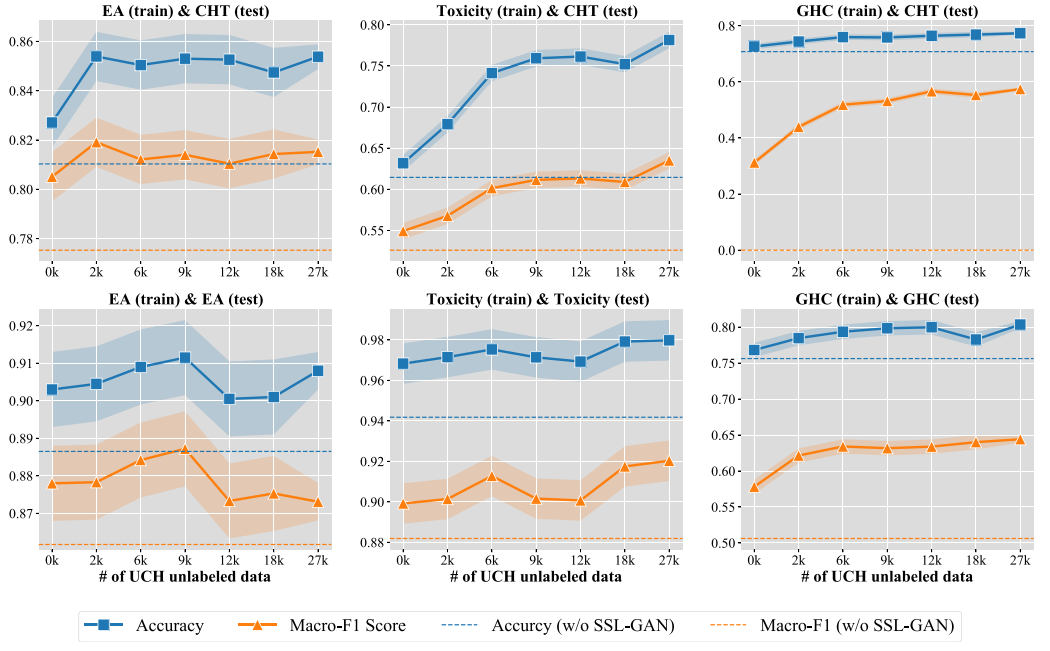

Figure 4. Comparison of the performance using accuracy and Macro-F1 with respect to different types of model configurations. Note: The dashed line in the figure represents the baseline without SSL-GAN and unannotated data settings.

and improve downstream task performance in the Cross-domain transfer scenario as displayed in Figure 3. In effect, we are only using unlabeled sequences from the test domain (CHT) and we are training the model exclusively on dataset from a related but yet different domain (EA).

Result analysis on EA test sequences: In this case, these experiments correspond to In-domain transfer scenario displayed in Figure 2, where we use a train and test set from the same domain (EA domain) and add unlabeled sequences from a different domain (UCH). The performance of the different models tested is displayed in Table 4 and Figure 4. We observe that similarly, using SSL-GAN-RoBERTa without unlabeled sequences presents performance gains on the EA test set when compared against the use of RoBERTa-large model. This shows that in this scenario the proposed algorithm without unlabeled sequences is a better option. However, after embedding the UCH unlabeled dataset, the performance results showed a higher degree of fluctuation. Still, it

https://doi.org/10.1017/S1351324923000396 Published online by Cambridge University Press

Table 5. Comparison of the results of six baseline models trained on EA datasets and tested on our target CHT test data.

Dataset	CHT dataset	
Metrics	Accuracy	Macro-F1
$T5_{large}$	$0.7895_{\pm 0.0132}$	$0.7341_{\pm 0.0241}$
$BERT_{large}$	$0.8113_{\pm 0.0173}$	$0.7615_{\pm 0.0148}$
$ALBERT_{large}$	$0.7328_{\pm 0.0218}$	$0.7141_{\pm 0.0231}$
$GPT2_{large}$	$0.8154_{\pm 0.0149}$	$0.7914_{\pm 0.0181}$
$DeBERTa_{large}$	$\mathbf{0.8311}_{\pm 0.0166}$	$0.7897_{\pm 0.0175}$
$RoBERTa_{large}$	$0.8291_{\pm 0.0168}$	$\mathbf{0.8008}_{\pm 0.0192}$

is clear that adding unlabeled sequences always has a positive effect on the model's performance. When the amount of unlabeled sequences is 9k, the results when testing on the EA test set are the best (accuracy: 0.9115 and Macro-F1: 0.8872). When the size of unlabeled sequences is between 12k and 27k, the overall performance of the model shows a lower result trend. We conjecture that this happened because in this scenario the train and test sets are from the same domain while the unlabeled sequences are not. Thus, when the model training procedure fits to one of the datasets (EA in this case), the performance will decline to a certain extent by adding small or large amounts of unlabeled sequences from a domain that is different from the test domain. This is why the model exhibits more fluctuations on the EA test set. Still, we verify that adding unlabeled sequences provided important performance gains when using intermediate levels of unlabeled sequences.

5.5 Choice of backbone model

The goal of the experiments in this section is to verify the effectiveness of selecting the RoBERTa model for integrating the Encoder Module we proposed in SSL-GAN-RoBERTa algorithm. We assessed the performance of six baseline models, including RoBERTa. More precisely, we considered the following alternative baselines: fine-tune related $BERT_{large}$, $T5_{large}$ (Raffel *et al.* 2020), $ALBERT_{large}$ (Lan *et al.* 2019), $GPT2_{large}$ (Radford *et al.* 2019), $DeBERTa_{large}$ (He *et al.* 2020), and $RoBERTa_{large}$ from HuggingFace repository with recommended hyperparameters. We assessed the performance of these six learning alternatives in turn using the same training parameters. We used EA as the training set and tested the performance on the CHT test set. We observed the average accuracy, Macro-F1, and F1 score evaluated on the minority/positive class, that is the F1 score of the hate class, over four runs.

According to Table 5, we observe that the Macro-F1 score of the $RoBERTa_{large}$ model is higher than other fine-tune-based models. Although the $DeBERTa_{large}$ model accuracy is higher than RoBERTa, our focus is on the model's classification effect on the positive class and its ability to extract features related to hate speech as an encoder. Our primary goal is to improve the fine-tune performance without introducing extra annotated datasets robustly. In this way, these experiments confirm the validity of selecting RoBERTa model to be included in our proposed solution.

The following experiments in Section 5.4 show that, in the case of limited labeled sequences available, our proposed SSL-GAN-RoBERTa solution is able to provide important improvements in the performance when classifying complex tasks such as racial discrimination classification.

https://doi.org/10.1017/S1351324923000396 Published online by Cambridge University Press

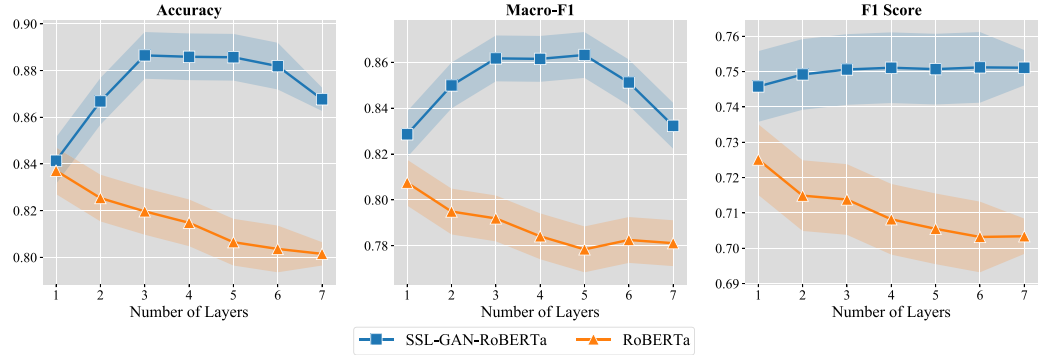

Figure 5. Performance comparison with the increment of linear layers.

5.6 Effect of extra parameters

In this set of experiments, our goal is to study the impact in the performance of the extra linear layers applied in the RoBERTa Encoder. Supported by previous findings on complementary generator theory in classification tasks which are described by Dai *et al.* (2017) and based on empirical experiments (Salimans *et al.* 2016; Ulyanov *et al.* 2018), to build a more robust classifier, we stacked multiple linear layers after RoBERTa's layers. In our proposed SSL-GAN-RoBERTa model, we use multiple RoBERTa layers to extract features from both the labeled and the unlabeled sequences. In Figure 1 we show our SSL-GAN-RoBERTa with three extra linear layers added to the RoBERTa Encoder for illustration purposes. Instead of directly feeding the features into the discriminator, we gradually add up to seven additional linear layers as an extra linear stack to extend the extracted features.

In these experiments, we tested the RoBERTa-large model and the SSL-GAN-RoBERTa with a number of extra linear layers in $\{1, 2, 3, 4, 5, 6, 7\}$. The parameters were set as follows: batch size = 16, optimizer = Adam, learning rate = 2e-6, dropout rate = 0.3, and epochs = 3. Figure 5 shows the results of these experiments.

We observe that, as the number of linear layers of RoBERTa encoder in SSL-GAN-RoBERTa algorithm increases, the model's performance also exhibits a corresponding growth trend. This growing performance trend is consistent for both Macro-F1 and Accuracy on SSL-GAN-RoBERTa (blue and red lines in Figure 5). Thus, this confirms the positive impact of adding these extra linear layers in SSL-GAN-RoBERTa. Based on our experiments, by adding up to five additional linear layers the model's Macro-F1 score is stably improved, and SSL-GAN-RoBERTa converges better on the overall task. However, when adding six and seven extra layers we observe a decreasing tendency in the performance. Still, we must highlight that the lowest performance scores are achieved with one added extra layer.

When observing the performance of RoBERTa model in this setting we verify a completely different tendency in the performance. In effect, we observe that RoBERTa classifier's Macro-F1 score and accuracy consistently decreased with the addition of extra linear layers while the opposite is verified for SSL-GAN-RoBERTa. We hypothesize that the reason for this observed behavior is related to a constraint associated with the classification performance on the RoBERTa pre-trained parameters. In RoBERTa case, adding an extra linear layer seems to lead to over-fitting on the EA training sequences due to the high model complexity which translates into a performance decrease.

However, when we embed the RoBERTa extra linear layers into the SSL-GAN-RoBERTa structure as an encoder, it shows an opposite effect boosting the overall performance. As it can be observed in Figure 5, SSL-GAN-RoBERTa reaches its peak Macro-F1 score with three extra linear

https://doi.org/10.1017/S1351324923000396 Published online by Cambridge University Press

layers. Our empirical experimental results validate the theory of complementary generator (Dai *et al.* 2017). Under the circumstance that the generator model parameters remain unchanged, by adding extra linear layers into the RoBERTa encoder (or discriminator to a certain extent), the performance of the model on the classification task can be improved. We observe that both the generator and the discriminator benefit from the extra linear layers stack.

5.7 Error analysis

In order to further analyze and understand the model errors, we investigated the results predicted by the SSL-GAN-RoBERTa model with 9k unlabeled sequences using CHT dataset as the test set. In our experiments, we found that there are 295 items that are misclassified cases, out of which 149 (50.29%) have the ground truth label of 1 (hate), which means the tweets that contain hate speech were incorrectly predicted to be 0 (non-hate), and the remaining 147 (49.71%) cases had the ground truth label of 0 (non-hate) and were incorrectly predicted as 1 (hate). Overall, this shows that the misclassified cases do not seem biased towards on type of error. We analyzed the misclassified tweets and summarized the three following possible reasons for the observed errors.

- **Misjudgment of high-frequency keywords like COVID, virus, racism**. Under normal circumstances, most hate speech contains the keywords: COVID, coronavirus, flu, or racism, so the model will misjudge sentences that contain these words but no discriminatory and hateful meanings. Examples of this type of error are displayed in Table 6.
- **Annotation error**. Errors in labeling the database caused the model's predicted results to be inconsistent with the actual results (the ground truth error). Table 6 provides some examples of this type of error.
- **Discussion about virus-related events**. This reason is similar to the first one, but the discussion mainly centered on China, COVID-19, and other incidents that do not have any racial discrimination or hatred. Examples of errors of this type are displayed in Table 6.
- **Prediction error.** Errors that are produced by SSL-GAN-RoBERTa where sentences have shown apparent Anti-Asian speech or malicious phrases towards East Asian groups but yet not being identified by the model are displayed in Table 6.

The observed errors show the complexity of this predictive task where the meaning of the sentences involving hate is difficult to capture.

5.8 Experimental evaluation on sentiment classification domain

In these experiments, to validate the effectiveness of SSL-GAN-RoBERTa algorithm in other tasks, we selected datasets from an entirely different domain associated with sentiment classification to validate the performance gains of our proposal. Table 2 provides the main characteristics of the used datasets, and their description is provided in Section 4.3.

To carry out these experiments, we considered the in-domain transfer scenario described in Section 4.1 where labeled sequences from the same domain are used in the training and testing stages and unlabeled sequences from a different related domain are used as a way of improving the generalization capability of the learned model. We selected this scenario because we had enough labeled sequences on the selected dataset domain which allowed to use the same domain dataset both for train and test. We used SST-2 as the labeled training set and the SST-2 test set as the final test set. Dataset IMDB was used as the source of unlabeled sequences.

Our experiments included training the original RoBERTa model as well as our proposed SSL-GAN-RoBERTa with different sentence amounts of unlabeled sequences including the particular case of not using unlabeled sequences. We experimented with the following sentence amounts

https://doi.org/10.1017/S1351324923000396 Published online by Cambridge University Press

Table 6. Examples of misclassified tweets caused by *Misjudgment of high-frequency keywords*, *Annotation error*, *Discussion about virus-related events,* and *Prediction error.*

Error type	Tweets	Reason analysis	Prediction	Label	Ratio
Annotation error	it s appalling that the media amp libtards bitch about the virus being referred to as the chinese virus but no one seems to give a shit about the poor lives of the dogs amp cats that are being eaten alive amp tortured for food why isn t peta raising hell about this just evil.	This is a discussion about virus, but the original annotation misclassified it to hate speech.	0	1	20.4%
	it s appalling that the media amp libtards bitch about the virus being referred to as the chinese virus but no one seems to give a shit about the poor lives of the dogs amp cats that are being eaten alive amp tortured for food why isn t peta raising hell about this just evil	This is a discussion about virus, but the original annotation misclassified it to hate speech.	1	0	
Misjudgment of high-frequency keywords	chinaliedpeopledied it s all because of this fucking country	Since there is no obvious keyword like ‘virus’ or ‘covid’, the model misclassified it into not hate.	0	1	37.5%
	damned if you do damned if you dont scientists question chinas decision not to report symptom free coronavirus cases	This is not a hate speech, but due to the keyword ‘coronavirus’ and ‘china’, the model misclassified it into hate.	1	0	
	ciaspygirl so blame china for the swine flu because after all in china they eat pork racismisavirus	This is not a hate speech, but due to the keyword ‘virus’ and ‘racism’, the model misclassified it into hate.	1	0	
Discussion about virus-related events	novel coronavirus many times deadlier then the flu abc news via googlenews how do you keep pedaling this lie and sleep at night china has a pop of 1 4 billion coronavirus started in 11 19 killed 3200 flu has killed 20 000 in america since jan 2020	This is a discussion about ‘covid’ but due to the keyword ‘virus’ and ‘china’ and ‘kill’, it was misclassed into hate speech.	1	0	18.7%
Prediction error	covid19 so this all started cuz some chinks want bat soup	This is an obvious anti-Asian speech but yet not being classified by the model	1	0	23.4%

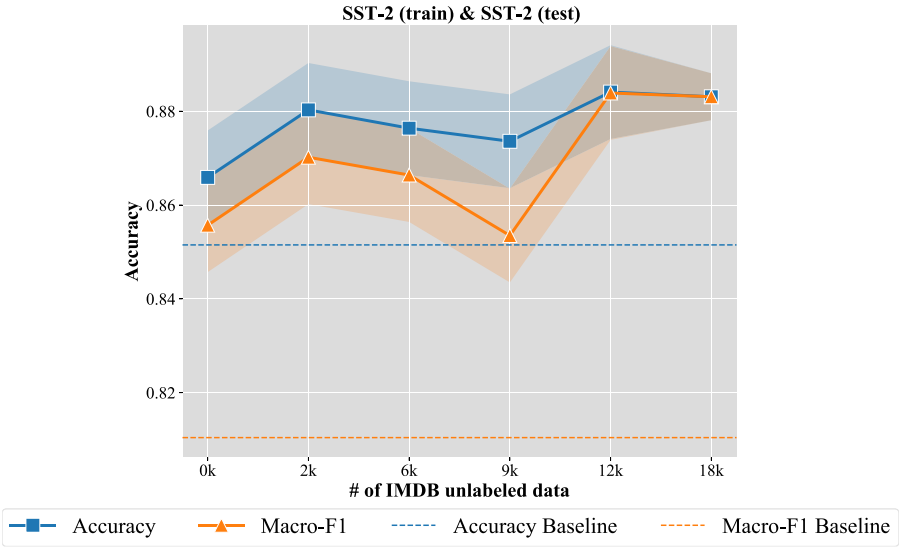

Figure 6. Comparison of the accuracy and Macro-F1 score results when using different model configurations on the SST-2 test data.

of unlabeled sequences on algorithm SSL-GAN-RoBERTa $\{0k, 2k, 6k, 9k, 12k, 18k\}$ and considered three extra linear layers on the RoBERTa Encoder module. The hyperparameters of the generator and discriminator on the SSL-GAN-RoBERTa were set as follows: batch size = 16, optimizer = Adam, learning rate = 2e-6, dropout rate = 0.3, and epochs = 3. We selected these parameters based on the previous experiments on the Anti-Asian COVID-19 hate speech detection.

The results of accuracy and Macro-F1 scores obtained in this experiment are summarized in Figure 6. We observe that the baseline RoBERTa model presents a reasonable performance of roughly 85% accuracy and 82% Macro-F1 score. However, we verify that training our SSL-GAN-RoBERTa even without the use of unlabeled sequences provides a boost in the performance of both metrics. In effect, the new proposed model is able to take advantage of the GAN embedded in the learning process and the extra linear layers used in the RoBERTa Encoder module. We can also confirm that introducing unlabeled sequences on a related knowledge domain in the SSL-GAN-RoBERTa structure can further improve the model's accuracy and Macro-F1 score on the original test set. The IMDB unlabeled dataset brings important benefits to the learning process.

Summary: The experiments carried out in this and previous subsections empirically validate the effectiveness of our SSL-GAN-RoBERTa in the knowledge transfer task on both hate speech and movie review domains. Overall, we observe that our proposed solution is effective for tackling the complex problem of Anti-Asian COVID-19 hate speech detection but is also useful for other domains. Based on the experiments displayed in Figure 6, we can observe the same trend as in the Anti-Asian COVID-19 hate speech domain dataset. Moreover, the extensive set of experiments carried out confirm the advantages of adding extra linear layers in the RoBERTa encoder module as well as the embedding of a GAN in the learning process.

6. Conclusion and future work

In this paper, we propose SSL-GAN-RoBERTa, a new learning algorithm for tackling the Anti-Asian COVID-19 hate speech detection problem. Our solution combines the transformer-based RoBERTa model and the GAN structure into a novel semi-supervised learning model. This allows

https://doi.org/10.1017/S1351324923000396 Published online by Cambridge University Press

the use of the vast amounts of unlabeled data available to improve the predictive performance in complex application domains. Our solution is especially relevant for application domains where the volume of available labeled data is scarce or heavily restricted and difficult to obtain. We show the effectiveness of our solution when using labeled and unlabeled datasets from related domains. With SSL-GAN-RoBERTa, we can accomplish transfer learning by using a related domain to obtain predictions in a different one. Our proposed solution confirms the complement generator theory in the GAN structure that helps the model to capture data features.

By extensively comparing the results of multiple experiments, we show that an appropriate amount of sentences of unlabeled data (the best size of unlabeled data is roughly half of the size of the labeled data) can help the model achieve better performance. Moreover, in the fine-tuning stage, we show that adding linear layers in our RoBERTa encoder module helps the model to capture data features more efficiently and accurately. This is a relevant step as it allows to better capture features from two different domains, which has an important impact on the performance results of the model.

Our experiments also have some limitations, such as being restricted to the English language. Since the data we used for all the experiments is in English language (tweets, movie reviews, etc.), the model's performance in languages other than English still needs to be investigated. Also, since the primary purpose of our experiments is to detect Anti-Asian and COVID-19-related negative remarks, the labeled data that can be used for model training are very limited. We believe there is an urgent need to develop more high-quality labeled datasets on this specific topic. We consider that this is a highly relevant future work direction that could strongly impact the advancement of research in this field. We also look forward to more high-quality datasets with a broader range of content, more language types, and high-quality annotations in the future. Other promising future research directions include carrying out experiments in a multilingual setting and extending our proposed algorithm from a binary to a multi-class scenario.

We hope to provide some inspiration and foundation for future experiments in many NLP classification tasks that could use our proposed SSL-GAN-RoBERTa architecture enabling the use of the vast amounts of frequently available unlabeled data. Another future direction of investigation is to explore new ways of combining the two deep learning frameworks (RoBERTa and GAN) in order to build more robust semi-supervised models. To facilitate the easy reproducibility of our work and further research, we made our code freely available at https://github.com/Jackline97/GAN_RoBERTa.

References

Abroshan H., Devos J., Poels G. and Laermans E. (2021). COVID-19 and phishing: effects of human emotions, behavior, and demographics on the success of phishing attempts during the pandemic. *IEEE Access* **9**, 121916–121929.

Agarwal S. and Chowdary C. R. (2021). Combating hate speech using an adaptive ensemble learning model with a case study on COVID-19. *Expert Systems with Applications* **185**, 115632.

Ali R., Farooq U., Arshad U., Shahzad W. and Beg M. O. (2022). Hate speech detection on Twitter using transfer learning. *Computer Speech & Language* **74**, 101365.

Berthelot D., Carlini N., Goodfellow I., Papernot N., Oliver A. and Raffel C. A. (2019). Mixmatch: a holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, vol. **32**.

Bigoulaeva I., Hangya V. and Fraser A. (2021). Cross-lingual transfer learning for hate speech detection. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pp. 15–25.

Conneau A., Khandelwal K., Goyal N., Chaudhary V., Wenzek G., Guzmán F., Grave E., Ott M., Zettlemoyer L. and Stoyanov V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 8440–8451.

Costello M., Cheng L., Luo F., Hu H., Liao S., Vishwamitra N., Li M. and Okpala E. (2021). COVID-19: a pandemic of Anti-Asian cyberhate. *Journal of Hate Studies* **17**(1), 108–118.

Croce D., Castellucci G. and Basili R. (2020). 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*. Online: Association for Computational Linguistics, pp. 2114–2119.

https://doi.org/10.1017/S1351324923000396 Published online by Cambridge University Press

Dai Z., **Yang Z.**, **Yang F.**, **Cohen W. W. and Salakhutdinov R.** (2017). Good semi-supervised learning that requires a Bad GAN. In *Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS'17*. Red Hook, NY: Curran Associates Inc., pp. 6513–6523.

Davidson T., **Warmsley D.**, **Macy M. and Weber I.** (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, vol. **11**, pp. 512–515.

Davidson T., **Bhattacharya D. and Weber I.** (2019). Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*. Florence: Association for Computational Linguistics, pp. 25–35.

de Gibert O., **Perez N.**, **García-Pablos A. and Cuadros M.** (2018). Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. Brussels: Association for Computational Linguistics, pp. 11–20.

Denton E. L., **Gross S. and Fergus R.** (2017). Semi-supervised learning with context-conditional generative adversarial networks. In *The International Conference on Learning Representations (ICLR)*. Palais des Congrès Neptune, Toulon: 5th International Conference on Learning Representations, p. 10.

Devlin J., **Chang M.-W.**, **Lee K. and Toutanova K.** (2019). 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, Volume 1 (Long and Short Papers)*. Minneapolis, MA: Association for Computational Linguistics, pp. 4171–4186.

Fan L., **Yu H. and Yin Z.** (2020). Stigmatization in social media: documenting and analyzing hate speech for COVID-19 on Twitter. *Proceedings of the Association for Information Science and Technology* **57**(1), e313.

Founta A., **Djouvas C.**, **Chatzakou D.**, **Leontiadis I.**, **Blackburn J.**, **Stringhini G.**, **Vakali A.**, **Sirivianos M. and Kourtellis N.** (2018). Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the International AAAI Conference on Web and Social Media*, vol. **12**.

He B., **Ziems C.**, **Soni S.**, **Ramakrishnan N.**, **Yang D. and Kumar S.** (2021). Racism is a virus: Anti-Asian hate and counterspeech in social media during the COVID-19 crisis. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 90–94.

He P., **Liu X.**, **Gao J. and Chen W.** (2020). *DeBERTa: Decoding-enhanced BERT with Disentangled Attention*. arXiv preprint arXiv:2006.03654

Johnson C. and John T. (2021). Atlanta spa attacks shine a light on anti-Asian hate crimes around the world. *CNN* (accessed 14 August 2021). https://edition.cnn.com/2021/03/21/world/anti-asian-hate-crime-intl/index.html

Kennedy B., **Atari M.**, **Davani A. M.**, **Yeh L.**, **Omrani A.**, **Kim Y.**, **Coombs K.**, **Portillo-Wightman G.**, **Havaldar S.**, **Gonzalez E.**, **Hoover J.**, **Azatian A.**, **Cardenas G.**, **Hussain A.**, **Lara A.**, **Omary A.**, **Park, C.**, **Wang X.**, **Wijaya C.**, **Zhang Y.**, **Meyerowitz B. and Dehghani M.** (2021). The Gab Hate Corpus. https://gwenythjpw.com/publication/2021-01-01-hate-annotation-paper.

Lan Z., **Chen M.**, **Goodman S.**, **Gimpel K.**, **Sharma P. and Soricut R.** (2019). ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *CoRR*, **abs/1909.11942**.

Lee V. L. S., **Gan K. H.**, **Tan T. P. and Abdullah R.** (2019). Semi-supervised learning for sentiment classification using small number of labeled data. *Procedia Computer Science* **161**, 577–584.

Li J. and Ning Y. (2022). Anti-Asian hate speech detection via data augmented semantic relation inference. In *Proceedings of the International AAAI Conference on Web and Social Media*, vol. **16**, pp. 607–617.

Li Y. and Ye J. (2018). Learning adversarial networks for semi-supervised text classification via policy gradient. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1715–1723.

Liu Y., **Ott M.**, **Goyal N.**, **Du J.**, **Joshi M.**, **Chen D.**, **Levy O.**, **Lewis M.**, **Zettlemoyer L. and Stoyanov V.** (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.

Lora J., **Palumbo D. and Brown D.** (2021). Coronavirus: How the pandemic has changed the world economy (accessed 17 August 2021). https://www.bbc.co.uk/news/business-51706225

Maas A. L., **Daly R. E.**, **Pham P. T.**, **Huang D.**, **Ng A. Y. and Potts C.** (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language*. Portland, OR: Association for Computational Linguistics, pp. 142–150.

Mathew B., **Saha P.**, **Yimam S. M.**, **Biemann C.**, **Goyal P. and Mukherjee A.** (2021). Hatexplain: a benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. **35**, pp. 14867–14875.

Miyato T., **Maeda S.-i.**, **Koyama M. and Ishii S.** (2018). Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **41**(8), 1979–1993.

Mozafari M., **Farahbakhsh R. and Crespi N.** (2020). A BERT-based transfer learning approach for hate speech detection in online social media. In *Complex Networks and Their Applications VIII: Volume 1 Proceedings of the Eighth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2019*. Cham: Springer, pp. 928–940.

Müller M., **Salathé M. and Kummervold P. E.** (2023). Covid-twitter-bert: a natural language processing model to analyse covid-19 content on twitter. *Frontiers in Artificial Intelligence* **6**, 1023281.

https://doi.org/10.1017/S1351324923000396 Published online by Cambridge University Press

Nguyen D. Q., **Vu T. and Tuan Nguyen A.** (2020). BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, pp. 9–14.

Radford A., **Wu J.**, **Child R.**, **Luan D.**, **Amodei D. and Sutskever I.** (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, **1**(8), 9.

Raffel C., **Shazeer N.**, **Roberts A.**, **Lee K.**, **Narang S.**, **Matena M.**, **Zhou Y.**, **Li W. and Liu P. J.** (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research* **21**(140), 1–67.

Salimans T., **Goodfellow I.**, **Zaremba W.**, **Cheung V.**, **Radford A. and Chen X.**, et al. (2016). Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, vol. **29**, pp. 2234–2242.

Socher R., **Perelygin A.**, **Wu J.**, **Chuang J.**, **Manning C. D.**, **Ng A. Y. and Potts C.** (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642.

Springenberg J. T. (2016). Unsupervised and Semi-supervised Learning with Categorical Generative Adversarial Networks. *CoRR*, **abs/1511.06390**.

Tekumalla R., **Baig Z.**, **Pan M.**, **Hernandez L. A. R.**, **Wang M. and Banda J.** (2022). Characterizing Anti-Asian rhetoric during the COVID-19 pandemic: a sentiment analysis case study on Twitter. In *Workshop Proceedings of the 16th International AAAI Conference on Web and Social Media*. Retrieved from https://doi.org/10.36190

Thain N., **Dixon L. and Wulczyn E.** (2017). *Wikipedia Talk Labels: Toxicity*. https://figshare.com/articles/dataset/Wikipedia_Talk_Labels_Toxicity/4563973

Ulyanov D., **Vedaldi A. and Lempitsky V.** (2018). It takes (only) two: adversarial generator-encoder networks. In *Thirty-Second AAAI Conference on Artificial Intelligence. Online: Thirty-Second AAAI Conference on Artificial Intelligence*.

Vidgen B. and Derczynski L. (2020). Directions in abusive language training data, a systematic review: garbage in, garbage out. *Plos One* **15**(12), e0243300.

Vidgen B., **Hale S.**, **Guest E.**, **Margetts H.**, **Broniatowski D.**, **Waseem Z.**, **Botelho A.**, **Hall M. and Tromble R.** (2020). Detecting East Asian prejudice on social media. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*. Online: Association for Computational Linguistics, pp. 162–172.

Wang A., **Singh A.**, **Michael J.**, **Hill F.**, **Levy O. and Bowman S.** (2018). GLUE: a multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Brussels: Association for Computational Linguistics, pp. 353–355.

Wang A., **Pruksachatkun Y.**, **Nangia N.**, **Singh A.**, **Michael J.**, **Hill F.**, **Levy O. and Bowman S.** (2019). Superglue: a stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, vol. **32**.

Wei B., **Li J.**, **Gupta A.**, **Umair H.**, **Vovor A. and Durzynski N.** (2021). Offensive language and hate speech detection with deep learning and transfer learning. arXiv preprint arXiv: 2108.03305.

Yuan L., **Wang T.**, **Ferraro G.**, **Suominen H. and Rizoiu M.-A.** (2019). Transfer learning for hate speech detection in social media. arXiv preprint arXiv: 1906.03829.

Cite this article: Su X, Li Y, Branco P and Inkpen D. SSL-GAN-RoBERTa: A robust semi-supervised model for detecting Anti-Asian COVID-19 hate speech on social media. *Natural Language Engineering* https://doi.org/10.1017/S1351324923000396