2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA) | 978-1-6654-4337-1/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICMLA52953.2021.00043

COVID-HateBERT: a Pre-trained Language Model for COVID-19 related Hate Speech Detection

Mingqi Li
School of Computing
Clemson University
Clemson, USA
mingqil@clemson.edu

Song Liao
School of Computing
Clemson University
Clemson, USA
liao5@clemson.edu

Ebuka Okpala
School of Computing
Clemson University
Clemson, USA
eokpala@clemson.edu

Max Tong*
School of Computing
Clemson University)
Clemson, USA
tongm23@ccescavs.org

Matthew Costello
Department of Sociology
Clemson University
Clemson, USA
mjcoste@clemson.edu

Long Cheng
School of Computing
Clemson University
Clemson, USA
lcheng2@clemson.edu

Hongxin Hu
Department of Computer Science and Engineering
University at Buffalo
Buffalo, USA
hongxinh@buffalo.edu

Feng Luo
School of Computing
Clemson University
Clemson, USA
luofeng@clemson.edu

***Abstract*—With the dramatic growth of hate speech on social media during the COVID-19 pandemic, there is an urgent need to detect various hate speech effectively. Existing methods only achieve high performance when the training and testing data come from the same data distribution. The models trained on the traditional hateful dataset cannot fit well on COVID-19 related dataset. Meanwhile, manually annotating the hate speech dataset for supervised learning is time-consuming. Here, we propose COVID-HateBERT, a pre-trained language model to detect hate speech on English Tweets to address this problem. We collect 200M English tweets based on COVID-19 related hateful keywords and hashtags. Then, we use a classifier to extract the 1.27M potential hateful tweets to re-train BERT-base. We evaluate our COVID-HateBERT on four benchmark datasets. The COVID-HateBERT achieves a 14.8%-23.8% higher macro average F1 score on traditional hate speech detection comparing to baseline methods and a 2.6%-6.73% higher macro average F1 score on COVID-19 related hate speech detection comparing to classifiers using BERT and BERTweet, which shows that COIVD-HateBERT can generalize well on different datasets.**

***Index Terms*—hate speech detection, language model, COVID-19, BERT**

I. Introduction

Hate speech is commonly defined as languages that instigate hate or violence to a group of people, usually targeting their nationalities, race, gender, religion, sexual orientation, or other [1]. Hate speech detection on social media has drawn attention to researchers in recent years. Especially with the dramatic growth of discussions about Coronavirus Disease (COVID-19) on social media such as Twitter, Facebook, etc., various hate speech has been generated. For instance, during the early stage of the COVID-19 pandemic, "kung flu" and "chop fluey", the terms against Asian Americans are shared more than 10,000 times on Twitter [2]. Simultaneously, the hashtags like #BoomerRemover derived during the pandemic show discrimination against older people [3]. Hate speech can not only be limited to words, but also can lead to real hate crimes. On the CAAA3PCON STOP AAPI HATE website, 1497 COVID-19 related incidents are reported in four weeks, even though AAPIs are not actively interacting with other people and most areas implement shelter-in-place policies [4]. Alshalan et al. [5] concluded that there might be a direct correlation between the spread of hate speech and real hate crimes during the COVID-19 pandemic. Therefore, the need to detect hate speech on social media effectively has never been more urgent. Even though some social platforms have their tools to detect traditional hate speech, a large amount of COVID-19 related hate speech remains, making COVID-19 related hate speech detection a challenging task.

For traditional hate speech detection, existing works achieved impressive performance. For example, Badjatiya et al. [6] conducted deep learning architectures that outperform baseline models by an 18% F1 score. Zhang et al. [7] introduced a Convolution-GRU architecture that outperformed state-of-the-art methods on several publicly available Twitter datasets. However, these works depended on the consistency of data distribution on the training and testing sets. In other words, their generalization ability is limited. Arango et al. [8] pointed that some prior works had methodological issues and demonstrated that their performance would be worse when using another testing set. Gröndahl et al. [9] concluded that the type of data was more important than model architecture. We speculate these existing models for traditional hate speech cannot perform well on COVID-19 related hate speech detection, since people use novel jargon and vocabularies related to COVID-19, which is unseen by traditional training set [2]. Most of the traditional hateful tweets target certain races, women, LGBTQ+, and some religions. Meanwhile, COVID-19 related hateful tweets generate new types of hate, such as hating masks, vaccines, and older people, which have different data distribution from traditional hate speech. Due to the limited generalization ability of existing works, it is imperative

*Intern from Christ Church Episcopal School, Greenville, SC, USA

978-1-6654-4337-1/21/$31.00 ©2021 IEEE
DOI 10.1109/ICMLA52953.2021.00043

to train a model using domain-specific data. Recently, Ziems et al. [10] collected a dataset of anti-Asian hate, including over 30 million tweets, and they annotated 2,400 tweets. Using this annotated data, they train a classifier with an average AUROC of 0.852. In addition, existing supervised learning methods need a large amount of annotated data, which is expensive and time-consuming. Here, we aim to train a model with limited labeled data that can generalize well on different datasets and detect COVID-19 related hate speech more effectively.

Language models that can learn general linguistic representations have been applied to many Natural Language Processing (NLP) downstream tasks and achieve state-of-the-art results. The pre-training process can use large amounts of unlabeled data, which is easier to access. Bidirectional Encoder Representations from Transformers (BERT) [11] is a language model pre-trained on large corpora whose variants made progress in various NLP tasks. Beltagy et al. [12] proposed SCIBERT using a large amount of unlabeled scientific publications to achieve state-of-the-art results on scientific NLP tasks. Nguyen et al. [13] proposed BERTweet that was pre-trained on English Tweets, and the performance was better than state-of-the-art models on several Tweet NLP tasks. Müller et al. [14] released CT-BERT, which is pre-trained on COVID-19 related tweets and improved 10-30% compared to BERT-large on five classification tasks. Here, we build a language model for COVID-19 related hateful tweets to increase the performance of detecting COVID-19 related hate speech and the model's generalization ability.

Our main contributions can be summarized as follows:

1) We collect a new Twitter dataset based on COVID-19 related hashtags. We explore six topics about COVID-19 and obtain 121 hashtags from these topics. Then, we collect 200M tweets between Jan 1, 2020, and Apr 1, 2021. Moreover, we use a classifier to extract 1.27M potential hateful data.
2) We build COVID-HateBERT, which is re-trained based on Bert-base using 1.27M potential hateful tweets. COVID-HateBERT is a pre-trained language model for COVID-19 related hate speech detection on English tweets.
3) We evaluate COVID-HateBERT on traditional hate speech datasets and COVID-19 related hate speech datasets. Our results show that COVID-HateBERT outperforms our baselines on both traditional hate and COVID-19 related hate datasets.

II. Related Work

A. Traditional hate speech detection

Due to a large amount of information, social media needs to detect and prevent hate speech effectively, which means manual detection can not meet the requirements. Currently, traditional machine learning methods and deep neural networks made progress in detecting hate speech. Davidson et al. [15] and Waseem et al. [16], [17] collected tweets with hateful keywords and labeled them using a list of criteria. Agrawal and Awekar [18] performed experiments on several hate speech datasets and compared the results between traditional machine learning models and DNN models. Badjatiya et al. [6] also experimented on different models with various tweet embeddings. However, Arango et al. [8] pointed out the weakness of prior work and proposed a novel method to solve bias issues of datasets. Recently, research focused on hate speech detection has sharply increased, and several tasks such as HatEval-2019 [19] and OffensivEval-2019 [20] were proposed to improve hate speech detection. Among teams that participated in HatEval-2019 [19], the Fermi team achieved the highest macro average F1-score (0.651) using Universal Sentence Encoder [21] as embeddings and an SVM model with RBF kernel. OffensivEval-2019 [20] reported that the best performing team, NULI [22], used pre-trained BERT model to achieve 0.829 F1-score. Additionally, Caselli et al. [23] applied their proposed annotation guidelines to OLID/OffensEval [20] to create a new English dataset, AbuseEval v1.0.

B. COVID-19 related hate speech detection

With the increase in people's discussion of the pandemic, the COVID-19 related hate speech on social media has also increased, so some researchers researched COVID-19 related hate speech. Ziems et al. [10] created anti-Asian hate during the pandemic and found that hateful users became more engaged after they posted their first anti-Asian tweet. Vidgen et al. [24] created a classifier to identify east Asian prejudice on Twitter. Alshalan et al. [5] used a CNN to identify hate speech on Arabic tweets and showed that most hate speeches targeted China and Iran. Fan et al. [25] collected over 3M tweets and identify 25,457 hate speech. They analyzed these hate speech based on demographics and emotions and found significant associations between them. Hardage et al. [26] aimed to train a model without using existing data in order to solve a real-world problem like COVID-19 hate speech detection. They proposed a novel algorithm that used global feature importance to penalize or reinforce predictions when there is a difference between local and global feature importance, trained the model on traditional hate datasets, and tested on COVID-19 related hate datasets. Our work also targeted unseen hateful data, but we used less traditional hateful data, evaluated with both same and different data distribution, and compared with state-of-the-art baselines. Since current research on hate speech related to the pandemic is not comprehensive, and COVID-19 related datasets are limited, we aim to further study it and fill the gap.

C. Language model

Transformer-based models like BERT [11] created a strong baseline in various NLP downstream tasks. In recent years, researchers proposed substantial language models such as GPT [27], RoBERTA [28], and XLNet [29] which were trained with large amounts of unlabeled data. However, these language models trained on the general domain yielded unsatisfactory results on specific domains, since the word distribution is different. Therefore, researchers focus on pre-training language

models with large amounts of domain-specific data to further make improvements. For instance, Lee et al. [30] proposed a domain-specific language model, BioBERT, which outperformed BERT on three representative biomedical text mining tasks. Gururangan et al. [31] pre-trained RoBERTA [28] on domain-specific text such as biomedical and computer science papers, news, and amazon reviews, and showed that domain-adaptive pre-training improved performance. Nguyen et al. [13] released BERTweet, which was pre-trained on a large-scale English tweets dataset. BERTweet yielded impressive results on three tweet NLP tasks that outperformed RoBERTA-base [28] and XLM-R-base [32]. Müller et al. [14] released CT-BERT pre-trained on COVID-19 related tweets. CT-BERT achieved 10-30% improvements compared to BERT-LARGE. Caselli et al. [33] proposed HateBERT, which was pre-trained for abusive language detection on a Reddit comments dataset. They explored abuse-inclined version evaluating on datasets for offensive, abusive language, and hate speech detection tasks. We also pre-train a language model for hate speech detection, but the difference is that our dataset is COVID-19 related, and we aim to improve the generalization ability of the model.

III. COVID-HATEBERT

A. Model configuration

BERT [11] is widely used on NLP downstream tasks and achieves state-of-the-art performance. Pre-training BERT through task-specific data and fine-tuning on downstream tasks can be an effective method [12] [13]. We train our model based on BERT-base and use masked language modeling tasks as an objective in this work.

B. Data collection

To collect tweets with potential hateful content related to COVID-19, first, we collected real-time tweets using Twitter Streaming API with two hashtags about COVID-19. We started with the essential hashtags "coronavirus" and "Covid-19". Then we found six hot topics with many discussions towards different groups or individuals, such as discussion about Asians, Trump, or "Boomerremover", which means old people who have a higher risk of being infected by a coronavirus. Other tweets might discuss Mask or Fauci. In each topic, we could find several hateful hashtags and other COVID-19 related hashtags. For example, "Chinavirus" is commonly used in tweets about Asians, while "Trumpvirus" is created for potential hate toward Trump. Next, we explored new hashtags in each topic by checking the frequency of new hashtags. For example, we began with hashtags "Chinavirus" and "Chinesevirus" in the Asian-hate topic and searched for other hashtags that frequently appeared with existing hashtags, such as "wuhanvirus" and "kungflu". Then we added them to our hashtags set and collected tweets that contained these hashtags. As a result, we obtained 41 hashtags about ten different types of hate towards individuals or groups and 80 COVID-19 related hashtags. Since Twitter would provide only one percent of real-time tweets, we could not get all these hashtags from real-time tweets. So we used a tool named "snscrape" [34] to collect all the past tweets id related to hashtags and then used the Twitter official API to get the content of the tweets. Based on this method, we collected 200M tweets with 121 COVID-19 related hashtags from Jan 1, 2020, to Apr 1, 2021. All the topics and hashtags are listed in Table I.

TABLE I
121 COVID-19 RELATED HASHTAGS FOR DATA COLLECTION

coronavirus, covid19, pandemic, virus, outbreak, plandemic, china, stayhome, covid-19, covid, corona, wuhan, covid_19, lockdown, coronavirusoutbreak, stayathome, socialdistancing, pandemic, coronaoutbreak, stayhomestaysafe, staysafestayhome, covid__19, covid-19, learntherisk, howwegothere, nonewnormal, gopsuperspreaders, herdimmunity, stopthecovidchaos, lockdown2, wuhancoronavirus, wuhanvirus, chinesevirus, chinavirus, coronaviruschina, ccpvirus, chinacoronavirus, chinaliedpeopledied, wuflu, kungflu, mask, antimask, maskless, maskfree, unmask, nomask, nomasks, unmaskarizona, unmaskamerica, nomasknancy, nomaskmandate, antimaskers, nomaskmandates, maskoff, maskoffamerican, maskdontwork, maskoffamerica, unmaskthetruth, unmasked, boomerremover, bommer, babyboomers, babyboomer, boomers, boomersooner, okboomer, vaccine, vaccines, coronavirusvaccine, russianvaccine, covidvaccine, covid19vaccine, vaccineinjury, fluvaccine, rnavaccines, vaccineswork, pfizer, pfizervaccine, covidiot, covidiots, covididiots, covidiotinchief, qanon, qanons, qanondon, killercuomo, trumpvirus, trumpkills, trumppandemic, trumpvirusdeathtoll193k, trumpvirusdeathtoll186k, trumpviruscatastrophe, trumpkillsamericans, trumplied200kdied, trumpliedpeopledied, trumphascovid, trumpcovid, trumpcovid19, trumpcovidhoax, covidcaughttrump, trumpcrimefamily, trumpisbroke, trumpvirusdeathtoll210k, trumpispathetic, trumpcrimefamilyforprison, trumpvirusdeathtoll225k, fauci, billgates, gates, gatesofhell, faucithefraud, drfauci, tonyfauci, drfaucitimecover, faucihero, faucifraud, firefauci, criminalfauci, exposebillgates, followthefauci, #who

C. Potential hate speech

We use a classifier to extract the potential hateful tweets from those 200M tweets to train the task-specific language model. The classifier, built by [10], was trained on their annotated dataset of 2,319 COVID-19 related hateful tweets. They represented tweets using linguistic features such as the number of characters and words, hashtags, and tweet embeddings (BERT). They trained Logistic Regression classifiers and conducted five-fold cross-validation on the three-class classification task. We select the tweets with the "Hate" label and finally get 1.27M tweets. We then use these 1.27M hateful tweets to train our COVID-HateBERT.

IV. EXPERIMENTS

A. Datasets

We use three publicly available Twitter hate datasets and one in-house annotated dataset to evaluate COVID-HateBERT. The datasets are listed in Table II. Since users delete some tweets, we can not retrieve all tweets in three publicly datasets through Twitter API. We acquire all possible data, and the number of tweets in each dataset is listed in the Table II.

We also annotate our in-house COVID-19 related hateful dataset. We use an open-source tool Perspective [35] to select

TABLE II
TWITTER DATASETS USED IN OUR WORK

Dateset	Target	Count
Waseem & Hovy [17]	Race and Gender	10612
HatEval 2019 [19]	Immigrants and Women	9000
COVID-HATE [10]	Asian	2319
In-house COVID-19 dataset	General and Asian	1679

the tweets whose score is greater than 0.8. These tweets totaling 1,679 are labeled as hate and non-hate. Our annotation code considers both the context and target of a tweet, since hateful tweets may not have slurs, and hateful keywords do not necessarily make tweets hateful. For example, tweets that combine an Asian location or a person's name with a virus are labeled as hateful tweets. Three graduate students label hundreds of tweets from different subsets of 1,679 tweets for three rounds and develop additional annotation code after each round during our annotation process. An expert will make the final classification if tweets are labeled differently by the students. Finally, our in-house dataset contains 554 hateful tweets and 1,125 non-hateful tweets.

In our experiments, some datasets have two classes, while others are multiple classes. For datasets with multiple classes, we convert them to hate/non-hate binary classification tasks. For example, Waseem & Hovy [17] has three classes: racist, sexist, and none. We combine the first two classes as hate class. Furthermore, the targets of the four datasets are quite different. HatEval 2019 [19] focuses on hateful to women and immigrants, while COVID-HATE [10] pays attention to Asian hate. Our dataset focuses on general hate tweets and Asian hate tweets.

B. Data preprocessing

The 1.27M potential hate tweets extracted by the classifier are preprocessed to meet the requirements for training a language model. The quality of tweets can affect the representations' generalization ability, thus affecting the model's predictions. It is crucial to clean our data and preprocess the tweets initially. In our experiments, we remove the retweets and tweets that are duplicated. Then, we normalize some special terms such as email, user, time, URL, number, and date. Also, we unpack the hashtags and contractions and correct spells for elongated words. Additionally, we convert each letter to lowercase and remove the emoji. The blank and extremely short tweets (less than five words) will be removed at the last step. Eventually, we have 1.21M tweets for our training.

C. Setups

We utilize the Hugging Face Transformers library [36] to train a language model based on BERT-base. Hugging Face Transformers is implemented via PyTorch and provides general-purpose architectures (BERT, GPT-2, XLM, XLNet, etc.) for NLP. Our model is optimized by Adam [37] and is pre-trained for 70 epochs in 4 days using 3 V100 GPUs. The learning rate is 5e-5, and the batch size is set to 128 per GPU.

TABLE III
EVALUATION RESULTS ON UNSEEN TRADITIONAL HATEFUL DATASET

Method	Precision	Recall	F1 score
Hate Detection			
[6]	68.8	15.4	23.5
[18]	75.3	3.5	6.7
BERT + GBDT	61.01	30.48	40.65
BERTweet + GBDT	59.89	30.50	40.42
COVID-HateBERT + GBDT	60.73	35.69	**44.96**
Non-hate Detection			
[6]	49.6	93.4	64.3
[18]	47.5	98.0	63.0
BERT + GBDT	63.01	85.87	**72.69**
BERTweet + GBDT	62.83	85.18	72.32
COVID-HateBERT + GBDT	64.10	83.27	72.44
Micro Average			
[6]	63.8	54.1	46.1
[18]	62.3	48.4	35.1
BERT + GBDT	62.17	62.59	59.22
BERTweet + GBDT	61.59	62.2	58.91
COVID-HateBERT + GBDT	62.68	63.27	**60.88**
Macro Average			
[6]	59.2	54.4	43.9
[18]	61.4	50.8	34.9
BERT + GBDT	62.01	58.18	56.67
BERTweet + GBDT	61.36	57.84	56.37
COVID-HateBERT + GBDT	62.41	59.48	**58.70**

D. Baselines

We compare our results with results in [8]. They fixed the problems in [6], and [18] and proposed a method to improve the performance. Badjatiya et al. [6] used an Embedding layer, an LSTM, and a fully connected layer as a feature extractor, and trained a Gradient-Boosted [38] Decision Tree as a predictor. Agrawal et al. [18] used DNN models, including an embedding layer, a BiLSTM layer, a fully connected layer, and a softmax layer. Arango et al. [8] added another dataset to alleviate the user-overfitting issue. Additionally, we also compare state-of-the-art language models such as BERT-base and BERTweet.

V. RESULTS

A. Traditional hate speech detection

To evaluate the generalization ability of COVID-HateBERT, we use different datasets as the training set and testing set, which means their data distribution is different. We use Waseem & Hovy dataset [17] as training set and HatEval 2019 dataset [19] as testing set to compare with replicated results in [8]. We do not compare with their improved results since they added an additionally labeled dataset contains 7,006 tweets. We present the results in Table III.

In our experiments, each word is represented by COVID-HateBERT as a 768 dimensions vector, and we use the average of each dimension to get the sentence representation. Then, these representations are fed into Gradient Boosted Decision Tree(GBDT) to train a classifier.

The generalization ability of original methods in [8] is poor, especially on hate detection. For hate detection, COVID-HateBERT improves the F1 score to 44.96%, and it is 4.31%-38.26% higher than other methods. COVID-HateBERT also

TABLE IV
EVALUATION RESULTS ON SINGLE COVID-19 RELATED HATEFUL DATASET

Dataset	Method	Precision	Recall	F1 score
Hate Detection				
COVID-HATE	BERT + GBDT	71.42	49.11	58.08
	BERTweet + GBDT	77.27	56.04	64.82
	COVID-HateBERT + GBDT	79.93	61.65	**69.59**
in-house COVID	BERT + GBDT	71.71	45.54	55.49
	BERTweet + GBDT	69.26	43.75	53.46
	COVID-HateBERT + GBDT	73.24	51.25	**60.24**
Non-hate Detection				
COVID-HATE	BERT + GBDT	81.37	91.77	86.25
	BERTweet + GBDT	83.72	93.17	88.18
	COVID-HateBERT + GBDT	85.53	93.60	**89.38**
in-house COVID	BERT + GBDT	76.97	90.88	83.32
	BERTweet + GBDT	76.26	90.26	82.65
	COVID-HateBERT + GBDT	78.81	90.62	**84.29**
Micro Average				
COVID-HATE	BERT + GBDT	78.46	79.30	78.01
	BERTweet + GBDT	81.84	82.32	81.35
	COVID-HateBERT + GBDT	83.89	84.26	**83.59**
in-house COVID	BERT + GBDT	75.21	75.76	74.04
	BERTweet + GBDT	73.93	74.75	72.92
	COVID-HateBERT + GBDT	76.95	77.49	**76.27**
Macro Average				
COVID-HATE	BERT + GBDT	76.40	70.44	72.17
	BERTweet + GBDT	80.50	74.61	76.50
	COVID-HateBERT + GBDT	82.73	77.63	**79.48**
in-house COVID	BERT + GBDT	74.34	68.21	69.41
	BERTweet + GBDT	72.76	67.00	68.06
	COVID-HateBERT + GBDT	76.02	70.93	**72.26**

TABLE V
EVALUATION RESULTS ON UNSEEN COVID-19 RELATED HATEFUL DATASET

Setting	Method	Precision	Recall	F1 score
Hate Detection				
1	BERT + GBDT	55.67	9.64	16.44
	BERTweet + GBDT	70.54	14.11	23.51
	COVID-HateBERT + GBDT	83.04	16.61	**27.68**
2	BERT + GBDT	43.60	75.81	55.36
	BERTweet + GBDT	44.08	74.19	55.31
	COVID-HateBERT + GBDT	46.48	71.09	**56.21**
Non-hate Detection				
1	BERT + GBDT	68.02	96.16	79.67
	BERTweet + GBDT	69.30	97.05	80.86
	COVID-HateBERT + GBDT	70.20	98.30	**81.91**
2	BERT + GBDT	85.61	59.48	70.20
	BERTweet + GBDT	85.14	61.12	71.16
	COVID-HateBERT + GBDT	84.71	66.18	**74.31**
Micro Average				
1	BERT + GBDT	63.90	67.30	58.58
	BERTweet + GBDT	69.72	69.39	61.74
	COVID-HateBERT + GBDT	74.48	71.05	**63.82**
2	BERT + GBDT	73.33	64.25	65.85
	BERTweet + GBDT	73.14	64.94	66.52
	COVID-HateBERT + GBDT	73.53	67.62	**69.02**
Macro Average				
1	BERT + GBDT	61.84	52.90	48.06
	BERTweet + GBDT	69.92	55.58	52.19
	COVID-HateBERT + GBDT	76.62	57.45	**54.79**
2	BERT + GBDT	64.61	67.64	62.77
	BERTweet + GBDT	64.61	67.66	63.23
	COVID-HateBERT + GBDT	65.60	68.64	**65.26**

outperforms other methods on micro and macro average F1 score [39]. Compared to BERT-base and BERTweet, COVID-HateBERT is on par with them on non-hate detection, but it improves 4.31%-4.54% on hate detection and achieves the best results on micro and macro average F1 score. The limited amount of data in hateful class harms the effectiveness of traditional classifiers for hate speech. On the other hand, our COVID-HateBERT trained on a large specific potential hateful dataset alleviates this problem, and thus achieving better results across different datasets.

B. COVID-19 related hate speech detection

We then evaluate COVID-19 related hate speech detection performance of COVID-HateBERT using the COVID-HATE [10] dataset. The original annotated dataset has three classes: Hate, Counterhate, and Neutral. Here we combine Counterhate and Neutral class as Non-hate. Due to the small amount of labeled data, 5-folds cross-validation is implemented during the evaluation process. The result is shown in Table IV.

Our COVID-HateBERT model outperforms the other two language models on all metrics. For hate detection, the F1 score improves from 58.08% and 64.82% to 69.59%. We can also observe a slight improvement in non-hate detection. Our micro average F1 score is 2.24%-5.58% higher than BERTweet and BERT-base, and the macro average F1 score is 2.98%-7.31% higher. COVID-HateBERT can help achieve impressive results on COVID-19 related hate speech detection.

To further evaluate the performance, we also train and test on our in-house dataset. We present the results in Table IV. Our COVID-HateBERT also outperforms BERT-base and BERTweet. The F1 score of COVID-HateBERT on hate detection improves 4.75%-6.78%. Although COVID-HATE and our dataset are both COVID-19 related hate datasets, the targets are different. COVID-HATE focused on Asian hate, but we annotate different kinds of hate. Our experimental results show that COVID-HateBERT generalizes well on different types of hate speech.

To further verify the generalization ability of COVID-HateBERT, we perform cross-classification using the above two datasets, which train on COVID-HATE dataset and test on the in-house dataset (setting 1), and train on in-house dataset and test on COVID-HATE dataset (setting 2). Table V shows the experimental results. In both two settings, our COVID-HateBERT outperforms BERT-base and BERTweet on all metrics. On micro average F1 score, COVID-HateBERT is 2.08%-5.24% higher than Bertweet and BERT-base in setting 1 and is 2.5%-3.17% higher than them in setting 2. It indicates that our COVID-HateBERT can generalize well on COVID-19 related hateful datasets, even if the type of hate and data distribution is different on the training and testing set.

VI. CONCLUSION

We collect 200M tweets during the COVID-19 pandemic and use a classifier to extract 1.27M potential hateful tweets. We pre-train a language model based on BERT-base targeting on COVID-19 related hate speech detection. Testing on traditional hate speech datasets, COVID-HateBERT outperforms all other methods on hate detection F1 score, micro and macro average F1 score without using extra labeled data. Compared to BERTweet, we use fewer data and time to achieve better results. We evaluate COVID-HateBERT on

COVID-HATE dataset and our in-house COVID-19 dataset. COVID-HateBERT outperforms BERT-base and BERTweet on both datasets, and the F1 score of HateBERT on hate detection significantly improves. Cross classification of COVID-19 related hateful datasets also shows that COVID-HateBERT outperforms its competitors BERT-base and BERTweet. We conclude that our proposed COVID-HateBERT can generalize well on unseen data and achieve impressive results on COVID-19 related hateful datasets.

In future work, we will explore more hateful keywords to track potential hateful tweets according to the shift on hot hateful topics. In addition, we will annotate more tweets to train our classifier to detect potential hateful tweets.

REFERENCES

[1] P. Fortuna and S. Nunes, "A survey on automatic detection of hate speech in text," *ACM Computing Surveys (CSUR)*, vol. 51, no. 4, pp. 1–30, 2018.

[2] N. Vishwamitra, R. R. Hu, F. Luo, L. Cheng, M. Costello, and Y. Yang, "On analyzing covid-19-related hate speech using bert attention," in *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2020, pp. 669–676.

[3] C. Monahan, J. Macdonald, A. Lytle, M. Apriceno, and S. R. Levy, "Covid-19 and ageism: How positive and negative responses impact older adults and society." *American Psychologist*, 2020.

[4] K. N. Russell Jeung, "Incidents of coronavirus-related discrimination: A report for a3pcon and caa," Website, 2020, http://www.asianpacificpolicyandplanningcouncil.org/wp-content/uploads/STOP_AAPI_HATE_MONTHLY_REPORT_4_23_20.pdf.

[5] R. Alshalan, H. Al-Khalifa, D. Alsaeed, H. Al-Baity, and S. Alshalan, "Detection of hate speech in covid-19–related tweets in the arab region: Deep learning and topic modeling approach," *Journal of Medical Internet Research*, vol. 22, no. 12, p. e22609, 2020.

[6] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in tweets," in *Proceedings of the 26th international conference on World Wide Web companion*, 2017, pp. 759–760.

[7] Z. Zhang, D. Robinson, and J. Tepper, "Detecting hate speech on twitter using a convolution-gru based deep neural network," in *European semantic web conference*. Springer, 2018, pp. 745–760.

[8] A. Arango, J. Pérez, and B. Poblete, "Hate speech detection is not as easy as you may think: A closer look at model validation," in *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*, 2019, pp. 45–54.

[9] T. Gröndahl, L. Pajola, M. Juuti, M. Conti, and N. Asokan, "All you need is" love" evading hate speech detection," in *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security*, 2018, pp. 2–12.

[10] C. Ziems, B. He, S. Soni, and S. Kumar, "Racism is a virus: Anti-asian hate and counterhate in social media during the covid-19 crisis," *arXiv preprint arXiv:2005.12423*, 2020.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[12] I. Beltagy, K. Lo, and A. Cohan, "Scibert: A pretrained language model for scientific text," *arXiv preprint arXiv:1903.10676*, 2019.

[13] D. Q. Nguyen, T. Vu, and A. T. Nguyen, "Bertweet: A pre-trained language model for english tweets," *arXiv preprint arXiv:2005.10200*, 2020.

[14] M. Müller, M. Salathé, and P. E. Kummervold, "Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter," *arXiv preprint arXiv:2005.07503*, 2020.

[15] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 11, no. 1, 2017.

[16] Z. Waseem, "Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter," in *Proceedings of the first workshop on NLP and computational social science*, 2016, pp. 138–142.

[17] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? predictive features for hate speech detection on twitter," in *Proceedings of the NAACL student research workshop*, 2016, pp. 88–93.

[18] S. Agrawal and A. Awekar, "Deep learning for detecting cyberbullying across multiple social media platforms," in *European conference on information retrieval*. Springer, 2018, pp. 141–153.

[19] V. Basile, C. Bosco, E. Fersini, N. Debora, V. Patti, F. M. R. Pardo, P. Rosso, M. Sanguinetti *et al.*, "Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter," in *13th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, 2019, pp. 54–63.

[20] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, "Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval)," *arXiv preprint arXiv:1903.08983*, 2019.

[21] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Céspedes, S. Yuan, C. Tar *et al.*, "Universal sentence encoder," *arXiv preprint arXiv:1803.11175*, 2018.

[22] P. Liu, W. Li, and L. Zou, "Nuli at semeval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers," in *Proceedings of the 13th international workshop on semantic evaluation*, 2019, pp. 87–91.

[23] T. Caselli, V. Basile, J. Mitrović, I. Kartoziya, and M. Granitzer, "I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language," in *Proceedings of the 12th language resources and evaluation conference*, 2020, pp. 6193–6202.

[24] B. Vidgen, A. Botelho, D. Broniatowski, E. Guest, M. Hall, H. Margetts, R. Tromble, Z. Waseem, and S. Hale, "Detecting east asian prejudice on social media," *arXiv preprint arXiv:2005.03909*, 2020.

[25] L. Fan, H. Yu, and Z. Yin, "Stigmatization in social media: Documenting and analyzing hate speech for covid-19 on twitter," *Proceedings of the Association for Information Science and Technology*, vol. 57, no. 1, p. e313, 2020.

[26] D. Hardage and P. Najafirad, "Hate and toxic speech detection in the context of covid-19 pandemic using xai: Ongoing applied research," 2020.

[27] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," 2018.

[28] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[29] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," *Advances in neural information processing systems*, vol. 32, 2019.

[30] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[31] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, "Don't stop pretraining: adapt language models to domains and tasks," *arXiv preprint arXiv:2004.10964*, 2020.

[32] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," *arXiv preprint arXiv:1911.02116*, 2019.

[33] T. Caselli, V. Basile, J. Mitrović, and M. Granitzer, "Hatebert: Retraining bert for abusive language detection in english," *arXiv preprint arXiv:2010.12472*, 2020.

[34] "snscrape," https://github.com/JustAnotherArchivist/snscrape.

[35] "Perspective api," Website, https://www.perspectiveapi.com/.

[36] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-demos.6

[37] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[38] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[39] J. Opitz and S. Burst, "Macro f1 and macro f1," *arXiv preprint arXiv:1911.03347*, 2019.