

Hate Speech Detection via Dual Contrastive Learning

Junyu Lu, Hongfei Lin, Xiaokun Zhang, Zhaoqing Li, Tongyue Zhang, Linlin Zong, Fenglong Ma, and Bo Xu*

Abstract—The fast spread of hate speech on social media impacts the Internet environment and our society by increasing prejudice and hurting people. Detecting hate speech has aroused broad attention in the field of natural language processing. Although hate speech detection has been addressed in recent work, this task still faces two inherent unsolved challenges. The first challenge lies in the complex semantic information conveyed in hate speech, particularly the interference of insulting words in hate speech detection. The second challenge is the imbalanced distribution of hate speech and non-hate speech, which may significantly deteriorate the performance of models. To tackle these challenges, we propose a novel dual contrastive learning (DCL) framework for hate speech detection. Our framework jointly optimizes the self-supervised and the supervised contrastive learning loss for capturing span-level information beyond the token-level emotional semantics used in existing models, particularly detecting speech containing abusive and insulting words. Moreover, we integrate the focal loss into the dual contrastive learning framework to alleviate the problem of data imbalance. We conduct experiments on two publicly available English datasets, and experimental results show that the proposed model outperforms the state-of-the-art models and precisely detects hate speeches.

Index Terms—Natural language processing, hate speech detection, contrastive learning, emotion analysis, data imbalance.

I. INTRODUCTION

THE widespread use of social media provides people with a broader space for communication and information exchange. People can freely express themselves on social media. While accelerating the dissemination of public opinions, social media also leads to the dissemination of undesirable speech, such as online hate speech. Nockleyby [1] described hate speech as any kind of communication in speech, writing, or behavior, that attacks or uses pejorative or discriminatory language concerning a person or a group based on their religion, nationality, race, gender, or other identity factors.

The ever-growing increase of online hate speech has become a pressing issue disturbing not only the groups which are humiliated and vilified but also the whole society due to the potential hate crimes [2]. Even at the risk of restricting the freedom of expression, some social platforms have taken action against the proliferation of hate speech in ways of sealing accounts and removing content.

Junyu Lu, Hongfei Lin, Xiaokun Zhang, Zhaoqing Li, Tongyue Zhang, and Bo Xu are with the school of computer science and technology, Dalian University of Technology, China. Linlin Zong is with the school of software, Dalian University of Technology, China. Fenglong Ma is with the college of information science and technology, Pennsylvania State University, USA. Corresponding Author: Bo Xu, e-mail: xubo@dlut.edu.cn

The increasing social issue caused by online hate speech has attracted considerable attention of researchers in the natural language processing (NLP) field, seeking efficient and appropriate solutions to detecting online hate speech [8], [10], [12], [27], [29]. As early attempts to detect online hate speech, Chen [3] proposed lexical syntactic features to distinguish whether a sentence is hate speech. Mehdad [4] detected hate speech using support vector machines (SVM) with sentiment features of a text.

The state-of-the-art work has incorporated sentiment information for hate speech detection. Zhou et al. [29] proposed the sentiment knowledge sharing (SKS) model integrated with an insulting word list and multi-task learning to detect hate speech. Although achieving promising performance in this task, the SKS model holds a strong assumption that *insulting and negative emotions can distinguish between hate speech and non-hate speech*. However, this assumption cannot be always true as both hate and non-hate speeches may contain large amounts of negative words. Therefore, the SKS model with an insulting word list of hate speech achieved limited performance by overly focusing on the token-level emotional semantics. To further explain this phenomenon, we provide two example sentences from the SemEval-2019 Task-5 dataset [23], a publicly available dataset for hate speech detection.

Exp. 1 *I can be a bitch and an asshole but I will love you and care about you more than any other person you have met.* (**Non-hate speech**)

Exp. 2 *Stop w 'we have to worry about the children' No we do not-many R >20yrs old Go home and make your country better or enter ours legally we can't afford them#NODACA* (**Hate speech**)

It can be observed that although containing two insulting words, “bitch” and “asshole”, the sentence in Exp. 1 is a non-hate speech as no attack is launched towards any social group. In contrast, Exp. 2 is a hate speech without any obvious abusive emotions, because it involves stereotypes of immigrant children. These two examples indicate that hate speech contains more complicated semantics and irregular expression patterns beyond negative emotions.

To precisely detect hate speech, compared with the lexical sentiment, the trained models should focus on contextual semantic information to avoid the misclassification of non-hate speech containing abusive and insulting words. For a sentence with abusive or insulting words, the sentence does not contain hate speech if it is not targeted at certain social groups. According to the statistics in TABLE I, the speeches with insulting words account for a considerable proportion of two widely used hate speech detection datasets, SemEval-2019 Task-5 and Davidson et al. [24]. However, *effectively*

TABLE I
PROPORTION OF SAMPLES CONTAINING INSULTING WORDS ON
SEM EVAL-2019 TASK-5 AND DAVIDSON DATASETS.

The SemEval-2019 Task-5 dataset		
Label	#Samples	Proportion
Hate speech	2812	55.85%
Non-hate speech	2730	39.36%
The Davidson dataset		
Label	#Samples	Proportion
Hate speech	1147	80.21%
Non-hate speech	19756	84.60%

detecting these speeches with insulting words remains an unsolved problem in hate speech detection.

Moreover, the datasets for hate speech detection mostly suffer from the problem of **data imbalance**. The imbalanced distribution of hate speech and non-hate speech would easily cause the detection model to pay too much attention to the class of non-hate speech with more samples, and ignore the class of hate speech with fewer samples, resulting in an imbalanced performance on data classification. Most existing methods are designed to optimize the overall performance, partly ignoring the data imbalance problem for hate speech detection.

To solve the above-mentioned problems, we propose a novel dual contrastive learning (DCL) framework for hate speech detection, which is tailored for the hate speech detection task by comprehensively considering the task-specific features, such as the subjectivity and contextualization of hate speech [46]. Specifically, our model integrates both self-supervised and supervised contrastive learning, enriching the semantic representations of hate speech with context information itself and supervised signals from labels, effectively mitigating the misclassification of non-hate speech containing abusive and insulting words. Furthermore, since self-supervised contrastive learning has stronger adaptability than supervised contrastive learning from labels [13], the representations learned from self-supervised contrastive learning can be considered as prior knowledge, facilitating the supervised classifications of hate speech by our DCL model. Therefore, we design the self-supervised contrastive learning before the supervised contrastive learning in DCL. In addition, we introduce the focal loss, a reshaped cross entropy loss, to alleviate the problem of data imbalance. The main contributions of this work are summarized as follows.

- We propose a dual contrastive learning framework for hate speech detection, particularly addressing the detection of hate speech containing insulting words by mining context information of data beyond the token-level emotional semantics.
- We integrate self-supervised and supervised contrastive learning into the focal loss to tackle the problem of data imbalance in hate speech detection.
- We examine the effectiveness of our model on two publicly used hate speech detection datasets, and demonstrate that our model can achieve state-of-the-art performance compared with the baseline models.

II. RELATED WORK

We discuss two categories of related work: hate speech detection methods and contrastive learning methods.

A. Hate Speech Detection Methods

Detecting hate speech is a challenging natural language processing (NLP) task. Early work has used machine learning methods in automatically detecting hate speech. Davidson et al. [24] presented a large-scale dataset and used Logistic Regression [6] and SVM [7] with effective n-gram features for hate speech detection. These machine learning based methods can obtain the token-level features but mostly ignore the contextual semantic information that is highly needed for precise detection of hate speech, leading to limited performance.

In recent years, with the development of deep learning and large-scale pre-training language models, many advanced models were proposed and achieved outstanding performance in hate speech detection. Several researchers use word embeddings obtained from unsupervised training on a large number of corpora to detect hate speech. Ding et al. [27] used the FastText [9] tools to acquire word representations and presented a stacked Bidirectional Gated Recurrent Units (BiGRUs). Mou et al. [8] proved the effectiveness of FastText and BERT [22] for exploiting word-level semantic information and sub-word knowledge to identify hate speech. [12] proposed a reinforcement learning model HateGAN to address the problem of imbalance class by data augmentation. [10] presented a hate speech detection dataset and used GPT-2 [11] to pre-train the detection model. [29] proposed the sentiment knowledge sharing (SKS) model combined with a negative word list and multi-task learning for hate speech detection. [45] evaluated the effectiveness of model to introduce infusing knowledge on out-of-distribution data. [47]–[49] facilitated the detection of implicit hate speech. Previous research shows that deep learning based models can better obtain contextual information. In addition, compared with the normative data in NLI tasks, hate speech crawled from social media is more nuanced, subjective, and contextual [46], which presents a huge challenge to natural language understanding. It is imperative to consider task-specific characteristics, such as the subjectivity and contextualization of hate speech, in designing effective detection models. Moreover, previous research has also demonstrated that the general methods of NLI task have limited performance in hate speech detection task [43]. Therefore, we propose a dual contrastive learning method for hate speech detection.

B. Contrastive Learning Methods

Contrastive learning learns representations by contrasting positive and negative samples [14] and it has been widely employed in computer vision tasks [35]–[42] for extracting in-depth supervision signals from the data itself. Nan et al. [15] introduced a dual contrastive learning approach to better align text and video. Han et al. [16] proposed a novel method based on contrastive learning and a dual learning setting (exploiting two encoders) to infer an efficient mapping between unpaired

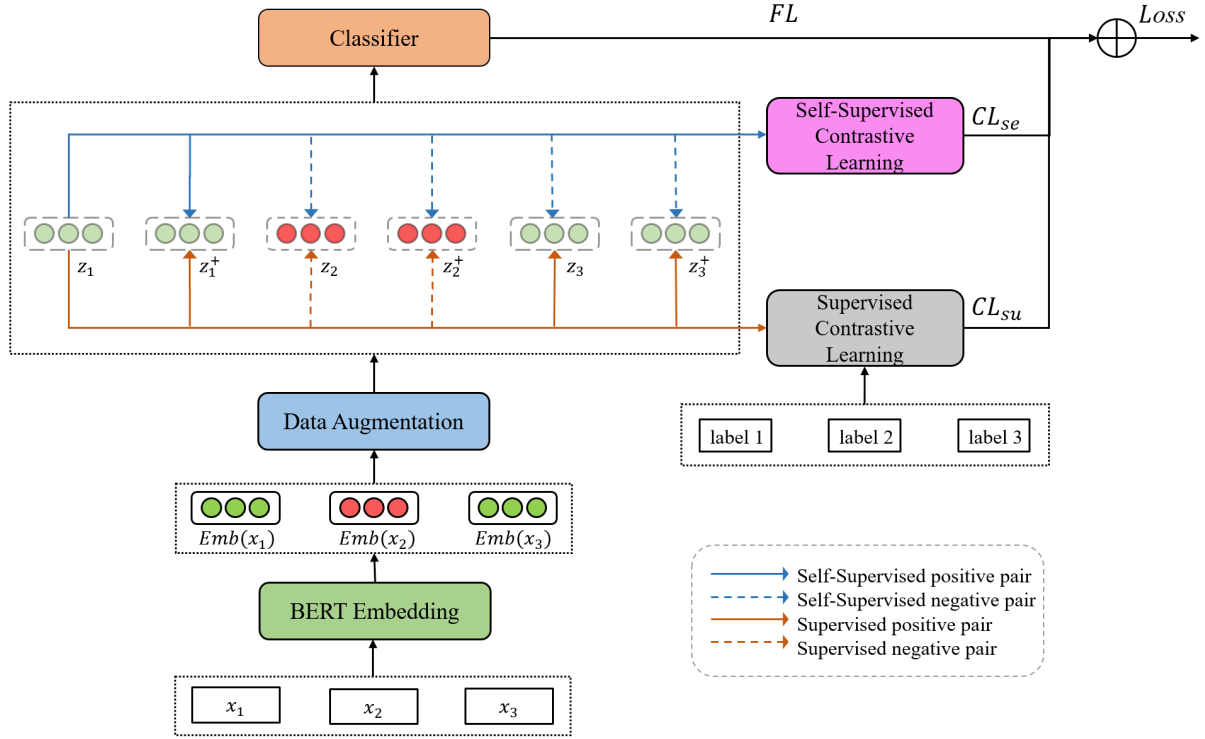


Fig. 1. The overall framework of our model. CL_{se} and CL_{su} are short for the self-supervised contrastive loss and the supervised contrastive loss, respectively. FL represents the focal loss. $Loss$ represents the final loss function. The colors of circles denote the labels of sentences, embedded as $Emb(x_i)$. Based on $Emb(x_i)$, two augmented samples z_j and z_j^+ can be generated using independently sampled dropout masks. Given z_j as the reference object, the solid blue/orange arrows point to the positive samples of z_j in CL_{se}/CL_{su} , while the dashed blue/orange arrows point to the contrastive samples in CL_{se}/CL_{su} .

data. Li et al. [17] proposed a contrastive learning framework to learn instance and cluster representations.

Contrastive learning has a wide range of applications in NLP, seeking for learning high-dimensional latent features of sentences by reducing reconstruction error. For example, Gao et al. [18] used standard dropout as noise twice for a sentence embedding to build contrastive samples and proposed SimCSE to calculate semantic similarity. [19] and [20] proposed supervised contrastive loss combined with cross-entropy to train a classification model for natural language understanding. [36] proposed a self-supervised clustering with contrastive learning for general NLI tasks. This method integrates both instance-level and cluster-level self-supervised contrastive learning to obtain pseudo labels, which are further used for representation learning. However, due to the subjectivity and contextualization of hate speech [46], pseudo labels generated by general self-supervised methods would become unreliable and difficult to use to determine whether a sentence contains hate speech. Totally different from [36], we propose a dual contrastive learning method for the task of hate speech detection. By considering the task-specific characteristics shown in Section II.A, our model integrates both self-supervised and supervised contrastive learning to enrich the semantic representations of hate speech.

III. METHODOLOGY

In this section, we introduce our model named DCL for hate speech detection. Our model seeks to learn adversarial samples

using dual contrastive learning mechanisms. We first illustrate the overall framework of our model and then introduce the self-supervised contrastive learning and the supervised contrastive learning used in our model. Besides, we provide more implementation details for easily reproducing our model.

A. Overall Framework

Fig. 1 shows the overall framework of our DCL model for hate speech detection. The input of our framework is a set of sentences including hate and non-hate speeches. Pre-trained BERT [22] is employed to represent the sentences, and data augmentation is performed for two-stage contrastive learning. The first stage adopts self-supervised contrastive learning to make our model learn representations that are invariant to different views of positive pairs of hate speech, which are generated from the same sample by strong data augmentation, while maximizing the distance between negative pairs of non-hate speech. In the second stage, supervised contrastive learning utilizes the label information to pull clusters of points belonging to the same class together in embedding space, while pushing apart clusters of samples from different classes. Finally, we integrate the dual contrastive learning objectives into the focal loss for model optimization to alleviate the problem of data imbalance in hate speech detection.

B. Self-Supervised Contrastive Learning

Considering the complicated expressions and ambiguous semantics in hate speech expressions, we propose to use

self-supervised contrastive learning for data augmentation and deep semantic information mining. By building positive and negative samples, self-supervised contrastive learning captures more comprehensive span-level features beyond token-level semantics for effectively distinguishing different samples. For hate speech detection, we propose a self-supervised contrastive learning method for mining potential useful semantic information of sentences in the model training process.

Our self-supervised contrastive objective intends to distinguish positive samples constructed by data augmentation for each input sample against a set of negative samples in each batch of data. Inspired by a simple yet powerful sampling strategy [18], we predict the input sentences itself with dropout noises [31] to retain the maximum semantic information of hate speech. Other sampling strategies can also be integrated in our framework, which remains as future work.

Specifically, for an input sentence x_i , we use standard dropout as noise twice for each sentence embedding, denoted as $Emb(x_i)$. Based on $Emb(x_i)$, two augmented samples z_j and z_j^+ with respect to x_i can be generated using independently sampled dropout masks placed on fully-connected layers. (z_j, z_j^+) is regarded as a pair of positive samples, and other samples in the same batch are treated as negative ones. Based on this idea, our self-supervised contrastive learning loss for hate speech detection can be formulated as follows.

$$CL_{se} = - \sum_{j=1}^{2N} \log \frac{e^{sim(z_j, z_j^+)/\tau_{se}}}{\sum_{k=1}^{2N} 1_{[j \neq k]} \cdot e^{sim(z_j, z_k)/\tau_{se}}} \quad (1)$$

where N denotes the batch size before data augmentation and τ_{se} is a non-negative temperature hyperparameter. $sim(\cdot)$ is the similarity scoring function between z_j and z_j^+ . In our implementation, we adopt the cosine similarity to capture the contextual semantic information by reconstructing the input samples, namely, $sim(z_j, z_j^+) = \frac{z_j^T z_j^+}{\|z_j\| \|z_j^+\|}$.

C. Supervised Contrastive Loss

Self-supervised contrastive learning augments the training data by highlighting the Span-level semantics of hate speech from the data itself. To further incorporate supervised signals for hate speech detection, we use supervised contrastive learning on the basis of the augmented data. In other words, our supervised contrastive learning method integrates label information into the embedding space of the input sentences. The learned sentence embedding contrasts a set of positive samples against a set of negative samples in the same batch. Compared with self-supervised contrastive learning, supervised contrastive learning incorporates more supervised information by considering more positive samples for each sampling batch. Specifically, for a batch of data with N samples, supervised contrastive loss can be formulated as follows:

$$CL_{su} = - \sum_{i=1}^N \frac{1}{N_{y_i} - 1} \sum_{j=1}^N 1_{[i \neq j]} \cdot 1_{[y_i = y_j]} \cdot \log \frac{e^{sim(z_i, z_j)/\tau_{su}}}{\sum_{k=1}^N 1_{[i \neq k]} \cdot e^{sim(z_i, z_k)/\tau_{su}}} \quad (2)$$

where (z_i, z_j) denotes a pair of positive samples, and (z_i, z_k) denotes a pair of randomly selected samples. y_i and y_j denotes the label of z_i and z_j , respectively. N_{y_i} is the number of samples with the same label as z_i . τ_{su} is the non-negative temperature coefficient of supervised contrastive loss. CL_{su} further guides the model with supervised information for building effective detection models. To jointly combine self-supervised and supervised information, we use an overall loss function of contrastive learning as follows:

$$CL = CL_{se} + CL_{su}. \quad (3)$$

D. DCL Integrating Focal Loss

We represent the input sentences using the pre-trained language model BERT [22]. Any sentence x_i is embedded as representations denoted as $Emb(x_i) \in R^{n \times d_{emb}}$, where n is the sequence length of x_i , and d_{emb} is the dimension of the embedding. A max-pooling layer is then applied to convert $Emb(x_i)$ into a vector representation $z_i \in R^{1 \times d_{emb}}$ that is treated as the sentence embedding of x_i . Given z_i , we can predict the target class of x_i using the softmax function:

$$p(c|z_i) = \text{softmax}(z_i W) \quad (4)$$

where $W \in d^{dim \times N_c}$ is a learnable parameter matrix. c is the target class of x_i . N_c is the number of classes. To estimate the inconsistency between the predicted label and the target label, we adopt the focal loss [21] that has been confirmed effective in imbalanced data classification. Since hate speech detection suffers from the problem of data imbalance, we introduce the focal loss to reshape the standard cross entropy loss such that the loss assigned to well-classified samples receives lower weights. The focal loss for hate speech detection is defined as:

$$FL = - \sum_{i=1}^N \alpha_i (1 - \hat{p}_i)^\gamma \log(\hat{p}_i) \quad (5)$$

where γ is a non-negative tunable focusing parameter to differentiate between easy and difficult samples. A smaller value of γ guides the learned model to focus more on the misclassified samples, and meanwhile reducing the relative loss for well-classified samples. $\alpha \in [0, 1]$ is a weighting factor to balance the importance of positive and negative samples, which is defined as:

$$\alpha_i = \begin{cases} \alpha & \text{if } y_i = 1 \\ 1 - \alpha & \text{otherwise} \end{cases} \quad (6)$$

\hat{p}_i in Eq. (5) reflects the relationship between the estimated probability and the target class.

$$\hat{p}_i = \begin{cases} p_i & \text{if } y_i = 1 \\ 1 - p_i & \text{otherwise} \end{cases} \quad (7)$$

where $p_i \in [0, 1]$ is the estimated probability for the class with the label $y_i = 1$ in each sentence embedding z_i . During the training phase, the focal loss and the contrastive learning loss are jointly optimized. To learn a more robust model, we introduce a weighting coefficient λ to balance the impact of these two loss functions. The final loss is defined as:

$$Loss = FL + \lambda \cdot CL, \quad (8)$$

where $\lambda \in [0, 1]$ is the weighting coefficient.

IV. EXPERIMENTS

In this section, we evaluate the performance of our model. We first introduce the two commonly-used datasets, experimental settings, and baselines, and then present the evaluation results of our model compared with other baseline models.

A. Datasets

We conduct our experiment on two publicly available datasets, which have been widely used in related research for comparison of hate speech detection models. The details of these datasets are introduced as follows:

SemEval-2019 Task-5 (SE) The SE dataset came from the Task-5 of SemEval-2019 [23]. The subtask A of this evaluation is hate speech detection. The hate speech of this dataset is against women and immigrants. The total number of data is 11,971, where 5,035 data are labeled as hate speech, and the remaining 6,936 data belong to the non-hate class. This dataset contains three subsets: The training set with 9000 samples, the validation set with 1000 samples, and the test set with 2971 samples.

Davidson Dataset (DV) The DV dataset was constructed by Davidson et al. [24]. The data were collected from tweets that contained hate speech including racist, sexist, homophobic, and offensive expressions in various ways. This dataset consists of 24,783 tweets, where only 1,430 ones are labeled as hate, and 23,353 data are non-hate. We can observe that this dataset is an extremely imbalanced dataset with relatively very few positive samples of hate speech.

B. Experimental Settings

We use BERT for representing the input sentences, which is fine-tuned on the downstream detection tasks. The pooling layer of bert-base-cased is taken as 768-dimensional sentence embedding. We use the 0.5 dropout rate and the AdamW optimizer [34] for model training. The learning rate is set to be $1e-4$ and the batch size as 128. We set $\tau_{se} = 0.1$ in the self-supervised contrastive loss, $\tau_{su} = 0.05$ in the supervised contrastive loss and $\alpha = 0.3$ and $\gamma = 2$ in the focal loss. All models were trained on NVIDIA GeForce GTX 1080 GPU.

To compare with baseline methods, we use accuracy (Acc) and F-measure (F1) as evaluation metrics and import the experimental results of baseline methods from the literature. Since the SE dataset is from an evaluation task, the reported experimental results are based on the performance of the test set of the official evaluation. We select the models and hyperparameters that perform best on the validation set and evaluate the performance on the test set. Results are evaluated based on the officially designated metrics, including accuracy (Acc) and macro F1. For the DV dataset, we adopt the mean accuracy and the weighted F1 after five-fold cross-validation, and save the parameters corresponding to the optimal model, which follows the settings in previous work [29]. We used the different F1 score metrics on two datasets following existing studies, such as the SOTA baseline model SKS [29]

TABLE II
COMPARISON WITH BASELINES ON SE AND DV. THE RESULTS WITH AN
ASTERISK (*) ARE IMPORTED FROM THE LITERATURE.

Dataset	SE		DV	
	Acc.	macro-F1	Acc.	weighted-F1
SVM*	49.2	45.1	-	87.0
LSTM*	55.0	53.0	94.5	93.7
GRU*	54.0	52.0	94.5	93.9
BiLSTM*	53.5	51.9	94.4	93.7
CNN-GRU*	62.0	61.5	-	94.0
BERT(BCE)	55.8	54.9	94.3	94.2
BERT(FL)	59.8	58.6	94.4	94.4
SKS*	65.9	65.2	95.1	96.3
DCL (R)	65.9	63.1	94.8	94.7
DCL (Ours)	67.8	67.2	95.9	95.6

for fair comparisons. In fact, the macro-F1 metric used for the SE dataset is a common choice in related tasks, while the weighted-F1 metric is a tailored version of macro-F1 for the DV dataset by considering that the DV dataset is very unbalanced with a ratio of hate to non-hate of about 1:15. If macro-F1 is used on DV, the performance of hate samples will dominate the overall performance. Therefore, to make more reasonable evaluations of different models on DV, weighted F1 is designed for this dataset, which considers the weights of hate and non-hate samples.

C. Baseline Methods

We compare our model with the following baselines:

SVM. The SVM-based hate speech detection model was proposed by Zhang et al. [25] and Mandl et al. [26]. The researchers extracted several statistical features, such as n-gram, insulting words, and the frequency of particular punctuation marks for learning SVM classifiers.

LSTM, GRU, Bi-LSTM. These methods were proposed by Ding et al. [27]. They employed word embedding and learned sentence representations using LSTM, GRU, and Bi-LSTM to detect hate speech, respectively.

CNN-GRU. Zhang et al. [25] applied convolution-GRU based deep neural network with word embedding to extract potential semantic features in detecting hate speech, which captures both word sequential and order information in tweets.

BERT. This baseline was proposed by Benballa et al. [28]. The final hidden state of [CLS] of BERT is used as the sentence embedding in hate speech detection. The classifier consists of a feed-forward layer and a softmax function. For a fair comparison, we train the model using cross-entropy loss and focal loss, respectively.

SKS. It was proposed by Zhou et al. [29]. This approach detected hate speech based on sentiment knowledge sharing and achieved state-of-the-art performance on the Davidson dataset and SemEval-2019 Task-5, which is a strong baseline for comparison.

D. Results and Discussions

TABLE II shows our evaluation results on the SE and DV datasets. From TABLE II, we can observe that:

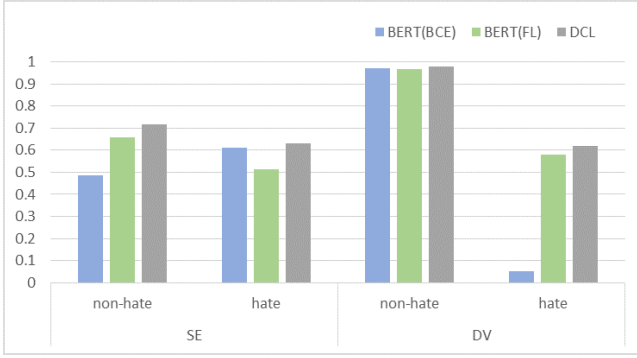


Fig. 2. F1-Score of hate and non-hate sentences on SE and DV. Blue: BERT trained with BCE, green: BERT trained with Focal loss, gray: DCL.

(1) Overall, the experimental performance on these two datasets is largely different. On the DV dataset, the values of the two used metrics are both above 93%. While on the SE dataset, the values are less than 70%. This is because the data distributions of these datasets differ a lot. Namely, subtle differences in data distributions can significantly affect the detection performance.

(2) The performance of neural network-based models is much better than the SVM-based models with manually crafted features. Compared with LSTM and its variants, hybrid neural networks, such as CNN-GRU achieved better performance, particularly on the SE dataset. Furthermore, SKS, benefiting from its sentiment knowledge-sharing mechanism and multi-task learning, achieved the best performance among all the baselines.

(3) Our model DCL outperformed all the baseline models on the SE dataset. The improvement of DCL over the BERT-based model is 13%, and the improvement over LSTM, GRU, and SVM is more 10% in terms of the macro-F1 and the accuracy. Compared with the best-performed baseline model SKS, DCL is superior in terms of both metrics.

(4) On the DV dataset, DCL achieved the best performance by accuracy, and better performance by weighted-F1 than all the baseline models except SKS. This is because **SKS used an external sentiment dataset** to enhance the performance. Although DCL does not use any external data, DCL achieved higher accuracy than SKS.

(5) We further analyze the impact of the sequence between the two stages. Specifically, we reverse the order of self-supervised and supervised contrastive learning, referred to as DCL(R). As the result shown in TABLE II, regardless of the order of DCL, it has a more competitive performance than baselines on the two datasets. Meanwhile, if self-supervised contrastive learning is before supervised contrastive learning, DCL has better detection effects. This is because the features learned from self-supervised contrastive learning represent the context information of the text itself and they are more adaptive than supervised training [13]. They can be considered as prior knowledge facilitating model decisions on downstream tasks. Therefore, it is more reasonable to employ self-supervised comparative learning as the first stage of DCL.

(6) Figure 2 shows the F1-Score of detection performance

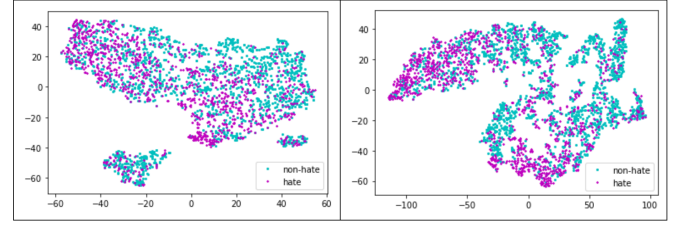


Fig. 3. t-SNE plots of the learned sentence-level embedding z_i on SemEval-2019 Task-5 test set using the BERT model (left) and our model (right). Cyan: non-hate examples; Pink: hate examples.

of hate and non-hate samples on SE and DV. From the figure, we observe that our model has the more advanced performance to distinguish whether the sentences contain hate speech than BERT trained with binary cross entropy or focal loss. This result indicates that the use of focal loss integrated with dual contrastive learning largely alleviates the data imbalance problem of hate speech detection. For DV, we find that the capability of hate speech detection is much lower than that of non-hate speech on a model trained using only cross-entropy. This is because the DV dataset is extremely imbalanced, which partly hinders the improvement of model performance.

To further validate the ability of dual contrastive learning in reconstructing text representation, we use t-Distributed Stochastic Neighbor Embedding (t-SNE) [33] to plot the learned sentence embedding z_i . t-SNE is utilized to reduce the dimension of representations from high-dimensional vector space to a two-dimensional plane. Since the number of hate speech on DV is fewer, we perform the t-SNE based plotting only on the test set of SE that contains 1180 hate speeches and 1625 non-hate speeches.

We illustrate the t-SNE plots of the learned sentence embeddings in Fig. 3. From the figure, we can observe that the distinction between hate speech and non-hate speech has been improved by introducing dual contrastive learning loss. Meanwhile, the vector space of the two classes still overlaps in certain dimensions, which indicates that some sentences with different labels have similar topical information such as immigrants. The vector representations of hate speech samples with the same topic tend to be closer than those with the same labels (i.e. hate and non-hate). This also leads to the fact that pseudo-labels generated by general self-supervised methods, such as [36], will become unreliable, making it difficult to determine whether a sentence contains hate speech or not. To further investigate the effectiveness of the loss functions used in our model, we provide an ablation study in the next section.

E. Ablation Experiments

In this section, we investigate the influence of contrastive learning and the choice of weighting coefficient λ in our model, respectively.

1) *The influence of contrastive learning.*: TABLE III shows the influence of different parts of our model, where "-self" is the proposed model without the self-supervised contrastive learning, and "-sup" is the proposed model without supervised contrastive learning.

TABLE III
THE RESULT OF ABLATION EXPERIMENTS.

Dataset	SE		DV	
Metrics	Acc.	macro-F1	Acc.	weighted-F1
-self	64.4	63.0	95.1	95.0
-sup	57.5	57.2	95.8	95.5
DCL	67.8	67.2	95.9	95.6

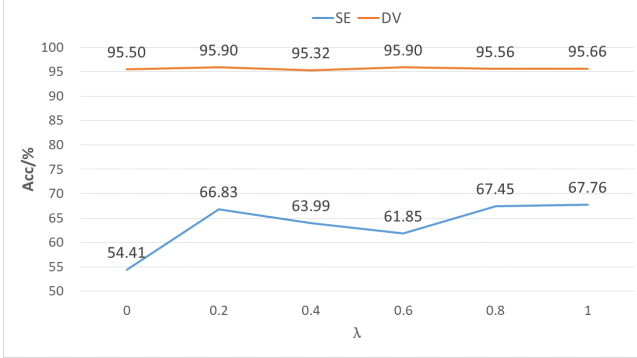


Fig. 4. The accuracy of the model under different λ .

Based on the results in TABLE III, we observe that: (1) The self-supervised contrastive learning loss contributes a lot on both datasets, which demonstrates that self-supervised contrastive learning can enhance the model’s ability in acquiring the high-level semantic features of potentially hate speech. (2) On different datasets, the performance based on supervised contrastive learning is quite different. The performance decreases more sharply on SE than that on DV. The reason for this phenomenon is that the proportion of hate speech on DV is much lower than SE, and our model hardly obtained enough positive samples for supervised contrastive learning. On SE, samples are relatively balanced and supervised contrastive learning can make the best of positive and negative samples for learning an effective detection model. This finding indicates that the label information is significant to supervised contrastive learning in our model.

2) *The choice of weighting coefficient λ* : To further examine the influence of contrastive learning in DCL, we tune the weighting coefficient λ and report the performance change in Fig. 4. From the figure, we observe that on DV, the best performance of DCL can be achieved when $\lambda = 0.2$ or $\lambda = 0.6$, while on SE, the best performance is achieved when $\lambda = 1.0$. The results indicate that contrastive learning exhibits higher performance on the balanced dataset SE, while the focal loss contributes more to the imbalanced dataset DV.

F. Performance of Detecting the Speeches Containing Insulting Words

In order to further verify whether our model has a stronger ability in detecting speech containing insulting words, we conducted this supplementary experiment. We first utilized an insulting vocabulary collected from Twitter¹ [32] and NoSwearing², a website listing swear words. The vocabulary

TABLE IV
PERFORMANCE OF MODELS TRAINED ON THE SAMPLES CONTAINING INSULTING WORDS.

Dataset	SE		DV	
Metrics	Acc.	macro-F1	Acc.	weighted-F1
BERT(BCE)	64.4	63.0	98.3	98.4
DCL	70.6	70.1	98.8	98.8

contains a total of 1060 frequently insulting words which are divided into six types of contexts: 1) Sexual 2) Appearance-related 3) Intellectual 4) Political 5) Racial 6) Combined. This resource is used to refine the samples with insulting words in the SE and DV datasets. The statistics of the refined datasets are illustrated in TABLE I, which indicates there is a large proportion of speeches containing insulting words in these two datasets. We then used the refined datasets to examine the detection performance of the learned model compared with the BERT-based model. The results on these refined datasets are reported in TABLE IV and Fig. 2.

From TABLE IV, we observe that the improvements on Acc. and macro-F1 are 6.2% and 7.1% on SE and 0.5% and 0.4% on DV, respectively. The experimental results showed that our model has a much stronger ability in detecting speeches containing insulting words than the BERT-based model. The dual contrastive learning and focal loss of our model unitedly contribute to the improved performance of hate speech detection.

G. Detection Examples and Error Analysis

1) *Detection Examples*: One advantage of our model is its capability in capturing span-level features. In this section, we provide four case studies to illustrate this capability of our model compared with the BERT-base-cased detection model. TABLE V shows the detection results. From the table, we observe that our model can precisely detect these examples, but the BERT-based model wrongly predicts their labels.

Although the first sentence has two negative words, “threats” and “lying”, that express somewhat insulting emotions, the sentence is not an attack towards certain social groups. Therefore, this sentence does not contain hate speech. On the contrary, the second sentence, as an example of hate speech, does not contain any insulting words but involves a stereotype of immigrant children. Our model correctly predicts that it is hate speech, which demonstrates the effectiveness of our model. Similarly, the third and the fourth sentences both contain an abusive word, “bitch”. By considering the context of each sentence, only the fourth sentence expresses hatred. For text containing the same insulting words, our model can also make correct predictions. This is because our model learns more contextual semantic information by dual contrastive learning, which helps effectively distinguish different kinds of samples, particularly hate speech containing insulting words.

To further verify the effectiveness of our model, we visually analyze the attention weights of the hidden layers of fine-tuned BERT encoder in the learned DCL model and the BERT-based model through two sentences. The result is shown in Fig. 5.

¹<https://github.com/Mrezvan94/Harassment-Corpus>

²<https://www.noswearing.com/>

TABLE V
EXAMPLES ON HATE SPEECH DETECTION OF OUR MODEL AND BERT.

Index	Sentence	Label	BERT	DCL
1	Like he ever kept out any threats. He's lying as usual. #BuildThatWall	non-hate	hate	non-hate
2	stop w 'we have to worry about the children' No we do not-many R>20yrs old Go home and make your country better or enter ours legally we can't afford them #NODACA	hate	non-hate	hate
3	I can be a bitch and an asshole but I will love you and care about you more than any other person you have met.	non-hate	hate	non-hate
4	my own sister even called and said bitch you pregnant now hoe know I can't have another one	hate	non-hate	hate

TABLE VI
MISCLASSIFIED INSTANCES ON HATE SPEECH DETECTION OF OUR MODEL AND BERT.

Index	Sentence	Label	BERT	DCL
1	>Harasses women and calls them bitch and crabby >They block him >Plays the victim	non-hate	hate	hate
2	Bitch how is you gonna claim to be a "real" nigga, yet still on that hoe shit? I'm not understanding the math, i'm ????	non-hate	hate	hate
3	He is 100% accurate. Diversity is our greatest weakness. Unity is our strength. United we stand, divided we fall. #buildthewall #deportthmall #stoptheinvasion #americafirst	hate	non-hate	non-hate

Exp. 1

stop w 'we have to worry about the children' No we do not-many R>20yrs old Go home and make your country better or enter ours legally we can't afford them #nodaca
stop w 'we have to worry about the children' No we do not-many R>20yrs old Go home and make your country better or enter ours legally we can't afford them #nodaca

Exp. 2

I can be a bitch and an asshole but I will love you and care about you more than any other person you have met
I can be a bitch and an asshole but I will love you and care about you more than any other person you have met

Fig. 5. Attention weights for each word of two sentences in the hidden layer of fine-tuned BERT encoder. For each sentence, the above one is trained with DCL, and the below one is trained with the BERT-based model. The depth of the background color indicates the weight of each word.

For each sentence, the above one is trained with DCL and the below one is trained with the BERT-based model.

In Fig. 5, the depth of red indicates the attention weight of the word. The darker the color, the more important the word is to the hate speech detection of the entire sentence. In Exp. 1, the word set $\{Go, home, can't, afford\}$ gets more attention from DCL than BERT. And in Exp. 2, the word set $\{I, will, love, you\}$ has a higher attention weight in sentences while the insulting words, such as "bitch" and "asshole", have a lower weight. The above sentences show that the model can better discover the key information of the context, which has a certain guiding significance for the hate speech detection task.

2) *Error Analysis*: To gain more insights into the performance of our model, a manual inspection has been performed on a set of misclassified sentences. Two main types of error have been identified:

Type I error refers to the sentences annotated as *non-hate*, but classified as *hate* by the detection models. Type I error is usually caused by colloquial and informal statements in tweets.

We enumerate two cases in TABLE VI as examples: The first case describes the scene in an informal flowchart-like fashion, while the second case contains many colloquial languages, such as "gonna", "yet still", which is not conducive to the model's understanding of text semantics. Therefore, both models wrongly predicted their labels.

Type II error refers to the sentence labeled as *hate*, but classified as *non-hate* by the detection models. Type II errors usually occur when there is a lack of necessary background knowledge. For the third case in TABLE VI, the meaning of this sentence is embodied by the information of hashtags, such as "#buildthewall", which reflect the hatred of opposition to racial diversity. Therefore, the stance contained in the hashtag needs to be considered as background knowledge in hate speech detection.

V. CONCLUSION

In this work, we propose a dual contrastive learning framework to tackle the problem of hate speech detection. Our framework integrates both self-supervised contrastive learning and supervised contrastive learning to capture high-level semantic information and complex language usage pattern in hate speech expressions. Furthermore, we integrate focal loss with dual contrastive learning to alleviate data imbalance for fine-grained hate speech detection. Experimental results on the SemEval-2019 Task-5 and Davidson dataset demonstrate the effectiveness of our model.

In the future, we will explore the following directions: (1) The analysis of Type I error shows that noises in text affect the model's performance. Therefore, we will further explore the impact of insulting words in informal contexts on hate speech detection. (2) The analysis of Type II error certifies the necessity of external knowledge in hate speech detection. We will explore how to introduce useful external knowledge to further improve detection performance.

VI. ACKNOWLEDGEMENTS

This research is supported by the Natural Science Foundation of China (No. 62076046, 62006034), Natural Science Foundation of Liaoning Province (No. 2021-BS-067).

REFERENCES

- [1] J. Nockleyby, “hate speech in encyclopedia of the american constitution,” *Electronic Journal of Academic and Special librarianship*, 2000.
- [2] M. Williams, “Hatred behind the screens: A report on the rise of online hate speech,” *J. Exp. Theor. Artif. Intell.*, 2019. [Online]. Available: <https://hatelab.net/wp-content/uploads/2019/11/Hatred-Behind-the-Screens.pdf>
- [3] Y. Chen, Y. Zhou, S. Zhu, and H. Xu, “Detecting offensive language in social media to protect adolescent online safety,” in *2012 International Conference on Privacy, Security, Risk and Trust, PASSAT 2012, and 2012 International Conference on Social Computing, SocialCom 2012, Amsterdam, Netherlands, September 3-5, 2012*. IEEE Computer Society, 2012, pp. 71–80. [Online]. Available: <https://doi.org/10.1109/SocialCom-PASSAT.2012.55>
- [4] Y. Mehdad and J. R. Tetreault, “Do characters abuse more than words?” in *Proceedings of the SIGDIAL 2016 Conference, The 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 13-15 September 2016, Los Angeles, CA, USA*. The Association for Computational Linguistics, 2016, pp. 299–303. [Online]. Available: <https://doi.org/10.18653/v1/w16-3638>
- [5] J. Qian, M. ElSherief, E. M. Belding, and W. Y. Wang, “Leveraging intra-user and inter-user representation learning for automated hate speech detection,” *CoRR*, vol. abs/1804.03124, 2018. [Online]. Available: <http://arxiv.org/abs/1804.03124>
- [6] R. E. Wright, “Logistic regression.” 1995.
- [7] T. Joachims, “Making large-scale svm learning practical,” Technical report, Tech. Rep., 1998.
- [8] G. Mou, P. Ye, and K. Lee, “SWE2: subword enriched and significant word emphasized framework for hate speech detection,” in *CIKM ’20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, M. d’Aquin, S. Dietze, C. Hauff, E. Curry, and P. Cudré-Mauroux, Eds. ACM, 2020, pp. 1145–1154. [Online]. Available: <https://doi.org/10.1145/3340531.3411990>
- [9] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching word vectors with subword information,” *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.
- [10] S. S. Tekiroglu, Y. Chung, and M. Guerini, “Generating counter narratives against online hate speech: Data and strategies,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, Eds. Association for Computational Linguistics, 2020, pp. 1177–1190. [Online]. Available: <https://doi.org/10.18653/v1/2020.acl-main.110>
- [11] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever et al., “Language models are unsupervised multitask learners,” *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.
- [12] R. Cao and R. K. Lee, “Hategan: Adversarial generative-based data augmentation for hate speech detection,” in *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, D. Scott, N. Bel, and C. Zong, Eds. International Committee on Computational Linguistics, 2020, pp. 6327–6338. [Online]. Available: <https://doi.org/10.18653/v1/2020.coling-main.557>
- [13] A. Anand, “Contrastive self-supervised learning,” 2020, <https://ankeshanand.com/blog/2020/01/26/contrastive-self-supervised-learning.html>.
- [14] M. U. Gutmann and A. Hyvärinen, “Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics,” *Journal of Machine Learning Research*, vol. 13, no. 2, 2012.
- [15] G. Nan, R. Qiao, Y. Xiao, J. Liu, S. Leng, H. Zhang, and W. Lu, “Interventional video grounding with dual contrastive learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 2765–2775.
- [16] J. Han, M. Shoeiby, L. Petersson, and M. A. Armin, “Dual contrastive learning for unsupervised image-to-image translation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2021, pp. 746–755.
- [17] Y. Li, P. Hu, Z. Liu, D. Peng, J. T. Zhou, and X. Peng, “Contrastive clustering,” in *2021 AAAI Conference on Artificial Intelligence (AAAI)*, 2021.
- [18] T. Gao, X. Yao, and D. Chen, “SimCSE: Simple contrastive learning of sentence embeddings,” in *Empirical Methods in Natural Language Processing (EMNLP)*, 2021.
- [19] B. Gunel, J. Du, A. Conneau, and V. Stoyanov, “Supervised contrastive learning for pre-trained language model fine-tuning,” in *International Conference on Learning Representations*, 2020.
- [20] Y. Moukafih, A. Ghanem, K. Abidi, N. Sbihi, M. Ghogho, and K. Smaïli, “SimSCL: A Simple fully-Supervised Contrastive Learning Framework for Text Representation,” in *AJCAI 2021 - 34th Australasian Joint Conference on Artificial Intelligence*, Sydney, Australia, Feb. 2022. [Online]. Available: <https://hal.archives-ouvertes.fr/hal-03367972>
- [21] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.
- [22] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *NAACL-HLT (1)*, 2019.
- [23] 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.
- [24] T. Davidson, D. Warmley, 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.
- [25] 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.
- [26] T. Mandl, S. Modha, P. Majumder, D. Patel, M. Dave, C. Mandlia, and A. Patel, “Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages,” in *Proceedings of the 11th forum for information retrieval evaluation*, 2019, pp. 14–17.
- [27] Y. Ding, X. Zhou, and X. Zhang, “Ynu_dyx at semeval-2019 task 5: A stacked bigru model based on capsule network in detection of hate,” in *Proceedings of the 13th International Workshop on Semantic Evaluation*, 2019, pp. 535–539.
- [28] M. Benballa, S. Collet, and R. Picot-Clemente, “Saagie at semeval-2019 task 5: From universal text embeddings and classical features to domain-specific text classification,” in *Proceedings of the 13th International Workshop on Semantic Evaluation*, 2019, pp. 469–475.
- [29] X. Zhou, Y. Yong, X. Fan, G. Ren, Y. Song, Y. Diao, L. Yang, and H. Lin, “Hate speech detection based on sentiment knowledge sharing,” in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 7158–7166.
- [30] L. van der Maaten and G. Hinton, “Visualizing data using t-sne,” *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.
- [31] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: a simple way to prevent neural networks from overfitting,” *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014. [Online]. Available: <http://dl.acm.org/citation.cfm?id=2670313>
- [32] M. Rezvan, S. Shekarpour, L. Balasuriya, K. Thirunarayan, V. L. Shalin, and A. Sheth, “A quality type-aware annotated corpus and lexicon for harassment research,” in *Proceedings of the 10th ACM Conference on Web Science*, 2018, pp. 33–36.
- [33] G. H. Laurens van der Maaten, “Visualizing data using t-sne,” *Journal of Machine Learning Research*, pp. 2579–2605, 2008.
- [34] I. Loshchilov and F. Hutter, “Fixing weight decay regularization in adam,” *CoRR*, vol. abs/1711.05101, 2017. [Online]. Available: <http://arxiv.org/abs/1711.05101>
- [35] W. V. Gansbeke, S. Vandenhende, S. Georgoulis, M. Proesmans, and L. V. Gool, “SCAN: learning to classify images without labels,” in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part X*, ser. Lecture Notes in Computer Science, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds., vol. 12355. Springer, 2020, pp. 268–285. [Online]. Available: https://doi.org/10.1007/978-3-030-58607-2_16
- [36] Y. Li, M. Yang, D. Peng, T. Li, J. Huang, and X. Peng, “Twin contrastive learning for online clustering,” *Int. J. Comput. Vis.*, vol. 130, no. 9, pp. 2205–2221, 2022. [Online]. Available: <https://doi.org/10.1007/s11263-022-01639-z>

- [37] Y. Lin, Y. Gou, X. Liu, J. Bai, J. Lv, and X. Peng, “Dual contrastive prediction for incomplete multi-view representation learning,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4447–4461, 2023. [Online]. Available: <https://doi.org/10.1109/TPAMI.2022.3197238>
- [38] P. Hu, H. Zhu, J. Lin, D. Peng, Y. Zhao, and X. Peng, “Unsupervised contrastive cross-modal hashing,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3877–3889, 2023. [Online]. Available: <https://doi.org/10.1109/TPAMI.2022.3177356>
- [39] M. Yang, Y. Li, P. Hu, J. Bai, J. Lv, and X. Peng, “Robust multi-view clustering with incomplete information,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 1055–1069, 2023. [Online]. Available: <https://doi.org/10.1109/TPAMI.2022.3155499>
- [40] Y. Li, P. Hu, J. Z. Liu, D. Peng, J. T. Zhou, and X. Peng, “Contrastive clustering,” in *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*. AAAI Press, 2021, pp. 8547–8555. [Online]. Available: <https://ojs.aaai.org/index.php/AAAI/article/view/17037>
- [41] X. Peng, Y. Li, I. W. Tsang, H. Zhu, J. Lv, and J. T. Zhou, “XAI beyond classification: Interpretable neural clustering,” *J. Mach. Learn. Res.*, vol. 23, pp. 6:1–6:28, 2022. [Online]. Available: <http://jmlr.org/papers/v23/19-497.html>
- [42] X. Peng, J. Feng, S. Xiao, W. Yau, J. T. Zhou, and S. Yang, “Structured autoencoders for subspace clustering,” *IEEE Trans. Image Process.*, vol. 27, no. 10, pp. 5076–5086, 2018. [Online]. Available: <https://doi.org/10.1109/TIP.2018.2848470>
- [43] P. Fortuna, M. Dominguez, L. Wanner, and Z. Talat, “Directions for NLP practices applied to online hate speech detection,” in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 11 794–11 805. [Online]. Available: <https://aclanthology.org/2022.emnlp-main.809>
- [44] I. Sen, M. Samory, C. Wagner, and I. Augenstein, “Counterfactually augmented data and unintended bias: The case of sexism and hate speech detection,” in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 4716–4726. [Online]. Available: <https://aclanthology.org/2022.naacl-main.347>
- [45] B. AlKhamissi, F. Ladhak, S. Iyer, V. Stoyanov, Z. Kozareva, X. Li, P. Fung, L. Mathias, A. Celikyilmaz, and M. Diab, “ToKen: Task decomposition and knowledge infusion for few-shot hate speech detection,” in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 2109–2120. [Online]. Available: <https://aclanthology.org/2022.emnlp-main.136>
- [46] B. Ross, M. Rist, G. Carbonell, B. Cabrera, N. Kurowsky, and M. Wozatki, “Measuring the reliability of hate speech annotations: The case of the european refugee crisis,” in *3rd Workshop on Natural Language Processing for Computer-Mediated Communication/Social Media*. Ruhr-Universitat Bochum, 2016, pp. 6–9.
- [47] M. ElSherief, C. Ziem, D. Muchlinski, V. Anupindi, J. Seybolt, M. D. Choudhury, and D. Yang, “Latent hatred: A benchmark for understanding implicit hate speech,” in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, M. Moens, X. Huang, L. Specia, and S. W. Yih, Eds. Association for Computational Linguistics, 2021, pp. 345–363. [Online]. Available: <https://doi.org/10.18653/v1/2021.emnlp-main.29>
- [48] T. Hartvigsen, S. Gabriel, H. Palangi, M. Sap, D. Ray, and E. Kamar, “Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection,” in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Association for Computational Linguistics, 2022, pp. 3309–3326. [Online]. Available: <https://doi.org/10.18653/v1/2022.acl-long.234>
- [49] J. Lu, B. Xu, X. Zhang, C. Min, L. Yang, and H. Lin, “Facilitating fine-grained detection of chinese toxic language: Hierarchical taxonomy, resources, and benchmarks,” *arXiv preprint arXiv:2305.04446*, 2023.