Information Fusion 96 (2023) 214–223

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Full length article

Finding hate speech with auxiliary emotion detection from self-training multi-label learning perspective

Changrong Min [a], Hongfei Lin [a,*], Ximing Li [b,c,**], He Zhao [d], Junyu Lu [a], Liang Yang [a], Bo Xu [a]

[a] School of Computer Science and Technology, Dalian University of Technology, China
[b] College of Computer Science and Technology, Jilin University, China
[c] Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, China
[d] School of Humanities, Jilin University, China

ARTICLE INFO

Keywords:
Hate speech detection
Emotion detection
Multi-label learning
Multi-task learning

ABSTRACT

Hate Speech Detection (HSD) aims to identify whether a text contains hate speech content, which often refers to discrimination and is even associated with a hate crime. The mainstream methods jointly train the HSD problem with relevant auxiliary problems, *e.g.,* emotion detection and sentiment analysis, under the paradigm of Multi-Task Learning (MTL). In this paper, we improve HSD by integrating it with emotion detection, since we take inspiration from the potential correlations between hate speech and certain negative emotion states, which have been studied theoretically and empirically. To be specific, we can concatenate their hateful labels and predicted emotion states as pseudo-multiple labels for hate speech samples, formulating a pseudo-Multi-Label Learning (MLL) problem. Beyond the existing MTL-HSD methods, we further incorporate this pseudo-MLL problem and solve it by capturing the correlations between hate speech and negative emotion states, so as to improve the performance of HSD. Based on these ideas, we propose a novel HSD method named the **E**motion-correlated **H**ate **S**peech Detect**OR** (**EHSor**). We conduct extensive experiments to evaluate EHSor, and the results show that it can consistently outperform the existing HSD methods across benchmark datasets.

1. Introduction

Emerging social media, *e.g.,* Twitter and Reddit, enables users to communicate with others freely, rapidly, and widely in our daily life. However, as a coin has two sides, social media can also be used to spread negative information, which may be fake, inactive, abusive, and even radical, due to its anonymous nature in some sense. Among various patterns of negative information, one of the most representative patterns is the so-called **hate speech**, which expresses hatred sentiment over social media. Unlike general emotion, hate speech is a complex phenomenon and essentially specific to relationships between groups. Formally, we introduce hate speech by borrowing the definition from the literature [1]:

"*Hate speech is a language that attacks or diminishes, that incites violence or hate against groups, based on specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity, or other, and it can occur with different linguistic styles, even in subtle forms or when humor is used.*"

According to the definition, hate speech often refers to discrimination and is even associated with hate crimes, as illustrated in Table 1. Thus detecting hate speech from massive social media posts becomes a significant topic and challenge, which can prevent the online environment in real-world applications. Naturally, **H**ate **S**peech **D**etection (**HSD**) has drawn more attention from the information retrieval community [2].

To address the task of HSD, a number of attempts have been made during the past decade [3]. Early shallow methods form features of hate speech samples by N-grams, TF–IDF, and part-of-speech, and further employ external information such as hate verbs [4], hateful terms [5], and inherent topics [6]. With the development of deep learning, a number of recent HSD methods employ neural models [7–9] and pre-trained language models [10] to extract strong latent features of hate speech samples, and they can empirically improve the performance on benchmark datasets.

Our story and contribution. The existing HSD methods mainly belong to content-based methods. Unfortunately, since hate speech can be regarded as a more complex psychological and cognitive phenomenon, it is insufficient to extract discriminative information from only text content as normal documents. Following the definition of hate speech, the

[*] Corresponding author at: School of Computer Science and Technology, Dalian University of Technology, China.
[**] Corresponding author at: College of Computer Science and Technology, Jilin University, China.
E-mail addresses: hflin@dlut.edu.cn (H. Lin), liximing86@gmail.com (X. Li).

https://doi.org/10.1016/j.inffus.2023.03.015
Received 12 January 2023; Received in revised form 8 March 2023; Accepted 22 March 2023
Available online 25 March 2023

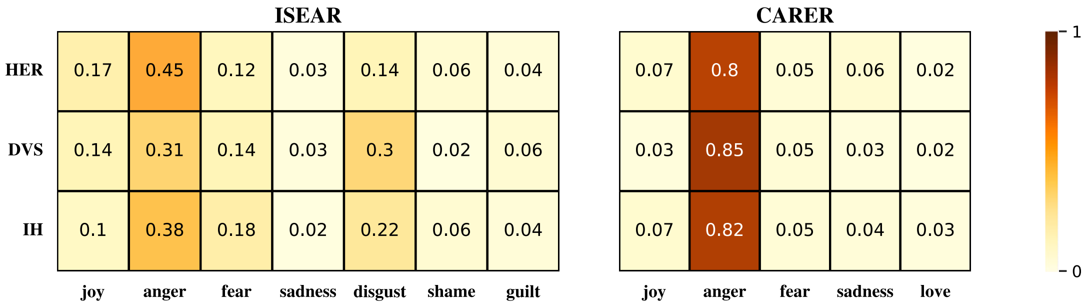

Fig. 1. Empirical emotion state distributions of hate speech. Specifically, we separately train emotion detectors on 2 public emotion datasets of ISEAR and CARER. We show the distributions of emotion states predicted by these two detectors over 3 hate speech datasets (each dataset per line), including HatEval-Relabeled (HER), DaVidSon (DVS), and Implicit Hatred (IH).

Table 1
Examples of hate speech.

Hate speech	Target
How is Mexico doing these days? people come here because you couldn't build it.	Mexicans
This fat faggot id such a child grow up porky	Homosexual

performance of hate speech is highly associated with human cognitive commonsense such as physical appearance, religion, sexual orientation, *etc.* Moreover, the emergence of hate speech is linked to the emotional and psychological states of speakers. More importantly, such hateful information is conveyed in an implicit and subtle way. Some hateful words or phrases have a cultural or historical context, which is challenging for general text classification models to correctly understand. Therefore, we can say that hate speech can be a more complex psychological and cognitive phenomenon than normal documents.To this end, we aim to improve the HSD performance by incorporating prior knowledge of the hate speech. We take inspiration from the **relationships between hate speech and emotion states** indicated by recent psychological and cognitive studies [11]. To be specific, as a negative phenomenon, hate speech tends to be associated with certain negative emotion states such as anger, contempt, humiliation, and disgust [12–14]. To further validate this relationship, we train emotion detectors on 2 public emotion datasets and then employ them to predict the emotion states over 3 benchmark hate speech datasets. As the results show in Fig. 1, we can clearly see that hate speech samples empirically tend to be associated with negative emotion states, which is consistent with the aforementioned psychological findings.

Inspired by this correlation knowledge between hate speech and human emotions, we can visit HSD from the perspective of Multi-Label Learning (MLL), where each training sample is assigned by hateful labels and emotion labels simultaneously. We can train an MLL classifier by capturing the correlations between hateful and emotion labels, which generates emotion-augmented hateful label predictions, so as to improve the performance of HSD. However, the emotion labels are exactly unknown. To solve this problem, we integrate HSD with emotion detection under the Multi-Task Learning (MTL) framework and employ the trained emotion detector to generate pseudo-emotion labels. Upon these ideas, we propose a novel MTL-HSD method with auxiliary emotion detection, namely **E**motion-correlated **H**ate **S**peech Detect**OR** (**EHSor**). Specifically, EHSor consists of three predictive heads, including a sub-hate detector, emotion detector, and super-hate detector. The sub-hate and emotion detectors are jointly trained with a shared BERT encoder over labeled hate speech and auxiliary emotion samples. For hate speech samples, we can predict their hateful labels and emotion states by using sub-hate and emotion detectors, respectively. The super-hate detector is trained in the self-training paradigm. Specifically, we can concatenate the hateful predictions with emotion predictions as latent features, and concatenate the ground-truth hateful labels with the sharpened emotion predictions as pseudo-multiple labels, so as to formulate the pseudo-MLL data. The super-hate detector is trained over the pseudo-MLL data by using a correlated module, which can capture the correlations between pseudo-multiple labels [15]. Eventually, the EHSor enables hate speech and human emotions to be fully exchanged from not only the low shared layer but also from a higher level of semantic interaction via multi-label learning. The EHSor can be trained in an end-to-end manner, and the super-hate detector is used to predict unseen hate speech samples finally. Extensive experiments have been performed to validate the effectiveness of EHSor on benchmark HSD datasets. The results demonstrate that EHSor can consistently outperform the existing HSD baseline methods.

In summary, the contributions of this paper can be summarized in three ways:

- We validate the correlations between hate speech and certain negative emotion states by synthetically considering the prior psychology studies and our prior experimental results.
- Inspired by the correlation finding, we propose a novel HSD method named EHSor, which jointly trains HSD and auxiliary emotion detection from the perspective of MLL.
- We conduct extensive experiments on benchmark datasets. The results indicate that EHSor is superior to existing HSD methods, especially those jointly training with the auxiliary emotion detection problem.

2. Related work

In this section, we review the related works on HSD, MTL, and MLL.

2.1. Hate Speech Detection

Traditional HSD methods are mainly manual feature-based, leveraging machine learning algorithms as classifiers and designing various strategies to extract hateful features from texts. These hand-crafted features include general features and hatred-specific features [1]. The former mainly collects hatred-related words or utterances via external resources [16,17] or extracts linguistic features [18–20] such as bag-of-words, n-grams, and TF–IDF. The latter considers "Our versus Theirs" features of hate speech. In addition to the other features, the superiority of the in-group [21] and the intersection of hatred [5] are also specified for the HSD task. Although the above feature engineering achieves good results on the HSD task, the task performance is still hindered by the quality of the hand-crafted features.

The emergence of deep learning allows various deep neural network (DNN) architectures to be applied to the HSD task [7,22–25]. Early DNN-based HSD methods take pre-trained context-free word embeddings [26,27] as features and then employ DNNs *e.g.,* Recurrent Neural Network (RNN) and Convolutional Neural Network, to learn deeper hateful features while performing the HSD [28]. Subsequently, with the emergence of contextual pre-trained language models [29,30], a series of BERT-induced hate speech detectors were proposed [31–33] due to the great success of BERT in the NLP community. For instance, Tran et al. [34] developed a BERT-based hate speech detector with adversarial learning for large-scale content. More recently, researchers have found that the appearance of hate speech is usually accompanied by a wealth of affective information, such as sentiment.

Additionally, there are other HSD methods integrating auxiliary problems under the paradigm of MTL [35–37]. For example, Zhou et al. [38] simultaneously conducted sentiment classification task and the HSD based on the relatedness of sentiment and hate speech [4]. Zhang et al. [39] formulated the HSD into word-level and sentence-level classification tasks and resolved the two tasks in a MTL framework. Orthogonal to the above HSD methods, we provide insight into the relationship between hate speeches and negative emotions. Furthermore, we treat the predicted emotion states of hate speech samples as their pseudo-emotion labels, naturally formulating the HSD task into the MLL problem.

2.2. Multi-Task Learning

The MTL is a learning paradigm that simultaneously learns multiple related tasks and exchanges knowledge across the tasks [40], and it has been applied to the scenarios of hate speech processing, dialogue systems, and sentiment analysis [35,36,41,42]. Existing MTL methods share parameters via either hard or soft ways [43]. The hard-sharing method contains a shared bottom and task-specific head modules upon the bottom. For example, Liu et al. [44] propose a Multi-Task Deep Neural Network based on the BERT and apply it to natural language understanding. More recently, inspired by ensemble learning, the Multi-gate Mixture-of-Experts (MMoE) [45] is proposed to improve the task conflicts. The shared bottom of the MMoE consists of a group of neural network-experts, and each task has a specific gate to control the injection of shared knowledge, better capturing task relationships. The soft-sharing way only shares a small part of the parameters among multiple tasks. The correlation of these tasks is implemented by sharing general representations [46] or imposing constraints. For instance, Duong et al. [47] exploit the ℓ_2 regularization term between two parameters belong two task-specific networks to tie the two tasks together. Compared to the hard parameter sharing, these methods need more parameters to train task-specific modules and can be more robust to task differences. Our work also leverages the MTL to exchange domain knowledge between emotions and hate speech. More importantly, we naturally and explicitly consider the relations between hate speech and emotions from the label level.

2.3. Multi-Label Learning

The MLL aims to simultaneously assign multiple labels for a single sample. A challenge faced by MLL is how to correlate multi-labels [48]. In recent years, different methods have been proposed to naturally correlate multi-labels of a single sample, especially in the community of natural language processing. Earlier approaches learn label correlation by encoding multiple labels sequentially [49,50]. Besides, Liu et al. [51] make use of matrix factorization to construct a label space in which multiple labels are correlated and apply it to the multi-label text classification. On the same task, Vu et al. [52] employ a Graph Neural Network to jointly learn label correlations and contextual representations. Orthogonal to the above methods, CorNet [15] is proposed to capture label correlations via a multi-layer feed-forward neural network, significantly improving the classification ability of MLL. This module can be easily adapted to various DNNs. Inspired by this work, we propose the EHSor, which resolves the task of HSD from the multi-label perspective.

Table 2
Summary of Notations.

Notation	Description
N_h	number of hate speech samples
N_e	number of emotion samples
C_h	number of hate speech labels
C_e	number of emotion states
$\mathbf{y}^h \in \{0,1\}^{C_h}$	label vector for hate speech samples
$\mathbf{y}^e \in \{0,1\}^{C_e}$	label vector for emotion samples
$\mathbf{p}^h$	hate prediction of HSD samples
$\mathbf{p}^{he}$	emotion prediction of hate speech samples
$\mathbf{p}^e$	emotion prediction of emotion samples
$\boldsymbol{\Phi}$	parameters of BERT
$\Theta_h = \{\mathbf{W}_h, \mathbf{b}_h\}$	parameters of the sub-hate detector
$\Theta_e = \{\mathbf{W}_e, \mathbf{b}_e\}$	parameters of the emotion detector
$\Pi = \left\{\left(\mathbf{W}_1^t, \mathbf{W}_2^t\right)\right\}_{t=1}^{T}$	parameters of the super-hate detector

3. The proposed method

Task definition of HSD. We now formulate the task of HSD, whose training dataset contains N_h labeled samples $\mathcal{D}_h = \{(x_i^h, \mathbf{y}_i^h)\}_{i=1}^{N_h}$. The notations x^h and $\mathbf{y}^h \in \{0,1\}^{C_h}$ denote the raw text and category label, respectively. In HSD, the dataset commonly contains 2 categories {`hateful`, `un-hateful`}, or at most 3 categories {`hateful`, `offensive`, `un-hateful`}. The goal of HSD is to train a hate speech detector over $\mathcal{D}_h$, enabling to predict whether a future text is hate speech or not (see Table 2).

3.1. Overview of ehsor

To resolve the task of HSD, a well-established method is to jointly train the hate speech detector over $\mathcal{D}_h$ and $\mathcal{D}_e = \{(x_i^e, \mathbf{y}_i^e)\}_{i=1}^{N_e}$, an auxiliary emotion dataset. Here, the notations x^e and $\mathbf{y}^e \in \{0,1\}^{C_e}$ are the raw text sample and the label of emotion state.

To further capture the correlations between hate speech and emotion states, we suggest a novel HSD method named EHSor that resolves the HSD task from the perspective of self-training MLL. As depicted in Fig. 2, our EHSor consists of 4 major components, including the **BERT encoder** f_{bert}, **sub-hate detector** f_{sub}, **emotion detector** f_{emo}, and **super-hate detector** f_{super}. To be specific, (1) the shared BERT encoder ingests text samples, including hate speech sample $x^h \in \mathcal{D}_h$ and emotion sample $x^e \in \mathcal{D}_e$, and then outputs their latent features $\mathbf{z}^h = f_{bert}(x^h)$ and $\mathbf{z}^e = f_{bert}(x^e)$. (2) The sub-hate detector is trained over $\{(\mathbf{z}_i^h, \mathbf{y}_i^h)\}_{i=1}^{N_h}$. (3) The emotion detector is trained over $\{(\mathbf{z}_i^e, \mathbf{y}_i^e)\}_{i=1}^{N_e}$, (4) The super-hate detector is trained in the self-training paradigm. Specifically, we predict each hate speech sample its hateful label $\mathbf{p}^h = f_{sub}(\mathbf{z}^h)$ with the sub-hate detector and emotion state $\mathbf{p}^{he} = f_{emo}(\mathbf{z}^h)$ with the emotion detector, respectively; and then use them to formulate a pseudo-MLL data $\{(\hat{\mathbf{z}}_i^h, \hat{\mathbf{y}}_i^h)\}_{i=1}^{N_h}$, where $\hat{\mathbf{z}}^h = \mathbf{p}^h \oplus \mathbf{p}^{he}$, $\hat{\mathbf{y}}^h = \mathbf{y}^h \oplus \mathbf{q}^{he}$, $\oplus$ is the concatenation operator, and $\mathbf{q}^{he}$ is the sharpened version of $\mathbf{p}^{he}$. The super-hate detector is trained over this pseudo-MLL data, so as to capture the correlations between hate speech and emotion states. Finally, the EHSor can be trained in an end-to-end manner, and the super-hate detector will be used to predict unseen hate speech samples. In the following, we introduce each component of EHSor in more detail.

3.2. BERT encoder

In EHSor, the BERT encoder is used to learn strong contextualized representations for text samples, and it is shared by the sub-hate and emotion detectors. Here, we employ the per-trained `bert-base-uncased` model[1] and use the embedding of the `[CLS]` token to represent the text samples. For each hate speech sample $x^h \in \mathcal{D}_h$ and

[1] https://huggingface.co/bert-base-uncased

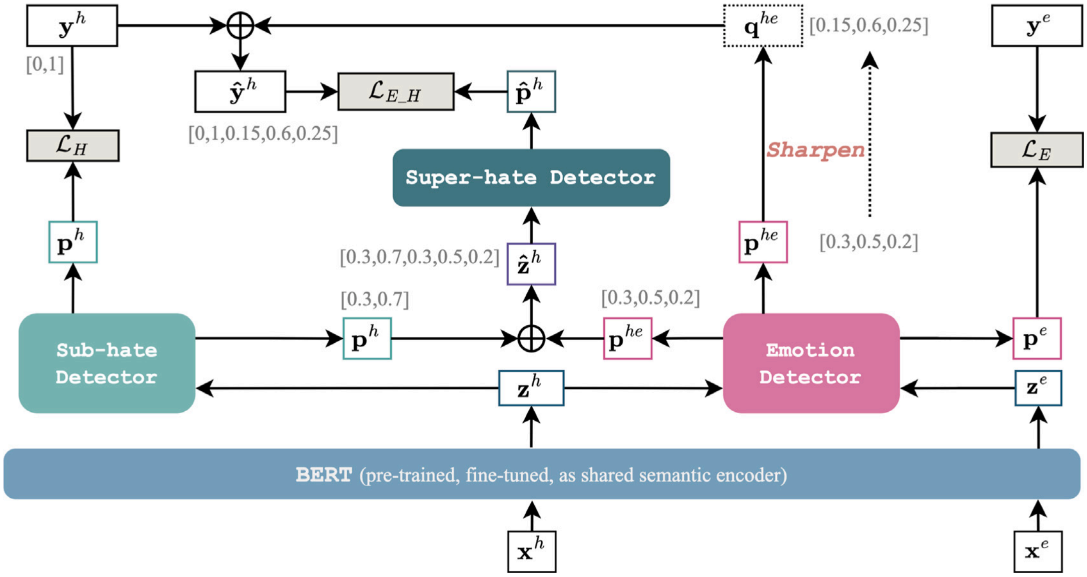

Fig. 2. The overall framework of EHSor. It consists of 4 major components, BERT-encoder, sub-hate detector, emotion detector, and super-hate detector. The sub-hate detector and emotion detector are jointly trained with the shared BERT-encoder under the paradigm of MTL. We train the super-hate detector in the self-training paradigm with a pseudo-MLL data, formed by the ground-truth hateful labels, and the predictions of the sub-hate detector and emotion detector. We train EHSor in an end-to-end manner. $\oplus$ is the concatenation operation.

emotion sample $x^e \in \mathcal{D}_e$, their latent features are formally presented as follows:

$$\mathbf{z}^h = f_{bert}\left(x^h; \boldsymbol{\Phi}\right) \tag{1}$$

$$\mathbf{z}^e = f_{bert}\left(x^e; \boldsymbol{\Phi}\right) \tag{2}$$

where $\boldsymbol{\Phi}$ denotes the trainable parameter of the BERT encoder.

3.3. Sub-hate detector

The sub-hate detector ingests $\mathbf{z}^h$, and outputs the hateful prediction $\mathbf{p}^h \in \Delta^{C_h-1}$. Here, we instantiate it as a one-layer fully connected neural network as follows:

$$\mathbf{p}^h = f_{sub}(\mathbf{z}^h; \Theta_h) = \mathbf{W}_h\mathbf{z}^h + \mathbf{b}_h \tag{3}$$

where $\Theta_h = \{\mathbf{W}_h, \mathbf{b}_h\}$ is the trainable parameter. We can train the sub-hate detector by minimizing the difference between the ground-truth hateful label $\mathbf{y}^h$ and $\mathbf{p}^h$, specifically formulated by the focal loss: [53]:

$$\mathcal{L}_H(\Theta_h) = -\sum_{(\mathbf{x}^h,\mathbf{y}^h)\in\mathcal{D}_h}\sum_{i=1}^{C_h} \alpha\mathbf{y}_i^h(1-\mathbf{p}_i^h)^\gamma log(\mathbf{p}_i^h) \tag{4}$$

where α and γ are hyperparameters of the focal loss[2]; and $\mathbf{p}_i^h$ denotes the probability for the ith hate speech category.

3.4. Emotion detector

The emotion detector ingests $\mathbf{z}^e$, and outputs the emotion prediction $\mathbf{p}^e \in \Delta^{C_e-1}$. Additionally, we instantiate it as a one-layer fully connected neural network as follows:

$$\mathbf{p}^e = f_{emo}(\mathbf{z}^e; \Theta_e) = \mathbf{W}_e\mathbf{z}^e + \mathbf{b}_e \tag{5}$$

where $\Theta_e = \{\mathbf{W}_e, \mathbf{b}_e\}$ is the trainable parameter. We can train the emotion detector by minimizing the difference between the ground-truth emotion label $\mathbf{y}^e$ and $\mathbf{p}^e$, also formulated by the focal loss:

$$\mathcal{L}_E(\Theta_e) = -\sum_{(\mathbf{x}^e,\mathbf{y}^e)\in\mathcal{D}_e}\sum_{i=1}^{C_e} \alpha\mathbf{y}_i^e(1-\mathbf{p}_i^e)^\gamma log(\mathbf{p}_i^e) \tag{6}$$

where $\mathbf{p}_i^e$ denotes the probability for the ith emotion category.

3.5. Super-hate detector

Based on f_{emo} and f_{sub}, we incorporate a super-hate detector that can predict each hate speech sample its hateful label and emotion labels simultaneously. That is, we formulate it as a pseudo-MLL problem in the self-training paradigm, so as to improve the HSD performance by capturing the correlations between hate speech and negative emotion states. Specifically, for each hate speech sample x^h, we use the emotion detector to predict its emotion label $\mathbf{p}^{he} = f_{emo}(\mathbf{z}^h; \Theta_e)$, and compute its sharpened version $\mathbf{q}^{he}$ as the pseudo emotion labels below:

$$\mathbf{q}^{he} = \mathrm{Sharpen}\left(\mathbf{p}^{he}, T\right) = \frac{\left(\mathbf{p}^{he}\right)^{\frac{1}{T}}}{\left\|\left(\mathbf{p}^{he}\right)^{\frac{1}{T}}\right\|_1} \tag{7}$$

where T is the temperature hyper-parameter.[3] We then concatenate the ground-truth hateful label with $\mathbf{q}^{he}$ to form its pseudo-multiple labels $\hat{\mathbf{y}}_i^h = \mathbf{y}^h \oplus \mathbf{q}^{he}$. Furthermore, we concatenate the hateful prediction $\mathbf{p}^h = f_{sub}(\mathbf{z}^h)$ with $\mathbf{p}^{he}$ to form its latent feature $\hat{\mathbf{z}}_i^h = \mathbf{p}^h \oplus \mathbf{p}^{he}$. Accordingly, we can formulate a pseudo MLL data $\{(\hat{\mathbf{z}}_i^h, \hat{\mathbf{y}}_i^h)\}_{i=1}^{N_h}$.

Inspired by the prior art [15], we instantiate the super-hate detector by using a correlated module. The correlation module utilizes correlation knowledge between hate speech and human emotions to enhance the raw label predictions and generate augmented label predictions for the task of hate speech detection. Given a latent feature $\hat{\mathbf{z}}_i^h$, its prediction $\hat{\mathbf{p}}^h$ of the super-hate detector is given below:

$$\hat{\mathbf{p}}^h = f_{super}(\hat{\mathbf{z}}^h; \Pi) = f_b^n\left(\cdots f_b^1(\hat{\mathbf{z}}_i^h; \Pi^1); \Pi^n\right), \tag{8}$$

where $\Pi = \{\Pi^1 \cdots \Pi^n\}$ denotes trainable parameters. The super-hate detector consists of n stacked basic blocks, and each block is a two-layer fully-connected network, described as follows:

$$f_b^l(x; \Pi^l) = \mathbf{W}_2^l\delta\left(\mathbf{W}_1^l\sigma(x) + \mathbf{b}_1^l\right) + \mathbf{b}_2^l, \tag{9}$$

[2] In this work, we empirically set α and γ to 0.3 and 4, respectively.

[3] This sharpening algorithm is a commonly used trick to form pseudo-labels with predictions [54]. In this work, we empirically set T to 0.5.

Table 3
Statistics of the HSD (top section) and emotion detection (bottom section) datasets. #Train: the number of training texts. #Test: the number of test documents.

Dataset	#Train	#Test	Category
DVS	19,826	4,957	*hate speech*, *non-hate speech*
HER	9,000	2,971	*hate speech*, *non-hate speech*
IH	15,290	3,832	*implicit hate*, *explicit hate*, *non-hate speech*
ISEAR	7,666	–	*anger*, *disgust*, *fear*, *sadness*, *shame*, *joy*, *guilt*
CARER	9,334	–	*anger*, *fear*, *sadness*, *surprise*

where x denotes the input; $\varPi^l = \{\mathbf{W}_1^l, \mathbf{W}_2^l, \mathbf{b}_1^l, \mathbf{b}_2^l\}$ is its trainable parameters; and $\delta(\cdot)$ and $\sigma(\cdot)$ are activation functions.

We can train the super-hate detector by minimizing the difference between the pseudo-multiple labels $\hat{\mathbf{y}}^h$ and $\hat{\mathbf{p}}^h$, which is formulated below:

$$\mathcal{L}_{E\_H}(\varPi) = \sum_{(\mathbf{x}^h,\mathbf{y}^h)\in\mathcal{D}_h} \left(\mathrm{FL}\left(\hat{\mathbf{p}}^h_{[C_e:]}, \mathbf{y}_h\right) + \beta\mathrm{MSE}(\hat{\mathbf{p}}^h_{[:C_e]}, \mathbf{q}^{he})\right) \tag{10}$$

where $\beta \in [0, 1]$ is a hyperparameter; FL is the focal loss; MSE is the mean square error; $\hat{\mathbf{p}}^h_{[C_e:]}$ and $\hat{\mathbf{p}}^h_{[:C_e]}$ denote the components of hateful prediction and emotion prediction, respectively.

3.6. Model training

By combining Eqs. (4), (6) and (11), we can formulate the final objective of EHSor as follows:

$$\mathcal{L}(\varPhi, \varTheta_h, \varTheta_e, \varPi) = \mathcal{L}_{E\_H} + \lambda\mathcal{L}_E + \mu\mathcal{L}_H \tag{11}$$

where coefficients λ and μ denote the coefficient parameters used to adjust the importance between the components. Eventually, we train EHSor in an end-to-end manner by optimizing Eq. (12) with gradient-based methods.

In addition, we declare that for any future text, we predict it by only using the outputs of the super-hate detector.

4. Experiment

In this section, we first describe the experimental settings, and then present the empirical evaluations of EHSor.

4.1. Experimental settings

HSD Datasets. To thoroughly evaluate EHSor, we employ three benchmark HSD datasets, including **H**at**E**val-**R**elabeled (**HER**), **D**a**V**id**S**on (**DVS**), and **I**mplicit **H**ate (**IH**). The details of the datasets are illustrated in Table 3. We briefly introduce them below.

HER [55]: The dataset is collected from Twitter and contains approximately 13,000 tweets labeled with *hate speech* or *non hate speech*. We follow settings of previous work and use 9,000 samples for training, 1,000 samples for validation, and the remaining 2,971 samples for testing.

DVS [56]: The dataset contains 24,783 samples with two labels, namely *hate speech* and *non hate speech* and has an unbalanced distribution. *Hate speech* samples account for 5.8% of the total, while the remaining 94.2% are labeled as *non hate speech.*

IH [57]: The dataset contains 19,112 samples. A total of 933 and 4,909 of these samples are labeled as *explicit hate* and *implicit hate* respectively, while the remaining 13,291 are *non hate speech* samples.

Auxiliary emotion detection datasets. We select two auxiliary emotion detection datasets **ISEAR**[4] and **CARER**.[5] [58]

Baseline methods. In the experiments, we select different types of baseline methods for comparisons, including traditional classification methods, traditional deep learning-based methods, BERT-based methods, and MTL-HSD methods. We briefly introduce the details of the baseline methods and our EHSor below.

- **SVM**: This method generates text features by using general linguistic features including n-grams, misspellings, and the number of derogatory words, and then employs SVM as the classifier.[6]
- **USE+SVM**: This method first employs the Universal Sentence Encoder [59] to generate text features and then employs SVM as the classifier.
- **RNN-based models** [60]: We develop four RNN variants: GRU, Bi-GRU, LSTM, and Bi-LSTM. The hidden layers of the models are set to 2, while the hidden state of each layer is set to 64. We implement in-house codes.
- **CNN-GRU** [61]: This method first leverage a convolutional layer with max-pooling to extract local n-gram features and further extract sequential features by a single GRU. Finally, a softmax layer is set to classify hate speech. The model is re-implemented by the PyTorch in our work. The hyper-parameter settings are as in [61].
- **BiGRU-Capsule** [60]: This method first applies two stacked bidirectional GRU to sequentially capture the context information of texts and then leverages a capsule layer to obtain high-level semantic features. Eventually, a fully connected layer is set to classify hate speech. In addition, it takes target dropout as the dropout layer.
- **BERT-based models**: To comprehensively compare with the BERT, we develop three variants named BERT,[7] RoBERTa,[8] and GPT.[9] We fine-tune these methods with a fully connected neural network as the HSD classifier.
- **SKS** [38]: The SKS is a state-of-the-art MTL-HSD method that takes sentiment analysis as the auxiliary task. We re-implement the in-house code with the same settings in the original work.
- **SKS w/ ISEAR** and **SKS w/ CARER**: For a more comprehensive comparison, we further extend the SKS by replacing the sentiment analysis task with emotion detection. We re-train the SKS with the ISEAR dataset (w/ ISEAR) and the CARER dataset (w/ CARER).
- **AbuseGNN** [39][10]: The AbuseGNN simultaneously conducts sentence-level and word-level classification tasks with a multi-task learning framework. For fairness, we take the BERT-based version of the AbuseGNN for comparison.
- **EHSor w/ ISEAR** and **EHSor w/ CARER**: We respectively train the EHSor over the CARER (EHSor w/ CARER) and the ISEAR (EHSor w/ ISEAR) datasets.

[4] https://www.unige.ch/cisa/research/

[5] https://github.com/dair-ai/emotion_dataset

[6] https://scikit-learn.org/stable/modules/classes.html#module-sklearn.svm

[7] https://huggingface.co/bert-base-uncased

[8] https://huggingface.co/roberta-base

[9] https://huggingface.co/gpt2

[10] https://github.com/random2719/MTL-Abusive-Detection

Table 4
The experimental results of all comparing methods on the three HSD datasets in terms of Accuracy (Acc), Precision (P), Recall (R), and Macro-F1 (F1). The best results are represented in **bold**.

Dataset	HER				DVS				IH			
Metric	Acc	P	R	F1	Acc	P	R	F1	Acc	P	R	F1
SVM	54.97*	50.02*	41.27*	40.88*	60.24*	51.90*	52.84*	50.38*	62.03*	52.17*	48.66*	45.79*
USE+SVM	67.45*	64.13*	62.00*	62.33*	**94.88**	**77.17**	52.97*	54.28*	72.67*	**70.89**	48.75*	49.98*
LSTM	65.39*	61.94*	55.12*	52.18*	93.75	64.93*	53.67*	55.08*	72.35*	68.78*	48.48*	50.14*
Bi-LSTM	67.17*	64.73*	58.32*	57.19*	94.31	47.16*	50.25*	48.58*	73.44*	64.22	52.16*	54.82*
GRU	64.75*	60.08*	55.74*	54.11*	94.16	71.40*	56.95*	59.85*	72.39*	64.94	47.07*	48.69*
Bi-GRU	65.84*	61.95*	57.44*	56.50*	94.42	76.69	55.26*	57.83*	72.87*	67.88*	51.37*	52.47*
CNN-GRU	65.28*	61.50*	60.11*	60.32*	94.30	72.43*	65.96*	65.54*	69.38*	43.67*	43.70*	43.02*
BiGRU-Capsule	65.58*	61.81*	56.17*	54.29*	94.01	75.71	63.78*	67.60*	73.27*	64.67	53.66*	56.36*
RoBERTa	63.26*	54.45*	50.82*	43.36*	94.47	73.57	61.84*	65.42*	72.33*	57.02*	53.31*	52.91*
GPT	67.85*	64.71*	63.35*	63.71*	94.71	47.35*	50.00*	48.64*	73.00*	64.71	54.34*	56.92*
BERT	68.78*	66.12*	62.09*	62.31*	94.47	76.32	58.29*	61.64*	73.90	68.71	54.25*	56.56*
SKS	67.09*	63.90*	58.46*	59.13*	93.65	72.59*	59.45*	62.78*	72.79*	66.77	54.59*	57.60*
SKS w/ ISEAR	68.24*	66.07*	60.17*	59.75*	94.13	75.08	61.21*	65.02*	72.49*	68.23	55.94*	57.87*
SKS w/ CARER	65.42*	58.04*	59.91*	58.70*	94.75	59.07*	**71.86**	65.57*	74.09	64.04*	57.12*	59.03*
AbuseGNN	67.09*	63.66*	60.88*	61.06*	94.81	74.10	63.16*	66.76*	73.70	61.55*	57.35*	58.77*
EHSor w/ CARER	**72.23**	**70.50**	**66.62**	**67.33**	93.78	72.27	67.94	68.57	**74.47**	65.19	58.49	60.43
EHSor w/ ISEAR	68.76	65.85	63.10	63.50	94.39	74.36	66.65	**69.34**	74.16	64.07	**58.90**	**60.57**

*Denotes that the performance improvement of the EHSor is statistically significant (paired sample t-tests) at 0.01 level.

Evaluation metrics. In the experiments, we employ *Accuracy*, *Precision*, *Recall* and *Macro-F1* to measure the performances of all comparing methods. Let TP_t, FP_t, TN_t, and FN_t respectively represent the true-positives, false-positive, true-negatives, and false-negatives of the tth class from the class set Ω. Then, the Accuracy is computed by:

$$\text{Accuracy} = \frac{\sum_{t\in\Omega} TP_t + TN_t}{\sum_{t\in\Omega} TP_t + FP_t + TN_t + FN_t} \tag{12}$$

The Precision and Recall are computed as follows:

$$\text{Precision} = \frac{\sum_{t\in\Omega} TP_t}{\sum_{t\in\Omega} TP_t + FP_t} \tag{13}$$

$$\text{Recall} = \frac{\sum_{t\in\Omega} TP_t}{\sum_{t\in\Omega} TP_t + FN_t} \tag{14}$$

The Macro-F1 is the average of F1-score of each class. It treats equally to each class and can be computed below:

$$\text{Macro-F1} = \frac{1}{|\Omega|} \sum_{t\in\Omega} \frac{2P_t R_t}{P_t + R_t} \tag{15}$$

Implementation details. We re-implement all the DNN-based models with the focal loss, due to the unbalanced distribution of the datasets. We employ the GloVe word embeddings[11] as the inputs for the non-BERT-based approaches. For all DNN-based methods, the batch size is set to 128, and dropout rate is set to 0.5. Moreover, we use the AdamW as the optimizer with a learning rate of 1e-4. The BERT-based baselines and the EHSor are trained for 5 epochs. For the other DNN-based methods, we set the number of epochs by following the original settings in their work. For the **DVS** and **IH** datasets, we apply 5-fold cross validation for training and report the mean results. For the **HER** dataset, we use the standard train/test split [55] and report the average results of 5 independent runs.

4.2. Comparing with baselines

Table 4 presents the experimental results of the two versions of our method (EHSor w/ CARER and EHSor w/ ISEAR) and the other baseline methods over the three HSD benchmark datasets. In light of the results, we can draw the following observations. **(1)** Compared with all baseline methods, our proposed EHSor can achieve very competitive performance on the four metrics. For example, the EHSor respectively gains 3.62%, 1.74%, and 1.54% with respect to the macro-f1 over the three datasets. This indicates that the EHSor effectively resolves the HSD task based on the correlations between emotion states and hate speech, modeling the intrinsic emotion states of hate speech by constructing and correlating hatred labels and emotion labels with MLL techniques. **(2)** Our proposed EHSor outperforms the strong MTL-based baseline SKS on the HER and IH datasets. This implies that fully and explicitly exploiting the co-occurrence of hatred labels and emotion labels guides our model to learn more credible hateful features for the HSD, in addition to sharing parameters across multiple tasks. **(3)** The EHSor w/ CARER significantly performs better than the EHSor w/ ISEAR on the HER dataset, but performs slightly worse on the DVS dataset. This may be because the ISEAR dataset provide much more positive emotion states than the CARER dataset, and over 90% of the samples in the DVS dataset are non-hateful. These non-hateful samples are more related to positive or neural emotions. Therefore, the EHSor can better distinguish non-hateful samples when training with the ISEAR dataset. On the HER dataset, the data distribution is much more balanced, which improves capability for discriminating hateful samples of the EHSor when increasing the proportion of negative emotions. **(4)** The EHSor gains inferior performance compared to the baselines on the DVS dataset, in which only 5.8% of samples have `hateful` labels. The reason for the performance degradation may be that correlating emotion states and non-hateful labels can be noise when classifying non-hateful samples, and the majority of the DVS dataset is non-hateful samples, which can heavily influence the results.

4.3. Sensitivity analysis

We experimentally analyze the impact of λ and μ in the Eq. (11) on the performance of the EHSor. The λ controls the regularization term of the emotion detector, while the μ controls the regularization term of the sub-hate detector. The value for each parameter is between 0 and 1. The results are reported in Figs. 3 and 4. From Fig. 3, we can observe that the results on all metrics do not continually increase as λ increases over either the ISEAR or CARER datasets. This justifies that appropriately introducing the emotion states can benefit the HSD task. However, excessively considering the emotion states can be noise and harmful, blocking the EHSor from learning hatred-specific features. In addition, from Fig. 4, the change in results in terms of the four metrics varies across the three HSD benchmarks. However, there is

[11] https://nlp.stanford.edu/projects/glove/

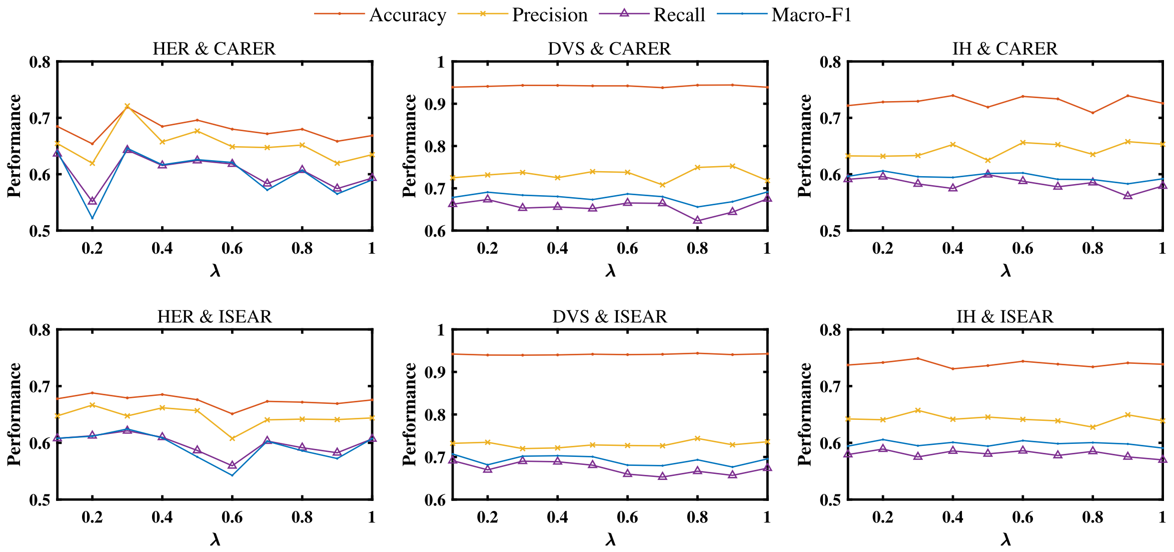

Fig. 3. Sensitivity analysis of the coefficient parameter λ across all three hate speech datasets with auxiliary datasets ISEAR (first line) and CARER (second line).

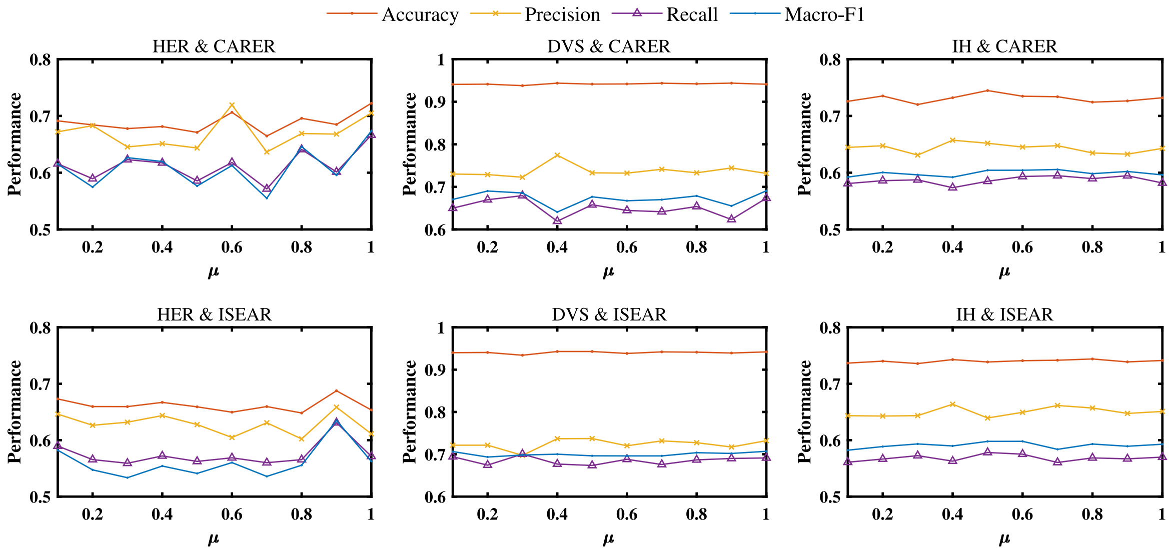

Fig. 4. Sensitivity analysis of the coefficient parameter μ across all three hate speech datasets with auxiliary datasets ISEAR (first line) and CARER (second line).

a common that the performance do not reach the peak when $\mu = 1$. This indicates that a small part of hateful samples may express hatred without negative emotions, which is hard for the super-hate detector to discriminate. As a result, we need to adjust the proportion of emotion information participating in the training process according to various data distributions.

4.4. The impact of emotion detector

We conduct extensive experiments to verify the impact of the effectiveness of emotion detection to the prediction of hate speech. Specifically, we train of the EHSor into two stages: First, we only train the emotion detector and the shared BERT. We control the capability of the EHSor to predict emotions, which can be regarded as the error rate of emotion detection, by gradually increasing the number of training epochs. We then freeze the emotion detector and the shared BERT, training the sub-hate detector and super-hate detector with HSD datasets. The results of the HSD are illustrated in Fig. 5. Overall, from the figure, we can observe that the predictive capability of the emotion detection part increases significantly as the number of training epochs increases, yet the capability of the EHSor to predict hate speech remains almost unchanged. This indicates that the EHSor is robust to the errors of emotion detection. Even though the emotion detector has insufficient training (only training with 1 epoch), the EHSor can still effectively exploit the emotion prediction results of hate samples.

4.5. Ablative study

To analyze the impact of the super-hate detector in the proposed EHSor on the performance, we conduct an ablative study and report the results in Table 5. Overall, we observe that the removal of the super-hate detector significantly hurts the performance of the EHSor over the three HSD datasets. This verifies that resolving the HSD task based on the relationship between hate speeches and emotion states from the MLL perspective is effective. To be specific, explicitly correlating the `hateful` and emotion labels of HSD samples can fully absorb knowledge from the auxiliary emotion detection task. In addition, in some way, this method can alleviate inefficient information exchange between the two tasks due to task conflicts in the MTL framework. However, for the ablation over the DVS dataset, we observe that the values of Accuracy and Precision increase, while the other two metrics decrease after removing the super-hate detector. The reason is that the

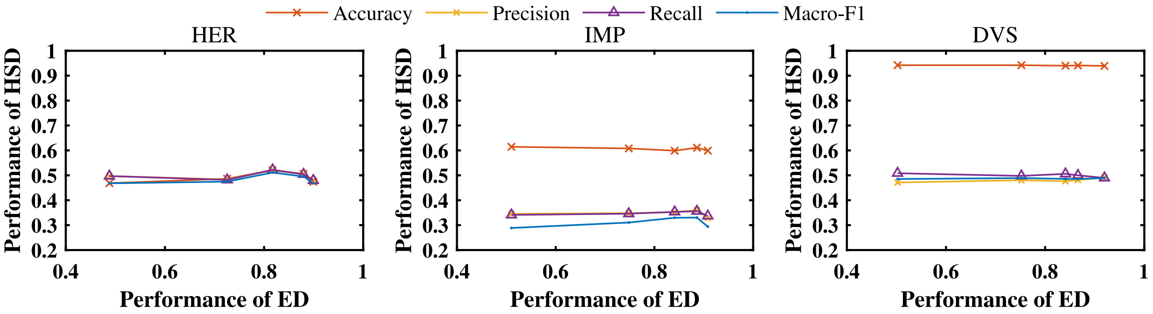

Fig. 5. Impact of the predictive capability of the emotion detector on the performance of HSD task. The *x*-axis represents the accuracy of the emotion detector over different training epochs (from 1 to 5). The *y*-axis represents the performance of EHSOR on the HSD task.

Table 5
Ablative study on the super-hate detector across all three hate speech datasets with auxiliary datasets ISEAR (top section) and CARER (bottom section). The better scores are indicated in bold.

Datasets	HER				DVS				IH			
Metrics	Acc	P	R	F1	Acc	P	R	F1	Acc	P	R	F1
EHSOR w/ ISEAR	**68.76**	**65.85**	**63.10**	**63.50**	**94.39**	**74.36**	**66.65**	**69.34**	**74.16**	64.07	**58.90**	**60.57**
w/o Super-hate	66.63	63.36	62.26	62.55	94.34	68.60	64.15	65.71	73.60	**64.27**	57.36	59.48
EHSOR w/ CARER	**72.23**	**70.50**	**66.62**	**67.33**	93.78	72.27	**67.94**	**68.57**	**74.47**	**65.19**	**58.49**	**60.43**
w/o Super-hate	69.26	66.48	63.75	64.21	**94.51**	**76.36**	63.33	66.51	72.73	62.83	55.03	57.25

capability of the EHSOR to predict "hate" samples is weakened after removing the super-hate detector. Its capability to predict "non-hate" ones is enhanced when not considering such correlations between hate speeches and human emotions from the label level. When evaluating over the DVS dataset, the data distribution is imbalanced, and only approximately 5% of samples are "hate". Therefore, the values of Accuracy and Precision show increases, while the other two metrics show decreases.

5. Conclusion

In this paper, we propose a novel solution for HSD entitled EHSOR that utilizes emotion states to improve HSD under the framework of MLL. Our story begins with the prior relationships between hate speech and negative emotion states indicated by psychological and empirical investigations. Rooted in these, any hate speech sample can be assigned with two kinds of labels, including a hateful label and an emotion label. The former is already given, while the latter is formatted by the emotion state inherent in the sample. By doing this, we address the HSD task by correlating the two labels with the MLL technique. The EHSOR fully exploits the emotion states and derives more accurate hateful predictions. In the future, we would like to investigate the interpretability of the HSD task.

CRediT authorship contribution statement

Changrong Min: Investigation, Conceptualization, Methodology, Software, Validation, Writing – original draft, Writing – review & editing. **Hongfei Lin:** Conceptualization, Validation, Investigation, Writing – original draft, Writing – review & editing. **Ximing Li:** Project administration, Writing – review & editing, Funding acquisition. **He Zhao:** Project administration. **Junyu Lu:** Project administration. **Liang Yang:** Project administration. **Bo Xu:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

We would like to acknowledge support for this project from the National Natural Science Foundation of China (NSFC) (No. 62276113), and the China Postdoctoral Science Foundation (No. 2022M721321).

References

[1] P. Fortuna, S. Nunes, A Survey on Automatic Detection of Hate Speech in Text, Vol. 51, No. 4, Association for Computing Machinery, New York, NY, USA, 2018, http://dx.doi.org/10.1145/3232676.

[2] S. MacAvaney, H.-R. Yao, E. Yang, K. Russell, N. Goharian, O. Frieder, Hate speech detection: Challenges and solutions, PLoS One 14 (8) (2019) e0221152.

[3] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, V. Patti, Resources and benchmark corpora for hate speech detection: a systematic review, Lang. Resour. Eval. 55 (2) (2021) 477–523.

[4] N.D. Gitari, Z. Zuping, H. Damien, J. Long, A lexicon-based approach for hate speech detection, Int. J. Multimed. Ubiquitous Eng. 10 (4) (2015) 215–230.

[5] P. Burnap, M.L. Williams, Us and them: identifying cyber hate on Twitter across multiple protected characteristics, EPJ Data Sci. 5 (2016) 1–15.

[6] G. Xiang, B. Fan, L. Wang, J. Hong, C. Rose, Detecting offensive tweets via topical feature discovery over a large scale twitter corpus, in: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, 2012, pp. 1980–1984.

[7] P. Badjatiya, S. Gupta, M. Gupta, 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.

[8] S. Khan, A. Kamal, M. Fazil, M.A. Alshara, V.K. Sejwal, R.M. Alotaibi, A.R. Baig, S. Alqahtani, HCovBi-Caps: Hate speech detection using convolutional and bi-directional gated recurrent unit with capsule network, IEEE Access 10 (2022) 7881–7894.

[9] M. Mozafari, R. Farahbakhsh, N. Crespi, Cross-lingual few-shot hate speech and offensive language detection using meta learning, IEEE Access 10 (2022) 14880–14896.

[10] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 4171–4186.

[11] A. Fischer, E. Halperin, D. Canetti, A. Jasini, Why we hate, Emot. Rev. 10 (4) (2018) 309–320.

[12] J. Van Doorn, Anger, feelings of revenge, and hate, Emot. Rev. 10 (4) (2018) 321–322.

[13] C. Cervone, M. Augoustinos, A. Maass, The language of derogation and hate: Functions, consequences, and reappropriation, J. Lang. Soc. Psychol. 40 (1) (2021) 80–101.

[14] M.A. Peters, Limiting the capacity for hate: Hate speech, hate groups and the philosophy of hate, Educ. Philos. Theory (2020) 1–6.

[15] G. Xun, K. Jha, J. Sun, A. Zhang, Correlation networks for extreme multi-label text classification, in: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 1074–1082.

[16] K. Dinakar, R. Reichart, H. Lieberman, Modeling the detection of textual cyberbullying, in: Proceedings of the International AAAI Conference on Web and Social Media, Vol. 5, No. 3, 2011, pp. 11–17.

[17] G.L. Clore, A. Ortony, M.A. Foss, The psychological foundations of the affective lexicon, J. Personal. Soc. Psychol. 53 (4) (1987) 751.

[18] E. Greevy, A.F. Smeaton, Classifying racist texts using a support vector machine, in: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2004, pp. 468–469.

[19] I. Kwok, Y. Wang, Locate the hate: Detecting tweets against blacks, in: Twenty-Seventh AAAI Conference on Artificial Intelligence, 2013.

[20] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, Y. Chang, Abusive language detection in online user content, in: Proceedings of the 25th International Conference on World Wide Web, 2016, pp. 145–153.

[21] W. Warner, J. Hirschberg, Detecting hate speech on the world wide web, in: Proceedings of the Second Workshop on Language in Social Media, 2012, pp. 19–26.

[22] J. Qian, M. ElSherief, E. Belding, W.Y. Wang, Hierarchical CVAE for fine-grained hate speech classification, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 3550–3559.

[23] R. Ali, U. Farooq, M.U. Arshad, W. Shahzad, M.O. Beg, Hate speech detection on Twitter using transfer learning, Comput. Speech Lang. 74 (2022) 101365.

[24] H.B. Zia, I. Castro, A. Zubiaga, G. Tyson, Improving zero-shot cross-lingual hate speech detection with pseudo-label fine-tuning of transformer language models, in: C. Budak, M. Cha, D. Quercia (Eds.), Proceedings of the Sixteenth International AAAI Conference on Web and Social Media, ICWSM 2022, Atlanta, Georgia, USA, June 6–9, 2022, 2022, pp. 1435–1439.

[25] P. Kazienko, J. Bielaniewicz, M. Gruza, K. Kanclerz, K. Karanowski, P. Miłkowski, J. Kocoń, Human-centered neural reasoning for subjective content processing: Hate speech, emotions, and humor, Inf. Fusion 94 (2023) 43–65.

[26] J. Pennington, R. Socher, C.D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP, 2014, pp. 1532–1543.

[27] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, 2013, arXiv preprint arXiv:1301.3781.

[28] Z. Zhang, D. Robinson, J.A. Tepper, Detecting hate speech on Twitter using a convolution-GRU based deep neural network, in: The Semantic Web - 15th International Conference. Vol. 10843, 2018, pp. 745–760.

[29] M. Ge, R. Mao, E. Cambria, Explainable metaphor identification inspired by conceptual metaphor theory, in: Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, the Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 – March 1, 2022, 2022, pp. 10681–10689.

[30] K. He, Y. Huang, R. Mao, T. Gong, C. Li, E. Cambria, Virtual prompt pre-training for prototype-based few-shot relation extraction, Expert Syst. Appl. 213 (Part) (2023) 118927.

[31] M. Mozafari, R. Farahbakhsh, N. Crespi, A BERT-based transfer learning approach for hate speech detection in online social media, in: International Conference on Complex Networks and their Applications, Springer, 2019, pp. 928–940.

[32] N.S. Samghabadi, P. Patwa, S. Pykl, P. Mukherjee, A. Das, T. Solorio, Aggression and misogyny detection using BERT: A multi-task approach, in: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, 2020, pp. 126–131.

[33] G.L.D. la Peña Sarracén, P. Rosso, Unsupervised embeddings with graph auto-encoders for multi-domain and multilingual hate speech detection, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, S. Piperidis (Eds.), Proceedings of the Thirteenth Lang. Resour. Eval. Conference, LREC 2022, Marseille, France, 20–25 June 2022, 2022, pp. 2196–2204.

[34] T. Tran, Y. Hu, C. Hu, K. Yen, F. Tan, K. Lee, S.R. Park, HABERTOR: An efficient and effective deep hatespeech detector, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP, 2020, pp. 7486–7502.

[35] H. Liu, P. Burnap, W. Alorainy, M.L. Williams, Fuzzy multi-task learning for hate speech type identification, in: The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13–17, 2019, 2019, pp. 3006–3012.

[36] F.M. Plaza-Del-Arco, M.D. Molina-González, L.A. Ureña-López, M.T. Martín-Valdivia, A multi-task learning approach to hate speech detection leveraging sentiment analysis, IEEE Access 9 (2021) 112478–112489.

[37] P. Kapil, A. Ekbal, A deep neural network based multi-task learning approach to hate speech detection, Knowl.-Based Syst. 210 (2020) 106458.

[38] X. Zhou, Y. Yong, X. Fan, G. Ren, Y. Song, Y. Diao, L. Yang, 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), Association for Computational Linguistics, 2021, pp. 7158–7166, http://dx.doi.org/10.18653/v1/2021.acl-long.556, Online, URL https://aclanthology.org/2021.acl-long.556.

[39] C. Zhang, X. Zhang, Q. Wang, J. Liang, G. Zhang, S. Guo, W. Zang, Y. Zhang, Abusive language detection with graph based multi-task learning, in: 2022 IEEE International Conference on Big Data, Big Data, 2022, pp. 675–684, http://dx.doi.org/10.1109/BigData55660.2022.10020761.

[40] Y. Zhang, Q. Yang, A survey on multi-task learning, IEEE Transactions on Knowledge and Data Engineering 34 (12) (2022) 5586–5609, http://dx.doi.org/10.1109/TKDE.2021.3070203.

[41] I. Bendjoudi, F. Vanderhaegen, D. Hamad, F. Dornaika, Multi-label, multi-task CNN approach for context-based emotion recognition, Inf. Fusion 76 (2021) 422–428.

[42] Y. Zhang, J. Wang, Y. Liu, L. Rong, Q. Zheng, D. Song, P. Tiwari, J. Qin, A multitask learning model for multimodal sarcasm, sentiment and emotion recognition in conversations, Inf. Fusion 93 (2023) 282–301.

[43] R. Mao, X. Li, Bridging towers of multi-task learning with a gating mechanism for aspect-based sentiment analysis and sequential metaphor identification, in: Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, 2021, pp. 13534–13542.

[44] X. Liu, P. He, W. Chen, J. Gao, Multi-task deep neural networks for natural language understanding, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 4487–4496.

[45] J. Ma, Z. Zhao, X. Yi, J. Chen, L. Hong, E.H. Chi, Modeling task relationships in multi-task learning with multi-gate mixture-of-experts, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018, pp. 1930–1939.

[46] I. Misra, A. Shrivastava, A. Gupta, M. Hebert, Cross-stitch networks for multi-task learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3994–4003.

[47] L. Duong, T. Cohn, S. Bird, P. Cook, Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), 2015, pp. 845–850.

[48] T.T. Nguyen, T.T.T. Nguyen, A.V. Luong, Q.V.H. Nguyen, A.W.-C. Liew, B. Stantic, Multi-label classification via label correlation and first order feature dependance in a data stream, Pattern Recognit. 90 (2019) 35–51.

[49] J. Read, B. Pfahringer, G. Holmes, E. Frank, Classifier chains for multi-label classification, Mach. Learn. 85 (3) (2011) 333–359.

[50] J. Nam, E. Loza Mencía, H.J. Kim, J. Fürnkranz, Maximizing subset accuracy with recurrent neural networks in multi-label classification, Adv. Neural Inf. Process. Syst. 30 (2017).

[51] H. Liu, G. Chen, P. Li, P. Zhao, X. Wu, Multi-label text classification via joint learning from label embedding and label correlation, Neurocomputing 460 (2021) 385–398.

[52] H. Vu, M. Nguyen, V. Nguyen, M. Tien, V. Nguyen, Label correlation based graph convolutional network for multi-label text classification, in: 2022 International Joint Conference on Neural Networks, IJCNN, 2022, pp. 01–08.

[53] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2980–2988.

[54] Q. Xie, Z. Dai, E.H. Hovy, T. Luong, Q. Le, Unsupervised data augmentation for consistency training, in: Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, 2020.

[55] C. Wang, M. Banko, Practical transformer-based multilingual text classification, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers, 2021, pp. 121–129.

[56] T. Davidson, D. Warmsley, M. Macy, 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, pp. 512–515.

[57] M. ElSherief, C. Ziems, D. Muchlinski, V. Anupindi, J. Seybolt, M. De Choudhury, D. Yang, Latent hatred: A benchmark for understanding implicit hate speech, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 345–363, URL https://aclanthology.org/2021.emnlp-main.29.

[58] E. Saravia, H.-C.T. Liu, Y.-H. Huang, J. Wu, Y.-S. Chen, Carer: Contextualized affect representations for emotion recognition, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 3687–3697.

[59] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R.S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, et al., Universal sentence encoder for english, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2018, pp. 169–174.

[60] Y. Ding, X. Zhou, 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.

[61] Z. Zhang, D. Robinson, 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.