

AngryBERT: Joint Learning Target and Emotion for Hate Speech Detection

Md Rabiul Awal¹, Rui Cao², Roy Ka-Wei Lee³, and Sandra Mitrović⁴

¹ University of Saskatchewan, Saskatoon, SK, Canada
mda219@usask.ca

² Singapore Management University, Singapore 188065
ruicao.2020@phdcs.smu.edu.sg

³ Singapore University of Technology and Design, Singapore 487372
roy_lee@sutd.edu.sg

⁴ Dalle Molle Institute for Artificial Intelligence, Lugano, Switzerland
sandra.mitrovic@idsia.ch

Abstract. Automated hate speech detection in social media is a challenging task that has recently gained significant traction in the data mining and Natural Language Processing community. However, most of the existing methods adopt a supervised approach that depended heavily on the annotated hate speech datasets, which are imbalanced and often lack training samples for hateful content. This paper addresses the research gaps by proposing a novel multitask learning-based model, **AngryBERT**, which jointly learns hate speech detection with sentiment classification and target identification as secondary relevant tasks. We conduct extensive experiments to augment three commonly-used hate speech detection datasets. Our experiment results show that **AngryBERT** outperforms state-of-the-art single-task-learning and multitask learning baselines. We conduct ablation studies and case studies to empirically examine the strengths and characteristics of our **AngryBERT** model and show that the secondary tasks are able to improve hate speech detection.

Keywords: Hate Speech Detection · Social Media · Multitask Learning.

1 Introduction

Motivation. The sharp increase in online hate speeches has raised concerns globally as the spread of such toxic content and misbehavior have not only sowed discord among individuals or communities online but also resulted in violent hate crimes. Therefore, it is a pressing issue to detect and curb hate speech in online social media. Researchers have proposed many traditional and deep learning hate speech classification methods to detect hate speeches in online social media automatically [9]. Specifically, the existing deep learning methods have achieved promising performance in the hate speech detection task [5]. However, most of these supervised methods depended heavily on the annotated hate speech datasets, which are imbalanced and often lack training samples for hateful

content [1]. A potential solution to address the challenges of imbalanced datasets is to perform data augmentation for the class with fewer training samples [4]. Nevertheless, the existing data augmentation methods have shown limited improvement in hate speech detection.

Research Objectives. In this paper, we adopt a different approach to address the research gaps. We propose a novel multitask learning-based model, **AngryBERT**⁵, which jointly learns hate speech detection with secondary relevant tasks. Multitask learning (MTL) [20] is a machine learning paradigm that aims to leverage useful information in multiple related tasks to help improve the generalization performance of all the tasks. Earlier studies have shown that MTL improved the performance of text classification tasks even when training with inadequate samples [20]. Similarly, the intuition of our **AngryBERT** model is that the auxiliary datasets from the secondary relevant tasks supplement the limited hateful samples of the datasets used for the main hate speech detection task. Specifically, we utilize emotion classification [13] and hateful target identification [8, 19] as the secondary tasks in our proposed model. Emotion classification is a relevant task as previous studies have demonstrated that sentiments are useful features in hate speech classification [5, 9]. Hateful target identification is an extension to the hate speech detection task where it aims to identify the target group or individual victim of the hateful content. Another key component in our **AngryBERT** model is the BERT transformer model [7], which is fine-tuned and used as the layer to share knowledge across various tasks. To the best of our knowledge, **AngryBERT** is the first model that uses a pre-trained and fine-tuned language model in a MTL framework for hate speech detection.

Contributions. We summarize our paper contribution as follows: (i) We propose a novel MTL and BERT-based model call **AngryBERT**, which jointly learns hate speech detection with secondary relevant tasks. (ii) We conduct extensive experiments on three commonly-used hate speech detection datasets. Our experiment results show that **AngryBERT** outperforms the state-of-the-art single-task and multitask baselines in hate speech detection. (iii) We identify case studies to demonstrate that **AngryBERT** is able to detect hate speeches accurately and identify the target of the hate speech and the emotion expressed. This showcases **AngryBERT**'s potential to provide some form of explainability to the hate speech detection task.

2 Related Work

In this section, we reviewed two groups of literature relevant to our study, namely, (i) existing studies on automated hate speech detection and (ii) multitask learning (MTL) for natural language processing (NLP) tasks.

Automatic detection of hate speech has received considerable attention from data mining, information retrieval, and NLP research communities. Earlier works have explored hand-crafted and canonical NLP features for automatic hate detection [6, 9, 16, 17]. In recent years, researchers have proposed deep learning methods

⁵ code implementation: https://gitlab.com/bottle_shop/safe/angrybert

to extract latent features more effectively for hate speech detection [3, 9, 14, 21]. Most of these methods adopt a supervised approach that heavily depends on labeled datasets for training, which is a challenge as existing hate speech datasets are highly imbalanced and lack training examples for hateful content.

MTL is a popular machine learning paradigm that has been explored and applied in various NLP problems, such as text classification [11, 12], etc. MTL has also been applied to abusive speech detection [15, 18]. Waseem et al. [18] proposed a fully-shared MTL model, which all tasks utilize the same fully shared features, to performed hate speech detection on three hate speech datasets. Unlike [18], our proposed AngryBERT model adopts the shared-private scheme, which model distinguishes between task-dependent and task-invariant (shared) features to perform the primary and secondary tasks. Furthermore, unlike [18] that only considered hate speech detection task and datasets, our proposed model used other relevant auxiliary tasks and dataset to improve the primary hate speech detection task. Closer to our study, Rajamanickam et al. [15] proposed a shared-private MTL framework that utilized a stacked BiLSTM encoder as the shared layer and attention mechanism for intra-task learning. The framework is trained on a hate speech detection dataset for the primary task and emotion detection as the secondary relevant task. Different from [15], our AngryBERT model adopted BERT [7] as the shared layer, and is trained on both emotion classification and hateful target identification as secondary tasks to aid hate speech detection.

3 Datasets and Tasks

Previous studies have shown that the relevance of tasks in an MTL framework affects the model’s stability of training and performance [20]. According to the definition of hate speech, there are two main characteristics of hate speech: (i) offensive language that (ii) targets individuals or groups. Considering the two aspects, we select two secondary tasks relevant to hate speech detection: emotion classification and target identification. Offensive language usually involves negative sentiments. Therefore, emotions in tweets can serve as complementary information for hate speech detection [5]. Our goal is to train a network that can extract emotions hidden in tweets using the emotion classification task. For the target identification task, we aim to train the model to identify targets in a text. Co-trained with these two secondary tasks, MTL models will be capable of extracting emotions and target groups or individuals in tweets, which facilitates hate speech detection indirectly. In the remaining parts of this section, we discuss the datasets involved to train the AngryBERT model and MTL baselines. Table 1 shows the statistical summary of the datasets.

3.1 Primary Task and Datasets

The primary task of AngryBERT is hate speech detection. Therefore, we train and evaluate our proposed model on three publicly available hateful and abusive speech datasets, namely, WZ-LS [14], DT [6], and FOUNTA [10].

Table 1: Statistic information about datasets in experiments

Dataset	#tweets	Classes (#tweets)
DT	24,783	hate(1,430), offensive(19,190), neither(4,163)
WZ-LS	16,035	racism(1923), sexism(3,079), neither(11,033)
FOUNTA	89,990	normal (53,011), abusive (19,232), spam (13,840), hate (3,907)
HateLingo	5,680	disability(257), ethnicity(351), gender(2841), religion(1590), sexual orientation(641)
SemEval_A	10,983	anger(2544), anticipation(978), disgust(2602), fear(1242), joy(2477), love(700), optimism(1984), pessimism(795), sadness(2008), surprise(361), trust(357)
OffensEval_C	4,089	individual(2,507), group(1,152), other(430)

WZ-LS [14]: Park et al. [14] combined two Twitter datasets [16,17] to form the WZ-LS dataset. We retrieve the tweets’ text using Twitter’s APIs and the tweet ids release in [14]. However, some of the tweets have been deleted by Twitter due to their inappropriate content. Thus, our dataset is slightly smaller than the original dataset reported in [14].

DT [6]: Davidson et al. [6] The researchers constructed the DT Twitter dataset, which manually labeled and categorized tweets into three categories: offensive, hate, and neither.

FOUNTA [10]: The FOUNTA dataset is a human-annotated dataset that went through two rounds of annotations. Awal et al. [2] found that there were duplicated tweets in FOUNTA dataset as the dataset annotators have included retweets in their dataset. For our experiments, we remove the retweets resulting in the distribution in Table 1.

3.2 Secondary Tasks and Datasets.

Three publicly available Twitter datasets are selected for the secondary tasks: SemEval_A [13], HateLingo [8], and OffensEval_C [19].

SemEval_A [13]: Mohammad et al. collected and annotated a Twitter dataset that supported array of subtasks on inferring the affectual state of a person from their tweet. We perform emotion classification task using this Twitter dataset.

HateLingo [8]: ElSherief et al. collected the HateLingo dataset that identifies the target of hate speeches. We perform hate speech target group identification task using the HateLingo dataset. Specifically, the task aims to identify the target group in a given hateful tweet.

OffensEval_C [19]: Zampieri et al. proposed the OffenEval_C dataset, which categorize the targets of abusive tweets into *individual*, *group*, or *other*. Similarly, our proposed model is trained on this dataset for target identification task.

4 Proposed Model

4.1 Problem Formulation

Essentially, hate speech detection (i.e., primary task) and the relevant secondary tasks can be generalized as text classification tasks. Therefore, we define a general

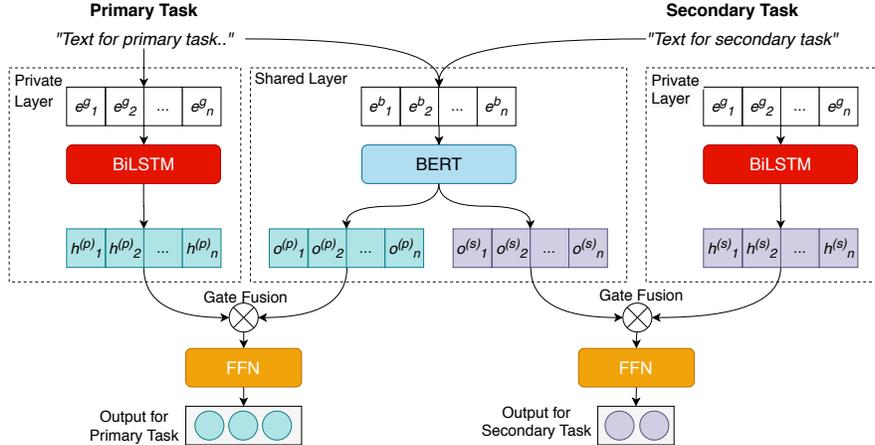


Fig. 1: The overall architecture of AngryBERT model

problem formulation of text classification tasks under the MTL setting. Assume we have K tasks and the input for the i -th task is: $S_i = \{s_1^i, s_2^i, \dots, s_n^i\}$, $i \in \{1, 2, \dots, K\}$, where n is the length of the sentence. For the i -th task, the goal is to correctly classify the input text into: $C = \{c_1^i, c_2^i, \dots, c_m^i\}$, where m is the number of classes of task i .

4.2 Architecture of AngryBERT

Figure 1 illustrates the overall architecture of AngryBERT model. We adopt a shared-private MTL setting where a shared layer is encouraged to learn the task-invariant features while the private layers aim to learn the task-specific representations. The gate fusion mechanism aggregates the shared and private information. Finally, the joint representation of each task is fed into their classification layer, respectively. To simplify our discussion, we ignore the superscript for each task in the rest of this section.

Shared Layer Here we exploit the pre-trained BERT model as the shared layer. Given a sentence, it is first tokenized using the default tokenizer of BERT then transformed into pre-trained BERT embeddings: $E_B = \{e_1^b, e_2^b, \dots, e_n^b\}$, $e_i^b \in R^{768}$. These embeddings are sent to a pre-trained BERT model. We use the output from the [CLS] token as the representation from the shared layer, denoted as $o_1 \in R^d$:

$$o_1 = BERT(E_B) \quad (1)$$

Private Layer For each task, a private layer is used to learn the task-specific representation. In order to fully exploit contexts of each word, a Bi-directional Long-Short Term Memory Network (Bi-LSTM) is applied. Each word of the sentence is first embedded using GloVe Embedding: $E_G = \{e_1^g, e_2^g, \dots, e_n^g\}$, $e_i^g \in R^{300}$. The embeddings are sent to the Bi-LSTM to learn the sequential information. The concatenation of final hidden states from forward and backward path is used as the latent representation learnt from the private layer, denoted

as $h_n \in R^d$:

$$h_n = Bi - LSTM(E_G) \quad (2)$$

Gate Fusion After learning the respective representations from shared and private layers, we exploit the gate mechanism for feature fusion. Instead of directly assign a weight for each vector, the gate fusion mechanism allows each position of vectors to have different contribution to the prediction. The joint representation from gate fusion is computed as below:

$$\alpha = \sigma(W_L o_1 + W_B h_n + b_g) \quad (3)$$

$$J = \alpha L + (1 - \alpha)B \quad (4)$$

where W_L , W_B and b_g are parameters to be learnt. $\alpha \in R^d$, which is of the same dimension as h_n and o_1 , is the attention vector. It controls the proportion of information from the private and shared flow.

Classification Layer For each task, we feed the joint representation after information aggregation to its classification layer. The classification layer is a Multi-Layer Perceptron (MLP) follow by a Softmax layer for normalization.

$$M = ReLU(W_f J + b_f) \quad (5)$$

$$O = softmax(W_e M + b_e) \quad (6)$$

where W_f , W_e and b_f , b_e are weights and biases to be learnt. The final prediction is $O \in R^m$ and each position of O denotes the confidence score for each class. Non linear activation function and weight normalization are used between two linear projection layers. Dropout is applied in order to avoid overfitting in the classification layers.

4.3 Training of AngryBERT

In this part, we describe the loss function of individual tasks and the training for AngryBERT under the MTL setting.

Single Task Loss For each task, cross entropy is used as the loss function. The loss of the i-th task is:

$$M_i = \sum_{t=1}^{N_i} Cross - Entropy(O_t^i, \hat{O}_t^i) \quad (7)$$

where \hat{O}_t^i is the ground-truth class for the t-th instance of task i and N_i is the number of training instances for the i-th task. For all tasks, we obtain: $M = \{M_1, M_2, \dots, M_K\}$.

Multi-task Loss There are several objectives involved: the primary task and secondary tasks. Rather than averaging all losses, we consider different speeds of divergence of tasks. The objective function is a weighted average of losses from different tasks:

$$\Phi = \sum_{i=1}^K \beta_i M_i \quad (8)$$

where weights β_i are learnt end-to-end, which represents the contribution from task i to the multitask loss. By exploiting multitask loss, tasks have different importance for parameter updating, which mitigates the issue of different speeds of convergence. All tasks are trained with the same number of epochs.

5 Experiments

In this section, we will first describe the settings of experiments conducted to evaluate our AngryBERT model. Next, we discuss the experiment results and assess how AngryBERT fares against other state-of-the-art baselines. We conduct more in-depth ablation studies on the various tasks co-trained with the primary hate speech detection task in the AngryBERT. We demonstrate interesting case studies where the tweets’ predicted labels for various tasks co-trained in AngryBERT presented.

5.1 Baselines.

We compare AngryBERT with the state-of-the-art hate speech classification baselines and multitask learning text classification models:

- **CNN**: Previous studies have utilized CNN to achieve good performance in hate speech detection [3]. We train a CNN model with word embeddings as input.
- **LSTM**: The LSTM model, is another model that was commonly explored in previous hate speech detection studies [3]. Similarly, we train a LSTM model with word embeddings as input.
- **HybridCNN**: We replicate the HybridCNN model proposed by Park and Fung [14] for comparison. The HybridCNN model trains CNN over both word and character embeddings for hate speech detection.
- **CNN-GRU**: The CNN-GRU model was proposed in a recent study by Zhang et al. [21] is also replicated in our study as a baseline. The CNN-GRU model takes word embeddings as input.
- **DeepHate**: The DeepHate model was proposed in a recent study by Cao et al. [5]. The DeepHate model trains on semantics, sentiment, and topical features for hate speech detection.
- **BERT**: BERT [7] is a contextualized word representation model that is based on a masked language model and pre-trained using bidirectional transformers. For our study, we fine-tune the pre-trained BERT model using the train set and subsequently perform classification on tweets in the test set.
- **SP-MTL**: Liu et al. [11] proposed the SP-MTL model, which is a Recurrent Neural Network (RNN) based multitask learning model for text classification tasks. We trained the SP-MTL model with the same tasks as our AngryBERT model.
- **MT-DNN**: Liu et al. [12] proposed the Multi-Task Deep Neural Network (MT-DNN), which combined multitask learning and language model pre-training for language representation learning. We replicated the MT-DNN

Table 2: Experiment results of AngryBERT and baselines on DT, WZ-LS, and FOUNTA datasets. “#” denotes MTL models that co-trained with other secondary tasks.

Model	DT			WZ-LS			FOUNTA		
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
CNN	89.32	90.07	89.35	80.63	78.35	78.21	79.97	80.35	79.84
LSTM	89.58	90.26	89.56	80.43	77.54	77.27	80.24	81.18	80.22
HybridCNN	88.65	89.91	88.85	80.71	78.91	78.3	79.86	80.52	79.86
CNN-GRU	88.89	89.80	88.91	80.85	77.05	77.12	79.96	80.73	79.99
DeepHate	89.97	90.39	89.92	77.95	79.48	78.19	78.95	80.43	79.09
BERT	90.35	90.53	90.34	83.25	80.05	79.95	79.69	80.03	79.79
SP-MTL#	89.44	90.22	89.44	81.11	81.59	80.68	80.46	81.65	80.66
MT-DNN#	90.29	90.69	90.31	83.05	80.25	80.18	80.66	81.64	80.72
MTL-GatedDEncoder#	89.20	89.55	89.22	81.33	78.62	78.18	80.00	81.33	80.08
AngryBERT#	90.71	91.14	90.71	83.19	81.45	81.25	81.00	81.82	81.08

as a baseline in our study. Similarly, we trained the MT-DNN with the same tasks as our AngryBERT model.

- **MTL-GatedDEncoder**: Rajamanickam et al. [15] proposed a shared-private MTL framework that utilized a stacked BiLSTM encoder as the shared layer and attention mechanism for hate speech detection. This is the state-of-the-art MTL baseline for hate speech detection.

5.2 Evaluation Metrics.

Similar to most existing hate speech detection studies, we use micro averaging precision, recall, and F1 score as the evaluation metrics. Five-fold cross-validation is used in our experiments, and the average results are reported.

5.3 Experiment Results.

Table 2 shows the experiment results on DT, WZ-LS, and FOUNTA datasets. In the table, the highest figures are highlighted in **bold**. We observe that AngryBERT outperformed the state-of-the-art single and multitask baselines in Micro-F1 scores. We observed that the single task BERT model is able to achieve good performance in hate speech detection, outperforming the other single task baselines for DT and WZ-LS datasets. Nevertheless, AngryBERT outperformed the BERT baseline by leveraging the BERT language model to learn shared knowledge across tasks.

Comparing the single-task baselines with the MTL-based models, we noted that the MTL-based models are able to outperform most single-task baselines across the three hate speech datasets. The observation shows the advantage to co-train the hate speech detection task with other secondary tasks in a multitask setting. AngryBERT is observed to outperform the state-of-the-art MTL hate speech detection model, MTL-GatedDEncoder, and other MTL text classification models. The good performance demonstrates BERT’s strength as the shared layer in the multitask learning architecture.

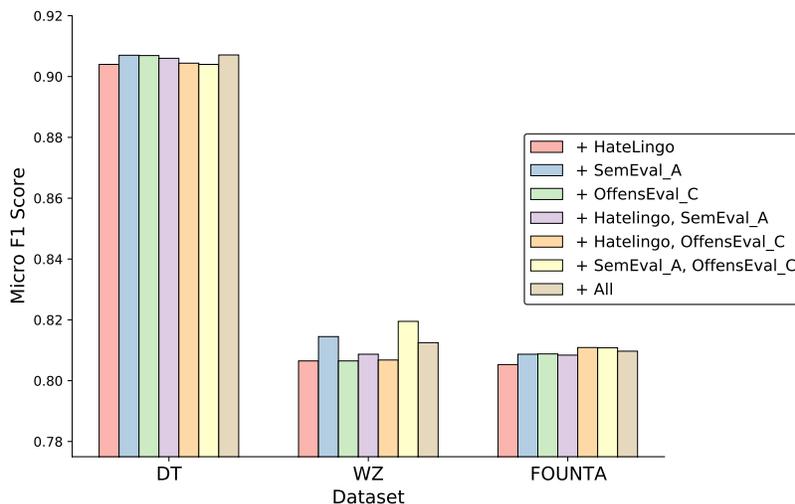


Fig. 2: Micro-F1 scores of AngryBERT model for various hate speech datasets co-trained with various combinations of secondary task datasets

It is worth noting that there are differences between HybridCNN and CNN-GRU models in our experiments and the results reported in previous studies [14, 21]. For instance, earlier studies for HybridCNN [14] and CNN-GRU [21] had conducted experiments on the WZ-LS dataset. However, we did not cite the previous scores directly as some of the tweets in WZ-LS have been deleted. Similarly, CNN-GRU was also previously applied to the DT dataset. However, in the previous work [21], the researchers have cast the problem into binary classification by re-labeling the offensive tweets as non-hate. In our experiment, we perform the classification based on the original DT dataset [6]. Therefore, we replicated the HybridCNN and CNN-GRU models and applied them to the updated WZ-LS dataset and original DT dataset.

5.4 Ablation Study.

The AngryBERT model is co-trained with several secondary tasks. In this evaluation, we perform an ablation study to investigate the effects of co-training the hate speech detection tasks with different secondary tasks.

Figure 2 shows the Micro-F1 scores of the AngryBERT model for different hate speech datasets co-trained with various secondary tasks. For example, the red bars show the AngryBERT co-training on hate speech detection tasks and the target identification task using the HateLingo dataset. We noted that the different hate speech datasets would require different task combination to achieve the best hate speech detection results. For instance, in the DT dataset, co-training the DT dataset with either the target identification task using OffensEval_C or SemEval_A will achieve similar performance as co-training all secondary tasks. For the FOUNTA dataset, co-training with the combinations of HateLingo + OffensEval_C or SemEval_A + OffensEval_C will achieve the best performance.

Table 3: Samples of AngryBERT predictions on DT dataset

Tweet	DT (Actual)	DT (Predict)	SemEval_A (Predict)	HateLingo (Predict)	OffenEval_C (Predict)
[USER] f*ck outta here and go put some more trash a*s ink on your faggot a*s self p*ssy	hateful	hateful	anger, disgust	sexual orientation	individual
RT [USER] We Muslims have no military honour whatsoever we are sub human savages that slaughter unarmed men women and children	hateful	hateful	anger, disgust	religion	group
RT [USER] I hate these Mone Davis commercials B*tch is gonna end up either a dyke or a loser like every other female	offensive	hateful	anger, disgust	gender	individual

Table 4: Samples of AngryBERT predictions on WZ-LS dataset

Tweet	WZ-LS (Actual)	WZ-LS (Predict)	SemEval_A (Predict)	HateLingo (Predict)	OffenEval_C (Predict)
[USER] Of course Muslim religious bigots like you think that is okay	racism	racism	disgust, fear	religion	group
[USER] And if you are going to follow a prophet that approved of collateral damage then do not complain about collateral damage	racism	racism	anger, disgust, fear	religion	individual
[USER] a lying taquiyya b*tch with zero followers opened an account to feed me bullshit	sexism	neither	disgust, anger	gender	individual

The WZ-LS dataset’s best performance is achieved by co-training with the SemEval_A + OffenEval_C, and co-training with only SemEval_A outperforms co-training with all secondary task datasets. Nevertheless, co-training with any combinations of the secondary tasks in AngryBERT outperforms the single-task methods in hate speech detection. These observations highlighted the intricacy of task selections for performing hate speech detection in a MTL setting. For future works, we will explore developing better approaches to automatically select the optimal combination of co-training tasks for hate speech detection.

5.5 Case Studies

To better understand how the secondary tasks could help in the hate speech detection task, we qualitatively examine some sample predictions of the **AngryBERT** model. Table 3, 4, and 5 shows example posts from DT, WZ-LS, and FOUNTA datasets respectively. In each example post, we display the actual label and predicted label from **AngryBERT** model. The correct predictions are marked in green font, while the incorrect predictions are represented in red font. Besides the hate speech predictions, we also display the predicted labels of various secondary tasks. Specifically, we highlighted the keywords in the given post that might have influenced the predicted target in HateLingo dataset.

From the example posts, we observed that the secondary tasks profoundly impact AngryBERT’s hate speech detection performance. For instance, we noted that most of the predicted hateful posts are also predicted to contain “anger” and “disgust” emotions using the secondary task emotion classifier co-trained

Table 5: Samples of AngryBERT predictions on FOUNTA datasets.

Tweet	FOUNTA (Actual)	FOUNTA (Predict)	SemEval_A (Predict)	HateLingo (Predict)	OffenEval_C (Predict)
RT [USER] I hope a tornado destroys your house you f*cking Jew	hateful	hateful	anger, disgust	ethnicity	individual
[USER] and added to his discomfort he is mightily pissed off at having to pay tampon tax the cheek	normal	normal	anger, disgust	gender	individual
RT [USER] They are F*CKING EVIL I DESPISE liberals They KILL RAPE ASSAULT	abusive	hateful	anger, disgust	gender	group

using the SemEval_A dataset. We postulate that the emotions captured by the AngryBERT model have helped the model in identifying hateful content as the two emotions are commonly exhibited in online hate speeches and abusive tweets. Another interesting observation is the identification of targets in hate speeches. We observe that the secondary task of target identification classifier co-trained using HateLingo dataset is able to predict the target in a hate speech reasonably. For example, the second tweet in Table 3 is a hateful tweet against Muslims, and the target identification classifier predicted “religion” as the target in this tweet. Although the AngryBERT has outperformed the state-of-the-art baselines in hate speech predictions, the model also made some incorrect predictions. However, we noted that as the ground truth labels of the incorrect predictions look contestable. For example, the last tweet in Table 5 seems hateful, but it was instead annotated as abusive.

The interesting predictions from secondary relevant tasks seems to provide a form of explanation that could help us understand the context when a tweet is predicted to be hateful. For future work, we will explore building explainable models that utilize the prediction of secondary tasks as supplementary information to aid explaining hate speech detection.

6 Conclusion

This paper proposed a novel multitask learning-based model, AngryBERT, which jointly learns hate speech detection with emotion classification and target identification as secondary relevant tasks. We evaluated AngryBERT on three publicly available real-world datasets, and our extensive experiments have shown that AngryBERT outperforms the state-of-the-art single-task and multitask baselines in the hate speech detection tasks. We identify case studies to demonstrate that AngryBERT is able to detect hate speeches accurately and identify the target of the hate speech and the emotion expressed. For future works, we will explore developing better approaches to automatically select the optimal combination of co-training tasks for hate speech detection. We will also explore developing explainable hate speech detection methods that utilized the predictions of secondary tasks as supplementary information.

References

1. Arango, A., Pérez, J., Poblete, B.: Hate speech detection is not as easy as you may think: A closer look at model validation. In: ACM SIGIR (2019)
2. Awal, M.R., Cao, R., Lee, R.K.W., Mitrović, S.: On analyzing annotation consistency in online abusive behavior datasets. arXiv preprint arXiv:2006.13507 (2020)
3. Badjatiya, P., Gupta, S., Gupta, M., Varma, V.: Deep learning for hate speech detection in tweets. In: WWW (2017)
4. Cao, R., Lee, R.K.W.: Hategan: Adversarial generative-based data augmentation for hate speech detection. In: COLING (2020)
5. Cao, R., Lee, R.K.W., Hoang, T.A.: Deephate: Hate speech detection via multi-faceted text representations. In: ACM WebSci (2020)
6. Davidson, T., Warmley, D., Macy, M., Weber, I.: Automated hate speech detection and the problem of offensive language. In: ICWSM (2017)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: NAACL (2019)
8. ElSherief, M., Kulkarni, V., Nguyen, D., Wang, W.Y., Belding-Royer, E.M.: Hate lingo: A target-based linguistic analysis of hate speech in social media. In: ICWSM (2018)
9. Fortuna, P., Nunes, S.: A survey on automatic detection of hate speech in text. ACM Computing Surveys (CSUR) **51**(4), 1–30 (2018)
10. Founta, A.M., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., Kourtellis, N.: Large scale crowdsourcing and characterization of twitter abusive behavior. In: ICWSM (2018)
11. Liu, P., Qiu, X., Huang, X.: Recurrent neural network for text classification with multi-task learning. In: IJCAI (2016)
12. Liu, X., He, P., Chen, W., Gao, J.: Multi-task deep neural networks for natural language understanding. In: ACL (2019)
13. Mohammad, S., Bravo-Marquez, F., Salameh, M., Kiritchenko, S.: Semeval-2018 task 1: Affect in tweets. In: SemEval (2018)
14. Park, J.H., Fung, P.: One-step and two-step classification for abusive language detection on twitter. In: Workshop on Abusive Language Online (2017)
15. Rajamanickam, S., Mishra, P., Yannakoudakis, H., Shutova, E.: Joint modelling of emotion and abusive language detection. In: ACL (2020)
16. Waseem, Z.: Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In: Workshop on NLPCSS (2016)
17. Waseem, Z., Hovy, D.: Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In: NAACL (2016)
18. Waseem, Z., Thorne, J., Bingel, J.: Bridging the gaps: Multi task learning for domain transfer of hate speech detection. In: Online harassment (2018)
19. Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., Kumar, R.: Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). In: SemEval (2019)
20. Zhang, Y., Yang, Q.: A survey on multi-task learning. arXiv preprint arXiv:1707.08114 (2017)
21. Zhang, Z., Robinson, D., Tepper, J.: Detecting hate speech on twitter using a convolution-gru based deep neural network. In: ESWC (2018)