Data Augmentation with Dual Training for Offensive Span Detection

Nasim Nouri
Raouf Medical Group
Tehran, Iran
nasimnouri@raoufmed.com

Abstract

Recognizing offensive text is an important requirement for every content management system, especially for social networks. While the majority of the prior work formulate this problem as text classification, i.e., if a text excerpt is offensive or not, in this work we propose a novel model for offensive span detection (OSD), whose goal is to identify the spans responsible for the offensive tone of the text. One of the challenges to train a model for this novel setting is the lack of enough training data. To address this limitation, in this work we propose a novel method in which the large-scale pre-trained language model GPT-2 is employed to generate synthetic training data for OSD. In particular, we propose to train the GPT-2 model in a dual-training setting using the REINFORCE algorithm to generate in-domain, natural and diverse training samples. Extensive experiments on the benchmark dataset for OSD reveal the effectiveness of the proposed method.

1 Introduction

It's no secret that social networks are growing in popularity. However, growth in popularity also brings some challenges, including the toxicity associated with the content posted by users. It may take different forms in social media, including insults, mockery, threats, discrimination, or swearing. The presence of offensive text in social networks can have a detrimental effect on their users, making it desirable to identify and remove them from the text.

Since this is an important requirement, the task of offensive language detection has been extensively studied in NLP community (Schmidt and Wiegand, 2017; Wulczyn et al., 2017; Feng et al., 2018; Borkan et al., 2019; Pavlopoulos et al., 2019; Sivanaiah et al., 2020; Yasaswini et al., 2021) Most existing works, however, only classify a text snippet as offensive or not, failing to provide further information on which specific words and phrases in the text contribute the most to its offensive tone. If the text snippet is lengthy, the moderators will need this information to decide how to proceed with the offenses flagged. As such, in this work, we fill this gap by proposing a novel model for the task of offensive span detection (OSD). As an example, in the given text "*This live streamer clearly has no brain; he is such a tool!*", the phrase "*has no brain*" and the slang word "*tool*" are two offensive spans responsible for the toxicity of the text. One of the barriers to this task is the lack of labeled data. Inspired by the recent advances in the application of pre-trained language models to augment training data for low-resources tasks (Zhang et al., 2020; Yang et al., 2020; Peng et al., 2020; Kumar et al., 2020; Anaby-Tavor et al., 2020), we propose to employ the GPT-2 model to overcome the data scarcity of OSD. To address this limitation, we propose a novel model in which the OSD training data are augmented with the synthetic samples generated by a transformer-based language model. In particular, the original labeled samples of OSD, with special markers before and after each offensive span, are employed to fine-tune the parameters of the GPT-2 model to generate sentences containing offensive spans. Moreover, in order to increase the quality of the generated samples, we propose to explicitly encourage the GPT-2 model to generate diverse sentences while keeping them similar to the original training samples. Also, the model is encouraged to generate sentences that will result in improvement of the performance of the OSD task. To fulfill these objectives, in a dual training setting, the REINFORCE algorithm (Williams, 1992) is exploited to train the GPT-2 model. We evaluate the proposed model on a recently released dataset for offensive span detection. Our extensive experiments show the effectiveness of the proposed model by outperforming the strong baselines.

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2569 - 2575

2 Model

Formal Task Description: The input to the model is the document $D = [w_1, w_2, \ldots, w_n]$ consisting of n words. The label provided for the document is also the sequence $Y = [y_1, y_2, \ldots, y_n]$ in which y_i is the label for the word w_i in BIO format. This problem is modeled as a sequence labeling task in which the model predicts the label of every word w_i in the document D. In this work, we propose a method to augment the original training samples $\mathcal{O}$, with synthetic labeled text $\mathcal{G}$ generated by a fine-tuned GPT-2 model. The rest of this section describes the base model and the data augmentation process.

2.1 Base Model

In our approach, we employ the pre-trained BERT$_{base}$ transformer as the base sequence labeling model which is trained on $\mathcal{D} = \mathcal{O} \bigcup \mathcal{G}$. Specifically, the document $D \in \mathcal{D}$ is fed into the BERT model in the form of $[CLS]w_1w_2 \ldots w_n[SEP]$ to obtain the word representations $X = [x_1, x_2, \ldots, x_n]$. Note that for the words consisting of multiple word pieces we take the average of their corresponding word-piece representations. Next, the representations x_i are sent to a feed-forward network to predict the label distribution $P(\cdot|D, \theta)$, where θ is the parameters of the BERT model. To train the model, we employ the negative log-likelihood:

$$\mathcal{L}_{base} = -\sum_{i=1}^{|\mathcal{D}|}\sum_{j=1}^{n} P(y_j|D_i, \theta) \qquad (1)$$

where y_j is the gold label for j-th word of the document D_i.

2.2 Data Augmentation

One of the limitations for OSD is the lack of enough labeled data. To address this limitation, inspired by the success of the generative language models to augment data for other tasks, we propose to employ GPT-2 to generate labeled synthetic data. We first discuss the generation process, then we provide details on how the generative model is encouraged to generate high-quality data.

Generation: Following prior works (Zhang et al., 2020), to generate synthetic data we employ GPT-2 (Radford et al., 2019) model. GPT-2 is a transformer-based language model pre-trained on 40 GB of textual data. In order to fine-tune GPT-2 for generating labeled data for OSD, we propose to employ the original labeled data $\mathcal{G}$. Specifically, the document $D \in \mathcal{G}$ is first augmented with special tokens at the beginning and the end of the document and also around the offensive spans: $D' = [BOS]w_1, w_2, \ldots [OFFENSIVE_S]w_i, w_{i+1}, \ldots, w_{i+t}[OFFENSIVE_E]w_{i+t+1}, \ldots, w_n[EOS]$, where t is the length of the offensive span in D. Note that there might be multiple offensive span in a document. Next, the GPT-2 model is trained in an auto-regressive manner on the labeled augmented documents D'. Specifically, the following loss is employed for the fine-tuning process:

$$\mathcal{L}_f = -\sum_{i=1}^{|\mathcal{O}|}\sum_{j=1}^{|D'_i|} P_G(w'_j|D'_{<j}, \alpha) \qquad (2)$$

where w'_j is the j-th word in the label augmented document D'_i, $D'_{<j}$ is the left context of the word w'_j in the document D'_i, and α is the parameters of the GPT-2 model.

Finally, the fine-tuned GPT-2 model is employed to generate $|\mathcal{O}|$ synthetic data. Specifically, the model is prompted with $[BOS]$ token and the generation is stopped by generating the $[EOS]$ token. In order to ensure that the generated data are labeled, we keep only the generated samples with at least one pair of $[OFFENSIVE_S]$ and $[OFFENSIVE_E]$ tokens. The generated samples, i.e., $\mathcal{G}$, are combined with the original samples $\mathcal{O}$, to obtain the final $\mathcal{D}$ dataset to train the base model.

Improving Quality of Generated Samples: While the fine-tuning process of GPT-2 is supposed to be effective to generate high-quality data, it has been shown that the generated data might be noisy or have repeated sentences (Pouran Ben Veyseh et al., 2021), providing harmful or less supervision signals to the base model training. As such, we propose to explicitly encourage GPT-2 model to generate documents that results in better performance on OSD task and satisfy the diversity requirements of the generated data. In particular, we propose to employ dual training with REINFORCE to ensure the following requirements are observed: (1) **Usefulness**: The generated documents should be helpful to improve the performance on the final task. As such, the F1 score of the base model, trained using $\mathcal{D}$, on the original data $\mathcal{O}$ is employed as a measure of usefulness of the generated data: $R_u(\mathcal{G}) = F1(\mathcal{O})$; (2) **Diversity**: If the generated

samples are identical or similar to the original data, they will not provide enough new training signals to the base model. As such, it is necessary to ensure that the generated data can increase the diversity of the data. To this end, using the representation of the $[CLS]$ token of each input document D obtained from the base model, we cluster the documents in the combined dataset $\mathcal{D}$[1]. The number of detected clusters are used as the diversity reward: $R_d(\mathcal{G}) = |C_{\mathcal{D}}|$

The overall reward for the generated documents $\mathcal{G}$ is computed as $R = \beta R_u(\mathcal{G}) + \gamma R_d(\mathcal{G})$, where β and γ are trade-off parameters. The REINFORCE algorithm is employed to update the parameters of the GPT-2 model. Concretely, the parameters of the generative model are updated by the estimated gradient: $\nabla \mathcal{L}_{\mathcal{G}} = -(R(\mathcal{G}))\nabla \log P(\mathcal{G}|\alpha, \mathcal{O})$, where $P(\mathcal{G}|\alpha, \mathcal{O})$ is the probability of the generated data $\mathcal{G}$ computed as the product of the probabilities $P(D'|\alpha, \mathcal{O}) = \sum_{t=1}^{|D'|} P_G(w'_j|D'_{<j}, \alpha)$.

Training Procedure: In order to simultaneously update the parameters of the base model and also the GPT-2 model, we propose a dual training procedure. Specifically, at the first epoch, the parameters of the GPT-2 model are updated using the loss $\mathcal{L}_f$. Next, GPT-2 model is employed to generate the labeled synthetic data to obtain the combined dataset $\mathcal{D}$. After one epoch of training the base model using the loss $\mathcal{L}_{base}$, the parameters of the GPT-2 model are updated using the REINFORCE algorithm. The updated GPT-2 model is employed to generate a new set of synthetic data to be replaced with the previously generated data in $\mathcal{D}$. The new combined data will be next employed to update the base model. This process is repeated until the convergence of training.

3 Experiments

In order to evaluate the effectiveness of the proposed model, called GAOSD (Generation-based Augmentation for Offensive Span Detection), in our experiments, we use the dataset of SemEval 2021 Task 5 (John Pavlopoulos and Laugier, 2021). This dataset contains annotations for 10,000 posts (comments) obtained from the archive of Civil Comment platform (a platform for community to share comments about various civility issues). We use the official splits with 7939/690/2000 documents in train/development/test sets. For each document, the word indices of offensive spans are provided. In our experiments, we create the BIO labels using the provided word indices of the offensive spans.

[1]We use K-means for clustering

Model	Precision	Recall	F1
BiLSTM-CRF	56.72	69.40	57.05
BERT-CRF	63.19	79.42	62.22
DUAL-MRC	62.89	80.21	64.75
SANER	63.09	82.21	65.19
HITSZ-HLT	75.01	89.66	70.83
GAOSD (Ours)	**78.92**	**92.37**	**73.27**

Table 1: Performance of the models in terms of averaged char-level Precision, Recall and F1 score on the test set of the SemEval 2021 Task 5 dataset

In our experiments, we use the BERT$_{base}$ to encode data; 2 layers for feed-forward neural networks with 250 hidden dimensions. The trade-off parameters β and γ are set to 0.1 and 0.05, respectively. The learning rate is set to 0.3 for the Adam optimizer and the batch size of 64 is employed during training. To evaluate the performance, we use the official evaluation metrics for the SemEval 2021 Task 5 (John Pavlopoulos and Laugier, 2021).

We compare the performance of GAOSD with the following baselines: (1) **BiLSTM+CRF**: The GloVe embedded document is encoded by BiLSTM and the labels are predicted by a CRF layer; (2) **BERT+CRF**: BERT$_{base}$ parameters are fine-tuned on OSD task and the task-specific head, i.e., CRF, is employed for label prediction; (3) **HITSZ-HLT** (Zhu et al., 2021): This baseline is the existing SOTA model on SemEval 2021 Task 5 dataset; (4) **SANER** (Nie et al., 2020): This baseline is the SOTA model for sequence labeling on user-generated text; (5) **DUAL-MRC** (Mao et al., 2021): This is the SOTA model for opinion and aspect term extraction. Note that since there are not target annotations in SemEval dataset, we skip the aspect term extraction task to train this baseline. To evaluate the performance we use the official metric, i.e. char-level F1-score, as the evaluation metric. Following prior work (Zhu et al., 2021), we also report the average of char-level precision and recall (Note that due to averaging, $F1 \neq 2(P * R)/(P + R)$).

Results: Table 1 shows the performance of the models on the test set. There are several observations from this table. First, the BiLSTM-CRF model significantly underperforms the other baselines that employ BERT embedding. It clearly shows that the background knowledge encoded in the BERT model is necessary for the task of offensive span detection. Second, both DUAL-

Model	Precision	Recall	F1
GAOSD	**78.39**	**93.82**	**74.21**
UR^{-}	73.29	88.22	68.99
DR^{-}	74.77	83.91	69.51
UDR^{-}	72.49	84.14	66.59
DT^{-}	70.03	79.58	61.72

Table 2: Ablation study on the development set of the SemEval 2021 Task 5 dataset

ID	Document
1	Such beautiful screen that will never turn on!!! Thanks [OFFENSIVE$_S$] stupid designer [OFFENSIVE$_E$] !
2	He constantly talks about his career [OFFENSIVE$_S$] without having any idea about what he says! [OFFENSIVE$_E$]
3	Never trusted this brand as it always deliver just [OFFENSIVE$_S$] crap [OFFENSIVE$_E$] products!

Table 3: Sample texts generated by the fine-tuned GPT-2 model. The toxic spans are also denoted by the special tokens [OFFENSIVE] generated by the model.

MRC and SANER baseline outperform the BERT-CRF model. This higher performance could be attributed to their capability to enhance the representation of the words obtained from the BERT model. Third, among all baselines, our proposed model achieves the highest performance. Our hypothesis for the achieved improvement is that in the proposed method we employ more diverse sets of patterns for expressing toxic. The increased diversity is realized by generating more diverse sentences. Also, this improvement proves that the generated sentences are in-domain and task specific, as such resulting in an improvement. The better performance of our model is impressive, especially considering that we use relatively simple base model compared to other baselines (in particular HITSZ-HLT which is an ensemble model).

Analysis: To study the contribution of the proposed techniques, we conduct an ablation study on the development set of the SemEval 2021 Task 5 dataset. Specifically, we ablate the quality improvement component which ensures the usefulness and diversity of the generated samples. In particular, we study the performance of the model when the Usefulness Reward (**UR**$^{-}$), the Diversity Reward (**DR**$^{-}$), or both of them (**UDR**$^{-}$) are ablated. Also, we study the performance of the model when no dual training is employed (**DT**$^{-}$). Specifically, we first pre-train the base model on the available original data. Next, we fix the parameters of the base model and we use it to compute the usefulness reward. The results are shown in Table 2. This table shows that all components are necessary, as removing each will hurt the performance. Specifically, the dual training has the largest effect on the final performance, indicating the importance of the proposed method. Also, among the two proposed rewards to improve the quality of the generated data, we observe that usefulness reward is more critical, indicating the importance of task-specific generation for data augmentation.

Finally, in order to provide more insight into the quality of the generated data, we provide some randomly selected text generated by the fine-tuned GPT-2 model. The results are shown in table 3. This table shows that the generated samples are natural and also they contain the offensive spans. The generative model is able to correctly locate the offensive spans in the generated text, thereby provided high-quality training samples for the base model. It is worth noting that the offensive spans generated by the fine-tuned GPT-2 model can be either short spans, as in samples 1 and 3 in table 3, or longer phrases, as in sample 2.

4 Related Work

Prior works related to this task can be categorized into two groups: (i) Toxicity Detection: These works aim to classify a piece of text as toxic or non-toxic (Wulczyn et al., 2017; Borkan et al., 2019; Schmidt and Wiegand, 2017; Pavlopoulos et al., 2017a,b, 2019; Zampieri et al., 2019). The main limitation of these works is that they cannot recognize the spans in the text that are responsible for the toxicity of the text. (ii) Opinion Word Extraction: In this group of prior works, models perform a sequence labeling task to identify the spans in the text that convey the sentiment (Liu et al., 2015; Xu et al., 2018; Yin et al., 2016; Wang et al., 2016, 2017; Li and Lam, 2017; Mao et al., 2021). The major limitation of all these models is that they require the existence of the target opinion (i.e., the word or phrase that the text has a sentiment polarity toward it).

5 Conclusion

In this work, we propose a novel method for augmenting data for offensive span detection tasks. Specifically, we employ the pre-trained language model GPT-2 to be fine-tuned on the available training samples for OSD. The fine-tuned model is able to generate in-domain texts with special tokens in-

dicating the offensive spans in them. Moreover, to improve the quality of the generated documents, we propose a novel dual training setting in which the feedback from the OSD model is employed to guide the GPT-2 model to generate more impact synthetic data. Together with a reward for encouraging the diversity of the generated data, the proposed method is effective to augment the training data for OSD, resulting in the state-of-the-art performance on the recent benchmark datasets.

Ethical Consideration

In this work, we present a method for automatically generating offensive spans using the pre-trained generative language model GPT-2. While the sole purpose of the proposed method is to enhance the performance of the offensive content detection systems in social networks, such a generative model can also be misused by someone to automatically make offensive posting continuously without much effort. Prior to our discussion on our measures to mitigate the potential harms of this research, we first justify the risk of this harm. First, as shown in the experiments, employing generation-based models can improve the offensive span detection performance by exposing the model to more diverse patterns of offensive content. Second and more importantly, automatically generating training data for this task reduces the need to expose annotators to a large amount of offensive content. More specifically, since the GPT-2 generated data is effective for training an OSD model, less offensive content is needed to be annotated by human. Thereby, the risk of harmful effects on the annotators is decreased. However, as mentioned before, there is still room for misuse of the findings of this research to automatically generate offensive content. As such, to mitigate the potential harms of this method, we take extra measures into account. In particular, first, we don't release the fine-tuned GPT-2 model on the offensive data, therefore, no one can directly use the artifacts of this research for harmful purposes. Second, since this research demonstrates the potential of the GPT-2 for generating natural-looking offensive content, in return, we also study the effectiveness of a defensive method in which a classifier is employed to identify contents generated by GPT-2 from contents posted by a human. More specifically, we train a BERT model on a dataset consisting of 7,939 human-generated and the same number, i.e., 7,939, automatically generated offensive posts[2]. The input content, i.e. $[CLS]w_1w_2 \dots w_n[SEP]$ where w_i is the $i-$th word of the post, is encoded using the BERT$_{base}$ model. The representation of the $[CLS]$ vector obtained from the final layer of the BERT$_{base}$ is sent to a binary classifier[3] to identify human-generated and automatically generated texts. We evaluate the performance of the trained binary classifier on a test set of 4,000 offensive posts, with a ratio of 50% human-generated content. The accuracy of the classifier on the test set is 92.7% (note that a random baseline would have an accuracy of 50%). Given the simplicity and the high performance of the classifier to recognize the automatically generated posts, we expect that one can directly use this defensive model to automatically and quickly identify the model-generated offensive contents in social networks, thereby mitigating the potential harms of the findings of this research. Also, in future work, with a more comprehensive classifier, better defensive performance is expected. One potential improvement is to incorporate the context of the postings. In particular, while this work shows that GPT-generated content is helpful to improve OSD performance, it does not show the degree to which the generated offensive content is related to the context of the posting. Finally, although this research is conducted on a publicly available dataset of offensive content, in order to prevent disclosing the identity of people mentioned in the dataset, both in the training of the GAOSD and GPT-2 models, we hire 5 undergrad students to double-check and anonymize the SemEval 2021 Task 5 dataset. We expect by anonymizing the data, fewer human subjects can be targeted by automatically generated offensive text.

References

Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. 2020. Do not have enough data? deep learning to the rescue! In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7383–7390. AAAI Press.

[2]Note that we use the fine-tuned GPT-2 model used in the final epoch of the dual training procedure.

[3]We use two-layer feed-forward neural network with 250 hidden states and a sigmoid activation function at the end.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pages 491–500.

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. Pathologies of neural models make interpretations difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3719–3728, Brussels, Belgium. Association for Computational Linguistics.

Jeffrey Sorensen John Pavlopoulos, Ion Androutsopoulos and Léo Laugier. 2021. Toxic span detection at semeval 2021. In *SemEval 2021 (To Appear)*.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. In *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*, pages 18–26, Suzhou, China. Association for Computational Linguistics.

Xin Li and Wai Lam. 2017. Deep multi-task learning for aspect term extraction with memory interaction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2886–2892, Copenhagen, Denmark. Association for Computational Linguistics.

Pengfei Liu, Shafiq Joty, and Helen Meng. 2015. Fine-grained opinion mining with recurrent neural networks and word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1433–1443, Lisbon, Portugal. Association for Computational Linguistics.

Yue Mao, Yi Shen, Chao Yu, and Longjun Cai. 2021. A joint training dual-mrc framework for aspect based sentiment analysis. *ArXiv preprint*, abs/2101.00816.

Yuyang Nie, Yuanhe Tian, Xiang Wan, Yan Song, and Bo Dai. 2020. Named entity recognition for social media texts with semantic augmentation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1383–1391, Online. Association for Computational Linguistics.

John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2017a. Deep learning for user comment moderation. In *Proceedings of the First Workshop on Abusive Language Online*, pages 25–35, Vancouver, BC, Canada. Association for Computational Linguistics.

John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2017b. Deeper attention to abusive user content moderation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1125–1135, Copenhagen, Denmark. Association for Computational Linguistics.

John Pavlopoulos, Nithum Thain, Lucas Dixon, and Ion Androutsopoulos. 2019. ConvAI at SemEval-2019 task 6: Offensive language identification and categorization with perspective and BERT. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 571–576, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Baolin Peng, Chenguang Zhu, Michael Zeng, and Jianfeng Gao. 2020. Data augmentation for spoken language understanding via pretrained models.

Amir Pouran Ben Veyseh, Viet Lai, Franck Dernoncourt, and Thien Huu Nguyen. 2021. Unleash GPT-2 power for event detection. 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)*, pages 6271–6282, Online. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Rajalakshmi Sivanaiah, Angel Suseelan, S Milton Rajendram, and Mirnalinee T.t. 2020. TECHSSN at SemEval-2020 task 12: Offensive language detection using BERT embeddings. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2190–2196, Barcelona (online). International Committee for Computational Linguistics.

Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. 2016. Recursive neural conditional random fields for aspect-based sentiment analysis. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 616–626, Austin, Texas. Association for Computational Linguistics.

Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. 2017. Coupled multi-layer tensor network for co-extraction of aspect and opinion terms. In *Proceedings of AAAI*, pages 3316–3322.

Ronald J. Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. In *Kluwer Academic*.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*, pages 1391–1399. ACM.

Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. 2018. Double embeddings and CNN-based sequence labeling for aspect extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 592–598, Melbourne, Australia. Association for Computational Linguistics.

Yiben Yang, Chaitanya Malaviya, Jared Fernandez, Swabha Swayamdipta, Ronan Le Bras, Ji-Ping Wang, Chandra Bhagavatula, Yejin Choi, and Doug Downey. 2020. G-daug: Generative data augmentation for commonsense reasoning. In *Findings of the Association for Computational Linguistics (EMNLP)*.

Konthala Yasaswini, Karthik Puranik, Adeep Hande, Ruba Priyadharshini, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. 2021. IIITT@DravidianLangTech-EACL2021: Transfer learning for offensive language detection in Dravidian languages. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 187–194, Kyiv. Association for Computational Linguistics.

Yichun Yin, Furu Wei, Li Dong, Kaimeng Xu, Ming Zhang, and Ming Zhou. 2016. Unsupervised word and dependency path embeddings for aspect term extraction. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, pages 2979–2985. IJCAI/AAAI Press.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Danqing Zhang, Tao Li, Haiyang Zhang, and Bing Yin. 2020. On data augmentation for extreme multi-label classification. volume abs/2009.10778.

Qinglin Zhu, Zijie Lin, Yice Zhang, Jingyi Sun, Xiang Li, Qihui Lin, Yixue Dang, and Ruifeng Xu. 2021. HITSZ-HLT at SemEval-2021 task 5: Ensemble sequence labeling and span boundary detection for toxic span detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 521–526, Online. Association for Computational Linguistics.