Fine-tuning for multi-domain and multi-label uncivil language detection

Kadir Bulut Ozler
University of Arizona
kbozler@email.arizona.edu

Kate M Kenski
University of Arizona
kkenski@email.arizona.edu

Stephen A Rains
University of Arizona
srains@email.arizona.edu

Yotam Shmargad
University of Arizona
yotam@email.arizona.edu

Kevin Coe
University of Utah
kevin.coe@utah.edu

Steven Bethard
University of Arizona
bethard@email.arizona.edu

Abstract

Incivility is a problem on social media, and it comes in many forms (name-calling, vulgarity, threats, etc.) and domains (microblog posts, online news comments, Wikipedia edits, etc.). Training machine learning models to detect such incivility must handle the multi-label and multi-domain nature of the problem. We present a BERT-based model for incivility detection and propose several approaches for training it for multi-label and multi-domain datasets. We find that individual binary classifiers outperform a joint multi-label classifier, and that simply combining multiple domains of training data outperforms other recently-proposed fine-tuning strategies. We also establish new state-of-the-art performance on several incivility detection datasets.

1 Introduction

In 2019, 93% of Americans identify incivility as a problem, with 68% classifying it as a "major" problem, and those who experienced incivility faced on average 10.2 uncivil interactions each week (Weber Shandwick et al., 2019). Of those who expect civility to get worse, "social media/the Internet" tops the list of what they blame, above "the White House", "politicians in general", "the news media", etc. Especially on social media and the Internet, this incivility often takes the form of *uncivil language*, features of discussion that convey an unnecessarily disrespectful tone toward the discussion forum, its participants, or its topics (Coe et al., 2014).

Uncivil language can range from name-calling (e.g., *Mark, you're some kind of special stupid*) to vulgarity (e.g., *Just build the damn mine already!*) to threats (e.g., *Fine. I will destroy you.*) and beyond. Different types of incivilities often appear in the same utterance (e.g., name-calling, vulgarity, and threats are all included in *SHUT UP, YOU FAT POOP, OR I WILL KICK YOUR ASS!!!*). Uncivil language appears in many places online, from microblogs like Twitter, to comments on online newspapers, to edit histories of resources like Wikipedia.

Uncivil language detection is thus a multi-label and multi-domain language processing problem. While there has been much research in natural language processing methods for identifying such incivility, especially in the subarea of *abusive language* (Wiegand et al., 2019; Zampieri et al., 2019; Basile et al., 2019; Sadeque et al., 2019; van Aken et al., 2018, etc.), the multi-label and multi-domain nature of incivility detection is understudied. We thus consider incivility detection on several datasets that (1) require the classification of incivility into several not-mutually-exclusive fine-grained categories, and (2) cover multiple genres of online interactions. Our contributions are:

- We achieved a new state-of-the-art on both the Coe et al. (2014) and Conversation AI (2018) datasets using BERT (Devlin et al., 2019).
- We compared several algorithms for training classifiers across the multiple domains in these datasets and showed that combining the training data from all domains outperforms other recently-proposed fine-tuning strategies.
- We compared several approaches for handling the multi-label nature of these datasets and showed that independent binary classifiers outperform jointly-trained models.

2 Task

We frame uncivil language detection as a multi-label text classification problem, where the input is a piece of text, and the outputs are the types of incivilities (*name-calling*, *vulgarity*, etc.) that are present. Formally, we aim to learn a function h such that for each piece of text x:

$$h(repr(x)) = \vec{y} \quad (1)$$

Proceedings of the Fourth Workshop on Online Abuse and Harms, pages 28–33
Online, November 20, 2020.
https://doi.org/10.18653/v1/P17

Annotation Scheme	Domain	Train	Dev	Test	Data split	aspersion	lying accusation	name-calling	pejorative	vulgarity	toxic	severe-toxic	obscene	threat	insult	identity-hate
Coe et al. (2014)	local news comments	3945	987	1233	standard	✓	✓	✓	✓	✓						
Coe et al. (2014)	local politics Tweets	3040	760	-	no standard			✓								
Coe et al. (2014)	Russian troll Tweets	1798	200	-	no standard			✓								
Conversation AI (2018)	Wikipedia comments	37902	312	-	only train available						✓	✓	✓	✓	✓	✓

Table 1: Statistics for the multi-domain and multi-label datasets considered. For data sets with no standard split, or where the test set is unavailable as in Conversation AI (2018), we created our own custom train/dev split.

where $repr(x)$ is a tensor representing that text (e.g., a series of word vectors), and $\vec{y}$ is a binary vector where $\vec{y}_i$ is 1 if x contains the i^{th} form of incivility and 0 otherwise.

We frame learning such h functions a multi-domain classifier training problem, where training and testing data are drawn from multiple domains (*news comments*, *politician tweets*, etc.). Formally, given a domain D_i, we aim to learn a function h_{D_i} that maximizes performance on test data $D_{i_{\text{test}}}$ by training on examples $(x, \vec{y})$ drawn from training data $D_{1_{\text{train}}} \bigcup D_{2_{\text{train}}} \bigcup \ldots \bigcup D_{n_{\text{train}}}$.

3 Data

We consider the following datasets for evaluating multi-label and multi-domain incivility detection.

Local news comments In this multi-label dataset, the following labels are defined and used to annotate online comments on local news articles by Coe et al. (2014):

- *aspersion*: "Mean-spirited or disparaging words directed at a person or group of people."
- *lying accusation*: "Mean-spirited or disparaging words directed at an idea, plan, policy, or behavior."
- *name-calling*: "Stating or implying that an idea, plan, or policy was disingenuous."
- *pejorative*: "Using profanity or language that would not be considered proper (e.g., pissed, screw) inprofessional discourse."
- *vulgarity*: "Disparaging remark about the way in which a person communicates."

Local politics Tweets Coe and colleagues also annotated a collection of microblog posts from the Twitter accounts of their local politicians, but only for *name-calling* incivility.

Russian troll Tweets Coe and colleagues also annotated a small subset of the 3 million English Tweets written by Russian trolls and collected by Linvill and Warren (2018)[1], again for just *name-calling* incivility.

Wikipedia comments In this multi-label dataset, also known as the Kaggle Toxic Comment Classification Challenge, Jigsaw/Google's Conversation AI team annotated comments from Wikipedia's talk page edits (Conversation AI, 2018) for the presence of the following types of abusive language, defined by Perspective AI (2020).

- *toxic*: "A rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion."
- *severe-toxic*: "A very hateful, aggressive, disrespectful comment or otherwise very likely to make a user leave a discussion or give up on sharing their perspective. This attribute is much less sensitive to more mild forms of toxicity, such as comments that include positive uses of curse words."
- *obscene*: "Swear words, curse words, or other obscene or profane language."
- *threat*: "Describes an intention to inflict pain, injury, or violence against an individual or group."
- *insult*: "Insulting, inflammatory, or negative comment towards a person or a group of people."
- *identity-hate*: "Negative or hateful comments targeting someone because of their identity."

Table 1 shows statistics for the different data sets.

[1] https://github.com/fivethirtyeight/russian-troll-tweets/

The three datasets annotated by Coe and colleagues can be used in multi-domain experiments, as they share the same annotation scheme. They share only the label *name-calling*, so our multi-domain experiments consider only binary classification. The local news comments and Wikipedia comments datasets can be used in multi-label experiments, as they have been annotated for multiple forms of incivility. They do not share annotation schemes, so our multi-label experiments consider each multi-label dataset separately.

4 Prior Work

There is much recent work on detecting incivility (also referred to as toxicity, abusive language, offensive language, etc.) in social media. Wiegand et al. (2019) presents an overview of such efforts and shows that many datasets constructed for this purpose have unintended bias because of how they have been sampled. We focus on the Coe et al. (2014) and Conversation AI (2018) datasets because they do not have the problems with topic-biased sampling that some other datasets do, where topic words are better predictors of incivility than uncivil words.

There have also been several recent shared tasks that consider incivility. Both the OffensEval shared task (Zampieri et al., 2019) and the HatEval (Basile et al., 2019) shared task ran as part of SemEval-2019 and considered detection of various forms of offensive and hate speech. Neither of these tasks focused on a multi-label or multi-domain problem.

A few models have been designed for and evaluated on the multi-label, multi-domain corpora we consider. Sadeque et al. (2019) considered the local news comments corpus, training recurrent neural network models, and focusing on only the top two most frequent labels for this dataset. They achieved 0.48 F_1 for *name-calling* and 0.53 F_1 for *vulgarity*. van Aken et al. (2018) presented multiple approaches to the Wikipedia comments dataset. They developed an ensemble of logistic regression, recurrent neural networks, and convolutional neural networks, achieving an AUC score of 0.983.

There are a few recent works in cross-domain abusive language detection. Wiegand et al. (2018); Karan and Šnajder (2018); Pamungkas and Patti (2019) all explore training models on one abusive language dataset and testing on another. They focus on binary predictions and bag-of-words support vector machine classifiers (though Pamungkas and Patti (2019) also explores a recurrent neural network). They do not consider multi-label problems, or modern pre-trained neural networks like BERT, which were more successful in recent shared tasks on abusive language (Zampieri et al., 2019). They also evaluate on several datasets that have been identified as problematic by Wiegand et al. (2019) due to their use of topic-biased sampling.

5 Experiments

We use BERT (Devlin et al., 2019) as the starting point for all experiments. BERT is a pre-trained transformer-based neural network that has shown impressive performance on a wide variety of NLP tasks. We follow the standard approach for fine-tuning BERT for text classification, placing a fully connected layer over BERT's `[CLS]` output. We use n sigmoids on this layer rather than a softmax activation, since we are performing multi-label classification. BERT is then fine-tuned as usual, with hyperparameters like learning rate, maximum sequence length, number of epochs, training batch size tuned on the development set. We explored each hyperparameter within the following ranges:

learning rate: 8e-6, 2e-5, 4e-5, 8e-5
maximum sequence length: 128, 256, 512
number of epochs: 2, 3, 4, 5, 6, 8
training batch size: 16, 32, 64, 128

5.1 Multi-domain models

We consider three methods for training classifiers for prediction in multiple domains:

Single One classifier is fine-tuned for each domain.

Joint One classifier is fine-tuned on the combined training data from all the domains.

Joint→Single First, a joint classifier is fine-tuned. Then, the joint classifier parameters are used to initialize n individual classifiers, one for each domain. This approach is inspired by Liu et al. (2019a), where for some natural language understanding problems, they found that multi-task fine-tuning followed by individual task fine-tuning outperformed multi-task fine-tuning alone.

Since our multi-domain datasets share only the label *name-calling*, we train our multi-domain classifiers only for binary classification (i.e., they are not also multi-label).

Data	Training method	Local news comments			Russian troll Tweets			Local politics Tweets		
		P	R	F_1	P	R	F_1	P	R	F_1
Dev	Single	0.63	0.52	0.57	0.67	0.63	0.65	0.65	0.76	0.70
Dev	Joint	**0.75**	**0.57**	**0.65**	0.81	**0.81**	**0.81**	0.75	**0.85**	0.80
Dev	Joint→Single	0.67	0.52	0.58	**0.91**	0.63	0.74	**0.81**	0.81	**0.81**
Test	Sadeque et al. (2019)	0.46	**0.51**	0.48	-	-	-	-	-	-
Test	Best Dev model: Joint	**0.62**	**0.51**	**0.56**	0.83	0.67	0.74	0.71	0.60	0.65

Table 2: Multi-domain results: Performance on the label *name-calling*, for different multi-domain training methods across different datasets. When results from two or more models are comparable, the highest performance is marked in bold. Sadeque et al. (2019) is the previous state-of-the-art on the local news comments. There is no prior state-of-the-art for the other datasets.

Table 2 shows the results of these experiments. The first three rows compare the different training procedures on the development sets. We find that simply combining all the data achieves the best F_1 for both the local news comments and Russian troll Tweets data, and similar F_1 to the more complicated Joint→Single procedure in the remaining dataset. When we evaluate this best model on the test data, we achieve a new state-of-the-art on the local news comments corpus, 0.56 F_1. We are the first to report results on the local politics Tweets and Russian troll Tweets domains, as Sadeque et al. (2019) did not evaluate on these.

These results did not replicate the findings of Liu et al. (2019a) when applied to our incivility datasets; the extra fine-tuning for each domain was unhelpful, and simply combining all the data was the best. This probably argues for exploring other approaches for domain adapatation, e.g., Kim et al. (2016), but it may also simply suggest that Coe et al. (2014)'s annotators were consistent across datasets, making it easy for BERT to learn the core linguistic phenomenon despite differences in domains.

5.2 Multi-label models

Similar to our approach for multi-domain models, we consider three methods for training classifiers for multi-label prediction:

Single One binary classifier is fine-tuned for each label. The output layer of the model is a single sigmoid unit.

Joint One joint classifier is fine-tuned for all labels. The output layer of the model is n sigmoid units, one for each label.

Joint→Single First, a joint classifier is fine-tuned. Then, the joint classifier parameters are used to initialize n binary classifiers, one for each label. This is again inspired by the multi-task training procedure of Liu et al. (2019a).

Since our multi-label datasets do not share an annotation scheme, we train the multi-label classifiers on only one dataset at a time (i.e., they are not also multi-domain).

Table 3 shows the results of these experiments[2]. We find that in most cases training individual binary classifiers (Single) is better than a jointly-learned multi-label classifier (Joint). This is somewhat surprising as the latter is the standard approach with neural networks (Adhikari et al., 2019).

Curious if the problem was some low-frequency classes, we tried training a multi-label model on just the three most frequent classes of the Wikipedia comments dataset (Joint top-3 classes), *toxic*, *obscene*, and *insult*. That slightly improved performance on those three classes, but of course at the cost of the classes now being ignored. Adding the staged training procedure (Joint→Single) on top of this classifier only decreased performance. This suggests that class imbalance may be part of the problem, but is not the full explanation.

Note that we are the first to report all individual label F_1s on both datasets. In the case of the local news comments data, this is because Sadeque et al. (2019), noting the class imbalance problem, decided to only train and evaluate on two classes. In the case of the Wikipedia comments data, this is because the official evaluation metric is AUC, so most systems focused on optimizing this measure. However, as Table 3 shows, while we achieve a state-of-the-art AUC, AUC is not a very discriminative measure for this dataset. For example, both the Single model that predicts all six classes and the Joint top-3 classes model that doesn't even try to predicts *severe-toxic*, *threat*, or

[2]Note that the Wikipedia comments dataset does not have a development split, so "Dev" experiments on that dataset are actually on the test set, following van Aken et al. (2018).

		Local news comments					Wikipedia comments						
		aspersion	lying accusation	name-calling	pejorative	vulgarity	toxic	severe-toxic	obscene	threat	insult	identity-hate	*official metric*
Data	Training method	F_1	F_1	F_1	F_1	F_1	F_1	F_1	F_1	F_1	F_1	F_1	AUC
Dev	Single	0.24	**0.52**	**0.59**	**0.52**	**0.46**	**0.86**	**0.50**	**0.88**	**1.00**	0.76	**1.00**	0.990
Dev	Joint all classes	**0.37**	0.44	0.54	0.46	**0.46**	0.80	0.33	0.83	**1.00**	0.80	**1.00**	0.988
Dev	Joint top-3 classes	-	-	-	-	-	0.83	0.00	**0.88**	0.00	**0.86**	0.00	0.990
Dev	Joint top-3→Single	-	-	-	-	-	0.83	0.00	0.77	0.00	0.80	0.00	0.983
Test	Sadeque et al. (2019)	-	-	0.48	-	**0.53**	-	-	-	-	-	-	-
Test	van Aken et al. (2018)	-	-	-	-	-	-	-	-	-	-	-	0.983
Test	Best Dev model: Single	0.05	0.36	**0.55**	0.28	0.50	0.86	0.50	0.88	1.00	0.76	1.00	0.990

Table 3: Multi-label results: Performance on each label, for different multi-label training methods across different datasets. When results from two or more models are comparable, the highest performance is marked in bold. The final column is the official evaluation measure for the Wikipedia comments dataset. Sadeque et al. (2019) is the state-of-the-art on the local news comments data, and van Aken et al. (2018) is the state-of-the-art on the Wikipedia comments data.

identity-hate achieve the same AUC of 0.990. The F_1 scores more clearly show that the Joint top-3 classes model is as good or better for all labels but *insult*.

6 Limitations

We focused on a BERT-based model due to its top-ranking performance in related shared tasks (Zampieri et al., 2019), but recent advances over BERT, e.g., RoBERTa (Liu et al., 2019b) might yield additional gains. We also focused on the limited number of datasets that could support multi-label and/or multi-domain experiments, but our results could be strengthened by creating new multi-label, multi-domain datasets. Finally, class imbalance only partly explains why a joint multi-label classifier failed to outperform independent binary classifiers, indicating that further investigation is needed into multi-label classification approaches for uncivil language.

7 Conclusion

We applied BERT on multi-label and multi-domain incivility detection tasks and achieved a new state-of-the-art on several different datasets. In exploring different training procedures, we found that it was better to directly combine data from multiple domains than other more complex procedures, and that it was better to train individual binary classifiers than to train a joint multi-label classifier.

References

Ashutosh Adhikari, Achyudh Ram, Raphael Tang, and Jimmy Lin. 2019. Rethinking complex neural network architectures for document classification. 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)*, pages 4046–4051, Minneapolis, Minnesota. Association for Computational Linguistics.

Betty van Aken, Julian Risch, Ralf Krestel, and Alexander Löser. 2018. Challenges for toxic comment classification: An in-depth error analysis. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 33–42, Brussels, Belgium. Association for Computational Linguistics.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Kevin Coe, Kate Kenski, and Stephen A. Rains. 2014. Online and Uncivil? Patterns and Determinants of Incivility in Newspaper Website Comments. *Journal of Communication*, 64(4):658–679.

Conversation AI. 2018. Toxic comment classification challenge. `https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge`.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mladen Karan and Jan Šnajder. 2018. Cross-domain detection of abusive language online. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 132–137, Brussels, Belgium. Association for Computational Linguistics.

Young-Bum Kim, Karl Stratos, and Ruhi Sarikaya. 2016. Frustratingly easy neural domain adaptation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 387–396, Osaka, Japan. The COLING 2016 Organizing Committee.

Darren L. Linvill and Patrick L. Warren. 2018. Troll factories: The internet research agency and state-sponsored agenda building. http://pwarren.people.clemson.edu/Linvill_Warren_TrollFactory.pdf.

Wei Liu, Lei Li, Zuying Huang, and Yinan Liu. 2019a. Multi-lingual Wikipedia summarization and title generation on low resource corpus. In *Proceedings of the Workshop MultiLing 2019: Summarization Across Languages, Genres and Sources*, pages 17–25, Varna, Bulgaria. INCOMA Ltd.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Endang Wahyu Pamungkas and Viviana Patti. 2019. Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 363–370, Florence, Italy. Association for Computational Linguistics.

Perspective AI. 2020. Available attributes and languages. https://support.perspectiveapi.com/s/about-the-api-attributes-and-languages.

Farig Sadeque, Stephen Rains, Yotam Shmargad, Kate Kenski, Kevin Coe, and Steven Bethard. 2019. Incivility detection in online comments. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 283–291, Minneapolis, Minnesota. Association for Computational Linguistics.

Weber Shandwick, Powell Tate, and KRC Research. 2019. Civilty in america 2019: Solutions for tomorrow. https://www.webershandwick.com/news/civility-in-america-2019-solutions-for-tomorrow/.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of Abusive Language: the Problem of Biased Datasets. 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)*, pages 602–608, Minneapolis, Minnesota. Association for Computational Linguistics.

Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg. 2018. Inducing a lexicon of abusive words – a feature-based approach. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1046–1056, New Orleans, Louisiana. Association for Computational Linguistics.

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.