Generalisability of Topic Models in Cross-corpora Abusive Language Detection

Tulika Bose, Irina Illina, Dominique Fohr
Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
tulika.bose, illina, dominique.fohr@loria.fr

Abstract

Rapidly changing social media content calls for robust and generalisable abuse detection models. However, the state-of-the-art supervised models display degraded performance when they are evaluated on abusive comments that differ from the training corpus. We investigate if the performance of supervised models for cross-corpora abuse detection can be improved by incorporating additional information from topic models, as the latter can infer the latent topic mixtures from unseen samples. In particular, we combine topical information with representations from a model tuned for classifying abusive comments. Our performance analysis reveals that topic models are able to capture abuse-related topics that can transfer across corpora, and result in improved generalisability.

1 Introduction

With the exponentially increased use of social networking platforms, concerns on *abusive* language has increased at an alarming rate. Such language is described as hurtful, toxic, or obscene, and targets individuals or a larger group based on common societal characteristics such as race, religion, ethnicity, gender, etc. The increased spread of such content hampers free speech as it can potentially discourage users from expressing themselves without fear, and intimidate them into leaving the conversation. Considering variations of online abuse, toxicity, hate speech, and offensive language as abusive language, this work addresses the detection of abusive versus non-abusive comments.

Automatic detection of abuse is challenging as there are problems of changing linguistic traits, subtle forms of abuse, amongst others (Vidgen et al., 2019). Moreover, the performance of models trained for abuse detection are found to degrade considerably, when they encounter abusive comments that differ from the training corpus (Wiegand et al., 2019; Arango et al., 2019; Swamy et al., 2019; Karan and Šnajder, 2018). This is due to the varied sampling strategies used to build training corpus, topical and temporal shifts (Florio et al., 2020), and varied targets of abuse across corpora. Since social media content changes rapidly, abusive language detection models with better generalisation can be more effective (Yin and Zubiaga, 2021). To this end, a cross-corpora analysis and evaluation is important.

Topic models have been explored for generic cross-domain text classification (Jing et al., 2018; Zhuang et al., 2013; Li et al., 2012), demonstrating better generalisability. Moreover, they can be learnt in an unsupervised manner and can infer topic mixtures from unseen samples. This inspires us to exploit topic model representations for cross-corpora abuse detection.

Recently, Caselli et al. (2021) have "retrained" BERT (Devlin et al., 2019) over large-scale abusive Reddit comments to provide the *HateBERT* model which has displayed better generalisability in cross-corpora experiments. Furthermore, Peinelt et al. (2020) show that combination of topic model and BERT representations leads to better performance at semantic similarity task. Taking these studies into account, we investigate if combining topic representation with contextualised HateBERT representations can result in better generalisability in cross-corpora abuse detection. Cross corpora evaluation on three common abusive language corpora supports and demonstrates the effectiveness of this approach. Besides, we bring some insights into how the association of unseen comments to abusive topics obtained from original training data can help in cross-corpora abusive language detection.

The rest of the paper is organised as follows: Section 2 describes the architecture of the combination of topic model and HateBERT. Section 3 presents our experimental settings. An analysis of the results obtained is present in Section 4, and Section 5 concludes the paper.

Proceedings of the 4th NLP4IF Workshop on NLP for Internet Freedom, pages 51–56
June 6, 2021.

2 Combining Topic Model and HateBERT

In this work, we leverage the Topically Driven Neural Language Model (TDLM) (Lau et al., 2017) to obtain topic representations, as it can employ *pre-trained* embeddings which are found to be more suitable for short Twitter comments (Yi et al., 2020). The original model of TDLM applies a Convolutional Neural Network (CNN) over word-embeddings to generate a comment embedding. This comment embedding is used to learn and extract topic distributions. Cer et al. (2018) show that transfer learning via sentence embeddings performs better than word-embeddings on a variety of tasks. Hence, we modify TDLM to accept the transformer based Universal Sentence Encoder (USE) (Cer et al., 2018) embeddings extracted from input comments, instead of the comment embeddings from CNN. The modified model is denoted as U-TDLM hereon. Refer to Appendix A.1 for the architecture of U-TDLM and also to Lau et al. (2017).

U-TDLM is trained on the train set from the source corpus and is used to infer on the test set from a different target corpus. The topic distribution per comment c is given by $T_c = [p(t_i|c)]_{i=1:k}$, where k is the number of topics. T_c is passed through a Fully Connected (FC) layer to obtain transformed representation $T_c^{'}$. Besides, we first perform supervised fine-tuning of HateBERT[1] on the train set of the source corpus. The vector corresponding to the [CLS] token in the final layer of this fine-tuned HateBERT model is chosen as the HateBERT representation for a comment. It is transformed through an FC layer to obtain the C vector. Finally, in the *combined model* (HateBERT+U-TDLM), the concatenated vector $[T_c^{'}; C]$ is passed through a final FC and a softmax classification layer. The readers are referred to Appendix A.2 for the architecture of the individual, and the combined models.

3 Evaluation Set-up

3.1 Experimental Settings

We perform experiments on three different publicly available abusive tweet corpora, namely, *HatEval* (Basile et al., 2019), *Waseem* (Waseem and Hovy, 2016), and *Davidson* (Davidson et al., 2017). We target a binary classification task with classes: *abusive* and *non abusive*, following the precedent of previous work on cross corpora analysis (Wiegand et al., 2019; Swamy et al., 2019; Karan and Šnajder, 2018). For *HatEval*, we use the standard partition of the shared task, whereas the other two datasets are randomly split into train (80%),development (10%), and test (10%). The statistics of the train-test splits of these datasets are listed in Table 1.

Datasets	Number of comments		Average comment length	Abuse %
	Train	Test		
HatEval	9000	3000	21.3	42.1
Waseem	8720	1090	14.7	26.8
Davidson	19817	2477	14.1	83.2

.

Table 1: Statistics of the datasets used (average comment length is calculated in terms of word numbers).

We choose a topic number of 15 for our experiments based on the results for in-corpus performance and to maintain a fair comparison. Besides, the best model checkpoints are selected by performing early-stopping of the training using the respective development sets. The FC layers are followed by Rectified Linear Units (ReLU) in the individual as well as the combined models. In the individual models, the FC layers for transforming T_c and the HateBERT representation have 10 and 600 hidden units, respectively. The final FC layer in the combined model has 400 hidden units. Classification performance is reported in terms of mean F1 score and standard deviation over five runs, with random initialisations.

3.2 Data Pre-processing

We remove the URLs from the Twitter comments, but retain Twitter handles as they can contribute to topic representations.[2] Hashtags are split into constituent words using the tool CrazyTokenizer[3], and words are converted into lower-case. U-TDLM involves prediction of words from the comments based on topic representations. In this part, our implementation uses stemmed words and skips stopwords.

4 Results and Analysis

Table 2 presents the in-corpus and cross-corpora evaluation of the HateBERT and U-TDLM models.

[1]Pre-trained model from https://osf.io/tbd58/

[2]Eg., the topic associated with @realDonaldTrump.

[3]https://redditscore.readthedocs.io/en/master/tokenizing.html

Train set	In-corpus performance		Cross-corpus test set	Cross-corpora performance		
	HateBERT	U-TDLM		HateBERT	U-TDLM	HateBERT + U-TDLM
HatEval	53.9±1.7	41.5±0.6	Waseem	66.5±2.2	55.5±2.6	**67.8**±2.4
			Davidson	59.2±2.5	**64.4**±2.3	60.4±1.4
Waseem	86.1±0.4	73.7±1.4	HatEval	**55.8**±1.4	36.7±0.0	55.4±0.7
			Davidson	59.8±3.6	28.2±2.4	**64.8**±1.8
Davidson	93.7±0.2	75.6±0.8	HatEval	**51.8**±0.2	50.5±1.3	**51.8**±0.3
			Waseem	66.6±3.0	48.7±3.3	**68.5**±2.1
Average	77.9	63.6		60.0	47.3	**61.5**

Table 2: Macro average F1 scores (mean±std-dev) for in-corpus and cross-corpora abuse detection. The best in each row for the cross-corpora performance is marked in bold.

All models are trained on the train set of the source corpus. The in-corpus performance of the models is obtained on the source corpora test sets, while the cross-corpora performance is obtained on target corpora test sets. It is shown in Table 2 that the cross-corpora performance degrades substantially as compared to the in-corpus performance, except for *HatEval* which indeed has a low in-corpus performance. *HatEval* test set is part of a shared task, and similar in-corpus performance have been reported in prior work (Caselli et al., 2021). Overall, comparing the cross-corpora performances of all models, we can observe that the combined model (HateBERT + U-TDLM) either outperforms HateBERT or retains its performance. This hints that incorporating topic representations can be useful in cross-corpora abusive language detection. As an ablation study, we replaced U-TDLM features with random vectors to evaluate the combined model. Such a concatenation decreased the performance in the cross-corpora setting, yielding an average macro-F1 score of 59.4. This indicates that the topic representations improve generalisation along with HateBERT.

4.1 Case-studies to Analyse Improvements from U-TDLM

We investigate the cases in Table 2 which report relatively large improvements, as compared to HateBERT, either with HateBERT+U-TDLM (train on *Waseem*, test on *Davidson*) or only with U-TDLM (train on *HateEval*, test on *Davidson*). Some of the prominent topics from *Waseem* and *HateEval* associated with abuse, and the top words corresponding to these topics are provided in Table 3 and Table 5, respectively. For better interpretation, topic names are manually assigned based on the top words and the knowledge of the individual corpora. We consider the abusive class as positive, and the non-abusive class as negative in the subsequent discussion.

Topic id	Names	Top words
4	Sexism in sports	football, sex, sport, feminist, drive, woman, call, sexist
9	Feminism	feminist, article, ebook, equality, patriarchy, abuse, freebsd, harass
12	Cooking show	katie, score, mkr, cook, c*nt, blond, less, strategic

Table 3: U-TDLM trained on Waseem's train set (topic names are assigned manually for interpretation).

Train on *Waseem* →Test on *Davidson*: In this case, U-TDLM shows poor performance due to the large number of *False Negatives* (#FN for U-TDLM: 1824), and less *True Positives* (#TP for U-TDLM: 266). The combined model, on the other hand, has higher *True Positives* compared to those obtained from HateBERT (#TP for HateBERT+U-TDLM: 1556, #TP for HateBERT: 1267). The count of *True Negatives* with the combined model remains similar to that in HateBERT (#TN for HateBERT + U-TDLM: 314, #TN for HateBERT: 340). This indicates that U-TDLM introduces some complementary information in the combined model.

We analyse a few abusive comments in the test set of *Davidson* (target) in Table 4, which are wrongly classified by HateBERT, but correctly detected as abusive by the combined model. The topical membership of these abusive comments from *Davidson* indicates that U-TDLM associates high

Source →Target	Abusive Comments in Target	Source topics
Waseem →Davidson	When women are so proud that they don't like to cook; clean b*tch stop being lazy..It's not cute.	4, 12
	ya girl is a slimy ass h*e. get her under control and tell her to stop spraying bullshit out her mouth all day.	4, 9, 12
HatEval →Davidson	No. Its wrong to try to change f*ggots; There is no "therapy"....sympathize like they are retards.	3, 7
	Naturally, when a shitty leftist rag talks trash about another shitty leftist rag, you better fall in line...	10

Table 4: Abusive comments in the target corpus, correctly classified by HateBERT+U-TDLM (Waseem →Davidson) and U-TDLM (HatEval →Davidson). "Source topics" : topics that are assigned high weights by U-TDLM trained on Source.

Topic id	Names	Top words
3	Explicit abuse 1	men, c*ck, d*ck, woman, picture, sl*t, s*ck, guy
7	Explicit abuse 2	b*tch, ho*, n*gger, girlfriend, f*ck, shit, s*ck, dumb
10	Politics related	therickwilson, anncoulter, c*nt, commies, tr*nny, judgejeanine, keitholbermann, donaldjtrumpjr

Table 5: U-TDLM trained on HatEval's train set (topic names are assigned manually for interpretation).

weights to the relevant abuse-related topics from *Waseem*. As indicated in the first example, an abusive comment against women that discusses cooking, in *Davidson*, is mapped to the topics 4 (sexism) and 12 (cooking show) from *Waseem*. Similarly, the second comment gets high weight in the three topics 4, 9 and 12 due to its sexist content and use of a profane word. Other pairs of corpora that yield improved performance with the combined model also follow similar trends as above.

Train on *HatEval* →Test on *Davidson*: In this case, while U-TDLM performs considerably well, the combined model only provides a slight improvement over HateBERT, as per Table 2. U-TDLM has a higher TP when compared to both HateBERT and the combined model (#TP for U-TDLM: 1924, #TP for HateBERT+U-TDLM: 1106, #TP for HateBERT: 1076), with lower TN (#TN for U-TDLM: 130, #TN for HateBERT+U-TDLM: 373, #TN for HateBERT: 374).

Few abusive comments from *Davidson* that are correctly classified by U-TDLM alone are presented in Table 4. The first comment for this case have high weights for the abuse-related topics 3 and 7 from *HatEval* due to the presence of the profane word "f*ggot". The second comment only gets a high weight for topic 10, which deals with politics. This is due to the word "leftist", which is associated with a political ideology. As per our analysis, we found that all of these source topics are highly correlated with the abusive labels in the source corpus of *HatEval*. As such, these comments from the target corpus of *Davidson* are correctly classified as abusive by U-TDLM.

5 Discussion and Conclusion

An in-corpus and cross-corpora evaluation of HateBERT and U-TDLM has helped us confirm our perspective on generalisation in the abusive language detection task. A contextualised representation model like HateBERT can achieve great levels of performance on the abusive language detection task, only when the evaluation dataset does not differ from the training set. The performance of this model degrades drastically on abusive language comments from unseen contexts. Topic models like U-TDLM, which express comments as a mixture of topics learnt from a corpus, allow unseen comments to trigger abusive language topics. While topic space representations tend to lose the exact context of a comment, combining them with HateBERT representations can give modest improvements over HateBERT or at the least, retain the performance of HateBERT. These results should fuel interest and motivate further developments in the generalisation of abusive language detection models.

Acknowledgements

This work was supported partly by the french PIA project "Lorraine Université d'Excellence", reference ANR-15-IDEX-04-LUE.

References

Aymé Arango, Jorge Pérez, and Barbara Poblete. 2019. Hate speech detection is not as easy as you may think: A closer look at model validation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 45–54, New York, NY, USA. Association for Computing Machinery.

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.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. Hatebert: Retraining bert for abusive language detection in english. *arXiv preprint arXiv:2010.12472*.

Daniel Cer, Yinfei Yang, Sheng yi Kong, Nan Hua, Nicole Lyn Untalan Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Céspedes, Steve Yuan, Chris Tar, Yun hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder. In *EMNLP demonstration*, Brussels, Belgium.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ICWSM '17, pages 512–515.

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.

Komal Florio, Valerio Basile, Marco Polignano, Pierpaolo Basile, and Viviana Patti. 2020. Time of your hate: The challenge of time in hate speech detection on social media. *Applied Sciences*, 10(12).

Baoyu Jing, Chenwei Lu, Deqing Wang, Fuzhen Zhuang, and Cheng Niu. 2018. Cross-domain labeled LDA for cross-domain text classification. In *IEEE International Conference on Data Mining, ICDM 2018, Singapore, November 17-20, 2018*, pages 187–196. IEEE Computer Society.

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.

Jey Han Lau, Timothy Baldwin, and Trevor Cohn. 2017. Topically driven neural language model. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 355–365, Vancouver, Canada. Association for Computational Linguistics.

Lianghao Li, Xiaoming Jin, and Mingsheng Long. 2012. Topic correlation analysis for cross-domain text classification. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, AAAI'12, page 998–1004. AAAI Press.

Nicole Peinelt, Dong Nguyen, and Maria Liakata. 2020. tBERT: Topic models and BERT joining forces for semantic similarity detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7047–7055, Online. Association for Computational Linguistics.

Steve Durairaj Swamy, Anupam Jamatia, and Björn Gambäck. 2019. Studying generalisability across abusive language detection datasets. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 940–950, Hong Kong, China. Association for Computational Linguistics.

Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019. Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

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.

F. Yi, B. Jiang, and J. Wu. 2020. Topic modeling for short texts via word embedding and document correlation. *IEEE Access*, 8:30692–30705.

Wenjie Yin and Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: a review on obstacles and solutions. *arXiv preprint arXiv:2102.08886*.

Fuzhen Zhuang, Ping Luo, Peifeng Yin, Qing He, and Zhongzhi Shi. 2013. Concept learning for cross-domain text classification: A general probabilistic framework. In *IJCAI International Joint Conference on Artificial Intelligence*, pages 1960–1966.

A Appendices

A.1 Topic Model U-TDLM

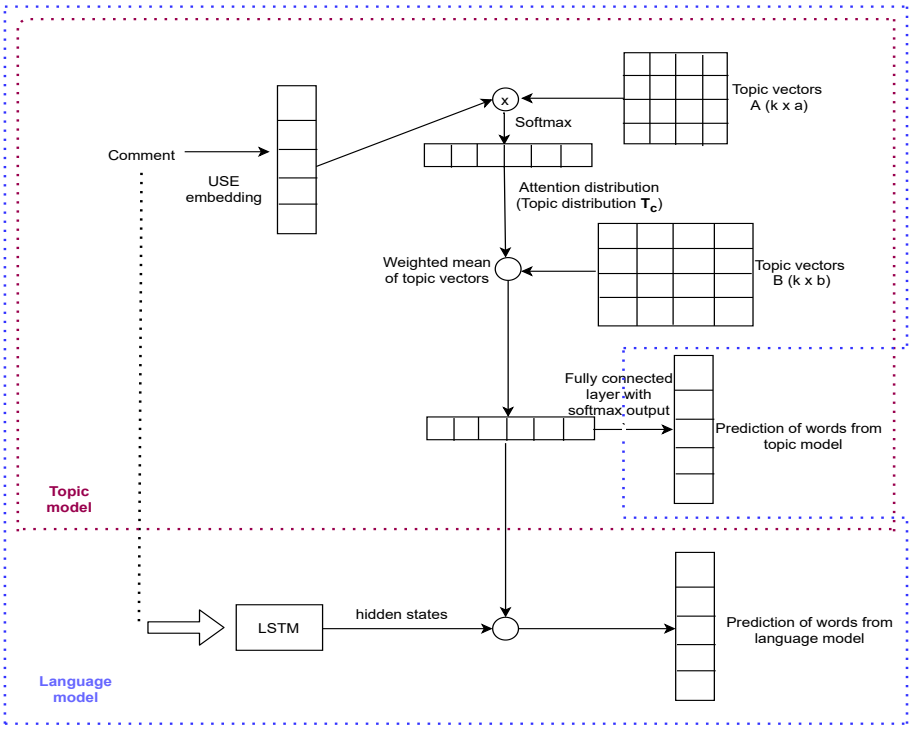

Figure 1: Architecture of U-TDLM. As compared to TDLM (Lau et al., 2017), CNN on comment is replaced by USE (Universal Sentence Embedding). k = number of topics.

A.2 Architecture of Combined Model

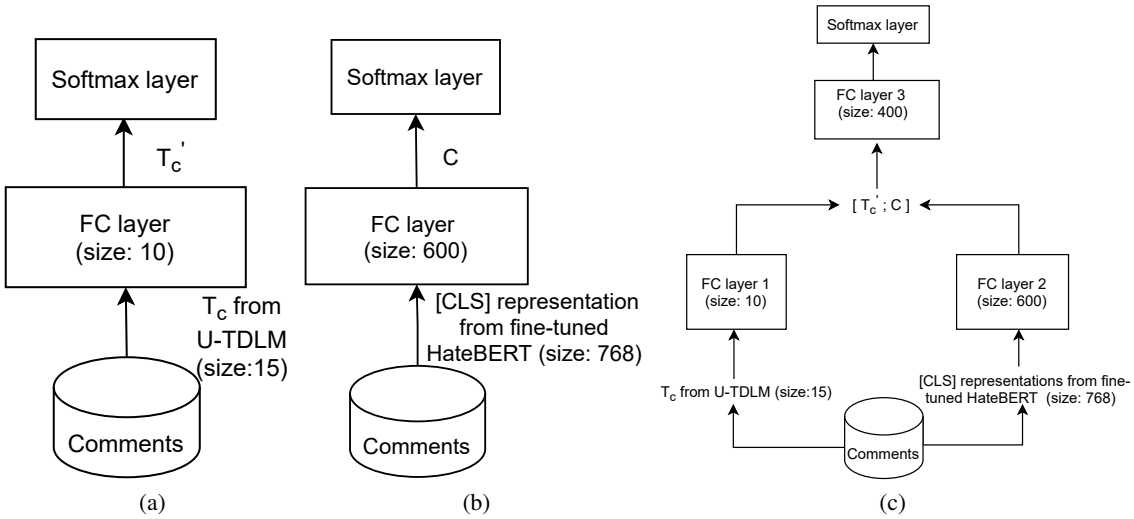

Figure 2: Architecture of classifier for individual models: (a) U-TDLM, (b) HateBERT, and the combined model (c) HateBERT + U-TDLM; FC: Fully Connected.