Hate Speech Detection Using Transformer Ensembles on the HASOC Dataset

Pedro Alonso[1], Rajkumar Saini[1], and György Kovács[1,2(✉)]

[1] Embedded Internet Systems Lab, Luleå University of Technology, Luleå, Sweden
{pedro.alonso,rajkumar.saini,gyorgy.kovacs}@ltu.se
[2] MTA-SZTE Research Group on Artificial Intelligence, Szeged, Hungary

Abstract. With the ubiquity and anonymity of the Internet, the spread of hate speech has been a growing concern for many years now. The language used for the purpose of dehumanizing, defaming or threatening individuals and marginalized groups not only threatens the mental health of its targets, as well as their democratic access to the Internet, but also the fabric of our society. Because of this, much effort has been devoted to manual moderation. The amount of data generated each day, particularly on social media platforms such as Facebook and twitter, however makes this a Sisyphean task. This has led to an increased demand for automatic methods of hate speech detection.

Here, to contribute towards solving the task of hate speech detection, we worked with a simple ensemble of transformer models on a twitter-based hate speech benchmark. Using this method, we attained a weighted F_1-score of 0.8426, which we managed to further improve by leveraging more training data, achieving a weighted F_1-score of 0.8504. Thus markedly outperforming the best performing system in the literature.

Keywords: Natural Language Processing · Hate speech detection · Transformers · RoBERTa · Ensemble

1 Introduction

There are many questions still surrounding the issue of hate speech. For one, it is strongly debated whether hate speech should be prosecuted, or whether free speech protections should extend to it [2,11,16,24]. Another question debated is regarding the best counter-measure to apply, and whether it should be suppression (through legal measures, or banning/blocklists), or whether it should be methods that tackle the root of the problem, namely counter-speech and education [4]. These arguments, however are fruitless without the ability of detecting hate speech en masse. And while manual detection may seem as a simple (albeit hardly scalable) solution, the burden of manual moderation [15], as well as the sheer amount of data generated online justify the need for an automatic solution of detecting hateful and offensive content.

A. Karpov and R. Potapova (Eds.): SPECOM 2020, LNAI 12335, pp. 13–21, 2020.
https://doi.org/10.1007/978-3-030-60276-5_2

1.1 Related Work

The ubiquity of fast, reliable Internet access that enabled the sharing of information and opinions at an unprecedented rate paired with the opportunity for anonymity [50] has been responsible for the increase in the spread of offensive and hateful content in recent years. For this reason, the detection of hate speech has been examined by many researchers [21,48]. These efforts date back to the late nineties and Microsoft research, with the proposal of a rule-based system named Smokey [36]. This has been followed by many similar proposals for rule-based [29], template-based [27], or keyword-based systems [14,21].

In the meantime, many researchers have tackled this task using classical machine learning methods. After applying the Bag-of-Words (BoW) method for feature extraction, Kwok and Wang [19] used a Naïve Bayes classifier for the detection of racism against black people on Twitter. Grevy et al. [13] used Support Vector Machines (SVMs) on BoW features for the classification of racist texts. However, since the BoW approach was shown to lead to high false positive rates-[6], others used more sophisticated feature extraction methods to obtain input for the classical machine learning methods (such as SVM, Naïve Bayes and Logistic Regression [5,6,39,40]) deployed for the detection of hateful content.

One milestone in hate speech detection was deep learning gaining traction in Natural Language Processing (NLP) after its success in pattern recognition and computer vision [44], propelling the field forward [31]. The introduction of embeddings [26] had an important role in this process. For one, by providing useful features to the same classical machine learning algorithms for hate speech detection [25,45], leading to significantly better results than those attained with the BoW approach (both in terms of memory-complexity, and classification scores [9]). Other deep learning approaches were also popular for the task, including Recurrent Neural Networks [1,7,10,38], Convolutional Neural Networks [1,12,32,51], and methods that combined the two [17,41,49].

The introduction of transformers was another milestone, in particular the high improvement in text classification performance by BERT [37]. What is more, transformer models have proved highly successful in hate speech detection competitions (with most of the top ten teams using a transformer in a recent challenge [46]). Ensembles of transformers also proved to be successful in hate speech detection [28,30]. So much so, that such a solution has attained the best performance (i.e. on average the best performance over several sub-tasks) recently in a challenge with more than fifty participants [35]. For this reason, here, we also decided to use an ensemble of transformer models.

1.2 Contribution

Here, we apply a 5-fold ensemble training method using the RoBERTA model, which enables us to attain state-of-the-art performance on the HASOC benchmark. Moreover, by proposing additional fine-tuning, we significantly increase the performance of models trained on different folds.

Table 1. Example tweets from the HASOC dataset [22].

Tweet	Label
@piersmorgan Dont watch it then. #dickhead	NOT
This is everything. #fucktrump https://t.co/e2C48U3pss	HOF
I stand with him ...He always made us proud 🙌#DhoniKeepsTheGlove	NOT
@jemelehill He's a cut up #murderer	HOF
#fucktrump #impeachtrump 😂😂😂😂😂😂😂 @ Houston, Texas https://t.co/8QGgbWtOAf	NOT

2 Experimental Materials

In this section we discuss the benchmark task to be tackled. Due to the relevance of the problem, many competitions have been dedicated finding a solution [3, 18, 23, 42, 46]. For this study, we consider the challenge provided by HASOC [22] data (in particular, the English language data) to be tackled using our methods. Here, 6712 tweets (the training and test set containing 5852 and 860 tweets, respectively) were annotated into the following categories:

- NOT: tweets not considered to contain hateful or offensive content
- HOF: tweets considered to be hateful, offensive, or profane

Where in the training set approximately 39% of all instances (2261 tweets) were classified in the second category, while for the test set this ratio was 28% (240 tweets). Some example tweets from the training set are listed in Table 1. As can be seen, in some cases it is not quite clear why one tweet was labelled as hateful, and others were not (#fucktrump). Other examples with debatable labeling are listed in the system description papers from the original HASOC competition [33]. This can be an explanation why Ross et al. suggested to consider hate speech detection as a regression task, as opposed to a classification task [34].

2.1 OffensEval

In line with the above suggestion, the training data published by Zampieri et al. for the 2020 OffensEval competition [47] were not class labels but scores, as can be seen in Table 2 below. Here, for time efficiency reasons we used the first 1 million of the 9,089,140 tweets available in the training set.

Table 2. Example tweets from the OffensEval corpus [47].

Tweet	Score
@USER And cut a commercial for his campaign	0.2387
@USER Trump is a fucking idiot his dementia is getting worse	0.8759
Golden rubbers in these denim pockets	0.3393
Hot girl summer is the shit!!! #period	0.8993

3 Experimental Methods

In this section we discuss the processing pipeline we used for the classification of HASOC tweets. This includes the text preprocessing steps taken, the short description of the machine learning models used, as well as the method of training we applied on said machine learning models.

3.1 Text Preprocessing

Text from social media sites, and in particular Twitter often lacks proper grammar/punctuation, and contains many paralinguistic elements (e.g. URLs, emoticons, emojis, hashtags). To alleviate potential problems caused by this variability, tweets were put through a preprocessing step before being fed to our model. First, consecutive white space characters were replaced by one instance, while extra white space characters were added between words and punctuation marks. Then @-mentions and links were replaced by the character series *@USER* and *URL* respectively. Furthermore, as our initial analysis did not find a significant correlation between emojis and hatefulness scores on the more than nine million tweets of the OffensEval dataset, all emojis and emoticons were removed. Hashtag characters (but not the hashtags themselves) were also removed in the process. Lastly, tweets were tokenized into words.

3.2 RoBERTa

For tweet classification and regression, in our study we used a variant of BERT [8], namely RoBERTa [20], from the SimpleTransformers library [43] (for a detailed description of transformers in general, as well as BERT and RoBERTa in particular, please see the sources cited in this paper). We did so encouraged by the text classification performance of BERT [37], as well as our preliminary experiments with RoBERTa. When training said model, we followed [43] in selecting values for our meta-parameters, with the exception of the learning rate, for which we used $1e-5$ as our value.

3.3 5-Fold Ensemble Training

In our experiments we used the following training scheme. First, we split the HASOC train set into five equal parts, each consisting of 1170 tweets (Dev_1, Dev_2, Dev_3, Dev_4, Dev_5). This partitioning was carried out in a way that the ratio of the two different classes was the same in each subset than it was in the whole set. Then, for each development set, we created a training set using the remaining tweets from the original training set ($Train_1$, $Train_2$, $Train_3$, $Train_4$, $Train_5$). After creating the five folds in this manner, we used each fold to train separate RoBERTa models. The final model was then defined as the ensemble of the five individual models, where the predictions of the ensemble model was created by averaging the predicted scores of the individual models.

Table 3. F_1-scores of different models on the test set of the HASOC benchmark. For each model, and each F_1-score, the best result is emphasized in bold.

Model	Fold	$HASOC_{only}$	$HASOC_{OffensEval}$
Macro F_1-score	1st	0.7586	0.7964
	2nd	0.7681	0.7855
	3rd	0.7688	0.7943
	4th	0.7914	0.7929
	5th	0.7758	**0.8029**
	Ensemble	**0.7945**	0.7976
Weighted F_1-score	1st	0.8125	0.8507
	2nd	0.8165	0.8402
	3rd	0.8244	0.8474
	4th	0.8415	0.8485
	5th	0.8327	**0.8537**
	Ensemble	**0.8426**	0.8504

Here, we examined two different ensembles. For one ($HASOC_{only}$), we used a pretrained RoBERTa model [43], and fine-tuned it on different folds, creating five different versions of the model. Then, we averaged the predictions of these models for the final classification. To examine how further fine-tuning would affect the results, we first fine-tuned the RoBERTa model using one million tweets from the OffensEval competition for training, and ten thousand tweets for validation. Then, we further fine-tuned the resulting model in the manner described above. However, since the first fine-tuning resulted in a regression model, when further fine-tuning these models, we first replaced NOT and HOF labels with a value of 0 and 1 respectively. In this case, the predicted scores before classification were first rescaled to the 0–1 interval by min-max normalization ($HASOC_{OffensEval}$).

4 Results and Discussion

We evaluated the resulting ensembles on the HASOC test set. Results of these experiments are listed in Table 3. As can be seen in Table 3, as a result of further fine-tuning using OffensEval data, the performance of individual models significantly increased (applying the paired t-test, we find that the difference is significant, at $p < 0.05$ for both the macro, and the weighted F_1-score). The difference in the performance of the ensembles, however, is much less marked. A possible explanation for this could be that the five models in the $HASOC_{OffensEval}$ case may be more similar to each other (given that here the original model went through more fine-tuning with the same data). Furthermore, while in the case of the $HASOC_{only}$ model the ensemble attains better F_1-scores using both metrics, this is not the case with the $HASOC_{OffensEval}$ model, where the best performance is attained using the model trained on the 5th fold. Regardless of this,

however, both ensemble methods outperform the winner of the HASOC competition [38] in both F_1-score measures (the winning team achieving a score of 0.7882 and 0.8395 in terms of macro F_1-score, and weighted F_1-score respectively).

5 Conclusions and Future Work

In this study we have described a simple ensemble of transformers for the task of hate speech detection. Results on the HASOC challenge showed that this ensemble is capable of attaining state-of-the-art performance. Moreover, we have managed to improve the results attained by additional pre-training using in-domain data. In the future we plan to modify our pre-training approach so that models responsible for different folds are first pre-trained using different portions of the OffensEval data. Furthermore, we intend to extend our experiments to other hate speech datasets and challenges, as well as other transformer models. Lastly, we also intend to examine the explainability of resulting models.

Acknowledgements. Part of this work has been funded by the Vinnova project "Language models for Swedish authorities" (ref. number: 2019-02996).

References

1. Badjatiya, P., Gupta, S., Gupta, M., Varma, V.: Deep learning for hate speech detection in tweets. In: Proceedings of the 26th International Conference on World Wide Web Companion, WWW '17 Companion, pp. 759–760 (2017)
2. Barendt, E.: What is the harm of hate speech? Ethic theory, moral prac., vol. 22 (2019). https://doi.org/10.1007/s10677-019-10002-0
3. Basile, V., et al.: 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, pp. 54–63 (2019). https://doi.org/10.18653/v1/S19-2007
4. Brown, A.: What is so special about online (as compared to offline) hate speech? Ethnicities **18**(3), 297–326 (2018). https://doi.org/10.1177/1468796817709846
5. Burnap, P., Williams, M.L.: Cyber hate speech on twitter: an application of machine classification and statistical modeling for policy and decision making. Policy Internet **7**(2), 223–242 (2015). https://doi.org/10.1002/poi3.85
6. Davidson, T., Warmsley, D., Macy, M., Weber, I.: Automated hate speech detection and the problem of offensive language. In: Proceedings of the 11th International AAAI Conference on Web and Social Media, ICWSM 2017, pp. 512–515 (2017)
7. Del Vigna, F., Cimino, A., Dell'Orletta, F., Petrocchi, M., Tesconi, M.: Hate me, hate me not: Hate speech detection on Facebook. In: ITASEC, January 2017
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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), pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota, June 2019. https://doi.org/10.18653/v1/N19-1423

9. Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., Bhamidipati, N.: Hate speech detection with comment embeddings. In: Proceedings of the 24th International Conference on World Wide Web, WWW 2015 Companion, pp. 29–30. Association for Computing Machinery, New York (2015). https://doi.org/10.1145/2740908.2742760
10. Do, H.T.T., Huynh, H.D., Nguyen, K.V., Nguyen, N.L.T., Nguyen, A.G.T.: Hate speech detection on vietnamese social media text using the bidirectional-LSTM model (2019), arXiv:1911.03648
11. Dworkin, R.: A new map of censorship. Index Censorship **35**(1), 130–133 (2006). https://doi.org/10.1080/03064220500532412
12. Gambäck, B., Sikdar, U.K.: Using convolutional neural networks to classify hate-speech. In: Proceedings of the First Workshop on Abusive Language Online, pp. 85–90. Association for Computational Linguistics, Vancouver, BC, Canada, August 2017. https://doi.org/10.18653/v1/W17-3013. https://www.aclweb.org/anthology/W17-3013
13. Greevy, E., Smeaton, A.F.: Classifying racist texts using a support vector machine. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2004, pp. 468–469. Association for Computing Machinery, New York (2004). https://doi.org/10.1145/1008992.1009074
14. Gröndahl, T., Pajola, L., Juuti, M., Conti, M., Asokan, N.: All you need is "love": evading hate speech detection. In: Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security, AISec 2018, pp. 2–12. Association for Computing Machinery, New York (2018). https://doi.org/10.1145/3270101.3270103
15. Hern, A.: Revealed: catastrophic effects of working as a Facebook moderator. The Guardian (2019). https://www.theguardian.com/technology/2019/sep/17/revealed-catastrophic-effects-working-facebook-moderator. Accessed 26 Apr 2020
16. Heyman, S.: Hate speech, public discourse, and the first amendment. In: Hare, I., Weinstein, J. (eds.) Extreme Speech and Democracy. Oxford Scholarship Online (2009). https://doi.org/10.1093/acprof:oso/9780199548781.003.0010
17. Huynh, T.V., Nguyen, V.D., Nguyen, K.V., Nguyen, N.L.T., Nguyen, A.G.T.: Hate speech detection on Vietnamese social media text using the bi-gru-lstm-cnn model. arXiv:1911.03644 (2019)
18. Immpermium: detecting insults in social commentary. https://kaggle.com/c/detecting-insults-in-social-commentary. Accessed 27 April 2020
19. Kwok, I., Wang, Y.: Locate the hate: detecting tweets against blacks. In: Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2013, pp. 1621–1622. AAAI Press (2013)
20. Liu, Y., et al.: Roberta: A robustly optimized bert pretraining approach (2019)
21. MacAvaney, S., Yao, H.R., Yang, E., Russell, K., Goharian, N., Frieder, O.: Hate speech detection: Challenges and solutions. PLOS ONE **14**(8), 1–16 (2019). https://doi.org/10.1371/journal.pone.0221152
22. Mandl, T., Modha, S., Mandlia, C., Patel, D., Patel, A., Dave, M.: HASOC - Hate Speech and Offensive Content identification in indo-European languages. https://hasoc2019.github.io. Accessed 20 Sep 2019
23. Mandl, T., Modha, S., Patel, D., Dave, M., Mandlia, C., Patel, A.: Overview of the HASOC track at FIRE 2019: Hate Speech and Offensive Content Identification in Indo-European Languages). In: Proceedings of the 11th Annual Meeting of the Forum for Information Retrieval Evaluation, December 2019

24. Matsuda, M.J.: Public response to racist spech: considering the victim's story. In: Matsuda, M.J., Lawrence III, C.R. (ed.) Words that Wound: Critical Race Theory, Assaultive Speech, and the First Amendment, pp. 17–52. Routledge, New York (1993)
25. Mehdad, Y., Tetreault, J.: Do characters abuse more than words? In: Proceedings of the SIGDIAL2016 Conference, pp. 299–303, January 2016. https://doi.org/10.18653/v1/W16-3638
26. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Proceedings of the NIPS, pp. 3111–3119 (2013)
27. Mondal, M., Silva, L.A., Benevenuto, F.: A measurement study of hate speech in social media. In: Proceedings of the 28th ACM Conference on Hypertext and Social Media, HT 2017, pp. 85–94. Association for Computing Machinery, New York (2017). https://doi.org/10.1145/3078714.3078723
28. Nina-Alcocer, V.: Vito at HASOC 2019: Detecting hate speech and offensive content through ensembles. In: Mehta, P., Rosso, P., Majumder, P., Mitra, M. (eds.) Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation, Kolkata, India, 12–15 December, 2019. CEUR Workshop Proceedings, vol. 2517, pp. 214–220. CEUR-WS.org (2019). http://ceur-ws.org/Vol-2517/T3-5.pdf
29. Njagi, D., Zuping, Z., Hanyurwimfura, D., Long, J.: A lexicon-based approach for hate speech detection. Int. J. Multimed. Ubiquitous Eng. **10**, 215–230 (2015). https://doi.org/10.14257/ijmue.2015.10.4.21
30. Nourbakhsh, A., Vermeer, F., Wiltvank, G., van der Goot, R.: sthruggle at SemEval-2019 task 5: an ensemble approach to hate speech detection. In: Proceedings of the 13th International Workshop on Semantic Evaluation, pp. 484–488. Association for Computational Linguistics, Minneapolis, Minnesota, June 2019. https://doi.org/10.18653/v1/S19-2086
31. Otter, D.W., Medina, J.R., Kalita, J.K.: A survey of the usages of deep learning for natural language processing. IEEE Trans. Neural Networks Learn. Syst., 1–21 (2020)
32. Park, J., Fung, P.: One-step and two-step classification for abusive language detection on Twitter. In: ALW1: 1st Workshop on Abusive Language Online, June 2017
33. Alonso, P., Rajkumar Saini, G.K.: The North at HASOC 2019 hate speech detection in social media data. In: Proceedings of the 11th Anual Meeting of the Forum for Information Retrieval Evaluation, December 2019
34. Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., Wojatzki, M.: Measuring the reliability of hate speech annotations: the case of the European refugee crisis. In: Beißwenger, M., Wojatzki, M., Zesch, T. (eds.) Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication, pp. 6–9, September 2016. https://doi.org/10.17185/duepublico/42132
35. Seganti, A., Sobol, H., Orlova, I., Kim, H., Staniszewski, J., Krumholc, T., Koziel, K.: Nlpr@srpol at semeval-2019 task 6 and task 5: linguistically enhanced deep learning offensive sentence classifier. In: SemEval@NAACL-HLT (2019)
36. Spertus, E.: Smokey: Automatic recognition of hostile messages. In: Proceedings of the 14th National Conference on Artificial Intelligence and 9th Innovative Applications of Artificial Intelligence Conference (AAAI-97/IAAI-97), pp. 1058–1065. AAAI Press, Menlo Park (1997. http://www.ai.mit.edu/people/ellens/smokey.ps
37. Sun, C., Qiu, X., Xu, Y., Huang, X.: How to fine-tune BERT for text classification? In: Sun, M., Huang, X., Ji, H., Liu, Z., Liu, Y. (eds.) CCL 2019. LNCS (LNAI), vol. 11856, pp. 194–206. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32381-3_16

38. Wang, B., Ding, Y., Liu, S., Zhou, X.: Ynu_wb at HASOC 2019: Ordered neurons LSTM with attention for identifying hate speech and offensive language. In: Mehta, P., Rosso, P., Majumder, P., Mitra, M. (eds.) Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation, Kolkata, India, 12–15 December, 2019, pp. 191–198 (2019). http://ceur-ws.org/Vol-2517/T3-2.pdf
39. Warner, W., Hirschberg, J.: Detecting hate speech on the world wide web. In: Proceedings of the Second Workshop on Language in Social Media, pp. 19–26. Association for Computational Linguistics, Montréal, Canada, June 2012. https://www.aclweb.org/anthology/W12-2103
40. Waseem, Z., Hovy, D.: Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In: Proceedings of the NAACL Student Research Workshop, pp. 88–93. Association for Computational Linguistics, San Diego, California, June 2016. https://doi.org/10.18653/v1/N16-2013. https://www.aclweb.org/anthology/N16-2013
41. Wei, X., Lin, H., Yang, L., Yu, Y.: A convolution-LSTM-based deep neural network for cross-domain MOOC forum post classification. Information **8**, 92 (2017). https://doi.org/10.3390/info8030092
42. Wiegand, M., Siegel, M., Ruppenhofer, J.: Overview of the germeval 2018 shared task on the identification of offensive language. In: Proceedings of the GermEval 2018 Workshop, pp. 1–11 (2018)
43. Wolf, T., et al.: Huggingface's transformers: State-of-the-art natural language processing. arXiv:1910.03771 (2019)
44. Young, T., Hazarika, D., Poria, S., Cambria, E.: Recent trends in deep learning based natural language processing (2017), arXiv:1708.02709 Comment: Added BERT, ELMo, Transformer
45. Yuan, S., Wu, X., Xiang, Y.: A two phase deep learning model for identifying discrimination from tweets. In: Pitoura, E., et al. (eds.) Proceedings of the 19th International Conference on Extending Database Technology, EDBT 2016, Bordeaux, France, March 15–16, 2016, Bordeaux, France, 15–16 March, 2016, pp. 696–697. OpenProceedings.org (2016). https://doi.org/10.5441/002/edbt.2016.92
46. 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: Proceedings of the 13th International Workshop on Semantic Evaluation, pp. 75–86 (2019)
47. Zampieri, M., et al.: SemEval-2020 Task 12: multilingual offensive language identification in social media (OffensEval 2020). In: Proceedings of SemEval (2020)
48. Zhang, Z., Luo, L.: Hate speech detection: a solved problem? the challenging case of long tail on twitter. Semantic Web Accepted, October 2018. https://doi.org/10.3233/SW-180338
49. Zhang, Z., Robinson, D., Tepper, J.: Detecting hate speech on Twitter using a convolution-GRU based deep neural network. In: Gangemi, A., Navigli, R., Vidal, M.-E., Hitzler, P., Troncy, R., Hollink, L., Tordai, A., Alam, M. (eds.) ESWC 2018. LNCS, vol. 10843, pp. 745–760. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-93417-4_48
50. Zimbardo, P.G.: The human choice: individuation, reason, and order versus deindividuation, impulse, and chaos. Nebr. Symp. Motiv. **17**, 237–307 (1969)
51. Zimmerman, S., Kruschwitz, U., Fox, C.: Improving hate speech detection with deep learning ensembles. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan, May 2018. https://www.aclweb.org/anthology/L18-1404