NLP-CIC at HASOC 2020: Multilingual Offensive Language Detection using All-in-one Model

Segun Taofeek Aroyehun[a], Alexander Gelbukh[a]

[a] CIC, Instituto Politécnico Nacional Mexico City, Mexico

Abstract

We describe our deep learning model submitted to the HASOC 2020 shared task on detection of offensive language in social media in three Indo-European languages: English, German, and Hindi. We fine-tune a pre-trained multilingual encoder on the combination of data provided for the competition. Our submission received a competitive macro- average F1 score of 0.4980 on the English Subtask A as well as comparatively strong performance on the German data.

Keywords

offensive content identification, deep learning, text classification, multilingual

1. Introduction

The impact of offensive content on web users range from subtle uneasiness to graver psychological and emotional distress which if go unchecked can result in violent actions to/from affected individuals. In order to make the web a safe place for all, platforms such as Twitter and Facebook pay close attention to content moderation. To aid in the arduous task of removal of objectionable content, it becomes necessary to build efficient and effective systems capable of identifying and classifying such content for automatic or human-assisted content moderation. A standard approach is to automatically flag such content for removal or review by human moderators. There are several studies on the English language due to availability of datasets and distributional representation with which models can be developed. While there is sizeable progress in the English language, the same cannot be said of other languages. With shared task series such as HASOC providing data in other languages, this provides avenue for further research in other languages. With the availability of datasets in several languages, it becomes expensive to design a robust system for each language. An alternative strategy will be to train a single model on languages for which annotated data is available.

We base our approach on the recent progress in the development of multilingual language models. In particular, the observation by Conneau et al. [1] that a multilingual model can reach the performance of several language-specific models, at least after pre-training. Can we say the same for fine-tuning on a downstream task? We examine whether jointly fine-tuning a multilingual model on a multilingual dataset is feasible for the task of offensive content

FIRE '20, Forum for Information Retrieval Evaluation, December 16–20, 2020, Hyderabad, India.

aroyehun.segun@gmail.com (S. T. Aroyehun)
www.gelbukh.com (A. Gelbukh)

identification and classification. We reckon that this approach will be more energy efficient and less computationally expensive.

Specifically, we examine the possibility of using a multilingual pre-trained language model (BERT) to train a single model for the three languages with datasets provided for the HASOC 2020 shared task [2] via transfer learning.

2. Related Work

The automatic detection of offensive content has been studied with several approaches. Traditionally, feature engineering in conjunction with classical machine learning models such as Support Vector Machines, Logistic Regression, and Naive Bayes have shown competitive performance [3]. In recent times, neural networks have outperformed the traditional approaches using architectures such as GRU, LSTM, and CNN in combination with word embeddings [4]. The introduction of contextual word embeddings based on pre-trained language models [5] and the transformers [6] architecture has led to state-of-the-art results on several NLP tasks including offensive content identification [7]. Typically, existing approaches rely on pre-trained language models which are adapted to the task at hand [8]. There has been significant progress on the detection of offensive content in English language and the same cannot be said of other languages. Recently, shared tasks such as TRAC 2018 [9], HASOC 2019 [10], TRAC 2020 [11], and Offenseval [12] have introduced datasets in languages other than English. However, the evaluation at those venues still proceeds on a monolingual level. It would be interesting to see evaluation settings that assess models on their multilingual and/or cross-lingual capabilities as exemplified in the work of Pamungkas and Patti [13] and Ranasinghe and Zampieri [14].

3. Methodology

Task. Given a text (tweets in this case) predict (1) For subtask A, whether it is offensive or not. and (2) For subtask B, categorize the text into one of the following classes: none, offensive, hate, and profane.

Data. The HASOC 2020 dataset includes annotated text data in English, German, and Hindi. The data has hierarchical labels at two levels. Level one has binary labels (Offensive VS. Not offensive) and level two has four mutually exclusive labels. Table 1 shows the details of the training set.

Approach. We train a single model using the combination of labeled datasets provided for each language by the organizers. So, we have a single model per task which covers the three languages covered by the competition.

We use as validation set the test set for the 2019 edition of HASOC (the gold labels). We observe that the application of language-agnostic pre-processing: URL removal, normalization of repeated characters, emoji to text conversion, and removal of punctuation marks resulted in performance drop on the validation set. Hence, we did not apply pre-processing to our submissions. It appears that a contextual model such as BERT is able to utilize the information

Table 1
Details of the dataset for subtasks A and B for each language. Total is the number of labeled examples per language. OFF – Offensive, and PRFN – profane

	A		B				Total
	OFF	NOT	OFF	HATE	PRFN	NONE	
EN	1856	1852	321	158	1377	1852	3708
DE	673	1700	140	146	387	1700	2373
HI	847	2116	465	234	148	2116	2963

Table 2
F1 score on the test set. Numbers in parenthesis represent the performance difference between our submission and the best model on the leaderboard.

Task	EN	DE	HI
A	0.4980 (−0.0172)	0.5177 (−0.0058)	0.5005 (−0.0332)
B	0.2537 (−0.0115)	0.2687 (−0.0256)	0.2374 (−0.0971)

that would have been removed. We experiment with both multilingual BERT [5] and XLM-R [1]. We find that the performance of the XLM-R model was unstable and inferior across runs. This is likely due to the size of the model and thus requires careful fine-tuning. Based on this observation, we select multilingual BERT for our submissions. We use as representation for the text the embedding of the [CLS] token, which is of dimension 768 and feed this to a single layer perceptron with softmax activation. This gives a probability distribution over the number of classes to be predicted (2 for subtask A and 4 for subtask B). Our training set up use the following hyperparameter settings: learning rate of $3e-5$, batch size of 128, Adam as optimizer, and a maximum of 5 epochs. We select the model with the best performance on the validation set for prediction on the unseen test set. For subtask B, we continue fine-tuning on the best model from subtask A using the same hyperparameter settings above. Our implementation uses the Flair library [15].

4. Results

Table 2 shows the scores received by our submissions per task on each language on the private test set maintained by the organizers. On the English subtask A, we recorded a macro-average F1 score of 0.4980, and 0.2537 for subtask B. These scores are within 2 F1 points of the highest ranked submission for English. On the German data, the performance gap on subtask A is the lowest, 0.0058, aproximately 1% F1 points. On the Hindi dataset, we observe the largest gap in performance on subtask B, about 10% F1 points. Also, the second largest gap is recoreded on the Hindi subtask A, around 3% F1 points less than the best submission on the leaderboard. We suspect that there is a negative transfer from either English or German to Hindi. This observation deserves a thorough investigation in the future. Overall, the scores on the subtask B are consistently lower than subtask A across languages. This indicates the difficulty of the task.

5. Conclusion

We examined the feasibility of using a single multilingual model to detect and classify offensive language in three Indo-European languages. We run fine-tuning experiments using multilingual BERT. We record a competitive macro-average F1 score of 0.4980 on the English subtask A. We observe that the performance gaps between our submission for tasks on Hindi and the best model on the leaderboard is the highest. In the future, we will like to experiment further with a mixture of more language-specific datasets and identify the limits of using a mixed-language dataset for fine-tuning multilingual encoders on the task of offensive content identification.

Acknowledgments

We thank the competition organizers for their support. The authors thank CONACYT for the computer resources provided through the INAOE Supercomputing Laboratory's Deep Learning Platform for Language Technologies.

References

[1] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 8440–8451. URL: https://www.aclweb.org/anthology/2020.acl-main.747. doi:10.18653/v1/2020.acl-main.747.

[2] T. Mandl, S. Modha, G. K. Shahi, A. K. Jaiswal, D. Nandini, D. Patel, P. Majumder, J. Schäfer, Overview of the HASOC track at FIRE 2020: Hate Speech and Offensive Content Identification in Indo-European Languages), in: Working Notes of FIRE 2020 - Forum for Information Retrieval Evaluation, CEUR, 2020.

[3] S. Malmasi, M. Zampieri, Detecting hate speech in social media, in: Proceedings of Recent Advances in Natural Language Processing (RANLP), Varna, Bulgaria, 2017, pp. 467–472.

[4] S. T. Aroyehun, A. Gelbukh, Aggression detection in social media: Using deep neural networks, data augmentation, and pseudo labeling, in: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), 2018, pp. 90–97.

[5] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://www.aclweb.org/anthology/N19-1423. doi:10.18653/v1/N19-1423.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural In-

formation Processing Systems 30, Curran Associates, Inc., 2017, pp. 5998–6008. URL: http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

[7] J. Risch, R. Krestel, Bagging bert models for robust aggression identification, in: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, 2020, pp. 55–61.

[8] C. Sun, X. Qiu, Y. Xu, X. Huang, How to fine-tune bert for text classification?, in: China National Conference on Chinese Computational Linguistics, Springer, 2019, pp. 194–206.

[9] R. Kumar, A. K. Ojha, S. Malmasi, M. Zampieri, Benchmarking aggression identification in social media, in: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), 2018, pp. 1–11.

[10] T. Mandl, S. Modha, P. Majumder, D. Patel, M. Dave, C. Mandlia, A. Patel, Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages, in: Proceedings of the 11th Forum for Information Retrieval Evaluation, 2019, pp. 14–17.

[11] R. Kumar, A. K. Ojha, S. Malmasi, M. Zampieri, Evaluating aggression identification in social media, in: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, European Language Resources Association (ELRA), Marseille, France, 2020, pp. 1–5. URL: https://www.aclweb.org/anthology/2020.trac-1.1.

[12] M. Zampieri, P. Nakov, S. Rosenthal, P. Atanasova, G. Karadzhov, H. Mubarak, L. Derczynski, Z. Pitenis, Ç. Çöltekin, Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020), arXiv preprint arXiv:2006.07235 (2020).

[13] E. W. Pamungkas, V. Patti, 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, Association for Computational Linguistics, Florence, Italy, 2019, pp. 363–370. URL: https://www.aclweb.org/anthology/P19-2051. doi:10.18653/v1/P19-2051.

[14] T. Ranasinghe, M. Zampieri, Multilingual offensive language identification with cross-lingual embeddings, arXiv preprint arXiv:2010.05324 (2020).

[15] A. Akbik, D. Blythe, R. Vollgraf, Contextual string embeddings for sequence labeling, in: COLING 2018, 27th International Conference on Computational Linguistics, 2018, pp. 1638–1649.