YNU_OXZ at HASOC 2020: Multilingual Hate Speech and Offensive Content Identification based on XLM-RoBERTa

Xiaozhi Ou, Hongling Li*

School of Information Science and Engineering, Yunnan University, Kunming, 650500, Yunnan, P.R. China

Abstract

This article introduces the submission of subtask A in three languages (English, German, Hindi) that we participated in the HASOC 2020 shared task, which aims to target hate speech and offensive language in multiple languages for identification. To solve this task, we propose a system based on the multilingual model XLM-RoBERTa and Ordered Neurons LSTM (ON-LSTM). When evaluated on the official test set, our system show the effectiveness of our method on subtask A of three languages. The Macro average F1 score of English subtask A is 0.5006, the Macro average F1 score of German subtask A is 0.5177, the Macro average F1 score of Hindi subtask A is 0.5200. This final leaderboard result is calculated with approximately 15% of the private test data.

Keywords

multilingual, hate speech, offensive language, identification, English, German, Hindi

1. Introduction

The existence and impact of hate speech and offensive language on social media platforms are becoming a major concern in modern society. Given the enormous amount of content created every day, automated methods are needed to detect and handle such content. So far, most studies have focused on solving the problem for the English language, while the problem is multilingual. Hate Speech and Offensive Content Identification in Indo-European Languages (HASOC[1]) was inspired by two evaluation forums OffensEval[2] and GermanEval 2018[3] and try to leverage synergies of both the forum. HASOC provides a forum and a data challenge for multilingual research on the identification of problematic content [1]. This year, the organizers once again provided two subtasks for English, German and Hindi, altogether more than 10,000 annotated tweets from Twitter.

- Subtask A: Identifying Hate, offensive and profane content.
- Subtask B: Discrimination between Hate, profane and offensive posts.

FIRE '20, Forum for Information Retrieval Evaluation, December 16–20, 2020, Hyderabad, India
xiaozhiou88@gmail.com (X. Ou); honglingli66@126.com (H. Li*)
0000-0001-6043-2348 (X. Ou)

[1]https://hasocfire.github.io/hasoc/2020/
[2]https://competitions.codalab.org/competitions/20011
[3]https://projects.fzai.h-da.de/iggsa/

We participate in subtask A for three languages: this task focuses on Hate speech and Offensive language identification offered for English, German, and Hindi. Subtask A is a coarse-grained binary classification in which participating systems are required to classify tweets into two classes, namely: Hate and Offensive (HOF) and Non-Hate and offensive (NOT). **HOF** refers to posts that contain Hate, offensive, and profane content, and **NOT** refers to posts that do not contain such content.

In order to effectively solve this task and achieve better results in low-resource languages, we focus on the approach to an effective strategy of combining the multilingual model XLM-RoBERTa with Ordered Neurons LSTM (ON-LSTM). Firstly, we are base on the pre-trained multi-language model XLM-RoBERTa, it not only inherits the training method of XLM also uses the ideas of RoBERTa for reference. Then, we use the ON-LSTM, It obtains hierarchical structure information by sorting neurons, which can express richer semantic information. In this paper, we present the related work, the details of our approach, our results and conclusion.

2. Related Work

From an NLP perspective, the topics of hate speech and offensive language and all its possible facets and related phenomena (such as profane/abusive language) and its identification have attracted great attention. His is shown by the proliferation, especially in the last few years, of contributions on this matter ([2], [3], [4], [5] to name a few), corpora and lexica (e.g. [6], [7], [8]), dedicated workshops, and shared tasks within national (GermEval[4], EVALITA[5], IberLEF[6]) and international (SemEval[7]) evaluation campaigns (see in particular [9]). In the literature on the offensive and hate language detection, many different subtasks have been considered, ranging from general offensive language detection to more refined tasks, such as hate speech detection [10] and cyberbullying detection [11]. Chen et al. applied the concept of NLP to develop sentence lexical and syntactic features for offensive language detection [12]. Huang et al. integrated the textual features with social network features, which significantly improved cyberbullying detection [13].

Unfortunately, other supervised methods to hate speech classification have conflated hate speech with offensive language, which makes it difficult to determine to what extent they actually recognize hate speech [14]. Neural language models show promise in the task but existing work has used training data that has a similarly broad definition of hate speech [15]. Non-linguistic features like gender or ethnicity of the author can help improve hate speech classification but this information is often unavailable or unreliable on social media [16]. Recently, Zampieri et al. provided an offensive language identification dataset, which aims to identify the type and target of offensive posts in social media [17]. This year they expanded the dataset to a multilingual version, thus promoting multilingual research in this field [9]. Pre-trained language models, such as BERT [18] and ELMo [19] have achieved great performance on a variety of tasks. Many recent papers have used basic methods of fine-tuning such pre-trained models in certain fields

[4]https://projects.fzai.h-da.de/iggsa/germeval/

[5]http://www.evalita.it/2020/tasks

[6]http://hitz.eus/sepln2019/?q=node/21

[7]http://alt.qcri.org/semeval2020/index.php?id=tasks

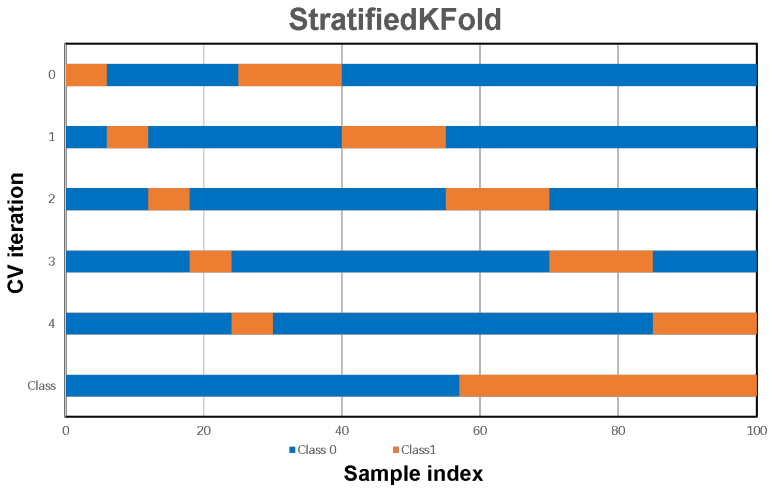

Figure 1: 5-fold stratified sampling to the training set (The color represents the label class.)

[20] or downstream tasks [21].

3. Approach

In this chapter, we will description of the data, system, and experimental parameters we use.

3.1. Data description

The HASOC organizers provided complete training datasets for the three languages, with about 3,708 in English, about 2,373 in German, and about 2,963 in Hindi. In our experiment, we use the multi-task learning method to combine the training datasets of the three languages and train the model to share the representations between related tasks to solve the subtask A of the three languages, but it did not achieve our expected effect. Finally, we use the HASOC 2019 dataset to merge the datasets of the three languages, respectively. Such as the final English training set is the combination of the HASOC 2019 English training dataset and the HASOC 2020 English training dataset (The same goes for other languages).

In our experiment, we use stratified sampling technology (StratifiedKFold) to randomly split all combined training datasets. As shown in Figure 1, we using StratifiedKFold cross-validation instead of ordinary k-fold cross-validation to evaluate a classifier. The reason is that StratifiedKFold can utilize stratified sampling to divide, which can ensure that the proportion of each category in the generated training set and validation set is consistent with the original training set so that the generated data distribution disorder will not occur. In the experiment, we use 5-fold stratified sampling.

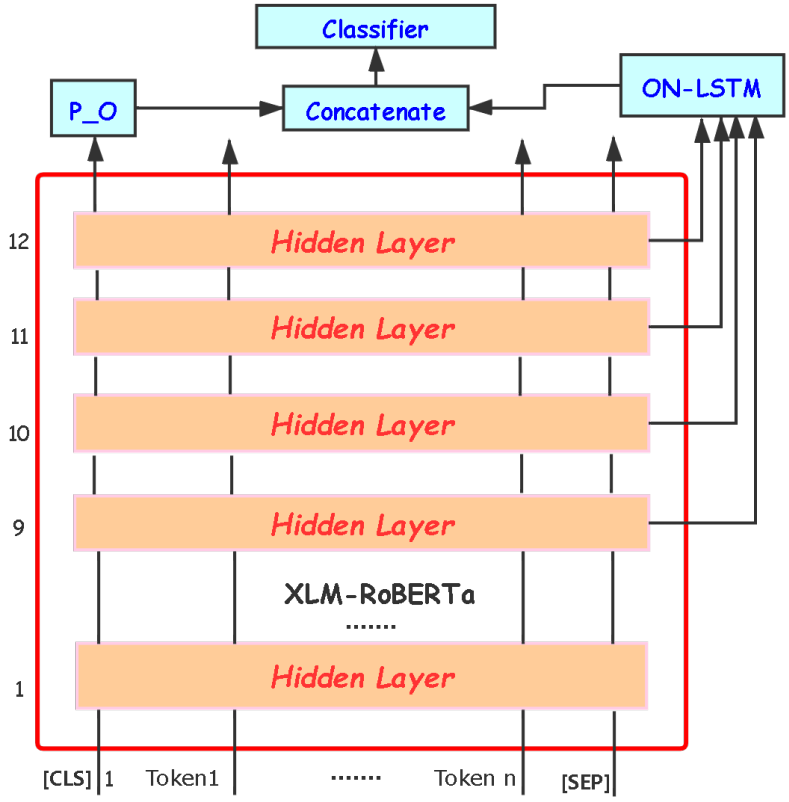

Figure 2: System overall architecture diagram

3.2. System description

The pre-training of XLM-RoBERTa is based on 100 languages, using more than 2TB of pre-processed CommonCrawl dataset to train cross-language representations in a self-supervised manner. XLM-RoBERTa [22] shows that the use of large-scale multi-language pre-training models can significantly improve the performance of cross-language migration tasks. In order to solve the subtask A of three languages at the same time, we propose a system architecture based on the multi-language model XLM-RoBERTa as shown in Figure 2. Firstly, we get pooler output (P_O), P_O is the pooler output of XLM-Roberta. It is obtained by its last layer hidden state of the first token of the sequence (CLS token) further processed by a linear layer and a tanh activation function. Then extract the hidden state of the last four layers of XLM-RoBERTa and input them into Ordered Neurons LSTM (ON-LSTM) [23]. Finally, we concatenate the P_O and output of ON-LSTM together input into the Classifier for the final classification.

Table 1
Test results of the three run systems for three language subtask A

Subtask A	System	Macro average F1
English	Run_1	0.81
	Run_2	0.88
	Run_3 (Final system)	**0.92**
German	Run_1	0.66
	Run_2	0.72
	Run_3 (Final system)	**0.77**
Hindi	Run_1	0.64
	Run_2	0.68
	Run_3 (Final system)	**0.73**

Table 2
Results on the official private test set (This final leaderboard is calculated with approximately 15% of the private test data)

Task	Our Score (Macro average F1)	Best Score	Rank
English Subtask A	0.5006	0.5152	12
German Subtask A	0.5177	0.5235	4
Hindi Subtask A	0.5200	0.5337	6

3.3. Experimental parameters

In our experiment, we did not clean the data. We use XLM-RoBERTa-base[8] pre-trained model. The batch size is set to 32 and the max sequence length is set to 150. We extract the last four hidden layer state of XLM-RoBERTa by setting the output hidden States is true. For the ON-LSTM, we set the hidden units to 512 and num levels to 1. We use binary cross-entropy, adam optimizer and learning rate to 5e-5. The model is trained in 10 epochs.

4. Result

This section will show the results and analysis of all three languages that we participated in subtask A on the test set and the official 15% private test set, the subtask A for the three languages is evaluated by following the macro average F1 of scikit-learn[9]. The test set results of subtask A in all three languages are shown in Table 1. For each language subtask A, we have performed three runs. Among them, Run_1 means we take the P_O of XLM-RoBERTa as the final output. Run_2 means that we extract the last four hidden layers of XLM-RoBERTa and input them into the convolution neural network (CNN) and K-max pooling. Run_3 means that we extract the last four hidden layers of XLM-RoBERTa and input them into ON-LSTM, which is also the system we finally submitted. Table 2 reports the official results of the best run

[8]https://huggingface.co/xlm-roberta-base
[9]https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html

of subtask A in the three languages we participated in, we can also see the best scores of the official leaderboard and our ranking.

5. Conclusion

In the experiment, we test the effects of using the external dataset and not using the external dataset. Our conclusion is that using data from the same language for training and test is a necessary condition for good performance. In addition, adding data from different languages can improve results.

References

[1] 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.

[2] A. Bohra, D. Vijay, V. Singh, S. S. Akhtar, M. Shrivastava, A dataset of hindi-english code-mixed social media text for hate speech detection, in: Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, 2018.

[3] D. Jurgens, E. Chandrasekharan, L. Hemphill, A just and comprehensive strategy for using nlp to address online abuse (2019).

[4] T. Caselli, V. Basile, J. Mitrović, I. Kartoziya, M. Granitzer, I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language, in: Proceedings of The 12th Language Resources and Evaluation Conference, 2020, pp. 6193–6202.

[5] P. Fortuna, J. R. da Silva, L. Wanner, S. Nunes, et al., A hierarchically-labeled portuguese hate speech dataset, in: Proceedings of the Third Workshop on Abusive Language Online, 2019, pp. 94–104.

[6] E. Bassignana, V. Basile, V. Patti, Hurtlex: A multilingual lexicon of words to hurt, in: 5th Italian Conference on Computational Linguistics, CLiC-it 2018, volume 2253, CEUR-WS, 2018, pp. 1–6.

[7] S. Rosenthal, P. Atanasova, G. Karadzhov, M. Zampieri, P. Nakov, A large-scale semi-supervised dataset for offensive language identification, arXiv preprint arXiv:2004.14454 (2020).

[8] M. Sanguinetti, F. Poletto, C. Bosco, V. Patti, M. Stranisci, An italian twitter corpus of hate speech against immigrants, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), 2018.

[9] 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).

[10] T. Davidson, D. Warmsley, M. Macy, I. Weber, Automated hate speech detection and the problem of offensive language, arXiv preprint arXiv:1703.04009 (2017).

[11] C. Van Hee, B. Verhoeven, E. Lefever, G. De Pauw, W. Daelemans, V. Hoste, Guidelines for the fine-grained analysis of cyberbullying, version 1.0, LT3 Technical Report Series (2015).
[12] Y. Chen, Y. Zhou, S. Zhu, H. Xu, Detecting offensive language in social media to protect adolescent online safety, in: 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing, IEEE, 2012, pp. 71–80.
[13] Q. Huang, V. K. Singh, P. K. Atrey, Cyber bullying detection using social and textual analysis, in: Proceedings of the 3rd International Workshop on Socially-Aware Multimedia, 2014, pp. 3–6.
[14] P. Burnap, M. L. Williams, Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making, Policy & Internet 7 (2015) 223–242.
[15] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, N. Bhamidipati, Hate speech detection with comment embeddings, in: Proceedings of the 24th international conference on world wide web, 2015, pp. 29–30.
[16] Z. Waseem, D. Hovy, Hateful symbols or hateful people? predictive features for hate speech detection on twitter, in: Proceedings of the NAACL student research workshop, 2016, pp. 88–93.
[17] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, R. Kumar, Predicting the type and target of offensive posts in social media, arXiv preprint arXiv:1902.09666 (2019).
[18] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
[19] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, arXiv preprint arXiv:1802.05365 (2018).
[20] N. Azzouza, K. Akli-Astouati, R. Ibrahim, Twitterbert: Framework for twitter sentiment analysis based on pre-trained language model representations, in: International Conference of Reliable Information and Communication Technology, Springer, 2019, pp. 428–437.
[21] Z. Liu, G. I. Winata, Z. Lin, P. Xu, P. Fung, Attention-informed mixed-language training for zero-shot cross-lingual task-oriented dialogue systems, arXiv preprint arXiv:1911.09273 (2019).
[22] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020.
[23] Y. Shen, S. Tan, A. Sordoni, A. Courville, Ordered neurons: Integrating tree structures into recurrent neural networks (2018).