IRIT-PREVISION AT HASOC 2020 : Fine-tuning BERT for Hate Speech and Offensive Content Identification

Josiane Mothe, Pratik Parikh, Faneva Ramiandrisoa

IRIT-PREVISION AT HASOC 2020 : Fine-tuning BERT for Hate Speech and Offensive Content Identification

Josiane Mothe[a,b], Pratik Parikh[a] and Faneva Ramiandrisoa[a]

[a]*IRIT, Université de Toulouse, France*
[b]*ESPE, UT2J*

Abstract

This paper describes the participation of the IRIT-PREVISION team at HASOC (Hate Speech and Offensive Content Identification in Indo-European Languages) 2020 shared task. Our approach is based on fine-tuning a pre-trained transformer based language model BERT (Bidirectional Encoder Representation from Transformer) [1]. We participated to the English sub-task A. We obtained a macro average F1 of 0.497 (self-computed).

Keywords

Information system, Hate Speech Detection, Social Media, BERT, Deep Learning

1. Introduction

Lu *et. al.* [2] report an increase in the number of users who are subject to or have witnessed hate speech, offensive languages, etc. online. This phenomenon is worrying and solutions should be found to detect it or even limit it. Identifying hate speech and aggression in social media is essential to protect users from such attacks, but manual detection can be very expensive. This finding has led many researchers to focus on building automatic systems for detecting hate speech, aggression, offensive languages, etc. on social networks. In the same vein, several shared tasks have also been organized, including "Abusive Language Online" (ALW) [3], "Trolling, Aggression and Cyberbullying" (TRAC) [4] and the "Semantic Evaluation" task (SemEval) on identifying offensive language in social media (OffensEval) [5].

HASOC [6, 7] is also a shared task where the goal is to detect hateful content in post published on social media. This paper describes our participation to the second edition of HASOC (2020). This second edition consists of two sub-tasks on three languages, namely English, German, and Hindi. We participated in one sub-task on English language. For this, we create a model by fine-tuning BERT (Bidirectional Encoder Representation from Transformer) model [1].

Josiane.Mothe@irit.fr (J. Mothe); parikh.er@gmail.com (P. Parikh); faneva.ramiandrisoa@irit.fr (F. Ramiandrisoa)
http://www.irit.fr/~Josiane.Mothe (J. Mothe)
0000-0001-9273-2193 (J. Mothe); 0000-0001-9386-3531 (F. Ramiandrisoa)

The rest of this paper is organized as follows: Section 2 presents related work in the area of hate speech detection; Section 3 describes the HASOC 2020 shared task as well as the data sets provided; Section 4 describes the methodology we proposed to answer the HASOC 2020 challenge as well as the preprocessing we developed; Section 5 presents the results we obtained; finally, Section 6 concludes this paper and presents some future work.

2. Related Work

Online detection of abusive language, hate speech, aggression, offensive content, etc. is an important topic. Indeed, the number of users who are subject to or have witnessed hate speech or offensive languages online is increasing [2] and can concern citizen, public persons, or organizations. Detecting this type of content on social media platforms such as Facebook and Twitter is an important challenge that authorities care about as illustrated for example by the recent French government decisions [1] to force social media to act.

In the past few years, a lot of research have been conducted to detect hate speech [7], offensive language [5], and aggression [8]. Supervised learning approaches are predominantly used to solve the problem ranging from deep learning based methods [9] such as convolutional neural networks (CNN) to traditional machine learning based methods [10] such as support vector machine (SVM).

During the first edition of HASOC (2019) [7], deep learning based methods were widely used and achieved the top performances. However, it can also be observed that some of the non-deep learning systems performances were quite close to the top ones. The team from IIIT-Hyderabad even obtained the best result on one sub-task in Hindi language with traditional machine learning [10].

3. Task Description and Data

HASOC is a shared task, where the goal is to detect hateful content in textual posts published on social media, namely Twitter. This shared task is a multilingual track joining English, German, and Hindi, and consists of two main sub-tasks:

1. Sub-task A (Identifying Hate, offensive and profane) : it focuses on Hate speech and Offensive language identification offered for English, German, and Hindi. This sub-task is a coarse-grained binary classification in which the objective is to classify tweets into two classes : HOF (Hate and Offensive) and NOT (Non- Hate and Offensive).
2. Sub-task B (Discrimination between Hate, profane and offensive) : it is a fine-grained classification for English, German, and Hindi. Here, the posts labeled as HOF from the sub-task A are further classified into three categories : HATE (Hate speech), OFFN (Offenive) and PRFN (Profane).

No context or meta-data like the users' network are provided [7].

[1] https://www.theguardian.com/world/2019/jul/09/france-online-hate-speech-law-social-media, accessed on October 15^{th}, 2020.

Table 1
Distribution of training data in HASOC 2020 shared task for English.

Sub-task		Train
A	HOF	1,856
	NOT	1,852
	Total	3,708
B	HATE	158
	OFFN	321
	PRFN	1,377

Table 1 presents the statistics of the training data set used during HASOC 2020 for English. More details on the test set can be found in [6]. During the shared task, the training set was released while test set was not. The training set were used to build models then these models were submitted to the organizers and tested on a non-shared data set from the organizers. In this edition, we participated to the English sub-task A. The associated training set is composed of 3,708 tweets. Looking at the distribution of the two classes for sub-task A, the training data set is balanced with half of the cases HOF, half non-HOF.

4. Methodology

Before building our model, we pre-processed the data set: we converted all the texts into lowercase, substituted all "URL" by "http", also substituted emoticon into their text equivalents by using the online emoji project on github[2]. Finally, we removed all non UTF-8 words.

During the HASOC 2020 shared task, we submitted only one model. Our model is based on BERT or Bidirectional Encoder Representations from Transformers [1]. More precisely, we fine-tuned a pre-trained BERT model called $BERT\_Large\_Uncased$ which contains 24 layers of size 1024, 16 self-attention heads and 340M parameters. Fine-tuning a pre-trained BERT model is less expensive than training a BERT model from scratch and fine-tuning is also very interesting and works well on small data sets which is the case of the HASOC data set. The pre-trained model we used was trained at Google on the corpus data composed of English Wikipedia (2,500M words) and BooksCorpus (800M words) [11]. This pre-trained model is publicly available on github[3].

During the fine-tuning, we used a batch size of 16, the Adam optimizer with a learning rate of 5e-5 and a number of epochs of 10 as parameters. Each sequence is truncated to max allowed sequence length of 40 characters. We used the library pytorch-pretrained-bert[4] for implementation. Training was carried out on a Nvidia Geforce GTX 1080TI GPU and took about 14 minutes in total.

[2]https://github.com/carpedm20/emoji, accessed on February, 04^{th} 2020.
[3]https://github.com/google-research/bert, accessed on February, 04^{th} 2020.
[4]https://github.com/shehzaadzd/pytorch-pretrained-BERT, accessed on February, 04^{th} 2020.

Table 2
The results of our model and the best three teams on English sub-task A of HASOC 2020. Bold value is the best performance.

Team	Macro average F1
IRIT Prevision	0.4969
IIIT_DWD	**0.5152**
CONCORDIA_CIT_TEAM	0.5078
AI_ML_NIT_Patna	0.5078

5. Results

This section reports the result our team obtained on the English sub-task A when participating to HASOC 2020. Table 2 reports the result as well as the three best teams' results. We obtained a macro average F1 of 0.4969 which places our team in the twenty-firstth place (out of thirty-five participants). The difference between our result and the best one is 0.0183. In general, the differences between participants' results are very small, where the difference between the best team and the thirty third team is 0.0581[5]. Only two teams out of thirty five got much lower results. The data set used for evaluation is non-shared by the organizers.

6. Conclusion and Future Work

In this paper, we presented our participation at the second edition of HASOC shared task in English language for sub-task A: identifying hate, offensive and profane content. We used a model that relies on BERT to tackle the problem. Our model achieved a macro average F1 of 0.4969 which ranked our team on the twenty-first place over thirty-five participants. However the difference with the best team is very small (0.0183).

For the short term future work, we plan to apply our model to the other HASOC 2020 sub-tasks and on the other two languages. For long term future work, we want to integrate or combine a keyphrase representation in our model [12]. We already created some models based on keyphrase lexicons but we did not submit it due to lack of time. We would also like to analyse the cross-domain transfer of some models that we developed to detect weak signals in social media [13].

Ethical issues. Working on online data raises ethical issues which are out of the scope of this paper. Training has been made on a publicly available data set while test was ran by the organizers them-selves.

Acknowledgments

This work is partially supported by the PREVISION project, which has received funding from

[5] https://competitions.codalab.org/competitions/26027#learn_the_details-results, accessed on October 14th, 2020.

the European Union's Horizon 2020 research and innovation programme under GA No 833115 (https://cordis.europa.eu/project/id/833115). The paper reflects the authors' view and the Commission is not responsible for any use that may be made of the information it contains.

References

[1] J. Devlin, M. 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), 2019, pp. 4171–4186. URL: https://doi.org/10.18653/v1/n19-1423. doi:10.18653/v1/n19-1423.

[2] Z. Lu, J. Nie, RALIGRAPH at HASOC 2019: VGCN-BERT: augmenting BERT with graph embedding for offensive language detection, in: Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation, Kolkata, India, December 12-15, 2019, 2019, pp. 221–228. URL: http://ceur-ws.org/Vol-2517/T3-6.pdf.

[3] S. T. Roberts, J. Tetreault, V. Prabhakaran, Z. Waseem (Eds.), Proceedings of the Third Workshop on Abusive Language Online, Association for Computational Linguistics, Florence, Italy, 2019. URL: https://www.aclweb.org/anthology/W19-3500.

[4] 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, TRAC@LREC 2020, Marseille, France, May 2020, 2020, pp. 1–5. URL: https://www.aclweb.org/anthology/2020.trac-1.1/.

[5] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, R. Kumar, Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval), in: Proceedings of the 13th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2019, Minneapolis, MN, USA, June 6-7, 2019, 2019, pp. 75–86. URL: https://doi.org/10.18653/v1/s19-2010. doi:10.18653/v1/s19-2010.

[6] 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.

[7] S. Modha, T. Mandl, P. Majumder, D. Patel, Overview of the HASOC track at FIRE 2019: Hate speech and offensive content identification in indo-european languages, in: Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation, Kolkata, India, December 12-15, 2019, 2019, pp. 167–190. URL: http://ceur-ws.org/Vol-2517/T3-1.pdf.

[8] F. Ramiandrisoa, J. Mothe, Aggression identification in social media: a transfer learning based approach, in: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, TRAC@LREC 2020, Marseille, France, May 2020, 2020, pp. 26–31. URL: https://www.aclweb.org/anthology/2020.trac-1.5/.

[9] S. Agrawal, A. Awekar, Deep learning for detecting cyberbullying across multiple social media platforms, in: Advances in Information Retrieval - 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceed-

ings, 2018, pp. 141–153. URL: https://doi.org/10.1007/978-3-319-76941-7_11. doi:10.1007/978-3-319-76941-7_11.

[10] V. Mujadia, P. Mishra, D. M. Sharma, Iiit-hyderabad at HASOC 2019: Hate speech detection, in: Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation, Kolkata, India, December 12-15, 2019, 2019, pp. 271–278. URL: http://ceur-ws.org/Vol-2517/T3-12.pdf.

[11] Y. Zhu, R. Kiros, R. S. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, S. Fidler, Aligning books and movies: Towards story-like visual explanations by watching movies and reading books, in: 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015, 2015, pp. 19–27. URL: https://doi.org/10.1109/ICCV.2015.11. doi:10.1109/ICCV.2015.11.

[12] J. Mothe, F. Ramiandrisoa, M. Rasolomanana, Automatic keyphrase extraction using graph-based methods, in: Proceedings of the 33rd Annual ACM Symposium on Applied Computing, SAC 2018, Pau, France, April 09-13, 2018, 2018, pp. 728–730. URL: https://doi.org/10.1145/3167132.3167392. doi:10.1145/3167132.3167392.

[13] F. Ramiandrisoa, J. Mothe, F. Benamara, V. Moriceau, IRIT at e-risk 2018, in: Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018, 2018. URL: http://ceur-ws.org/Vol-2125/paper_102.pdf.