Multi-task Learning for Hate Speech and Aggression Detection

Faneva RAMIANDRISOA[1,†]

[1]*IRIT, Univ. de Toulouse, Toulouse, France*

Abstract

In recent studies, multi-task learning (MTL) has achieved remarkable success in natural language processing applications. In this paper, we present the application of MTL with transformer-based models (RoBERTa [1]) on two different but related, shared tasks: Hate Speech and Offensive Content Identification (HASOC) [2, 3], and Trolling, Aggression and Cyberbullying (TRAC) [4, 5]. The MTL model performs slightly better than RoBERTa on two datasets, slightly worse on one dataset and they have the same perfomance on another one. The MTL model performs better than the participants' systems only on the HASOC 2019 dataset.

Keywords

Information Retrieval, Social Media Analysis, Text Mining, Aggression Detection, Hate Speech Detection, Transfer Learning, Multi-task Learning

1. Introduction

Multi-task learning (MTL) is attracting increasing interest, especially in the era of deep learning [6]. It is widely used in natural language processing [6, 7], computer vision, recommendation [8], tasks, etc.

MTL has been used in different ways: considering a single task, but on multi corpora [7], multi-tasks on a single corpus [9], and finally multi-tasks on multi corpora [6]. Our work is related to the latter.

We investigate the use of MTL with transformer-based models (RoBERTa [1]) on two different, but related, shared tasks: Hate Speech and Offensive Content Identification (HASOC) [2, 3], and Trolling, Aggression and Cyberbullying (TRAC) [4, 5]. We hypothesize that the performance of models on individual tasks can be improved via joint learning. Our empirical experiments show that the MTL results are only slightly better than RoBERTa results on two datasets out of four, slightly worse on one dataset, and the same on one dataset. Furthermore, The MTL model performs better than the participants' systems only on HASOC 2019 dataset.

The rest of the paper is organized as follows. First, Section 2 presents related work. Then, we describe the multi-task learning model we used in Section 3, followed by the dataset description in Section 4 and results presentation in Section 5. We conclude with future work in Section 6.

CIRCLE (Joint Conference of the Information Retrieval Communities in Europe), July 04–07, 2022, Samatan, Gers, France

faneva.ramiandrisoa@irit.fr (F. RAMIANDRISOA)
0000-0001-9386-3531 (F. RAMIANDRISOA)

2. Related Work

2.1. Hate Speech and Aggression Detection

Detecting online abuse, hate speech, aggression, offensive content, etc are important issues. In recent years, much research has been conducted to detect hate speech [10, 11], offensive language [12], and aggression [13, 14]. Several European projects and workshops are addressing this challenge and a number of evaluation forums dealing with offensive content, hate speech and aggression have been organised recently. In order to solve these challenges, participants heavily rely on deep learning techniques which achieve the best results. Transfer learning using transformer such as BERT [15], RoBERTa [1], etc have been used a lot recently and often achieved the best results. This is the case in GermEval [16], SemEval-2019 Task 6 [12], TRAC [4, 5] and HASOC [2, 3].

2.2. Multi-task learning

Multi-task learning (MTL) aims to improve the learning of a model for a given task by using the knowledge contained in tasks where all or a subset of tasks are related [17].

A MTL framework is similar to that of transfer learning, but with significant differences. In MTL, the goal is to improve performance on all tasks (there is no distinction between different tasks) while in transfer learning, the target task is more important than the source tasks. Indeed, the objective of transfer learning is to improve the performance of a target task using source tasks [17]. In other word, MTL treats all the tasks equally while transfer learning gives more attention to the target task.

MTL and transfer learning can also be combined, i.e. considering the target tasks in transfer learning as MTL tasks for joint learning [6].

3. MTL model

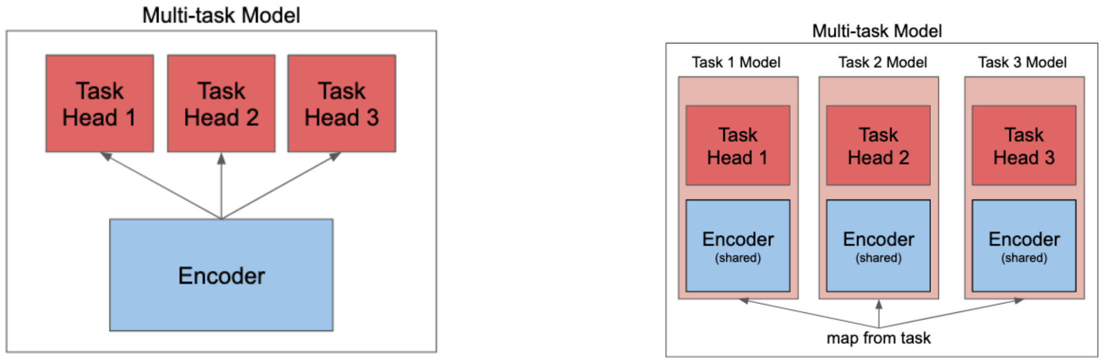
In this paper, we study the effectiveness of an MTL with transformer-based models (RoBERTa [1]) for Hate Speech and Aggression Detection.

In the BERT [15] era, a multi-task model works by having one shared encoder transformer, and several task head, one for each task (see in Figure 1a). Note that a multi-task model is trained on different tasks in parallel and not sequentially as in the original BERT.

The idea of the MTL model we used is to create separate models for each task, but these models will share the encoder weights (see Figure 1b). This allows us to have different forms of input for each task; this is not the case with a single encoder transformer. This model is also easy to implement. This will achieve the same objective as joint encoder trained for multiple tasks, while maintaining the independent implementation for each model.

For the multi-task learning, we used the architecture presented by Jason Phang on github[1] as well as the same hyperparameters.

[1] https://github.com/zphang/zphang.github.io/blob/master/files/notebooks/Multi_task_Training_with_Transformers_NLP.ipynb



(a) Model with one encoder.

(b) Model with shared encoder weights.

Figure 1: Two multi-task model architectures: (a) MTL with one encoder and several task heads, (b) MTL with shared encoder weights (model we use)[2].

4. Datasets

For our experiments, we use four datasets in total, two for each of the two shared tasks HASOC (Hate Speech and Offensive Content Identification) [2, 3] and TRAC (Trolling, Aggression and Cyberbullying) [4, 5].

4.1. HASOC

The aim of the HASOC shared task is to automatically detect hateful content in text messages posted on social media, especially Twitter. It is a multilingual track combining English, German and Hindi, and consists of two main sub-tasks:

1. Sub-task A: it focuses on the identification of hate speech and offensive language for English, German and Hindi. The goal is to classify texts into two classes: HOF (hateful and offensive) and NOT (not hateful and offensive).
2. Sub-task B: it is a fine-grained classification for English, German and Hindi. Here, messages labelled as HOF in subtask A are further classified into three categories: HATE (hate speech), OFFN (offensive) and PRFN (profane).

In this work, we focused only on the English datasets and on subtask A. We did not consider subtask B because both sub-tasks (A and B) use the same texts and only the labels change. As our model is a multi-task learning one, we did not want to feed the model twice with the same input. We hypothesize that applying a multi-task learning on the both sub-tasks will lead to an over-fitting model. This hypothesis will be studied in future work.

We used two English datasets from HASOC 2019 and HASOC 2020. Table 1 presents the statistics of these training and test datasets.

[2] Source : https://github.com/zphang/zphang.github.io/blob/master/files/notebooks/Multi_task_Training_with_Transformers_NLP.ipynb

Table 1
Distribution of datasets in HASOC 2019 and 2020 shared task for English.

HASOC		Train	Test
2019	HOF	2,261	288
	NOT	3,591	865
	Total	5,852	1,153
2020	HOF	1,856	807
	NOT	1,852	785
	Total	3,708	1,592

4.2. TRAC

The aim of TRAC is to identify aggression, trolling, cyberbullying and other related phenomena in both speech and text from social media. The shared task goal is to distinguish between three levels of text aggressiveness: overtly aggressive (OAG), covertly aggressive (CAG) and non-aggressive (NAG). Overtly aggressive means that there is a direct expression of aggression with specific words while covert aggression expresses aggression in a subtle way such as indirect attack or by polite expressions.

Here we focused on English language (the dataset also has an Hindi part). We used two English datasets from TRAC 2018 and TRAC 2020. The 2020 edition of TRAC has another challenge, but we did not consider it in this work for the same reason as for HASOC subtask B.

TRAC 2018 comprises two test sets. We consider here the one that contains texts from the same social media as the training data texts. We will study the generalisation of our model in future work.

Table 2 presents the statistics of the TRAC 2018 and 2020 English training and test datasets.

Table 2
Distribution of texts in TRAC 2018 and 2020 datasets - English.

TRAC		Train	Validation	Test
2018	CAG	4,240	1,057	142
	OAG	2,708	711	144
	NAG	5,051	1,233	630
	Total	11,999	3,001	916
2020	CAG	453	117	224
	OAG	435	113	286
	NAG	3,375	836	690
	Total	3,375	1,066	1,200

5. Results

This section reports the results of our MTL model on the English datasets of HASOC (2019 and 2020), and TRAC (2018 and 2019) shared tasks.

As an evaluation measure, we use the Macro-F1 and Weighted-F1 which are the official measures of the HASOC and TRAC shared tasks.

To train our MTL model, we used the training parts of the four datasets presented in Section 4 all together. As a baseline, we consider a RoBERTa, that is to say a single model, that we fine-tuned individually on each dataset. Table 3 reports the results on each test dataset.

Table 3
MTL outperforms the baseline model or has similar results on each shared task test dataset. Best results are in bold for each data sets. The difference between MTL and baseline results are not statistically significant (t-student with p=0.05)

Task	Edition	Model	Macro-F1	Weighted-F1
HASOC	2019	MTL	**0.80**	**0.85**
		baseline	0.77	0.82
	2020	MTL	0.91	0.91
		baseline	0.91	0.91
TRAC	2018	MTL	**0.55**	0.63
		baseline	0.54	0.63
	2020	MTL	0.64	0.73
		baseline	**0.68**	**0.75**

The MTL model outperforms or achieves the baseline results, except on TRAC 2020 dataset. Our hypothesis for this result is the dataset distribution. Indeed, the TRAC 2020 dataset is more unbalanced than the others. A deeply analysis has to be conducted for in-depth understanding.

We also compare the MTL results to HASOC and TRAC shared task participants' results, except HASOC 2020 because we do not know how the organizers computed the participants results. We observed that MTL outperforms the best participant's results in HASOC 2019 where best Macro-F1 is 0.79 and weighted-F1 0.84. Concerning TRAC, according to weighted-F1 measure, the MTL achieved the fifth best score compared to 2020 edition's results (best: 0.80) and the third best score compared to 2018 edition's results (best: 0.64). Table 4 reports these results.

Table 4
MTL outperforms best participant's result on HASOC 2019 test dataset and achieves third and fifth best score respectively on TRAC 2018 and 2020. Best results are in bold for each data sets.

Task	Edition	Model	Macro-F1	Weighted-F1
HASOC	2019	MTL	**0.80**	**0.85**
		YNU_wb [18]	0.79	0.84
TRAC	2018	saroyehun [19]	-	**0.64**
		EBSILIAUNAM [20]	-	0.63
		MTL	0.55	0.63
	2020	Julian [21]	-	**0.80**
		sdhanshu [22]	-	0.76
		Ms8qQxMbnjJMgYcw [23]	-	0.76
		zhixuan	-	0.74
		MTL	0.64	0.73

The results show the efficiency of using MTL for Hate Speech and Aggression detection given the fact that we only used a simple approach (architecture) of MTL with transformer-based models. These results lead us to believe that if we improve our MTL architecture or approach, the better results we will have.

6. Conclusion

In this paper, we presented the use of MTL for Hate Speech and Aggression detection. For this, we trained an MTL model on two different but related shared tasks: Hate Speech and Offensive Content Identification (HASOC) [2, 3], and Trolling, Aggression and Cyberbullying (TRAC) [4, 5]. Our experiments show the efficiency of MTL on both shared tasks, where the MTL model outperforms or achieves the simple fine-tuned model (consider as baseline) results. The results are also promising when compared to shared tasks participants' results where MTL outperforms the best participant's results in HASOC 2019, achieves the third best score in TRAC 2018 and the fifth best score in TRAC 2020.

There are some limitations to this work. Our results on MTL training show that MTL is not always effective as we have seen with HASOC 2020. This may be due to the high imbalance of the dataset. It is however promising since we used a simple MTL architecture with transformer-based models. As future work, we would like to investigate the following:

- Improving the model architecture by using a more complex one that would be able to lean more.
- Testing other transformer based model such as XLNet [24] which should handle dependencies between tasks well.
- In-depth analysis of the datasets and the impact of their characteristics on the model effectiveness.

References

[1] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, CoRR abs/1907.11692 (2019). URL: http://arxiv.org/abs/1907.11692. arXiv:1907.11692.

[2] T. Mandl, S. Modha, A. K. M, B. R. Chakravarthi, Overview of the HASOC track at FIRE 2020: Hate speech and offensive language identification in tamil, malayalam, hindi, english and german, in: P. Majumder, M. Mitra, S. Gangopadhyay, P. Mehta (Eds.), FIRE 2020: Forum for Information Retrieval Evaluation, Hyderabad, India, December 16-20, 2020, ACM, 2020, pp. 29–32. URL: https://doi.org/10.1145/3441501.3441517. doi:10.1145/3441501.3441517.

[3] T. Mandl, S. Modha, P. Majumder, D. Patel, M. Dave, C. Mandalia, A. Patel, Overview of the HASOC track at FIRE 2019: Hate speech and offensive content identification in indo-european languages, in: P. Majumder, M. Mitra, S. Gangopadhyay, P. Mehta (Eds.), FIRE '19: Forum for Information Retrieval Evaluation, Kolkata, India, December, 2019, ACM, 2019, pp. 14–17. URL: https://doi.org/10.1145/3368567.3368584. doi:10.1145/3368567.3368584.

[4] R. Kumar, A. K. Ojha, S. Malmasi, M. Zampieri, Evaluating aggression identification in social media, in: R. Kumar, A. K. Ojha, B. Lahiri, M. Zampieri, S. Malmasi, V. Murdock,

D. Kadar (Eds.), Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, TRAC@LREC 2020, Marseille, France, May 2020, European Language Resources Association (ELRA), 2020, pp. 1–5. URL: https://aclanthology.org/2020.trac-1.1/.

[5] R. Kumar, A. K. Ojha, S. Malmasi, M. Zampieri, Benchmarking aggression identification in social media, in: R. Kumar, A. K. Ojha, M. Zampieri, S. Malmasi (Eds.), Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying, TRAC@COLING 2018, Santa Fe, New Mexico, USA, August 25, 2018, Association for Computational Linguistics, 2018, pp. 1–11. URL: https://aclanthology.org/W18-4401/.

[6] Y. Peng, Q. Chen, Z. Lu, An empirical study of multi-task learning on BERT for biomedical text mining, in: D. Demner-Fushman, K. B. Cohen, S. Ananiadou, J. Tsujii (Eds.), Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing, BioNLP 2020, Online, July 9, 2020, Association for Computational Linguistics, 2020, pp. 205–214. URL: https://doi.org/10.18653/v1/2020.bionlp-1.22. doi:10.18653/v1/2020.bionlp-1.22.

[7] X. Wang, Y. Zhang, X. Ren, Y. Zhang, M. Zitnik, J. Shang, C. P. Langlotz, J. Han, Cross-type biomedical named entity recognition with deep multi-task learning, Bioinform. 35 (2019) 1745–1752. URL: https://doi.org/10.1093/bioinformatics/bty869. doi:10.1093/bioinformatics/bty869.

[8] S. Liu, E. Johns, A. J. Davison, End-to-end multi-task learning with attention, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, Computer Vision Foundation / IEEE, 2019, pp. 1871–1880. URL: http://openaccess.thecvf.com/content_CVPR_2019/html/Liu_End-To-End_Multi-Task_Learning_With_Attention_CVPR_2019_paper.html. doi:10.1109/CVPR.2019.00197.

[9] K. Xue, Y. Zhou, Z. Ma, T. Ruan, H. Zhang, P. He, Fine-tuning BERT for joint entity and relation extraction in chinese medical text, in: I. Yoo, J. Bi, X. Hu (Eds.), 2019 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2019, San Diego, CA, USA, November 18-21, 2019, IEEE, 2019, pp. 892–897. URL: https://doi.org/10.1109/BIBM47256.2019.8983370. doi:10.1109/BIBM47256.2019.8983370.

[10] 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.

[11] J. Mothe, P. Parikh, F. Ramiandrisoa, IRIT-PREVISION AT HASOC 2020: Fine-tuning BERT for hate speech and offensive content identification, in: P. Mehta, T. Mandl, P. Majumder, M. Mitra (Eds.), Working Notes of FIRE 2020 - Forum for Information Retrieval Evaluation, Hyderabad, India, December 16-20, 2020, volume 2826 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020, pp. 260–265. URL: http://ceur-ws.org/Vol-2826/T2-21.pdf.

[12] 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.

[13] 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/.

[14] F. Ramiandrisoa, J. Mothe, IRIT at TRAC 2020, in: R. Kumar, A. K. Ojha, B. Lahiri, M. Zampieri, S. Malmasi, V. Murdock, D. Kadar (Eds.), Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, TRAC@LREC 2020, Marseille, France, May 2020, European Language Resources Association (ELRA), 2020, pp. 49–54. URL: https://aclanthology.org/2020.trac-1.8/.

[15] 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.

[16] J. M. Struß, M. Siegel, J. Ruppenhofer, M. Wiegand, M. Klenner, Overview of germeval task 2, 2019 shared task on the identification of offensive language, in: Proceedings of the 15th Conference on Natural Language Processing, KONVENS 2019, Erlangen, Germany, October 9-11, 2019, 2019. URL: https://corpora.linguistik.uni-erlangen.de/data/konvens/proceedings/papers/germeval/GermEvalSharedTask2019Iggsa.pdf.

[17] Y. Zhang, Q. Yang, A survey on multi-task learning, CoRR abs/1707.08114 (2017). URL: http://arxiv.org/abs/1707.08114. arXiv:1707.08114.

[18] B. Wang, Y. Ding, S. Liu, X. Zhou, Ynu_wb at HASOC 2019: Ordered neurons LSTM with attention for identifying hate speech and offensive language, in: P. Mehta, P. Rosso, P. Majumder, M. Mitra (Eds.), Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation, Kolkata, India, December 12-15, 2019, volume 2517 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019, pp. 191–198. URL: http://ceur-ws.org/Vol-2517/T3-2.pdf.

[19] S. T. Aroyehun, A. F. Gelbukh, Aggression detection in social media: Using deep neural networks, data augmentation, and pseudo labeling, in: R. Kumar, A. K. Ojha, M. Zampieri, S. Malmasi (Eds.), Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying, TRAC@COLING 2018, Santa Fe, New Mexico, USA, August 25, 2018, Association for Computational Linguistics, 2018, pp. 90–97. URL: https://aclanthology.org/W18-4411/.

[20] I. Arroyo-Fernández, D. Forest, J. Torres-Moreno, M. Carrasco-Ruiz, T. Legeleux, K. Joannette, Cyberbullying detection task: the EBSI-LIA-UNAM system (ELU) at coling'18 TRAC-1, in: R. Kumar, A. K. Ojha, M. Zampieri, S. Malmasi (Eds.), Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying, TRAC@COLING 2018, Santa Fe, New Mexico, USA, August 25, 2018, Association for Computational Linguistics, 2018, pp. 140–149. URL: https://aclanthology.org/W18-4417/.

[21] J. Risch, R. Krestel, Bagging BERT models for robust aggression identification, in: R. Kumar, A. K. Ojha, B. Lahiri, M. Zampieri, S. Malmasi, V. Murdock, D. Kadar (Eds.), Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, TRAC@LREC 2020, Marseille, France, May 2020, European Language Resources Association (ELRA), 2020, pp. 55–61. URL: https://aclanthology.org/2020.trac-1.9/.

[22] S. Mishra, S. Prasad, S. Mishra, Multilingual joint fine-tuning of transformer models for identifying trolling, aggression and cyberbullying at TRAC 2020, in: R. Kumar, A. K. Ojha, B. Lahiri, M. Zampieri, S. Malmasi, V. Murdock, D. Kadar (Eds.), Proceedings of

the Second Workshop on Trolling, Aggression and Cyberbullying, TRAC@LREC 2020, Marseille, France, May 2020, European Language Resources Association (ELRA), 2020, pp. 120–125. URL: https://aclanthology.org/2020.trac-1.19/.

[23] D. Gordeev, O. Lykova, BERT of all trades, master of some, in: R. Kumar, A. K. Ojha, B. Lahiri, M. Zampieri, S. Malmasi, V. Murdock, D. Kadar (Eds.), Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, TRAC@LREC 2020, Marseille, France, May 2020, European Language Resources Association (ELRA), 2020, pp. 93–98. URL: https://aclanthology.org/2020.trac-1.15/.

[24] Z. Yang, Z. Dai, Y. Yang, J. G. Carbonell, R. Salakhutdinov, Q. V. Le, Xlnet: Generalized autoregressive pretraining for language understanding, in: H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, 2019, pp. 5754–5764. URL: https://proceedings.neurips.cc/paper/2019/hash/dc6a7e655d7e5840e66733e9ee67cc69-Abstract.html.