UHH-LT at SemEval-2020 Task 12: Fine-Tuning of Pre-Trained Transformer Networks for Offensive Language Detection

Gregor Wiedemann and **Seid Muhie Yimam** and Chris Biemann

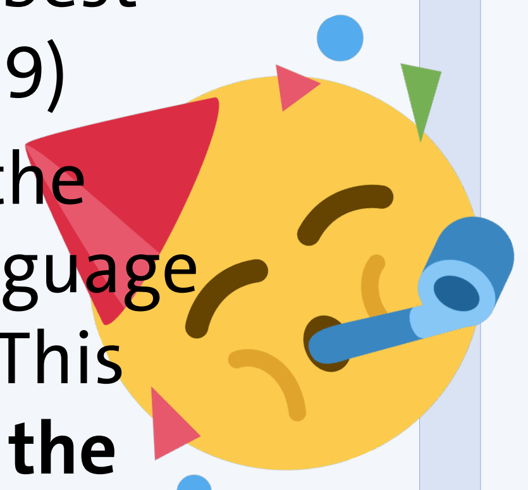
OffensEval 2020: Multilingual Offensive Language Identification in Social Media at SemEval 2020 (Task 12), 12-13 December 2020

INTRODUCTION

- NLP Shared Task for automatic detection of offensive language in Twitter
- OffensEval 2020 provides test set along with a weakly labeled large set of ca. 9 million Tweets for training
- OffensEval 2019 training data is still valid to use since it is the base for the weakly labeled training data
- We use 2020 training data for unsupervised pre-training and 2019 training + test data for supervised training

Two main contributions

1. Evaluation, which pre-trained transformer-based neural network model performs best. Winner models: best individual model: RoBERTa (Liu et. al 2019), best ensemble; ALBERT (Lan et al., 2019)
2. Study on further pre-training of the RoBERTa model with masked language for performance improvements. This model achieved **the first place in the OffensEval 2020 Shared Task A** for English.



DATASET

	Training			Test		
Language	OFF	NOT	Total	OFF	NOT	Total
English	1 448 861	7 640 279	9 089 140	1 090	2 807	3 897

Table 1: Subtask A (all languages): statistics about the data.

	Training			Test		
Language	TIN	UNT	Total	TIN	UNT	Total
English	149 550	39 424	188 974	850	1 072	1 922

Table 2: Subtask B (English): statistics about the data.

	Training				Test			
Language	IND	GRP	OTH	Total	IND	GRP	OTH	Total
English	120 330	22 176	7 043	149 549	580	190	80	850

Table 3: Subtask C (English): statistics about the data.

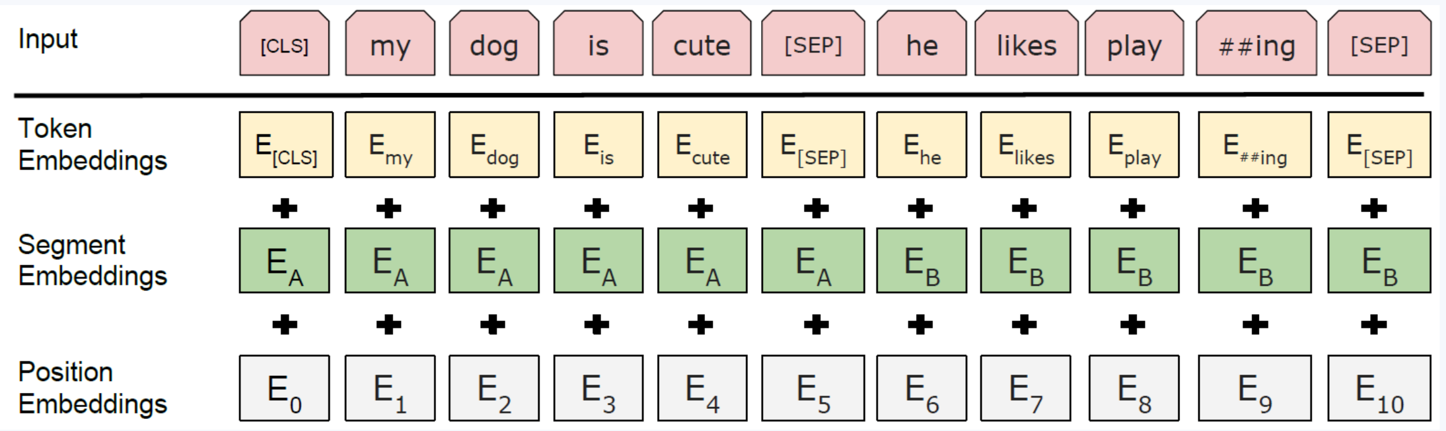
OFF→ offensive, NOT→ not offensive
TIN→ Targeted Insult, UNT → untargeted
IND → Individual, GRP → Group, OTH → Other

cp. Zampieri et al. (2020)

EVALUATING TRANSFORMERS

Pretraining with masked language modeling (MLM)

- BERT (Devlin et. al 2019)
- RoBERTa (Liu et al. 2019)

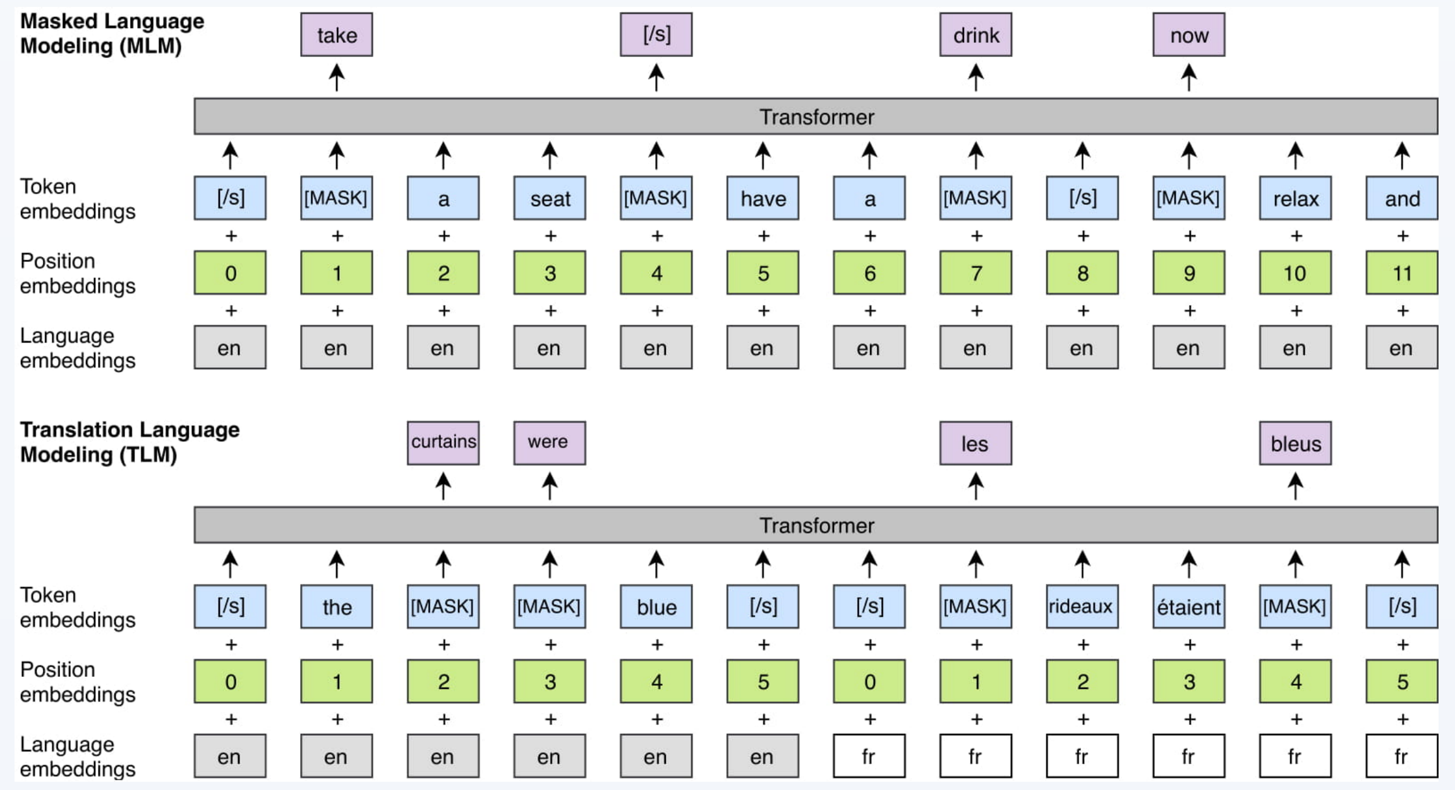


Input	[CLS]	my	dog	is	cute	[SEP]	he	likes	play	##ing	[SEP]
Token Embeddings	$E_{[CLS]}$	E_{my}	E_{dog}	E_{is}	E_{cute}	$E_{[SEP]}$	E_{he}	E_{likes}	E_{play}	$E_{\#\#ing}$	$E_{[SEP]}$
	+	+	+	+	+	+	+	+	+	+	+
Segment Embeddings	E_A	E_A	E_A	E_A	E_A	E_A	E_B	E_B	E_B	E_B	E_B
	+	+	+	+	+	+	+	+	+	+	+
Position Embeddings	E_0	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}

cp. Devlin et al. (2019)

Multi-lingual masked language modeling

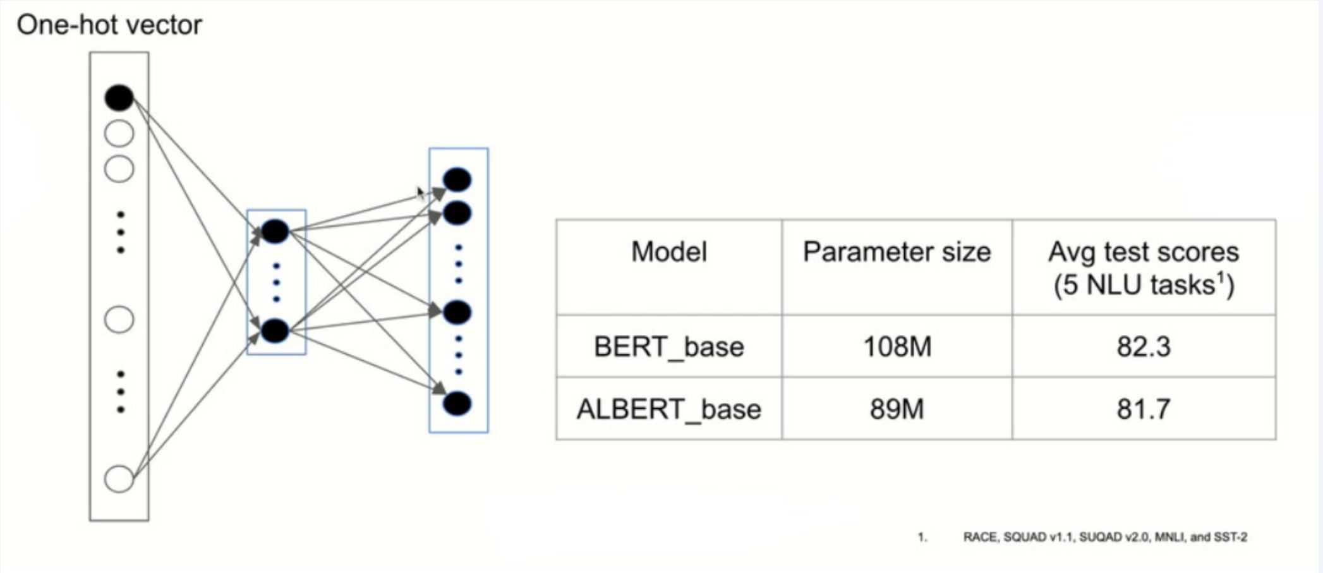
- XLM-RoBERTa (Conneau et al. 2019)


Masked Language Modeling (MLM)
take
[/s]
drink
now
Transformer
Token embeddings
[/s] [MASK] a seat [MASK] have a [MASK] [/s] [MASK] relax and
Position embeddings
0 1 2 3 4 5 6 7 8 9 10 11
Language embeddings
en en en en en en en en en en en en
Translation Language Modeling (TLM)
curtains
were
les
bleus
Transformer
Token embeddings
[/s] the [MASK] [MASK] blue [/s] [/s] [MASK] rideaux étaient [MASK] [/s]
Position embeddings
0 1 2 3 4 5 0 1 2 3 4 5
Language embeddings
en en en en en en fr fr fr fr fr fr

cp. Lample & Conneau (2019)

One-hot word piece compression by low-dimensional projection + parameter sharing across layers

- ALBERT (Lan et al. 2020)



Model	Parameter size	Avg test scores (5 NLU tasks[1])
BERT_base	108M	82.3
ALBERT_base	89M	81.7

[1] RACE, SQUAD v1.1, SUQAD v2.0, MNLI, and SST-2

Further MLM-pretraining

- continue unsupervised MLM pre-training of RoBERTa on OffensEval 2020 Twitter sample with before fine-tuning on OffensEval 2019 target data

RESULTS

- Pre-trained transformers (individual + ensemble predictions)

	NOT			OFF				
Model	**P**	**R**	**F1**	**P**	**R**	**F1**	**Macro F1**	**Acc.**
				Baselines				
All NOT	72.21	100.00	41.93	-	0.00	0.00	41.93	72.21
All OFF	-	0.00	0.00	27.78	100.00	43.49	21.74	27.79
				Single pre-trained transformer models				
BERT-base	99.06	90.2	94.42	79.34	97.78	87.60	91.01	92.31
BERT-large	99.65	90.35	94.77	79.81	99.17	88.44	91.60	92.80
RoBERTa-base	99.45	90.70	94.88	80.33	98.70	88.57	91.73	92.93
RoBERTa-large	99.53	90.92	95.03	80.73	98.89	88.89	91.96	93.13
XLM-RoBERTa	99.03	91.31	95.01	81.22	97.69	88.69	91.85	93.08
ALBERT-large-v1	98.87	90.24	94.36	79.32	97.31	87.40	90.88	92.20
ALBERT-large-v2	98.87	90.20	94.34	79.26	97.31	87.36	90.85	92.18
ALBERT-xxlarge-v1	98.35	91.09	94.58	80.57	96.02	87.62	91.10	92.46
ALBERT-xxlarge-v2	98.47	91.73	94.98	81.76	96.30	88.44	91.71	93.00
				Ensembles of pre-trained transformer models				
All	**99.65**	90.95	95.10	80.83	**99.17**	89.06	92.08	93.23
BERT	99.42	91.16	95.11	81.11	98.61	89.01	92.06	93.23
RoBERTa	99.57	90.84	95.01	80.62	98.98	88.86	91.93	93.11
ALBERT-all	98.23	**92.66**	**95.36**	**83.37**	95.65	89.00	92.23	**93.49**
ALBERT-xxlarge	98.70	92.16	95.32	82.62	96.85	**89.17**	**92.25**	93.47

Table 1: Performance (in %) of baselines, single models, and ensemble models on the OLID test set.

- Improvements through further MLM pre-training on Tweets

	NOT			OFF				
Model	**P**	**R**	**F1**	**P**	**R**	**F1**	**Macro F1**	**Acc.**
RoBERTa-large	98.96	91.49	95.07	81.54	97.49	88.76	91.93	93.15
RoBERTa-large MLM-ft	**99.15**	**91.53**	**95.18**	**81.66**	**97.96**	**89.06**	**92.12**	**93.31**

Table 2: Performance (in %) of MLM fine-tuned models on the OLID test set (average of 10 runs).

Final results for all three OffensEval 2020 sub-tasks:

- A – Offensive Language Detection
- B – Offensive Language Categorization
- C – Offensive Language Target Identification

Team	**UHH-LT**	
	Macro F1 (%)	Rank
Task A	92.04	1 out of 84
Task B	65.98	6 out of 42
Task C	66.83	3 out of 38

A qualitative look into a sample of false predictions: False positives (FP) and false negatives (FN)

1. FP: *@USER the gov and nyc mayor are the biggest joke except maybe for the idiots who elected them*
2. FP: *@USER What the fuck*
3. FP: *@USER I know it was bad but I used to love it*
4. FP: *men who voted yes are childish, sorry are you 17??Men, would you have a problem if a girl said if she's not receiving head she's not giving head?*
5. FN: @USER It's as if every single one of her supporters are just as stupid as her. Wehdone..
6. FN: I'm gonna say he doesn't really. You should check the zip code demographics of his various properties Only liars could deny that Al Sharpton hates white people
7. FN: @USER Fuck the chargers
8. FN: Last night I watched the Democrats throwing shit up against the wall, and none of it stuck.

CONCLUSIONS

- Multiple pre-trained transformer architectures among the best performing models for the OffensEval 2020 Shared Task
- ALBERT ensemble performs best for Task A (with less model parameters than BERT or RoBERTa)
- Still room for improvement on Tasks B and C (e.g. through measures against negative effects from class imbalance)
- Weak labels for training are not helpful
- Further pre-training with MLM on unlabelled in-domain data improves offensive language detection

REFERENCES

- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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), pages 4171–4186, MN, USA. ACL.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. RoBERTa: A robustly optimized BERT pretraining approach. CoRR, abs/1907.11692.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzman, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross- ´ lingual representation learning at scale. CoRR, abs/1911.02116.
- Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A lite BERT for self-supervised learning of language representations. CoRR, abs/1909.11942.
- Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and C¸ agrı C¸ ˘ oltekin. 2020. SemEval-2020 Task 12: Multilingual Offensive Lan- ¨ guage Identification in Social Media (OffensEval 2020). In Proceedings of The 14th International Workshop on Semantic Evaluation (SemEval), Barcelona, Spain. ACL.

ACKNOWLEDGEMENTS

The work was supported by BWFG Hamburg (Germany) within the "Forum 4.0" project as part of the ahoi.digital funding line, and the DFG project "FAME" (WI 4949/2-1, grant no. 406289255).

ACKNOWLEDGEMENTS

yimam@informatik.uni-hamburg.de
g.wiedemann@leibniz-hbi.de