# FINE-TUNING OF PRE-TRAINED TRANSFORMERS FOR OFFENSIVE LANGUAGE DETECTION

**Xuanting Chen, Zhitong Su**
Duke University
{xc174,zs134}@duke.edu

## ABSTRACT

SemEval-2020 Task 12 on Offensive Language Detection raises great interest of research community in this topic and attracts a large number of participants. The subtask A is aimed to automatically classify a substantial number of tweets to either offensive or non-offensive. We reproduce the model as well as experiments created by the team UHH-LT (Wiedemann et al., 2020) who ranked 1st in the Subtask A. The state-of-the-art result was produced by fine-tuning on a further pre-trained RoBERTa ensemble.The dataset contains over 14,000 labeled English tweets and more than nine million unlabeled ones. The winning team utilized the nine million unlabeled data to do further pre-training with MLM (masked language modeling) objective and later used the labeled data to do fine-tuning. After the competition organizer released the test data, the team found that a pre-trained ALBERT-xxlarge ensemble even outperforms the submitted model. In this project, we also try to implement the potentially best model suggested by the team UHH-LT.

## 1 INTRODUCTION

With the wide use of social media platforms nowadays, offensive language has become ubiquitous in our life. Given the variety and multitude of related terms and definitions, a great amount of research and contests have been carried to identify offensive language within bulky texts collected from social media. In order to create a friendly online social environment, an automatic and precise way of detecting and filtering out offensive language is in desperate need. One well-known competition concerning such issues is *SemEval-2019/2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2019/2020)*. The task can be divided into three subtasks: offensive language identification (subtask A), categorization of offense types (subtask B), and offense target identification (subtask C). In the OffensEval 2019 dataset (2019 OLID), subtask A contains over 14,000 Tweeter tweets that are manually labeled as *offensive* (OFF) or *not offensive* (NOT). For subtask B, the tweets labeled as offensive are further classified into *targeted* (TIN) or *untargeted* (UNT). Subtask C further separates the targeted tweets into three categories: *groups* (GRP), *individuals* (IND), and *others* (OTH). The OffensEval 2020 complemented the existing labeled data with more than nine million weakly-labeled ones, as well as datasets in other languages (2020 SOLID). The labels of the nine million data are predicted by an ensemble of the five different classifiers such as LSTM and BERT trained on the OffensEval 2019 labeled dataset. UHH-LT (Wiedemann et al., 2020) was the winning team on OffenseEval 2020 subtask A with an F1 score of 0.9204. With the same modeling strategies, they ranked 6th in subtask B, and third in subtask C. In this project, We conduct all the experiments the team went through, replicate the result they achieved on subtask A, and implement a new approach they claimed that even outperforms their result of submission.

## 2 RELATED WORK

### 2.1 OFFENSIVE LANGUAGE DETECTION

Offensive language detection is a popular topic in the field of text classification. Many public datasets are available for detecting offensive language. However, these datasets differ notably

in terms of data sources, label defintions, sampling strategies, and annotation guidelines. Therefore, models created from different datasets are not comparable, and also we cannot combine these datasets to train an alpha model. Efforts have been done to unify the datasets in this research area. Mandl et al. (2019), StruB et al. (2019), and Basile et al. (2019) have organized shared tasks on this topic. The area of offensive langauge detection is still filled with problems and challenges, even though the current systems are able to generate a high prediction accuracy. One challenge is that the current system is highly vulnerable to adversarial data. Grondahl et al. (2018), for instance, show that adding a word "love" to an offensive tweet can fool the classifier to believe that it is not offensive. The current datasets containing text solely might have issues as well. StruB et al. (2019) claim that linguistic information alone is not enough to determine whether a text is offensive or not. Wiedemann et al. (2018) suggest additional features such as user information of users mentioned in the tweet.

### 2.2 Pre-training for Text Classification

In the area of text classification, especially when the training dataset is small, transfer learning has been dominating in almost every competition, because it solves the problem that neural network need to consume a considerable amount of data to be powerful. For example, liu et al.(2019) won the 2019 SemEval Offensive Language Detection task by fine-tuning the pre-trained BERT model. Mozafari et al. (2009) successfully used BERT for hate speech recognition.The advent of the BERT model (Devlin et al.) has revolutionized a wide range of tasks in NLP. The BERT-base model trains on a large English corpus and learns a representation of the English language that can then be later used to extract features useful for downstream tasks. After BERT, a number of variants improve the base model with less training time, or larger datasets, or different network architectures. In this project, we are going to test a selection of pre-trained models available online such as RoBERTa and ALBERT, and ensemble some of them to generate the best result,

## 3 Approach

### 3.1 Background

The subtask A of offensive language detection is essentially a binary classification problem over text data similar to sentiment analysis. Almost all top teams employ some kinds of pre-trained transformers either individually or in an ensemble to approach this classification problem (Marcos et al., 2020). For the additional nine million unlabeled data, most teams found it useful and deployed the new dataset in their final systems (Marcos et al., 2020).

### 3.2 Method

The team UHH-LT did not trust the effectiveness of the predicted labels and decided to use the 9 million data in an unsupervised manner for further pre-training. They first compared the performance of different pre-trained transformer models available online such as ALBERT-large-v1 and ALBERT-large-v2 individually on the 2019 OLID dataset. Then, they picked the best individual model, RoBERTa-large in their experiment, to do further pre-training with MLM objective on the nine million raw texts. In fact, they randomly sample merely 5 percent of the nine million data to reduce the training time. Next, they fine-tuned the double pre-trained model on the 2019 labeled dataset. At the end, they obtained the majority vote from an ensemble of MLM pre-trained RoBERTa-large trained on 10 different subsets (9 folds per subset) of the training data using 10-fold division. In a nutshell, their result of submission was produced by an MLM RoBERTa-large ensemble in a 10-fold division manner.

### 3.3 Extensions

After the contest organizer released the test data, the team UHH-LT, ran experiments over the text data, and found that an ensemble of ALBERT-xxlarge-v1 and ALBRET-xxlarge-v2 (without further pre-training) even outperforms the results they submitted. Based on this intuition, we experiment over different ways of ensembling including an ensemble of all models, ensembles of a single type of transformer with different sizes or structures (e.g. BERT-base and BERT-large).

| Tweet | A | B | C |
|---|---|---|---|
| This account owner asks for people to think rationally | NOT | - | - |
| this job got me all the way fucked up real shit | OFF | UNT | - |
| wtf ari her ass toooo big | OFF | TIN | IND |
| @USER We are a country if morons | OFF | TIN | GRP |

Table 1: Examples in 2019 OLID dataset.

# 4 EXPERIMENTS

## 4.1 DATASETS AND EVALUATION

Both the OffensEval 2019 dataset (2019 OLID) and the OffensEval 2020 dataset (2020 SOLID) are going to be used in the final model in the MLM RoBERTa-large ensemble model. Other ensembles in our experiment use 2019 OLID solely. 2019 OLID contains over 14,000 manually labeled Tweeter tweets with three-level labels: identification of offensive language (”offensive” or ”not offensive”), categorization of offensive language (”targeted” or ”untargeted”), and targets of offensive language (”group” or ”individuals” or ”others” ). Examples of 2019 OLID are presented above. 2020 SOLID contains nine million weak labels of identification of offensive language solely, where the labels are predicted by an ensemble of PMI, Fast-Text, LSTM, and BERT. In effect, 2020 SOLID provides average probabilities of the four models and the variance. Whether converting them to actual labels is up to the user. In terms of data preprocessing, we strip URLs and user mentions and remove duplicates. The performance of subtask A is evaluated in Macro F1 score, because both false positives and false negatives are equally important for this task. In a desirable system, we want to correctly identify all offensive tweets as much as possible; meanwhile, we also do not want to take the risk of misclassifying a normal tweet into the offensive group.

## 4.2 MODEL AND TRAINING DETAILS

For the further pretraining of the best individual model RoBERTa-large, we train for one epoch with batch size of 4 and learning rate of 2e-5. For the fine-tuning, we train for six epoch with batch size of 4 and learning rate of 5e-6. We also use gradient clipping to avoid gradient exploding. The baselines of F1 scores are the F1 scores derived when all predictions are ”NOT” or ”OFF”. In our experiment, we first fine-tune each transformer-based pre-trained model individually on the labeled dataset and obtain its F1 score. The candidate pre-trained models include BERT-base, BERT-large, RoBERTa-base, RoBERTa-large, ALBERT-large-v1, ALBERT-large-v2, ALBERT-xxlarge-v1, ALBERT-xxlarge-v2. Then, we are going to further pretrain the presumably best model RoBERTa-large with MLM objective on the unlabeled tweets. We will denote this model as RoBERTa-large MLM in the result section. Finally, we will try ensembles of a single type of transformer with different sizes or structures including an ensemble of BERT-base and BERT-large, an ensemble of RoBERTa-base and RoBERTa-large, an ensemble of all four ALBERT models, and an ensemble of ALBERT-xxlarge-v1 and ALBERT-xxlarge-v2 which is the potentially best model proposed by the team.

## 4.3 RESULTS

In the first phase of our experiment, we evaluate different transformer-based pre-trained models individually on the test data. As the team claims, the best individual model is indeed RoBERTa-large with Macro F1 of 91.86. We then further pre-train this model on the unlabeled data, and obtain a f1 score of 91.99. Further pre-training with MLM objective turns out to be indeed, but slightly useful. For the ensemble, however, the best ensemble we obtain is not ALBERT-xxlarge, the one they claim to be even better than RoBERTa-large MLM. The best ensemble we obtain is an ensemble of all four ALBERTs with a f1 score of 91.90. Generally, ensembles work better than individual models. The best model among all is the ensemble of all four ALBERTs.

| Model | Macro F1 | Accuracy |
|---|---|---|
| All NOT | 41.93 | 72.21 |
| All OFF | 21.74 | 27.79 |
| | Single models | |
| BERT-base | 90.93 | 92.26 |
| BERT-large | 91.42 | 92.74 |
| RoBERTa-base | 91.70 | 92.87 |
| RoBERTa-large | **91.86** | **93.10** |
| RoBERTa-large MLM | 91.99 | 93.21 |
| ALBERT-large-v1 | 91.50 | 92.15 |
| ALBERT-large-v2 | 91.49 | 92.13 |
| ALBERT-xxlarge-v1 | 91.39 | 92.42 |
| ALBERT-xxlarge-v2 | 91.55 | 92.91 |
| | Ensembles | |
| BERT | 91.60 | 93.15 |
| RoBERTa | 91.83 | 93.03 |
| ALBERT-all | **91.90** | **93.49** |
| ALBERT-xxlarge | 91.58 | 93.27 |

Table 2: Performance (in percentage) on the 2020 SOLID test data.

## 5 Conclusion

Generally, our results have proved the feasibility of the methods applied by UHH-LT. In terms of evaluating a single model, most of our experiments produce similar results compared to their claimed results. Further pre-training and applying ensemble will indeed improve the f1 score, but the improvements are not significant. The ensemble of albert-xxlarge-v1 and albert-xxlarge-v2 in our experiment gives an F1 score of only 91.58, which is far less than their claimed value. An ensemble of all ALBERT is the best ensemble and outperforms an ensemble of ALBERT-xxlarge, which is the opposite of their results. In our opinions, the high score of ALBERT-xxlarge might be a result of randomness, but the effectiveness of further pre-training as well as the ensemble method have been proved effective.

### Author Contributions

Basically we work everything together in the same room each time. If we have to distinguish who did what, Xuanting Chen is the one who was typing code and running code since he has GPU installed on his PC, Zhitong Su is the one who was typing the report.

## References

Gregor Wiedemann, Seid Yimam, and Chris Biemann. 2020. UHH-LT at SemEval-2020 Task 12: Fine-tuning of pre-trained transformer networks for offensive language detection. In Proceedings of the International Work-shop on Semantic Evaluation (SemEval).

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. 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, pages 14–17, Kolkata, India.

Julia Maria Struß, Melanie Siegel, Josep Ruppenhofer, Michael Wiegand, and Manfred Klenner. 2019. 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), pages 354–365, Erlangen, Germany.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In Proceedings of the 13th International Workshop on Semantic Evaluation, pages 54–63, MN, USA. ACL.

Tommi Gro¨ndahl, Luca Pajola, Mika Juuti, Mauro Conti, and N. Asokan. 2018. All you need is "love": Evading hate speech detection. In Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security, page 2–12, NY, USA. ACM.

Ping Liu, Wen Li, and Liang Zou. 2019a. NULI at SemEval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers. In Proceedings of the 13th International Workshop on Semantic Evaluation, pages 87–91, MN, USA, June. ACL.

Marzieh Mozafari, Reza Farahbakhsh, and Noe¨l Crespi. 2019. A BERT-based transfer learning approach for hate speech detection in online social media. In Hocine Cherifi, Sabrina Gaito, Jose´ Fernendo Mendes, Esteban Moro, and Luis Mateus Rocha, editors, Proceedings of the 8th International Conference on Complex Networks and their Applications, pages 928–940, Lisbon, Portugal.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirec- tional 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.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and C¸ ag˘rı C¸ o¨ltekin. 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.