A Toxic Comment Classification Model Based on Ensemble

Jian Xu' and Yuqing Zhai*  
1 School of Cyber Science and Engineering, Southeast University, Nanjing, Jiangsu  211111, China  
*Corresponding author’s e-mail: yqzhai@seu.edu.cn

Abstract. Accurate classification of toxic comments in different languages is an important task in today's international social networking platforms. In order to improve the classification of toxic comments in different languages, this paper combines the advantages of high accuracy of monolingual model and strong generalization ability of multi-language model, and adopts the ensemble of multilingual model and monolingual model to classify toxic comments. For the monolingual model, the monolingual pre-training model is fine-tuned with labeled task data; for the multilingual model, before fine-tuning the model with labeled task data, a further pre-training is applied to train model on unlabeled data, which aims to make full use of the large amount of unlabeled data and reduce the dependency of amount of labeled comment data while improving the classification effect. Comparative experiments on Conversaison AI's multilingual toxic comment dataset show that the model in this paper has improved results on different evaluation metrics compared to the XLM-RoBERTa multilingual fine-tuning model, illustrating the validity of the model.

1. Introduction

People obtain and share information on social platforms almost every day[1]. However, toxic comments on the Internet severely disrupt the online information exchange experience of users and even lead to tension and disruption in the social environment[2]. Building models that can automatically classify toxic comments has become a focus of researchers and industry[3].

Earlier toxic comment classification models are generally based on manual rule building and simple classifiers. Gitari et al[4] built a toxic vocabulary to record and classify toxic comments, while Rdulesuc et al[5] distinguished toxic comments from normal comments by calculating the similarity between sentences. Ravi et al[6] used machine learning models such as Naive Bayes, Logistic Regression and Support Vector Machines for toxic comment classification.

In recent years, since deep learning models allow researchers to avoid the limitations of manual feature engineering and their good results in toxic comment classification, models represented by Convolutional Neural Networks (CNN)[7] and Recurrent Neural Networks (RNN)[8] are gradually used. Spiros[9] and Li, Siyuan[10] et al. used CNN and RNNs on toxic comment classification tasks to obtain better performance than simple classifiers. However, since CNNs and RNNs rely on static word embedding vectors and cannot dynamically adjust the word vector representation based on different contexts, Ashwin et al [11] used task data to fine-tune BERT[12] and then used the model for toxic comment classification, obtaining better results than the previous model. However, along with the exchange and integration of cultures around the world, people from different countries communicate with each other on international social platforms such as Twitter, Ins, and Facebook, resulting in multilingual comments. Although BERT is effective in toxic comment classification in a single language, it lacks some generality when facing multilingual comments.
Although deep learning models are good at capturing contextual relationships, their drawback is that they rely excessively on a large annotated corpus, and the performance of the models is influenced by the quantity and quality of the annotated corpus. In fact, a large amount of unlabeled data is not fully used, and if the information of these unlabeled data can be fully utilized, it can further improve the effectiveness of deep learning in classifying toxic comments [13].

To solve the problem of lack of generality of monolingual models in the face of multilingual comments, this paper combines the advantages of high classification accuracy of monolingual models and strong generalization ability of multilingual models, and proposes a toxic comment classification model that combines both models, which can classify a single comment (only words in one language in each comment) at a time. The comments to be classified are shown in Figure 1, where E1-E2 are normal comments and E3, E4 and E5 are toxic comments. Taking the English comment of E2 as an example, as shown in Figure 2, it is input into the multilingual model and the monolingual model of English respectively during classifying, and the output probability results of the two models are weighted and summed to obtain the final probability results.

![Figure 1. Multilingual comments](image1)

In order to make full use of the information of both labeled and unlabeled data to improve the classification effect, for the multilingual models, this paper first uses Masked Language Modeling to achieve further pre-training, so that the multilingual pre-trained models are unsupervisedly retrained on unlabeled task data to learn the contextual information of the task data, and then use labeled task data. For each monolingual model, the model is fine-tuned using the labeled task data of the corresponding language. Collectively, the model proposed in this paper has higher accuracy and stronger generalization ability in multilingual toxic comment classification tasks.

This paper is then presented in four main sections: firstly, Section 2 introduces the related work on monolingual pre-training model, multilingual pre-training model and Masked Language Modeling; Section 3 introduces the toxic comment classification model based on ensemble proposed in this paper; Section 4 introduces the related experimental data and evaluation metrics, and compares and analyzes the experimental results; finally, the work of this paper is summarized.
2. Related Works

2.1. Monolingual Pre-trained model
With the development of deep learning, the number of parameters for models increases rapidly, which requires larger training datasets to train the models to prevent overfitting. However, building large-scale labeled datasets is a huge challenge for most natural language processing (NLP) tasks, but unlabeled data is relatively easy to obtain. To take advantage of the huge unlabeled text data, researchers of pre-trained models first use different tasks to let the models learn good representations on large amounts of unlabeled text data, and then use these representations for other tasks[14-15]. Recent studies have shown that pre-training a model on a large text corpus allows it to learn generic text representations, which helps in downstream tasks. In order to reduce the training dataset required for toxic comment classification and speed up the convergence of model training, this paper uses the BERT as the monolingual model.

2.2. Multilingual Pre-trained model
In the field of multilingual pre-training model research, XLM[16] proposed by Facebook is a optimized model based on BERT and uses Byte-Pair Encoding(BPE) encoding[17] instead of the original word input and character input to increase the shared vocabulary. In addition, it mixes different language texts and uses new training targets, allowing the model to learn more linguistic information and improve the prediction of different tasks. Alexis et al[14] proposed XLM-RoBERTa, an improved model based on the XLM. This model fully combines the advantages of XLM and RoBERTa[18], and is trained on a large-scale multilingual corpus, which improves the performance of multilingual migration tasks and becomes one of the best performing multilingual pre-training models in multilingual classification[18]. To meet the task of multilingual toxic comment classification and to improve the effectiveness, this paper uses XLM-RoBERTa as the multilingual model.

2.3. Masked Language Modeling
Unsupervised learning is a learning approach that combines supervised learning with unsupervised learning and is a key research problem in the field of machine learning[19]. As an unsupervised learning task, Masked Language Modeling[12] has similarities with Denoising Self-encoder Modeling proposed by Vincent et al[20], both of which achieve model learning of input data by partially masking and re-predicting input information, but in contrast to Denoising Self-encoder Modeling, in the Masked Language Modeling task, models only learn the part of words that are masked, rather than the entire input sequence[16]. Chi Sun[21] et al. experimentally showed that the use of unsupervised retraining on the task dataset helps to improve the performance of BERT pre-trained models in downstream tasks. In the training process of the multilingual pre-training model XLM-RoBERTa, this paper first retrained XLM-RoBERTa using a Masked Language Modeling task to allow the model to learn textual representations relevant to the toxic comment detection task on the available unlabeled data to improve the convergence speed of the model's objective function and its ability to fit the data in the fine-tuning task on labeled data.

3. Toxic Comment Classification Model Based on Ensemble

3.1. Model Structure
The overall structure of the model is shown in Figure 3. In the Data Processing part, the given corpus is firstly pre-processed, including removal of stopwords and numeric symbols and converted into the one-hot vector. Then the data is split into multilingual unlabeled data and multilingual labelled data according to whether they are labeled or not. Finally, the multilingual labeled comment data is classified into different language labeled comment data by language and split into test sets. In the Training part, the labeled data of different languages are provided to the corresponding monolingual BERT for fine-tuning (corresponding to fine-tuning(2) in Figure 3) and the multilingual data are provided to XLM-
RoBERTa for further pre-training and fine-tuning(1). Among them, further pre-training uses multilingual unlabeled data and Masked Language Modeling task to train the XLM-RoBERTa so that the model is initially adapted to the distribution of the task data, and fine-tuning(1) uses multilingual labeled data to fine-tune XLM-RoBERTa. In the Classification part, the output of XLM-RoBERTa and the monolingual BERT model for multiple corresponding languages are weighted and summed (see 3.4 for details) to obtain the final predicted output results.

3.2. Further Pre-training based on Masked Language Modeling
In the Further Pre-training shown in Figure 3, in order to make full use of the large amount of unlabeled data for the model to learn a generic representation of the task data, this paper uses Masked Language Modeling task and unlabeled data for the unsupervised retraining of XLM-RoBERTa. Similar to the pre-training of BERT[13], the procedure is as follows.

a. 15% of the input words are randomly selected for masking each time.

b. For all masked words, 80% of them are marked as [MASK], 10% are randomly replaced with arbitrary words, and 10% are replaced with the original words.

c. For each masked word T in step a, this paper uses it as a prediction target for the output and trains the model with the cross-entropy loss function.

3.3. Fine-tuning
As shown in Figure 4, the structures of both BERT and XLM-RoBERTa are mainly composed of Embedding Layers, Transformer Encoders[22], and a downstream structure. In fine-tuning(1) of Figure 3, the multilingual labeled data is used to fine-tune XLM-RoBERTa; in fine-tuning(2), BERT is fine-tuned using the monolingual labeled data. The specific approach is shown in Figure 4. The output vector C corresponding to the starting identifier [CLS] in a single comment is taken as the output of the model, the Fully Connected Layer is connected, and then the Sigmoid activation function is used to map the output results into probability values, and the cross-entropy loss is calculated by combining the true label of the sentence as the objective function to train XLM-RoBERTa and BERT.
3.4. Ensemble
Since the monolingual model is trained to classify data in a single language only, its advantage is that
the model fits the data better, while the multilingual model is trained to classify data in a mixture
of multiple languages, and the Masked Language Modeling task makes the model learn more contextual
information about words, and its advantage is better generalization performance. Since monolingual
and multilingual models have their own advantages, this paper uses a combination of both to improve
the classification effects. In the process of Ensemble, for each sample, in order to determine the respective
contribution share of monolingual and multilingual models, this paper uses cross-validation[23] for
effects evaluation, and the F1 scores of whose are used as the final contribution share of model prediction
results, respectively. Specifically, if the F1 score of the multilingual model for the prediction result of
the validation dataset is $F_1^{\text{multi}}$ and the F1 score of the monolingual model is $F_1^{\text{single}}$, the final
probability to determine the positive class is:
\[
\frac{F_1^{\text{multi}}}{F_1^{\text{multi}} + F_1^{\text{single}}} P_{\text{multi}} + \frac{F_1^{\text{single}}}{F_1^{\text{multi}} + F_1^{\text{single}}} P_{\text{single}}
\]
(1)
where $P_{\text{multi}}$ denotes the probability of determination of positive class by the multilingual model, and
$P_{\text{single}}$ denotes the probability of determination of positive class by the monolingual model.

4. Experiment and Analysis

4.1. Dataset
Currently available toxic comment datasets often have only a single category of language, such as
Wikipedia dataset[24], Twitter dataset, etc.[25]. While in current Internet environment, commenters
may come from different countries, so in this paper, we choose the Multilingual Toxic Comment
Classification dataset provided by Conversasion AI. In this paper, the dataset is used to construct a
binary classification task for toxic comment classification.

Because of the unbalanced of language types in this dataset, we obtain data as follows:
- A training set of 16941 comments, split by language types, including 8941 comments in English,
  3000 comments in Turkish, 2500 comments in Spanish, and 2500 in Italian;
- A validation set of 2542 comments, including 1342 comments in English, 450 comments in
  Turkish, 375 comments in Italian and 375 in Spanish;
- A unlabeled data set consists of 30,932 comments, including 14,000 comments in Turkish, 8494
  comments in Italian, and 8438 in Spanish.

4.2. Evaluation Metrics
The use of a single evaluation metric leads to a possible inability to evaluate the model in a
comprehensive manner. In order to evaluate the model comprehensively and objectively, accuracy,
precision, recall, and F1 score are chosen as the evaluation metrics in this paper.

The F1 score is calculated as follows.
\[
P_{\text{na}} = \frac{\sum_{i=1}^{N} R_i}{N}
\]
(2)
\[
R_{\text{na}} = \frac{\sum_{i=1}^{N} R_i}{N}
\]
(3)
\[
F1_{\text{na}} = \frac{2 \times P_{\text{na}} \times R_{\text{na}}}{P_{\text{na}} + R_{\text{na}}}
\]
(4)
Where, $P_{ma}$ denotes the value of precision; $R_{ma}$ denotes the value of recall rate; and $F_{1ma}$ is the combination of precision and recall.

4.3. Compared Models
In order to verify the effectiveness of the proposed model, multiple sets of comparison experiments are designed in this paper. Among them, The XLM-RoBERTa, which is fine-tuned using only the full data, will be used as the baseline model, which is compared with other models respectively, denoted as follows.

- **Ensemble**: a fusion model for multilingual malicious comment detection after fusion of XLM-R_MLM and BERT_base.
  - Baseline: XLM-RoBERTa obtained by fine-tuning on the multilingual labeled comment data;
  - BERT_base: BERT obtained by fine-tuning on the monolingual labeled comment data;
  - XLM-R_MLM: XLM-RoBERTa obtained by further pre-training on the multilingual unlabeled comment data and fine-tuning on the multilingual labeled comment data;
  - Ensemble: Ensemble model combining BERT_base and XLM-R_MLM.

4.4. Environment and Parameter Settings
In this paper, we use TensorFlow2 framework for model construction, and use TPU server on google cloud platform for training. The dataset used is the dataset described in 4.1. Combining the sentence length distribution of the existing comment dataset, the actual hardware conditions and the experimental efficiency, the corresponding parameters are set as follows: the batch size for training is 16, the word vector dimension is 768, the maximum input sentence length for all models is 224, the insufficient part is complemented by <PAD> markers, and the excess part is truncated. The optimal learning rate, the number of training rounds and the size of training batches are determined experimentally, and the appropriate truncation length of the input sentence length is determined according to the length of the text in the dataset, while the word vector dimension and the number of fully connected layers follow the common parameters of XLM-RoBERTa and BERT.

4.5. Results and Analysis
In order to verify the effectiveness of further pre-training and the effect of the Ensemble model, this paper sets up the comparison experiments of the training process of Baseline and XLM-R_MLM models mentioned in 4.3 and the comparison experiments of the testing effect of each comparison model on the validation set, respectively.

4.5.1. Comparison of Fine-tuning Process. In order to verify the effectiveness of further pre-training, this paper tracks and records several metrics of Baseline and XLM-R_MLM during the training process, including the training loss, the validation loss, and F1 score. The overall trend of the curves in Figure 5 shows that the unsupervised further pre-training not only accelerates the convergence speed of fine-tuning, but also reduces the value of the loss function after convergence to a certain extent. The training efficiency, the ability to fit the task data and the generalization ability of the model are all improved to some extent.

4.5.2. Comparison of Results. To verify the improvement in the effects of ensemble, the performance of Baseline, XLM-R_MLM, Ensemble under the four metrics on the multilingual test dataset and the performance of BERT_base under each metric on the test dataset of their respective corresponding languages are shown in Figure 6. It can be seen that ensemble improves the accuracy by 3.1% and the macro-F1 score by 6.2% relative to the Baseline model. It indicates that the ensemble model combines the features of high accuracy of the monolingual model on the corresponding languages and the strong generalization ability of the multilingual model to improve the classification of multilingual toxic comments. The training time of each comparison model under the current hardware conditions described
in 4.4 is shown in Figure 7. From the results, the fusion model increases the training time by 56% compared to Baseline, but it can be kept low as both are based on fine-tuned training.

![Figure 5. Comparison of Fine-tuning process](image)

![Figure 6. Comparison of results on the four evaluation metrics](image)

![Figure 7. Comparison of convergence time](image)

5. Conclusion

In this paper, we propose a toxic comment classification model based on ensemble that combines monolingual and multilingual models. First, we use unsupervised further pre-training and supervised fine-tuning for the multilingual pre-trained model and supervised learning for the monolingual pre-trained model, which enhances the adaptability of the model to the task. In addition, this paper also uses ensemble to improve the generalization performance on toxic comment classification task, which combines the respective advantages of monolingual models and multilingual model. However, the model in this paper has not yet solved the problem of classifying toxic comments containing multiple language words in a single comment at the same time, which can be further optimized in the future.

References

[1] Wulczyn, E., Thain, N., & Dixon, L. (2017). Ex machina: Personal attacks seen at scale. In: Proceedings of the 26th International Conference on World Wide Web. pp. 1391-1399.

[2] Parekh, P., & Patel, H. (2017). Toxic Comment Tools: A Case Study. International Journal of Advanced Research in Computer Science, 8(5).

[3] Risch, J., & Krestel, R. (2020). Toxic comment detection in online discussions. In: Deep Learning-Based Approaches for Sentiment Analysis. Springer. pp. 85-109.

[4] Gitari, N. D., Zuping, Z., Damien, H., & Long, J. (2015). A lexicon-based approach for hate speech detection. International Journal of Multimedia and Ubiquitous Engineering, 10(4), 215-230.
[5] Rădulescu, C., Dinsoreanu, M., & Potolea, R. (2014). Identification of spam comments using natural language processing techniques. In: 2014 IEEE 10th International Conference on Intelligent Computer Communication and Processing (ICCPP). pp. 29-35.

[6] Bhaskaran, J., Kamath, A., & Paul, S. (2017). DISCo: Detecting insults in social commentary.

[7] Mikolov, T., Kombrink, S., Burget, L., Černocký, J., & Khudanpur, S. (2011). Extensions of recurrent neural network language model. In: 2011 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 5528-5531.

[8] Dos Santos, C., & Gatti, M. (2014). Deep convolutional neural networks for sentiment analysis of short texts. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. pp. 69-78.

[9] Georgakopoulos, S. V., Tasoulis, S. K., Vrahatis, A. G., & Plagianakos, V. P. (2018). Convolutional neural networks for toxic comment classification. In: Proceedings of the 10th Hellenic Conference on Artificial Intelligence. pp. 1-6.

[10] Li, S. (2018). Application of recurrent neural networks in toxic comment classification (Doctoral dissertation, UCLA).

[11] d'Sa, A. G., Illina, I., & Fohr, D. (2020). Bert and fasttext embeddings for automatic detection of toxic speech. In: 2020 International Multi-Conference on:“Organization of Knowledge and Advanced Technologies”(OCTA). pp. 1-5.

[12] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[13] Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. Journal of Big Data, 2(1), 1-21.

[14] Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., & Huang, X. (2020). Pre-trained models for natural language processing: A survey. Science China Technological Sciences, 1-26.

[15] Erhan, D., Courville, A., Bengio, Y., & Vincent, P. (2010, March). Why does unsupervised pre-training help deep learning?. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. pp. 201-208.

[16] Lample, G., & Conneau, A. (2019). Cross-lingual language model pretraining. arXiv preprint arXiv:1901.07291.

[17] Shibata, Y., Kida, T., Fukamachi, S., Takeda, M., Shinohara, A., Shinohara, T., & Arikawa, S. (1999). Byte Pair encoding: A text compression scheme that accelerates pattern matching. Technical Report DOI-TR-161, Department of Informatics, Kyushu University.

[18] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

[19] Lison, P. (2015). An introduction to machine learning. Language Technology Group (LTG), 1(35).

[20] Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P. A. (2008). Extracting and composing robust features with denoising autoencoders. In Proceedings of the 25th International Conference on Machine Learning. pp. 1096-1103.

[21] Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to fine-tune BERT for text classification?. In: China National Conference on Chinese Computational Linguistics. pp. 194-206.

[22] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is All you Need. In: NIPS.

[23] Browne, M. W. (2000). Cross-validation methods. Journal of mathematical psychology, 44(1), 108-132.

[24] Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In: Proceedings of the International AAAI Conference on Web and Social Media (Vol. 11, No. 1).

[25] Mohammad, S., Bravo-Marquez, F., Salameh, M., & Kiritchenko, S. (2018). Semeval-2018 task 1: Affect in tweets. In: Proceedings of the 12th international workshop on semantic evaluation. pp. 1-17.