ISSN 1330-3651 (Print), ISSN1848-6339(Online)
https://doi.org/10.17559/TV-20210325125414
Original scientific paper

A Hybrid Model for Monolingual and Multilingual Toxic Comment Detection

Guizhe SONG, Degen HUANG, Yanping ZHANG*

Abstract: Social media provides a public and convenient platform for people to communicate. However, it is also open to hateful behavior and toxic comments. Social networks, like Facebook, Twitter, and many others, have been working on developing effective toxic comment detection methods to provide better service. Monolingual language model focuses on a single-language and provides high accuracy in detection. Multilingual language model provides better generalization performance. In order to improve the effectiveness of detecting toxic comments in multiple languages, we propose a hybrid model, which fuses monolingual model and multilingual model. We use labeled data to fine-tune the monolingual pre-trained model. We use masked language modeling to semi-supervise the fine-tuning of multilingual pre-trained model on unlabeled data and then use labeled data to fine-tune the model. Through this way, we can fully utilize the large amount of unlabeled data; reduce dependence on labeled comment data; and improve the effectiveness of detection. We also design several comparative experiments. The results demonstrate the effectiveness and advantage of our proposed model, especially compared to the XLM-RoBERTa multilingual fine-tuning model.

Keywords: masked language modelling; model fusion; toxic commenting; XLM-RoBERTa

1 INTRODUCTION

With the rapid development of Internet technology, social networks become more and more popular. According to [1], social media users worldwide increased by 13% over the past year and reached around 4.2 billion. Every second, 15.5 users join social media [1]. Social media has made it convenient for users to share their experience, opinions, beliefs, and daily life [1, 3]. However, it is also widely used by abusers. Users could be bullied for their political opinions, religious beliefs, race, color, or even daily pictures for no apparent reason. Toxic comments seriously impact user experience on social media, which could cause mental issues and even lead to tension and chaos [2]. Therefore, toxic comment detection, including identifying toxic language [4, 5], online harassment [6, 7] and cyberbullying [8, 9], has attracted more and more research efforts. It is critical to build a model that automatically detects toxic comments.

Researchers have worked on manual rule construction and simple classifiers for toxic comment detection. Gitari et al. [10] have constructed a vocabulary list, which records hate comments and creates a classifier. Rdulesuc et al. [11] distinguished spam comments from normal comments by calculating the similarity between the comments and related topics in "weibo body". Ravi et al. [12] used machine learning models such as naive Bayes, logistic regression and support vector machines to detect insulting comments.

In recent years, deep learning models overcome the limitations of artificial feature engineering and provide an effective way of detecting toxic comments. Convolutional neural networks (CNN) [13] and recurrent neural networks (RNN) [14] have been widely used as the representative model in toxic comment detection. Spiros [15] and Li, Siyuan [16] used CNN and RNN to obtain a better performance on toxic comment detection than shallow classification. However, CNN and RNN rely on static word embedding vectors. They lack the flexibility for dynamically adjusting the word vector representation based on contexts. Ashwin et al. [17] used toxic comment data to fine-tune pre-trained BERT model [18] and used the model to detect toxic comments. The model has demonstrated its advantage, however, multilingual comments on global social platforms (such as Twitter, Ins, Facebook, etc.) bring new challenges. Even though pre-trained BERT model after fine-tuned is effective to detect malicious comments in a single-language, it lacks versatility when dealing with multilingual comments [19].

Meanwhile, deep learning has the advantage in capturing contextual information, but it relies heavily on the huge labeled corpus and its performance is greatly affected by the quantity and quality of the labeled corpus. In reality, a large amount of unlabeled data has not been utilized. With better utilization of these unlabeled data, the effectiveness of deep learning in detecting toxic comments can be further improved [20]. In order to address the challenges of dealing with multilingual comments, we propose a hybrid detection model, which takes advantage of high accuracy of monolingual model and great generalization of multilingual model. Our model analyzes one comment at a time. Each comment contains vocabulary in the same language. The comments to be analyzed are shown in Fig. 1. E1-E2 are normal comments. E3, E4, and E5 are toxic comments. Take E2, a comment in English, as an example. As shown in Fig. 2, the comment will be the input for multilingual model and monolingual (English) model. Both models will produce their own output probabilities, which will be used to calculate a weighted sum that is the final probability.

E1.不喜欢，但是又不敢说，怕你们攻击。
E2. Oh, he's so handsome. I love him to death.
E3.他也太垃圾了吧，就这样也敢出来混?
E4. She's a shameless bitch.
E5. Und du hast keine angst, dass ich dich umbringe?

Figure 1 Example of multilingual comments

In our proposed model, we fully utilize both labeled and unlabeled data in order to improve the effectiveness of multilingual detection. We use masked language modeling to fine-tune multilingual model, and use unlabeled data to semi-supervise the multilingual pre-trained model to learn the contextual information, and then use labeled data to fine-tune the model. For each monolingual model, the labeled data in the corresponding language are used to fine-

tune the model. Overall, our proposed model provides high accuracy and outstanding generalization performance for multilingual toxic comment detection.

The three contributions of this article are summarized as follows:

1) We use masked language modeling to semi-supervise the fine-tuning of multilingual pre-trained model on unlabeled data and then use labeled data to fine-tune the model.
2) We develop a hybrid model for monolingual and multilingual toxic comment detection that effectively combines monolingual model and multilingual model from different BERT model.
3) We compare various fusion models and determine the optimal hybrid scheme from these models.

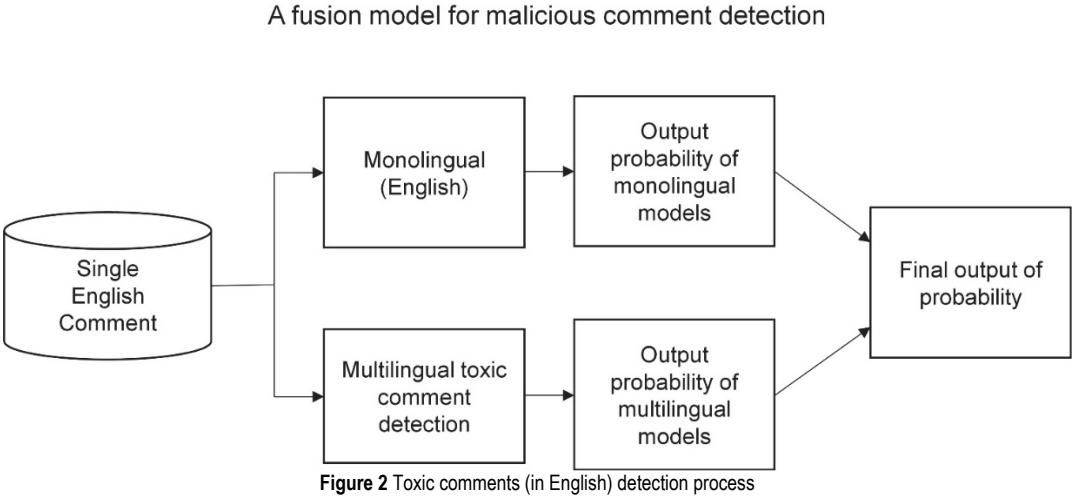

Figure 2 Toxic comments (in English) detection process

The rest of this paper is organized as follows. Section 2 lists related work on monolingual pre-trained models, multilingual pre-trained models, and masked language modeling. Section 3 presents our proposed model which integrates both monolingual model and multilingual model for toxic comment detection. Section 4 provides experiment design and evaluation. Finally, Section 5 concludes the paper.

2 RELATED WORK

2.1 Monolingual Language Pre-trained Model

With the development of deep learning, the number of parameters for models increases rapidly. It will require a larger training data set to train the model and to prevent overfitting [21]. However, it brings huge challenge to build a large-scale labeled data set, while it is relatively easy to obtain unlabeled data. In order to utilize the huge amount of unlabeled data, researchers [22] used different tasks to pre-train models to learn effective representations on a large amount of unlabeled text data. The advantages of pre-trained model can be summarized as follows [23]: (1) Pre-training the model on a huge text corpus enables it to learn general text representations and to solve downstream tasks; (2) The pre-trained model provides better model initialization parameters, which usually leads to better generalization performance and accelerates the convergence; (3) The pre-trained model can be considered as a regularization way to avoid overfitting on a small amount of data. In order to reduce the training data set required for toxic comment detection and accelerate the convergence of model training, in this paper, we use pre-trained BERT model as monolingual model. The model has learned a general text representation through a large amount of single-language text data, which reduces the requirements for training data, speeds up the training process and ensures the effectiveness of detection.

2.2 Multilingual Language Pre-trained Model

In the field of multilingual pre-trained model, the XLM model proposed by Facebook outperforms other models. It is a multilingual model presented as an improved version of BERT [24]. It uses a pre-processing technique Byte-Pair Encoding (BPE) [25] and a dual-language training mechanism with BERT. It enhances BERT for cross-lingual language model [26].

In [27], Alexis et al. proposed an improved model XLM-RoBERTa, which is inspired by XLM model and RoBERTa model. It uses large-scale multilingual corpus such as Wikipedia for model training. It improves the performance of multilingual migration tasks and becomes one of the best pre-trained models in the field of multilingual classification. In this paper, we choose XLM-RoBERTa model as multilingual pre-trained model.

2.3 Masked Language Modelling

Semi-supervised learning is a learning method that combines supervised learning and unsupervised learning [28]. It is an important research topic in machine learning. In practice, we sometimes need a small amount of labeled data and a large amount of unlabeled data to improve the accuracy of classification, to reduce the cost of manually labeling data, and to improve the generalization performance of the model. Masked language modeling [18], a semi-supervised learning method, partially masks and re-predicts the input information to achieve input data learning, which is similar to the denoising autoencoder modeling proposed by Vincent et al. [29]. The difference is that masked language modeling only learns the part that

is masked instead of the entire input sequence [24]. In this paper, we use masked language modeling to fine-tune XLM-Roberta, so that the model can learn text related to toxic comment detection based on unlabeled data, and to improve the convergence speed of the objective function in fine-tuning on labeled data and the performance of data fitting.

3 A HYBRID METHOD FOR MONOLINGUAL AND MULTILINGUAL TOXIC COMMENT DETECTION

In this section, we present the proposed hybrid method for monolingual and multilingual comment detection. Fig. 3 shows the overall framework.

3.1 The Overall Framework of the Method

For data processing, we first preprocess the given corpus, including filtering stop words, numbers and symbols, and converting the text into one-hot vector encoding index required by the pre-trained model, to facilitate subsequent model training. Then, the data will be divided into multilingual unlabeled comment data and multilingual labeled comment data. Finally, the multilingual labeled comment data will be classified based on languages.

For model training, the labeled comment data in different languages are provided to the corresponding single-language pre-trained BERT model for fine-tuning, as explained in Fig. 3 fine tuning (3). Then, the single-language toxic comment detection model in multiple corresponding languages is obtained. For example, if the data set includes Chinese, English and German comments, three monolingual models: BERT-Chinese, BERT-English and BERT-German will be trained. The multilingual comment data will be provided to the pre-trained XLM-Roberta model for fine-tuning (1) and fine-tuning (2). Fine-tuning (1) uses multilingual unlabeled comment data and masked language modeling to fine-tune the XLM-Roberta model. It will preliminarily adjust the parameters of the model to obtain a preliminary fine-tuned XLM-Roberta, as shown in Fig. 3 fine-tuning (1). Fine-tuning (2) uses multilingual labeled comment data to fine-tune the multilingual toxic comment detection model (1), and get the XLM-Roberta model after the second fine-tuning, as shown in Fig. 3.

In the prediction phase, the output probability results from fine-tuning (2) and fine-tuning (3) are weighted and added (as explained in Section 3.4) to obtain the final predicted result.

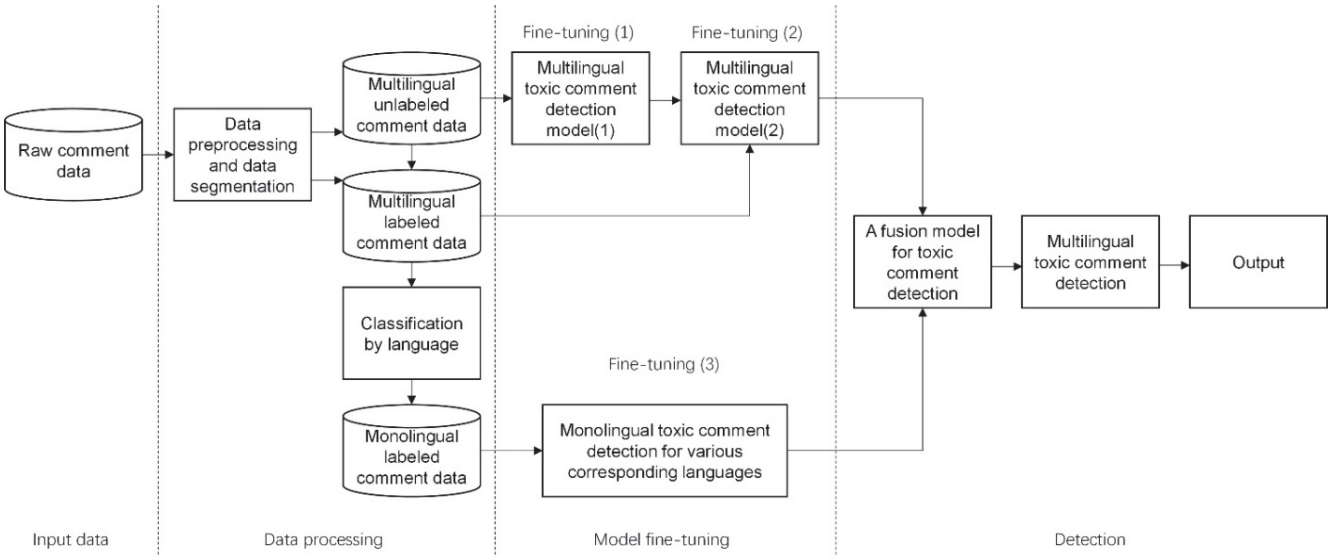

Figure 3 The overall framework of the method

3.2 Semi-supervised Fine-tuning based on Masked Language Modeling

For fine-tuning (1), in order to make the model learn a general representation of data, we use masked language modeling and a large amount of unlabeled data to do semi-supervised training on pre-trained XLM-Roberta model. Similar to the training process of BERT using masked language modeling on large-scale corpus [20], the detailed process is:

(1) Each time, randomly select 15% of the input words to block;

(2) For all blocked words, 80% of the words are marked as [MASK], 10% of the words are randomly replaced with any words, and 10% of the words are replaced with the original words;

(3) We use each mask word T in the previous step as the prediction target, train the model with the cross-entropy loss function, and use masked language modeling to fine-tune the XLM-Roberta model, which is the multilingual toxic comment detection model (2) in Fig. 3.

3.3 Supervised Fine-tuning

As shown in Fig. 4, the supervision and fine-tuning of BERT single-language model and XLM-Roberta monolingual model are mainly composed of word embedding layer, encoder in the Transformer [30], and the downstream structure. For fine-tuning (2) in Fig. 3, after fine-tuned using masked language modeling, XLM-Roberta is fine-tuned using monolingual labeled comment data. For fine-tuning (3), the single-language labeled

comment data is used to fine-tune BERT. The details of the method are shown in Fig. 4. Using the supervised fine-tuning method in BERT as a reference, the output vector C corresponding to [CLS] in a single comment is used as output. The fully connected layer uses sigmoid function to map output to probabilities. The probability distribution predicted by the model using the true label of the sentence together with the cross entropy loss function of the true label distribution is used as the objective function to train the model to obtain the Monolingual detection model (2), and use this as the objective function to train the model to obtain the multilingual model (2) and monolingual model.

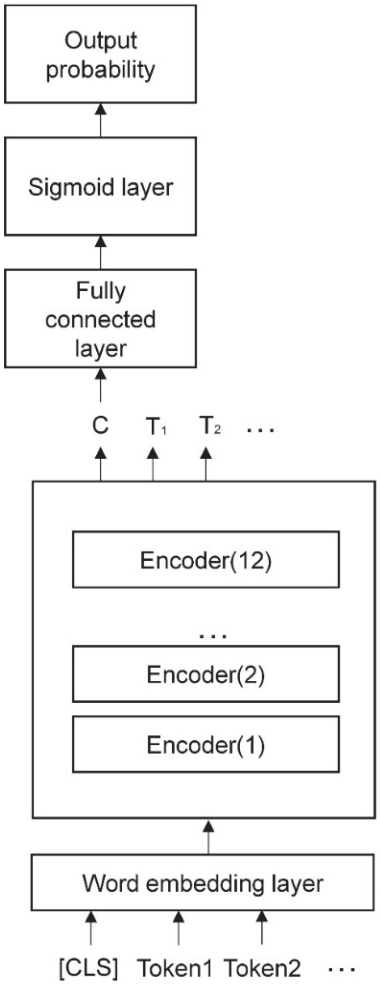

Figure 4 The structure of the model to be supervised and fine-tuned

3.4 Model Fusion

Monolingual model performs classification training on the data of a single-language so it fits the data better. Multilingual model performs mixed classification training with data in multiple languages, and masks language modeling makes the model learn more contextual information. Its advantage lies in better generalization performance. Both monolingual model and multilingual model have their own advantages. In this paper, we fuses monolingual model and multilingual model to improve the overall classification.

In the process of model fusion, for samples in each language, we use cross validation [31] for classifying data multiple times, model training and evaluation, and use F1 scores of monolingual model and multilingual model to determine their contribution ratio. Specifically, on the verification data set, if the F_1 score of the prediction result from multilingual model is F_1^{multi}, and the F_1 score of the prediction result from the monolingual model is F_1^{single}, then the probability of determining the positive class by the fusion model is:

$$\frac{F_1^{\text{multi}}}{F_1^{\text{multi}}+F_1^{\text{single}}}P_{\text{multi}}+\frac{F_1^{\text{single}}}{F_1^{\text{multi}}+F_1^{\text{single}}}P_{\text{single}} \tag{1}$$

In this formula, P_{multi} is the judgment probability of the monolingual model for the positive class, and P_{single} is the judgment probability of the monolingual model for the positive class.

4 EXPERIMENT AND ANALYSIS

4.1 Experimental Data

The existing hate comment data sets are often in a single-language, such as Wikipedia dataset [32], Twitter data set [33], etc. As social networks become a platform for users worldwide, a single-language dataset is not enough. In this paper, we choose the jigsaw multilingual Toxic Comment Classification competition dataset provided by Conversation AI, collected from Wikipedia, including comments that are rude, disrespectful, or offensive (insults, threats, obscenity, racial discrimination, etc.). We classify the above comments as toxic comments, and build a two-classification task for toxic comment detection. After sampling, we got a total of 16941 texts in the training set, including 8941 in English, 3000 in Turkish, 2500 in Spanish, and 2500 in Italian. In the validation set, there are 2542 pieces in total, including 1342 in English, 450 in Turkish, 375 in Italian data, and 375 in Spanish. There are in total 30932 texts of unlabeled data, including 14000 in Turkish, 8494 in Italian, and 8438 in Spanish.

4.2 Experimental Evaluation Index

In order to evaluate the model comprehensively [34], we select accuracy, precision, recall and F_1 score as the evaluation metrics [35].

The F_1 score is calculated as follows:

$$P_{\text{ma}}=\frac{\sum_{i=1}^{N}P_i}{N} \tag{2}$$

$$R_{\text{ma}}=\frac{\sum_{i=1}^{N}R_i}{N} \tag{3}$$

$$F_{1\text{ma}}=\frac{2\times P_{\text{ma}}\times R_{\text{ma}}}{P_{\text{ma}}+R_{\text{ma}}} \tag{4}$$

P_{ma} represents the value of precision rate. R_{ma} represents the value of recall rate. $F_{1\text{ma}}$ score is the combined expression of precision rate and recall rate.

4.3 Experimental Comparison Model

In order to verify the effectiveness of the proposed model, we design multiple sets of comparative experiments. The pre-trained model XLM-RoBERTa is chosen to be the basic model for malicious comment detection, and compared with other models, as follows:

Baseline: A multilingual toxic comment detection model obtained by fine-tuning the pre-trained model XLM-RoBERTa on multilingual annotated comment data;

BERT_base: A single-language toxic comment detection model obtained by fine-tuning the pre-trained model BERT on single-language labeled comment data;

XLM-R_MLM: The pre-trained model XLM-RoBERTa after fine-tuning on multilingual unlabeled comment data and multilingual labeled comment data;

Ensemble: We name our model as Ensemble, which is a fusion model for multilingual malicious comment detection after fusing XLM-R_MLM and BERT_base.

4.4 Experimental Environment and Parameter Settings

This article uses Tensorflow2.3 and the Keras framework for model construction, and uses the TPU server on the Google cloud for training. The dataset is the jigsaw multilingual Toxic Comment Classification competition data set as described in step 4.1.

Based on sentences' length distribution, hardware conditions and experimental efficiency, we set parameters as follows. The training batch size is 16. The word vector dimension is 768. The maximum length of input sentences for all models is 224. If an input is longer than the limit, the extra part is truncated. Otherwise, if an input is shorter, <PAD> mark is used to complete the shortness. The main parameters used in the model are shown in Tab. 1.

Table 1 Main parameters of the model

Parameter Name	Parameter value
Fully Connected Layers Number	1
Learning Rate	$9e^{-6}$
Training Round Times	3
Training Batch Size	16
Input Sentence Length	224
Word vector dimension	768

4.5 Experimental Results and Analysis

In order to evaluate the performance of our proposed model, we design the following comparison experiments.

4.5.1 Model Training Process Comparison Before and After Semi-supervised Fine-Tuning

To examine the effectiveness of masked language modeling tasks, we use every 10 training batches as a unit to track and record multiple metrics of Baseline and XLM-R_MLM, including function loss value (Loss), function loss value on the validation set (Validation Loss), and F_1 value on the validation set (F_1 Score). As shown in Fig. 5, the model after semi-supervised fine-tuning not only accelerates the convergence speed of training, but also reduces the value of the loss function after convergence to a certain extent. In terms of Loss and F_1 value after the function converges, the semi-supervised fine-tuned model also has better performance on the verification data, which indicates the improvement on the training effectiveness of the model, the ability to fit task data and the generalization.

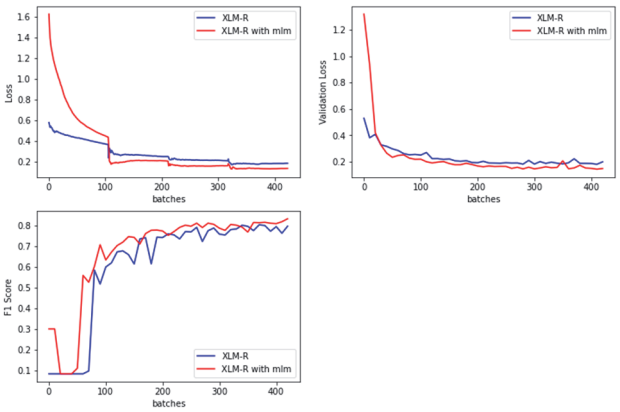

Figure 5 Comparison of the model training process before and after semi-supervised fine-tuning

4.5.2 Comparison of Test Results of Each Comparison Model

To examine the effectiveness of the fusion model (Ensemble), the performance of under the four evaluation metrics on the multilingual review test set and the performance of under each evaluation metrics on the verification set of their corresponding languages are plotted in Fig. 6.

Fig. 6 shows the performance of the fusion model, XLM-R_MLM, Ensemble, and BERT based models. For each model, we evaluate four metrics as shown in the figure. Compared with XLM-R_MLM, the fusion model Ensemble has a better (0.39%) accuracy of the macro, and higher (0.0062 higher) average F_1 score. Moreover, the prediction of the fusion model in Italian and Spanish outperforms the monolingual model, and the prediction effect in English and Turkish is close to the monolingual model. This result may be due to the fact that English and Turkish have relatively more training and testing samples than Italian and Spanish. It indicates that the fusion model provides high accuracy and strong generalization capability, which improves the effectiveness of multilingual toxic comments detection.

5 CONCLUSION

In this paper, we propose a hybrid model using both monolingual model and multilingual model. We use semi-supervised learning and supervised learning to fine-tune multilingual pre-trained model. We also use supervised learning to fine-tune the monolingual pre-trained model. Through this way, our model reduces the dependence on labeled data and improves flexibility. In addition, our model combines the advantages of monolingual model and multilingual model, and uses model fusion to improve generalization performance and the effectiveness of detection.

As our future work, the toxic comment detection can be refined to be multiple classifications. We will also refine our model to handle more complicated dataset, for example, a single comment contains words in multiple languages.

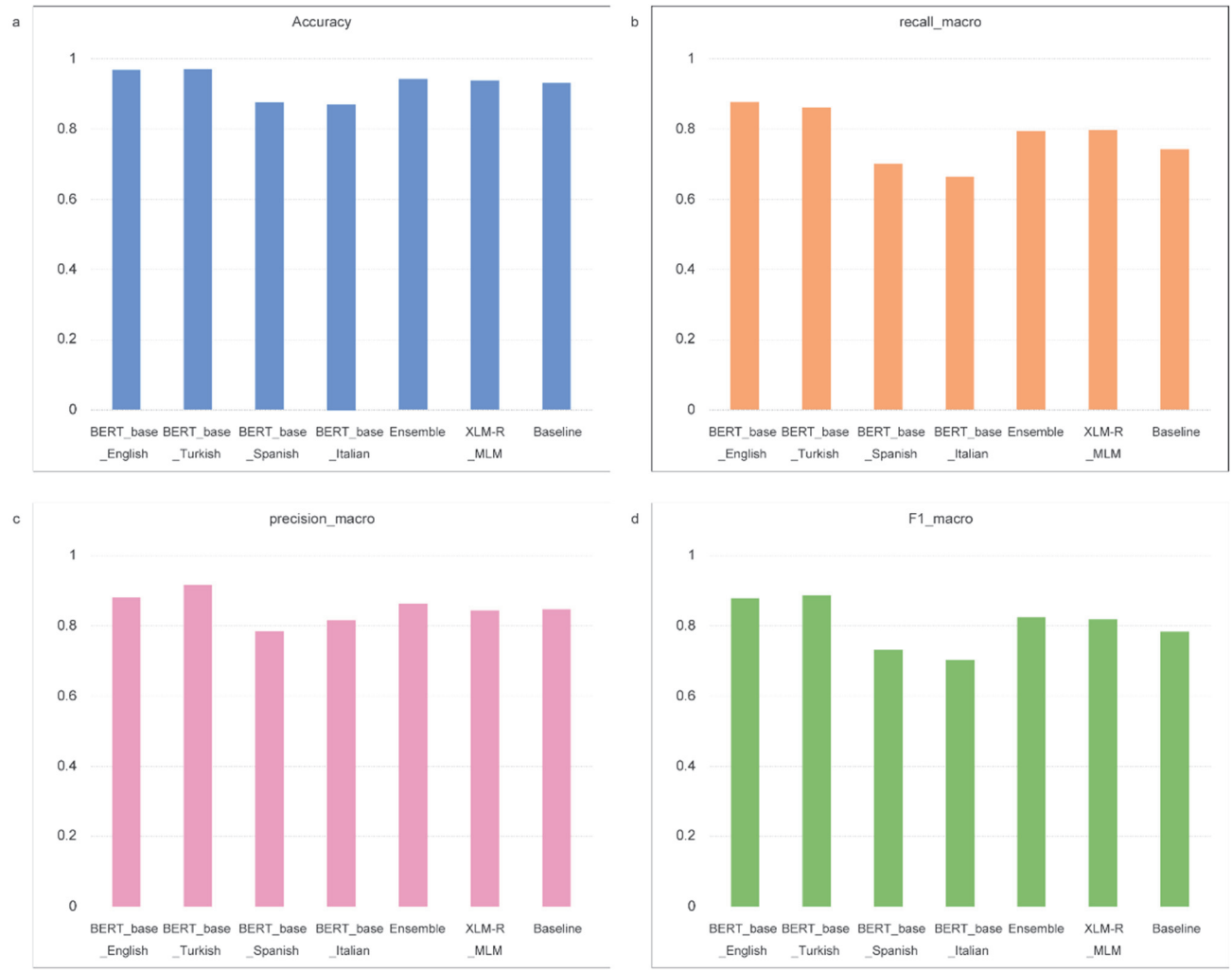

Figure 6 Comparison of prediction effects of various comparison models

6 REFERENCES

[1] See https://blog.hootsuite.com/simon-kemp-social-media/#:~:text=Social%20media%20user%20numbers%20jumped,new%20users%20every%20single%20second

[2] Parekh, P. & Patel, H. (2017). Toxic comment tools: A case study. *International Journal of Advanced Research in Computer Science, 8*(5).

[3] See https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

[4] Leite, J. A., Silva, D. F., Bontcheva, K., & Scarton, C. (2020). Toxic language detection in social media for brazilian portuguese: New dataset and multilingual analysis. *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 914-924.

[5] Karan, M. & Šnajder, J. (2019). Preemptive toxic language detection in Wikipedia comments using thread-level context. *Proceedings of the Third Workshop on Abusive Language Online*, 129-134. https://doi.org/10.18653/v1/W19-3514

[6] Blackwell, L., Chen, T., Schoenebeck, S., & Lampe, C. (2018). When online harassment is perceived as justified. *Proceedings of the International AAAI Conference on Web and Social Media, 12*(1).

[7] Kennedy, G., McCollough, A., Dixon, E. et al. (2017). Technology solutions to combat online harassment. *Proceedings of the first workshop on abusive language online*, 73-77. https://doi.org/10.18653/v1/W17-3011

[8] Song, T. M. & Song, J. (2021). Prediction of risk factors of cyberbullying-related words in Korea: Application of data mining using social big data. *Telematics and Informatics, 58*, 101524. https://doi.org/10.1016/j.tele.2020.101524

[9] Van Hee, C., Jacobs, G., Emmery, C. et al. (2018). Automatic detection of cyberbullying in social media text. *PloS one, 13*(10). https://doi.org/10.1371/journal.pone.0203794

[10] Gitari, N. D., Zhang, 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. https://doi.org/10.14257/ijmue.2015.10.4.21

[11] Rădulescu, C., Dinsoreanu, M., & Potolea, R. (2014). Identification of spam comments using natural language processing techniques. *2014 IEEE 10th International Conference on Intelligent Computer Communication and Processing (ICCP)*, 29-35. https://doi.org/10.1109/ICCP.2014.6936976

[12] Ravi, P. (2012). *Detecting Insults in Social Commentary*.

[13] Mikolov, T., Kombrink, S., Burget, L., Cernocky, J., & Khudanpur, S. (2011). Extensions of recurrent neural network language model. *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 5528-5531. https://doi.org/10.1109/ICASSP.2011.5947611

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

[15] Georgakopoulos, S. V., Tasoulis, S. K., Vrahatis, A. G., & Plagianakos, V. P. (2018). Convolutional neural networks for toxic comment classification. *Proceedings of the 10th Hellenic Conference on Artificial Intelligence*, 1-6. https://doi.org/10.1145/3200947.3208069
[16] Li, S. (2018). *Application of recurrent neural networks in toxic comment classification*. Thesis, UCL.
[17] D'Sa, A. G., Illina, I., & Fohr, D. (2020). BERT and fastText Embeddings for Automatic Detection of Toxic Speech. *SIIE 2020-Information Systems and Economic Intelligence.* https://doi.org/10.1109/OCTA49274.2020.9151853
[18] Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding.
[19] Fan, A., Bhosale, S., Schwenk, H. et al. (2021). Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research, 22*(107), 1-48.
[20] Najafabadi, M., Villanustre, F., Khoshgoftaar, T., & Seliya, N. (2015). Deep learning applications and challenges in big data analytics. *Journal of Big Data, 2*(1). https://doi.org/10.1186/s40537-014-0007-7
[21] Taylor, L. & Nitschke, G. (2018). Improving deep learning with generic data augmentation. *IEEE Symposium Series on Computational Intelligence (SSCI).* https://doi.org/10.1109/SSCI.2018.8628742
[22] Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., & Huang, X. (2020). Pre-trained models for natural language processing: A survey. *China Technology Science, 63*, 1872-1897. https://doi.org/10.1007/s11431-020-1647-3.
[23] Erhan, D., Bengio, Y., Courville, A., Manzagol, P., & Vincent, P. (2010). Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research, 11*, 625-660.
[24] Lample, G. & Conneau, A. (2019). Cross-lingual language model pretraining.
[25] Shibata, Y., Kida, T., Fukamachi, S., & Takeda, M. (1999). *Byte Pair encoding: A text compression scheme that accelerates pattern matching*. Technical Report DOI-TR-161, Department of Informatics, Kyushu University.
[26] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *Roberta: A robustly optimized bert pretraining approach*.
[27] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Myle Ott, L., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8440-8451. https://doi.org/10.18653/v1/2020.acl-main.747
[28] Van Engelen, J. E. & Hoos. H. H. (2020). A survey on semi-supervised learning. *Machine Learning, 109*(2), 373-440. https://doi.org/10.1007/s10994-019-05855-6
[29] Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P. (2008). Extracting and composing robust features with denoising autoencoders. *Proceedings of the 25th international conference on Machine learning*, 1096-1103. https://doi.org/10.1145/1390156.1390294
[30] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *The 31st Conference on Neural Information Processing Systems (NIPS 2017).*
[31] Browne, M. (2000). Cross-validation methods. *Journal of mathematical psychology, 44*(1). 108-132. https://doi.org/10.1006/jmps.1999.1279
[32] Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). *Automated hate speech detection and the problem of offensive language*.
[33] Mohammad, S., Bravo-Marquez, F., Salameh, M., & Kiritchenko, S. (2018). Semeval-2018 task 1: Affect in tweets. *Proceedings of the 12th international workshop on semantic evaluation*, 1-17. https://doi.org/10.18653/v1/S18-1001
[34] Ganapathi, A., Kuno, H., Dayal, U., Wiener, J. L., Fox, A., Jordan, M., & Patterson, D. (2009). Predicting multiple metrics for queries: Better decisions enabled by machine learning. *2009 IEEE 25th International Conference on Data Engineering*, 592-603. https://doi.org/10.1109/ICDE.2009.130
[35] Hossin, M. & Sulaiman, M. (2015). A Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining & Knowledge Management Process, 5*(2), 1-11. https://doi.org/10.5121/ijdkp.2015.5201

Contact information:

Guizhe SONG, PhD
Dalian University of Technology,
School of Computer Science and Technology,
No. 2 Linggong Road, Ganjingzi District, Dalian City,
Liaoning Province, P. R. China, 116024
E-mail: labyrinth369@126.com

Degen HUANG, PhD, Professor
Dalian University of Technology,
School of Computer Science and Technology,
No. 2 Linggong Road, Ganjingzi District, Dalian City,
Liaoning Province, P. R. China, 116024

Yanping ZHANG, PhD, Associate Professor
(Corresponding author)
Gonzaga University,
Department of Computer Science,
502 East Boone Avenue,
Spokane, WA 99258-0102, Canada
E-mail: zhangy@gonzaga.edu