yasuo at HASOC2020: Fine-tune XML-RoBERTa for Hate Speech Identification

Li Xu, Jun Zeng and Shi Chen

School of Information Science and Engineering, Yunnan University, Kunming, P.R. China

Abstract

In recent years, people are more concerned about hate speech identification and identification than ever. This paper describes our system for English and German Sub-Task A in HASOC2020. For these subtasks, we fine-tune the XLM-RoBERTa pre-training model for sentence embedding and extract the layer with the best performance for slicing and splicing. In order to make full use of both English and German corpus, we propose a multi-task method to optimize two classification tasks at the same time. Our model has achieved 0.9076 for F1 score in English Sub-Task A and 0.8165 in German Sub-Task A.

Keywords

Hate speech, XLM-RoBERTa, fine-tune, slicing and splicing

1. Introduction

With the rapid development of the Internet, communication between humans becomes more convenient through social media. Individuals can publish their own opinions freely on the Internet. But every coin has two sides, free speech often leads to sexism, racism or other aggressive behaviors [1] and cyberbullying [2, 3]. Over the past decade, as global hatred and bigotry spread through social media, ethnic minorities around the world are facing new and growing threats. Measures should be taken to reduce the spread of hate speech on social [4] .

However, social media has encountered many difficulties in detecting hate speech because of its close association with other forms of abusive language [5]. The multiplicity of languages and slang adds to the complexity. With the increasing number of tweets posted online, manually monitoring hate speech is not a viable solution, HASOC 2020 provides a forum and a competition for multilingual research on identification of hate speech to solve the problem that human surveillance always lacks of scalability by automatically identifying hate speech content[6].

In this paper, we concentrate on detecting multilingual hate speech which are written in English and German and detail our solution. We fine-tune the XLM-RoBERTa pre-training model for sentence embedding and extract the layer with the best performance for slicing and splicing. To make full use of both English and German corpus, we further propose a multi-task method to optimize two classification tasks at the same time.

2. Related Works

Detecting abusive language in a sea of data on social media is a difficult and arduous work, and researches have only conducted in recent years. Some research shows that the deep learning model with word-embedding can achieve better results in text classification tasks. As a result, Word2Vec is commonly used to obtain semantic information and attributes in the text through an unsupervised word embedding method. Common machine learning algorithms include Logistic Regression(LR), Support Vector Machine(SVM), Random Forest, etc. For deep learning methods, most of them are based on Long-Short Term Memory network (LSTM), convolutional neural network (CNN) or a recursive neural network (RNN). Some of the early work use features like bag of words, word and character n-grams with relatively machine learning classifiers for detection (Dinakar et al.[7]; Waseem and Hovy[8]; Nobata et al.[5]). Kim et. al[9] use CNNs for sentiment classification, It requires very few hyperparameter adjustments and static vectors to achieve good results on multiple baseline. MacAvaney et al.[10] propose a multi-view SVM approach that achieves near state-of-the-art performance, while being simpler and producing more easily interpretable decisions than neural methods.

More recently, BERT released by Google gains more attention in the research community, as it can capture both long-distance reliance and true bidirectional context information compared with the traditional way of deep learning.

BERT [11] and other models have achieved good results in monolingual NLP tasks, but for NLP in addition to English, researchers have cultivated more and more monolingual models for a variety of different languages. At the same time, there appears to be an alternative approach that has received little attention: the multilingual model. XLM[12] is a cross-language pre-training model that extends the full-length and training strategy MLM (Masked Language Model) proposed in BERT to multiple languages, and has been experimentally proven to take effect.

RoBERTa [13] is an upgraded version of BERT. Compared to BERT, it uses a larger number of model parameters, a larger batch size and more training data. It is built on the BERT language masking strategy, which modifies key hyperparameters in BERT, including deleting BERT's next sentence prediction task, which enables RoBERTa representation to be better expansibility to downstream tasks than BERT.

As for the XLM-RoBERTa model, it combines these advantages of both. We fine-tune XLM-RoBERTa and use multitask training to improve the performance of prediction. This enables our model to achieve good results in the category of hate text.

3. Dataset and Task description

The dataset of HASOC task A contains over 10,000 annotated tweets respectively composed of userID, tweets, and labels from Twitter. Through our analysis on the dataset, English dataset has a uniform distribution, while German dataset is uneven.

We took part in the English and German Sub-Task A, which involves building a coarse-grained binary categorization model to test whether a text is offensive or insulting (HOF). If a text contains any form of unacceptable verbal, aggression and profanity, it will be considered as

Table 1
The initial statistics about training and valid data

Language	Type	Not Hate/Offensive/Profane	Hate/Offensive/Profane	Total
English	Train	3040	2968	6008
English	Valid	760	742	1502
English	Test	457	423	880
German	Train	3164	1206	4370
German	Valid	791	302	1093
German	Test	396	145	541

HOF.

4. Our Solution

Our solution is specifically divided into data preprocessing, feature extraction and model structure.

4.1. Data Preprocessing

Because there was noise in tweets in official dataset, which would affect the performance of model training, we cleaned the data before training. The steps are as following:

- Keep the label
- Handle user name and @ beginning uniformly as "username"
- Separate conjunction
- Turn Emojis expression into the corresponding phrase
- Remove words that have no emotional meaning
- Remove all URLs
- Convert text to lowercase
- Numbers are normalized to strings as "number"

After data preprocessing, we combined the English and German datasets as one dataset. In order to improve the generalization ability of model, let the model learn the real data characteristic distribution, the method of stratified sampling is adopted to ensure the training set and the valid set have the same data distribution.

Besides, we divide 20% as the valid set, 80% as the train set. Table 1 shows the distribution of HOF and NOT in the dataset. The test set is given by official.

4.2. Feature vectors

XLM-RoBERTa embeddings: We utilized XLM-RoBERTa[14] for embedding, which is a multilingual model with Transformer for the major structure. There are 12 layers of the model, output with 768 dimensions.
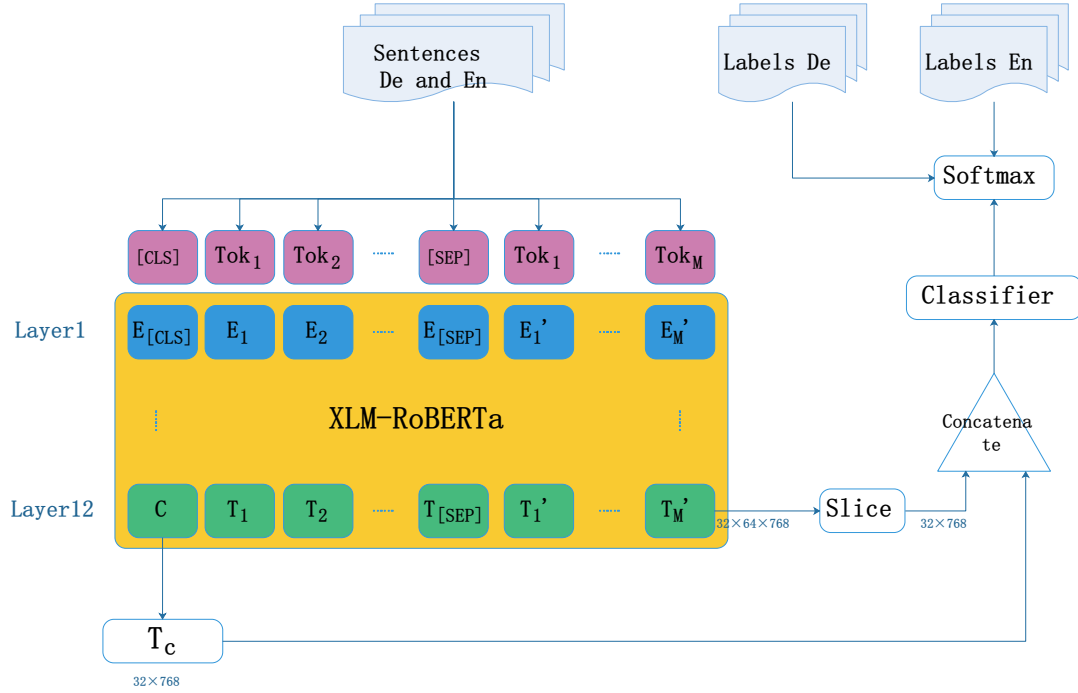
Figure 1: The architecture of our model

4.3. Our Model

After getting the embedded vectors of the texts, we fine-tuned XLM-RoBERTa to make it more suitable for the downstream task of hate speech identification. XLM-roBERTa has a total of 12 layers which learn different semantic information. Generally speaking, the shallower the layers, the more word level semantic information is learned. The deeper layers, the more generalized semantic information is learned. Whereas T_c([CLS]vectors) contains the semantic information for classification of the entire sentence, we try to combine T_c with vectors at a certain layer to improve training performance. The influence of each layer is given in Table 2.

Global semantic information is more helpful for binary classification such as sub-Task A. The hidden layer of the XLM-RoBERTa model is 768 dimensions, with 12 layers of Transformer. Because the shape of T_c is [32,768] and the hidden vector of the 12th layer's shape is [32,60,768], we take out 12th layer as dim= (0,2) and splice together with T_c, then passed it to classifier and send the result to Softmax. We trained both English and German tasks by feeding the processed data into the model.

Multitask learning can learn useful information from similar tasks. All tasks will share XLM-RoBERTa's layers. In order to take full advantage of each classification corpus, two classification tasks are combined to carry out multi-task training. In this way, the number of datasets can also be expanded.

Table 2
Performance of each layer

Layer	Test error rates(%)
Layer-1	11.07
Layer-2	9.81
Layer-3	9.29
Layer-4	8.66
Layer-5	7.83
Layer-6	6.83
Layer-7	6.83
Layer-8	6.41
Layer-9	6.04
Layer-10	5.70
Layer-11	5.46
Layer-12	**5.42**
First 4 Layers	8.78
Last 4 Layers	5.43
All 12 Layers	6.88

5. Result

5.1. Baslines

LR is a machine learning method to solve the problem of categorization (0 or 1). We use it as the baseline classifier for both English and German datasets. The configuration is as follows: we use L2 regularization with the hyper parameter C=1.2 (Inverse of regularization strength) and use TF-IDF features of word n-grams(1,6) for training the classifier.

SVM is a binary classification model [15]. The basic method is to solve the separated hyperplane that can correctly divide the training dataset and has the maximum geometric spacing. We use it as the baseline classifier for both English and German datasets. The configuration is as follows: we uses the 'linear' kernel, L2 regularization with the hyper parameter C=1.0 (Inverse of regularization strength) , and the same TF-IDF features of word n-grams(1,6) to train the classifier.

BiLSTM[16] model is implemented with 100 units, adopts sigmoid activation. For training, *binary cross entropy loss function* and adam optimizer are used. As for regularization, 50% dropout is configured. For English and German subtasks, we respectively use 300 dimensional English fastText embeddings, 300 dimensional German fastText embeddings to initialize the word vectors.

5.2. Comparison

XLM-RoBERTa model and other model's accuracy and F1 macro-average score in English and German are showed in Table 3 and Table 4. To verify the effectiveness of the fine-tuning strategies, ablation experiments are conducted. We also using BERT base model and $XLM-RoBERTa_{base}$ as the baseline.

Table 3

The result of the English dataset under the test set

Approach	Features	Acc(%)	F1(%)
LR	Char n-grams (1,6)	63.65	63.25
SVM	Char n-grams (1,6)	67.83	64.32
BiLSTM	pre-trained fastText	69.79	67.36
BERT base	-	88.96	88.52
$XLM - RoBERTa_{base}$	-	89.78	89.65
$XLM - RoBERTa_{fine-tuned}$	-	**90.79**	**90.76**

Table 4

The result of the German dataset under the test set

Approach	Features	Acc(%)	F1(%)
LR	Char n-grams (1,6)	62.80	62.25
SVM	Char n-grams (1,6)	66.79	63.13
BiLSTM	pre-trained fastText	73.39	69.16
BERT base	-	79.49	79.36
$XLM - RoBERTa_{base}$	-	80.93	80.61
$XLM - RoBERTa_{fine-tuned}$	-	**81.67**	**81.65**

In the two subtasks, BERT and XLM-RoBERTa models performance better than LR, SVM and BiLSTM models. It may because deep learning has great advantages over traditional machine learning methods in document classification. For traditional machine learning methods, it is not easy to extract text features. Moreover, these features cannot well represent the semantics and syntax of the document, and a large part of useful information is lost. Deep learning is to hand over the feature extraction to deep network for automatic completion. Higher computational costs in exchange for more comprehensive and better text features. So deep learning methods performance better in our hate speech identification tasks.

These experiments prove that our fine-tune XLM-RoBERTa model is effective for both German and English tasks. Table 3 and 4 show the comparison between different models where deep learning models performs better than traditional machine learning models.

Our system ranked 16th in the German subtask and 34th in the English subtask, F1 macro-average Score for German subtask was 0.4968 (Top team was 0.5235) under the official private dataset, F1 macro-average score for English subtask was 0.4856 (Top team was 0.5152) under the official private dataset.

6. Conclusions

We have proposed a neural solution with fine-tuning XML-RoBERTa for hate speech identification. Particularly, the output and hidden layers are slicing and splicing, which solves the sparsity of data and increases the generalization ability of the model in a multi-task way. Experiments have proved the competitiveness of our method.

Acknowledgments

References

[1] R. Kumar, A. K. Ojha, S. Malmasi, M. Zampieri, Benchmarking aggression identification in social media, in: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), 2018, pp. 1–11.

[2] C. Chelmis, D.-S. Zois, M. Yao, Mining patterns of cyberbullying on twitter, in: 2017 IEEE International Conference on Data Mining Workshops (ICDMW), IEEE, 2017, pp. 126–133.

[3] M. Yao, C. Chelmis, D.-S. Zois, Cyberbullying detection on instagram with optimal online feature selection, in: 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE, 2018, pp. 401–408.

[4] J. Waldron, The harm in hate speech, Harvard University Press, 2012.

[5] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, Y. Chang, Abusive language detection in online user content, in: Proceedings of the 25th international conference on world wide web, 2016, pp. 145–153.

[6] T. Mandl, S. Modha, G. K. Shahi, A. K. Jaiswal, D. Nandini, D. Patel, P. Majumder, J. Schäfer, Overview of the HASOC track at FIRE 2020: Hate Speech and Offensive Content Identification in Indo-European Languages), in: Working Notes of FIRE 2020 - Forum for Information Retrieval Evaluation, CEUR, 2020.

[7] K. Dinakar, R. Reichart, H. Lieberman, Modeling the detection of textual cyberbullying, in: In Proceedings of the Social Mobile Web, Citeseer, 2011.

[8] Z. Waseem, D. Hovy, Hateful symbols or hateful people? predictive features for hate speech detection on twitter, in: Proceedings of the NAACL student research workshop, 2016, pp. 88–93.

[9] Y. Kim, Convolutional neural networks for sentence classification, arXiv preprint arXiv:1408.5882 (2014).

[10] M. Sap, D. Card, S. Gabriel, Y. Choi, N. A. Smith, The risk of racial bias in hate speech detection, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 1668–1678.

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

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

[13] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, Albert: A lite bert for self-supervised learning of language representations, arXiv preprint arXiv:1909.11942 (2019).

[14] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave,

M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, arXiv preprint arXiv:1911.02116 (2019).

[15] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, N. Bhamidipati, Hate speech detection with comment embeddings, in: Proceedings of the 24th international conference on world wide web, 2015, pp. 29–30.

[16] A. Baruah, F. Barbhuiya, K. Dey, Abaruah at semeval-2019 task 5: Bi-directional lstm for hate speech detection, in: Proceedings of the 13th International Workshop on Semantic Evaluation, 2019, pp. 371–376.

A. Online Resources

- GitHub