Evaluate BERT and XLM-R models for Chinese News Topic Prediction

Luna Liu
yuli8896@colorado.edu
Matt Niemiec
matthew.niemiec@colorado.edu

Qiuyang Wang
qiwa8995@colorado.edu
Xinyu Jiang
xiji6874@colorado.edu

Abstract

Multi-lingual text classification gives brings a variety of challenges, but can have many benefits. Among the top-performing models in this field, BERT gives state-of-the-art results, and XLM-R, while less tested, shows great promise. We seek to analyze the two models side by side by training on and classifying Chinese news articles. Our findings reveal that BERT yields an impressive 95% accuracy, while XLM-R struggles to reach 91%. In addition to highlighting the importance of proper use of a model, one compelling reason for these results is BERT's robustness, even on medium-resource languages such as Chinese.

1 Introduction

News topic prediction is crucial for news industries around the world and society as a whole. For example, when the public health sector needs to communicate to the public, it is useful to be able to classify a news article so that the general public can readily access that information. Besides the application in public health, transforming unstructured data into distinct categories is also useful when compiling a list of relevant data and information on a subject. In the field of natural language processing (NLP), news topic prediction is always an important task. For example, a large volume of studies use news content to predict the trend in the stock market (Nikfarjam et al., 2010; Vargas et al., 2017), as well as monitor public health (Ng et al., 2020; Mahabaleshwarkar et al., 2019).

Furthermore, much of the data that exists in the world today is not in English, but another language. Therefore, it is important to be able to have NLP models that can analyze data in different languages, though this can prove challenging in more complex languages like Chinese. There is an abundance of data written in Chinese, and to make any use of it, it is important to analyze the top-performing models in Chinese text classification.

In this study, we explore text classification by implementing two pre-trained models, mBERT and XLM-R, on Chinese news articles. We first fine-tune each model on the training data. Then, we conduct a comparison by analyzing the performance of each model on the validation data. With these results, we compare the accuracy, F1 score, and confusion matrices of each model for deeper analysis.

2 Related Work

2.1 BERT and Chinese BERT-wwm

Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) is one of the highest-performing pre-trained models in NLP. The key technical innovation behind BERT is the bidirectional pre-training for language representations, which has helped overcome some of the previous limitations with standard, unidirectional pre-training. In order to conduct bidirectional pre-training, BERT uses two unsupervised tasks, masked language model (MLM) and next sentence prediction (NSP). These innovations greatly facilitate the fine-tune process, and has helped BERT achieve state-of-the-art performance for a large number of NLP tasks.

Recently, an updated version of BERT, BERT Whole Word Masking (BERT-wwm) was released. In the original BERT model, 15% of the WordPiece token was masked randomly in each sentence (Devlin et al., 2019). By contrast, BERT-wwm masks all WordPieces that belong to a whole word together (Cui et al., 2019). This is especially important for Chinese because most Chinese words are made up of several characters. Cui et al. (2019) then adapted the whole word masking strategy in Chinese BERT and re-trained the Chinese BERT model. Their experiment results have shown that compared to mBERT, Chinese BERT-wwm improved significantly on long sequences datasets.

2.2 XLM-R

XLM-R, or XLM-RoBERTa, is a cross-language model that emphasises pre-training multilingual language models (Conneau et al., 2019). It was pre-trained from more than two terabytes of CommonCrawl data in 100 languages and has been shown to keep critical information without sacrificing per-language performance. XLM-R uses self-supervised training technology to achieve state-of-the-art performance for cross-language understanding,. It even performs much higher than multilingual BERT (mBERT), the previous state-of-the-art, on many tasks. XLM-R performs especially well on low-resource languages, such as Swahili and Urdu (Conneau et al., 2019).

Because of XLM-R's strong performance on a variety of tasks, it should be considered when running non-English NLP tasks, as we do in this paper. XLM-R is trained on 46.9GB of Chinese data alone, which is a lot, but still less than Bulgarian, Persian, Finnish, Norwegian, Romanian, Thai, and 11 other languages (Conneau et al., 2019). Therefore, while Chinese may have more resources than Urdu and Swahili, it is certainly not one of the highest-resource languages, which makes XLM-R a strong candidate for our task.

2.3 Use of Chinese Language in BERT and XLM-R

The Chinese language provides us with a strong opportunity to perform these analyses. For example, some variations of BERT mask whole words for the analysis. This makes for an interesting comparison with XLM-R, a strong multilingual model that excels without knowing a lot about a language. This isn't the first time the two models have been compared in a non-English context. Putra and Purwarianti (2020) compare the performance of XLM-R and mBERT when adding English data to classify Indonesian text for sentiment analysis and hate speech. When showing that adding English data did improve performance, they also demonstrated that XLM-R performs higher on the classifications than mBERT, and that XLM-R benefited more from the added cross-linguistic data.

Thus, this study proposes the following hypothesis and research question:

Hypothesis 1: *The XLM-R model will perform better than Chinese BERT-wwm on text classification.*

Research Question 1: *How does the XLM-R model (or Chinese BERT-wwm model) perform better?*

3 Method

3.1 Process

In order to evaluate the quality of the two models, we perform a controlled experiment, classifying a data set of Chinese news articles. The two models perform exactly the same classification task, so the results are comparable. Given the abundance of text and useful keywords in each news article, it is very likely that both models will perform extremely well. However, we are more interested in taking a deep dive into the error analysis, and exploring the details regarding how does one model performs better.

3.2 Data Used In Experiments

We use a dataset from Sina News that contains a variety of long news articles from different topics. The dataset is broken up into two columns - the category, of which there are ten, and then the text of the article itself. The ten types of articles are *sports, entertainment, home, real estates, education, fashion, politics, game, technology*, and *finance*. This dataset including 65,000 news text in total. Table 1 shows the distribution of categories in the dataset. We allocated 77% of the overall dataset for training data, 15% for test data, and 8% for validation data.

Sports	**Entert.**	**Home**	**Real.E.**	**Edu.**
6,500	6,500	6,500	6,500	6,500
Fashion	**Politics**	**Game**	**Tech.**	**Finance**
6,500	6,500	6,500	6,500	6,500

Table 1: Distribution of Categories in Data

We cleaned and shuffled the dataset for better performance. Although the data files we acquired from GitHub are csv files, the labels and sentences were not separated by comma, and the data was not properly formatted. So, we cleaned it and converted the original csv dataset to tsv for better efficiency. Because unshuffled data results in extremely low accuracy and high loss (Si et al., 2019), we randomly shuffle all the articles in the dataset. To do this, we first split up the articles according to their labels.

3.3 Experiment of Chinese BERT-wwm

When applying BERT to an NLP task, we first need to fine-tune BERT for the specific task. To do this, we adopt the method introduced in Devlin et al. (2019) paper, which adds the final classification layer weights $W \in \mathbf{R}^{K \times H}$, where K represents the number of labels.

We use a batch size of 32, 3 epochs, and a learning rate of 3e-5 to train on the classification task.

3.4 Experiment of XLM-R

When performing experiments with XLM-R, we first apply vectorization and word embedding into the training, validation and test data using XLM-R.

Next, we apply Logistic Regression and Random Forest classifiers to fine-tune XLM-R for the text classification task. Finally, in order to check if there is overfitting during the training, we print out the confusion matrix and the F1 score. The results show that there is no overfitting in our training process.

3.5 Comparison experiment of Chinese BERT-wwm and XLM-R

The above experiments have tested each model's performance on Chinese text classification separately. However, BERT-wwm and XLM-R use Tensorflow word embedding and Pytorch word embedding, respectively. Although the results were encouraging, we cannot compare them and draw final conclusion.

We conduct a variable control experiment between BERT-wwm and XLM-R by applying the same word embedding to each. Specifically, we use bert seriving client to generate the word embedding file, and apply it to BERT-wwm. We then apply logistic regression and a random forest classifier on both XMLR and BERT-wwm's word embedding file in order to fine-tune the model for the classification task. This procedure ensures the results can be compared.

4 Results

4.1 Findings

Table 2 shows the overall results for each experiment. The results for the Chinese BERT-wwm experiment are heartening. BERT-wwm with weighted final classification layer shows the best performance among all models, with an accuracy of 96.8%. While looking at the equivalent results from BERT-wwm + LR and XLM-R + LR, BERT-wwm still outperforms XLM-R. Therefore, Hypothesis 1 is rejected. We conclude that Chinese BERT-wwm performs better than XLM-R on Chinese news text classification.

The following findings further demonstrate that BERT-wwm performs better, with a breakdown by label.

Method	Acc.	F1
BERT-wwm + weighed final layer	96.80%	0.9677
BERT-wwm + LR	95.08%	0.9504
BERT-wwm + RF	90.30%	0.8976
XLM-R + LR	90.50%	0.9002
XLM-R + RF	89.77%	0.8914

Table 2: Experiment Results for Testset

Table 3 demonstrates the prediction results of Chinese BERT-wwm with a weighted final classification layer. Overall, the *sports* and *game* labels show the highest precision, recall, and F1 score, while *real estate* has the lowest.

Label	P	R	F1
Sports	0.9990	0.9990	0.9990
Entert.	0.9850	0.9860	0.9855
Home	0.9766	0.8780	0.9247
Real.E.	0.9052	0.9070	0.9061
Edu.	0.9796	0.9590	0.9692
Fashion	0.9652	0.9990	0.9818
Politics	0.9632	0.9680	0.9656
Game	0.9900	0.9930	0.9915
Tech.	0.9698	0.9960	0.9827
Finance	0.9485	0.9950	0.9712

Table 3: Prediction results of BERT-wwm + weighted final classification layer

Table 4 shows the prediction results of Chinese BERT-wwm with a logistic regression classifier. Overall, it shows similar results to BERT-wwm with a weighted final classification layer, with *sports*, *entertainment*, and *game* showing the highest precision, recall, and F1 score. The *real estate* label still performs the worst. Compared with BERT-wwm + weighted final classification layer, all its scores are lower.

Table 5 shows the prediction results of XLM-R with logistic regression classifier. Once again, *sports* and *entertainment* have the best performance, while *real estate* and *finance* do not perform well.

Label	P	R	F1
Sports	0.9970	0.9920	0.9945
Entert	0.9899	0.9790	0.9844
Home	0.9537	0.8240	0.8841
Real.E	0.8500	0.8780	0.8637
Edu.	0.9619	0.9350	0.9483
Fashion	0.9601	0.9860	0.9729
Politics	0.9455	0.9550	0.9502
Game	0.9763	0.9870	0.9816
Tech.	0.9461	0.9830	0.9642
Finance	0.9330	0.9890	0.9602

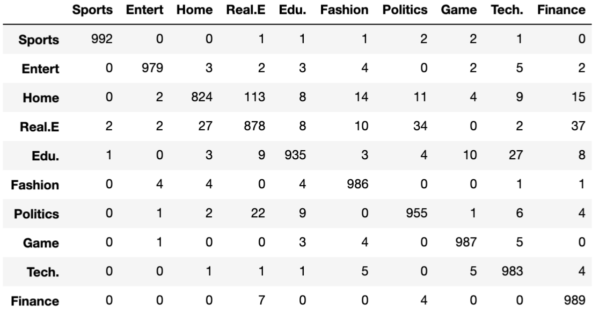
Table 4: Prediction results of BERT-wwm with Logestic Regression

Label	P	R	F1
Sports	0.9990	0.9820	0.9904
Entert	0.9649	0.9610	0.9629
Home	0.9103	0.4870	0.6345
Real.E	0.6874	0.8640	0.7656
Edu.	0.9237	0.9320	0.9278
Fashion	0.9329	0.9730	0.9525
Politics	0.9180	0.9410	0.9294
Game	0.9448	0.9590	0.9519
Tech.	0.9326	0.9690	0.9505
Finance	0.8944	0.9820	0.9361

Table 5: Prediction results of XLM-R with Logestic Regression

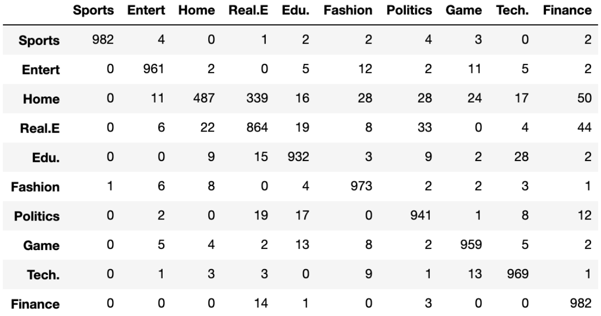
4.2 Qualitative findings

Figure 1 shows the confusion matrix for Chinese BERT-wwm. After carefully reviewing the results, we find that most of the errors occurred in *home* and *real estate* labels. In addition, the errors largely appear between these two labels. Since there aren't any similar errors in other labels, we can assume that the errors are due to the high similarity in text between the *home* and *real estate* labels. BERT-wwm can't tag the text it reads from the data into correct labels. Our theory is proven by going through the text of these two labels in the training data manually. Words like "living room", "houses", and "home decorations" are constantly mentioned in both sets of labels, even though the contexts of the words are completely different. In short, word usages are highly similar between the *home* and *real estate* topics, so it is understandable why such errors take place.



	Sports	Entert	Home	Real.E	Edu.	Fashion	Politics	Game	Tech.	Finance
Sports	992	0	0	1	1	1	2	2	1	0
Entert	0	979	3	2	3	4	0	2	5	2
Home	0	2	824	113	8	14	11	4	9	15
Real.E	2	2	27	878	8	10	34	0	2	37
Edu.	1	0	3	9	935	3	4	10	27	8
Fashion	0	4	4	0	4	986	0	0	1	1
Politics	0	1	2	22	9	0	955	1	6	4
Game	0	1	0	0	3	4	0	987	5	0
Tech.	0	0	1	1	1	5	0	5	983	4
Finance	0	0	0	7	0	0	4	0	0	989

Figure 1: Confusion Matrix for BERT-wwm



	Sports	Entert	Home	Real.E	Edu.	Fashion	Politics	Game	Tech.	Finance
Sports	982	4	0	1	2	2	4	3	0	2
Entert	0	961	2	0	5	12	2	11	5	2
Home	0	11	487	339	16	28	28	24	17	50
Real.E	0	6	22	864	19	8	33	0	4	44
Edu.	0	0	9	15	932	3	9	2	28	2
Fashion	1	6	8	0	4	973	2	2	3	1
Politics	0	2	0	19	17	0	941	1	8	12
Game	0	5	4	2	13	8	2	959	5	2
Tech.	0	1	3	3	0	9	1	13	969	1
Finance	0	0	0	14	1	0	3	0	0	982

Figure 2: Confusion Matrix for XLM-R

Figure 2 illustrates the confusion matrix for XLM-R. The model yields a similar error pattern. One reason for error might be that the dataset was not well-preprossessed. For example, there are a lot of numbers and meaningless words contained in the dataset. Specifically, after manually reviewing the *real estate* category, we found a large volume of street names and shop names. Since both BERT and XLM-R cut the description at 500 words during the word embedding process, too many meaningless characters might decrease the learning power for each model. Moreover, sentences in the *real estate* category tend to have less in common than those of other categories.

5 Conclusion

While we are unable to prove our hypothesis that XLM-R would outperform BERT-wwm, we are still able to provide valuable insight into the strengths and adaptability of each. While both models are greatly confused by word similarities between label sets, BERT proved to be more robust on a medium-resource language like Chinese.

However, this study is not without limitations. On one hand, this dataset is lack of preprocessing. In the future, we would like to do more preprocessing on the data. For example, we could remove stop words, delete the nouns that would confuse the prediction result like movie, street, and product names, and so on. On the other hand, we would like to conduct more experiments in order to improve the models' performances. For instance, we could potentially improve the results by using more advanced machine learning methods such as Long Short-Term Memory, Convolutional Neural Network, and Multilayer Perceptron, as well as by adjusting the hyper-parameters of the classifiers.

References

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. 2019. Pre-Training with Whole Word Masking for Chinese BERT.

Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1(Mlm):4171–4186.

Ameya Mahabaleshwarkar, Pranav Gupta, and Shamla Mantri. 2019. DeepDiseaseInsight: A Deep Learning & NLP based Novel Framework for generating useful Insights from Disease News Articles. In *2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA)*, pages 1–6. IEEE.

Victoria Ng, Erin E Rees, Jingcheng Niu, and Abdelhamid Zaghlool. 2020. Application of natural language processing algorithms for extracting information from news articles in event-based surveillance. *Canada Communicable Disease Report*, 46(6):186–191.

Azadeh Nikfarjam, Ehsan Emadzadeh, and Saravanan Muthaiyah. 2010. Text mining approaches for stock market prediction. *2010 The 2nd International Conference on Computer and Automation Engineering, ICCAE 2010*, 4:256–260.

Ilham Firdausi Putra. 2020. Improving Indonesian Text Classification Using Multilingual Language Model. pages 0–4.

Chenglei Si, Shuohang Wang, Min Yen Kan, and Jing Jiang. 2019. What does BERT learn from multiple-choice reading comprehension datasets? *arXiv*.

Manuel R. Vargas, Beatriz S.L.P. De Lima, and Alexandre G. Evsukoff. 2017. Deep learning for stock market prediction from financial news articles. *2017 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications, CIVEMSA 2017 - Proceedings*, pages 60–65.