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 rise to a variety of problems, but can have many benefits. Among the top performing models in this field, BERT gives state-of-the-art results, and XLM-R has been tested less, but 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 97% accuracy, while XLM-R struggles to reach 91%. We conclude that the reason for such findings is that...

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), the 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, especially 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 are going to explore text classification by implementing two pre-trained models—mBERT and XLM-R on Chinese news articles. Rather than optimizing accuracy for the main goal, we'll be comparing the performance of these two language models in order to test which one perform best, and what makes it do so well.

2 Related Work

2.1 BERT and Chinese BERT-wwm

Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) is undeniably one of the most popular pre-trained models in the NLP field. The key technical innovation BERT applied is the bidirectional pre-training for language representations, which has helped overcome the previous limitations within standard language models–unidirectional pre-training. In order to conduct the bidirectional pre-training, BERT used two unsupervised tasks, including masked language model (MLM) and Next Sentence Prediction (NSP). These innovations greatly facilitated fine-tune process, and helped BERT achieve state-of-the-art performance for a large volume 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). In the contrast, BERT-wwm masks all WordPieces that belong to a whole word altogether (Cui et al., 2019). This is especially important for Chinese language because most of the Chinese words are made up of several characters. Cui et al. (2019) then adapted the whole word masking strategy in Chinese BERT and retrained the Chinese BERT model. Their experiment results have shown that compared to mBERT, Chinese BERT-wwm improved significantly on long sequences datasets. Although the accuracy improvement for document classification experiment was limited, in the competitive NLP field, any small gains in performance can have large implications.

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, performing 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, such as what we do in this paper. XLM-R has been 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 it a stronger 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. Because the same Chinese character may be used in entirely different words and contexts, relation extraction is particularly difficult. Similarly, because Chinese sentences don't contain any spaces, it can be difficult for a model to interpret a word in context. 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 performed a controlled experiment, classifying a data set of Chinese news articles. The two models performed exactly the same classification task, so the results will be extremely 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, at such high performance measures, as previously noted, even small gains in performance can have large implications.

3.2 Data Used In Experiments

We use a Sina News dataset that contains a variety of news data with long news articles. The dataset is broken up into two columns - the category, of which there are ten, and then the text of the article itself. The 10 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. In terms of train/test/dev split, we use 77% of this dataset as the training set, 15% as the test set, and 8% as the validation set.

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 dataset we acquired from GitHub are csv files, the labels and sentences are not separated by comma, and the data format was not setting up correctly. Therefore, we cleaned it and converted the original csv dataset to tsv for better reading efficiency. As for data shuffling, the sentences were split up according to labels. Given the fact that unshuffled data always results in extremely low accuracy and high loss (Si et al., 2019), we randomly shuffled all the sentences in the dataset.

3.3 Analysis of Chinese BERT-wwm

When adapting BERT to NLP tasks, we need to fine-tune BERT for the specific task, and fit the dataset for BERT training.

Regarding fine-tuning BERT for the text classification task, we adopt the method introduced in Devlin et al. (2019) paper, which is adding the final classification layer weights $W \in \mathbf{R}^{K \times H}$, where K represents the number of labels.

For actually running BERT for text classification, we implement a DataProcessor class for processing data we read. The DataProcessor reads three datasets(train, test, val), and returns an array type parameters containing label and text information. The DataProcessor class also has get_label () function that returns all the labels in the datasets.

We use a 32 batch size, 3 epochs, and 3e-5 learning rate for the classification task.

*In the next experiment for BERT, we will (1) fine-tune learning rate; and (2) run the same task on original mBERT and analyze the results with current experiment.

3.4 Analysis of XLM-R

We first apply vectorization and word embedding into the training, validation and test data by using XLM-R language model.

We then apply Logistic Regression, Random Forest and Naive Bayes classifier to fine-tuning XLM-R for the text classification task. Also, in order to check if there is over fitting during the machine learning period, we print out the confusion matrix and the result has shown that there is no potential over fitting in our model fitting process.

Sports	Entert.	Home	Real.E.	Edu.	Fashion	Politics	Game	Tech.	Finance
982	4	0	1	2	2	4	3	0	2
0	961	2	0	5	12	2	11	5	2
0	11	487	339	16	28	28	47	17	50
0	6	22	864	19	8	33	0	4	44
0	0	9	15	932	3	9	2	28	2
1	6	8	0	4	973	2	2	3	1
0	2	0	19	17	0	941	1	8	12
0	5	4	2	13	8	2	959	5	2
0	1	3	3	0	9	1	13	969	1
0	0	0	14	1	0	3	0	0	982

Confusion Matrix for XLM-R + LR

4 Results *still working on it*

4.1 Findings

Table 2 shows the overall results for each experiment. The result for our Chinese BERT-wwm experiment is heartening. The accuracy of evaluation was below 11% before we shuffled the data but it boosted to 97% after all the data had been shuffled.

Method	Acc.	F1
mBERT	00%	00
Chinese BERT-wwm	97%	00
XLM-R + LR	00%	00
XLM-R + RF	00%	00

Table 2: Experiment Results

Table 3 demostrates the prediction results of Chinese BERT-wwm, break down to each label. Overall, label Sports and Technology show the highest precision, recall, and F1 score, while Real Estates label 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 Chinese BERT-wwm

4.2 The Difference Between the Models

Is there anything that we could do to improve these results? Are our results surprising? Why do we believe our models performed the way they did?

5 Conclusion

Error analysis. What do the results mean?

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.