# 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

## 1 Introduction

Coronavirus Disease 2019 (COVID-19) is no doubt the biggest threat to the whole human society in 2020. During such a pandemic, effective communication from the public health as well as the government is extremely important. News media is one of the most significant channels that the public use to gain new information. During the COVID-19 pandemic, this is especially important for China, where COVID-19 was first detected. Therefore, the news topic prediction is crucial for Chinese news industry and the whole society. Even before the pandemic, the news topic prediction is always an important task in the field of natural language processing (NLP). 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).

Evidence has shown that text topic prediction is usually accomplished through text generation and text classification, while text classification is more popular in recent years (Liu et al., 2020). In this study, we are going to explore text prediction by implementing two pre-trained models—BERT 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 Model and XLM-R Model

Bert (Bidirectional Encoder Representations from Transformers) is a language representation model (Devlin et al., 2019), which considers the context on both sides of the word with previous methods only considering one side of the text. In addition, BERT can perform multi-task learning so that it could perform different NLP tasks at the same time (Devlin et al., 2019). We believe these approaches can help the model better analyze the Chinese language context.

XLM-R is a language training model which stands for Cross-lingual language model pretraining. It is a new cross language model that keeps major information without sacrificing per-language performance (Ruder et al., 2019). XLM-R uses self-supervised training technology to achieve state-of-the-art performance for cross-language understanding.

### 2.2 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 (Hou et al., 2020). Similarly, because Chinese sentences don't contain any spaces, models must make special accommodations when processing text. For example, BERT must first identify the words before masking them for the analysis (Devlin et al., 2019). This makes for an interesting comparison with XLM-R, a strong multilingual model that excels without knowing a lot about a language (Ruder et al., 2019).

## 3 Method

Building upon the study conducted by Liu et al. (2020), we will use a dataset that contains COVID-19 related Chinese news from January 2020 to May 2020 as raw input, and implement news topic prediction through text classification. We will collect the dataset by retrieving the news articles from Xinhua Agency's online archives. Finally, we will compare accuracy and efficiency of the models according to the generated output.

## References

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.

Jiaqi Hou, Xin Li, Haipeng Yao, Haichun Sun, Tianle Mai, and Rongchen Zhu. 2020. BERT-Based Chinese Relation Extraction for Public Security. *IEEE Access*, 8:132367–132375.

Jingang Liu, Chunhe Xia, Xiaojian Li, Haihua Yan, and Tengteng Liu. 2020. A BERT-based ensemble model for chinese news topic prediction. *ACM International Conference Proceeding Series*, pages 18–23.

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.

Sebastian Ruder, Anders Søgaard, and Ivan Vulić. 2019. Unsupervised Cross-Lingual Representation Learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 31–38, Stroudsburg, PA, USA. Association for Computational Linguistics.

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.