Rethinking News Text Classification from a Timeliness Perspective under the Pre-training and Fine-tuning Paradigm

Anonymous ACL submission

Abstract

Pre-trained language models (PLMs) have made significant progress in NLP. News text classification is one of the most fundamental tasks in NLP, and various existing works have shown that fine-tuned on PLMs could score up to the accuracy of 98% on the target task. It seems that this task has been well-addressed. However, we discover that news timeliness can cause a massive impact on the news text classification, which drops nearly 20% points from the initial results. In this paper, we define timeliness issues in news classification and design the experiment to measure the influence. Moreover, we investigate several methods to recognize and replace obsolete vocabularies. However, the results show that it is difficult to eliminate the impact of news timeliness from the words' perspective. In addition, we propose a set of large-scale, time-sensitive news datasets to facilitate the study of this problem.

1 Introduction

Pre-trained language models (PLMs) like BERT (Devlin et al., 2019) and GPT (Radford et al., 2019) have achieved remarkable success in various NLP applications (Qiu et al., 2020; Devlin et al., 2019; Liu et al., 2019). Massive news articles are generated and posted online every day (Wu et al., 2020a), which contain rich textual information (Wu et al., 2021), and PLMs have the potentials to enhance news text modeling(Miao et al., 2018; Cecchini and Na, 2018) for various intelligent news applications like news text classification. Substantial work (Nugroho et al., 2021; Liu et al., 2021; Wu et al., 2021) has shown that on large corpus PLMs are beneficial for news text classification. Fine-tuned method could score up to the accuracy of 98% on the target task. It seems that recent algorithms (Xu et al., 2020; Meng et al., 2019) are approaching the ceiling of this task.

ChannelName **AUTOMOBILE**
Title **Leading 5G CARDINAL companies gather in Jinqiao, Shanghai GPE "Oriental Smart City" is coming.**
Content **On March 12 DATE , Shanghai Pudong Jinqiao 5G Ecological Park officially opened. Dozens of representatives from the government and enterprises gathered in the Shanghai Jinqiao Economic Development Zone ORG wearing masks. At present, 48 enterprises and key projects have settled and signed contracts. With an investment of 13 billion yuan MONEY , focusing on the three major industries of "5G+Future Cars", "5G+Intelligent Manufacturing", and "5G+Data Port", the goal of Shanghai Jinqiao Economic Development Zone is to build an "Oriental Intelligent Manufacturing City."**

Table 1: An example from our dataset.

However, we found that news classification remains various issues worth exploring. We attempt in a simple experimental setting: training our model on outdated news datasets and testing on new-updated news datasets which crawls from the same source. After experiments, we surprisingly discover the accuracy of result drops nearly 20 points from the initial results. We tested on different pre-trained models in the same setting. The experiment results all demonstrate that different PLMs bring a slight improvement to this problem.

We distribute this problem to the impact of news timeliness on text classification. Although PLMs have achieved amazing results in many natural language understanding (NLU) tasks, there is little research to explore whether large-scale pre-trained models can relieve the news timeliness influence.

In this paper, we investigate several ways to recognize and replace the time-sensitive vocabulary to improve its performance on news classification task. However, these methods do not seem to be helpful to this phenomenon. We believe there are many aspects worth exploring in this issue. In summary, our contribution points can be summarized as the following:

- We found that the news timeliness can cause a huge impact on the news text classification.

- We propose a set of large-scale time-sensitive news datasets to facilitate the study of this problem.

- We reveal that it is difficult to eliminate the influence of news timeliness on the words' perspective and provide a reference value for future work.

2 Related Work

2.1 News Text Classification

Previous work on text representation can be categorized into three main types (Zheng et al., 2020): statistics-based (Joachims, 1998; Zhang et al., 2015; Robertson, 2004), neural-network-based (Chen, 2015; Lai et al., 2015; Socher et al., 2013) and pretraining-based embeddings (Devlin et al., 2019; Liu et al., 2019; Yang et al., 2019). Nowadays, with the prevalence of pretraining techniques, recent algorithms (Xu et al., 2020) are approaching the ceiling of these datasets with accuracy scores up to 98%. Different from any of existing models, our study involves the impact of news timeliness on the target task.

2.2 News Datasets

We have compiled several datasets for news text classification and summarized them in Table 2. Most datasets are in Chinese (SogouCS (Wang et al., 2008), THUCNews (Sun et al., 2016), ChinaNews (Zhang and LeCun, 2017)) and English (Kaggle (Fuks, 2018), MIND (Wu et al., 2020b), N15News (Wang et al., 2021)). Since news is time-sensitive text, most of them are outdated datasets. Some of them are small in scale. Different from any of the existing datasets, our datasets are more timeliness, providing a new stage to test the performance of future algorithms.

Dataset	Lang.	# Doc	# Class
SANAD (2019)	AR	200k	7
ATCD (2021)	AM	50K	6
Kaggle (2018)	EN	125K	31
MIND (2020b)	EN	128K	4
N15News (2021)	EN	200K	15
SogouCS (2008)	ZH	577K	5
THUCNews (2016)	ZH	740K	14
ChinaNews (2017)	ZH	1.51M	7
Our dataset	ZH	192K	3

Table 2: Comparison of news classification datasets.

3 Dataset

3.1 Data Collection and Cleaning

We crawl our datasets from Sina news website[1] and People's Daily Online[2], and collect news from January 1th, 2021 to June 30th. However, the quality of the crawled data is definitely not high, and we need to clean the news data. Since the headlines of the news have already summarized the news content to a certain extent, we intend to deal with the news content mainly. Firstly, we use a feature-based approach to remove the words that are not related to the classification in the news. Secondly, we de-duplicate the repetitive news to get higher quality data.

3.2 Data Statistics

In this dataset, each piece of data consists of five parts: namely title, content, title entity, content entity and category. The dataset consists of ten categories, namely FINANCE, TECHNOLOGY, GAMES, etc. Among them, in addition to 75,572 other categories, it consists of various news categories other than the first nine categories. An example is shown in Table 1, and the data statistics and average length are reported in Table 3.

TYPE	STATISTICS	TIT.	CON.
FINANCE	14,877	21	1,219
REAL ESTATE	12,912	20	1,076
EDUCATION	11,953	18	1,185
MILITARY	11,476	21	1,055
TECHNOLOGY	22,578	20	645
AUTOMOBILES	23,117	22	1,019
SPORTS	14,506	20	487
GAMES	21,784	19	564
ENTERTAINMENT	15,831	21	770
OTHERS	75,572	19	531
TOTAL	224,606	20	748

Table 3: Size overview of our dataset.

3.3 Extractive Strategies

We follow the traditional ChineseNLP tools[3] to recognize the entity in the content and title, which contains 35 types: PERSON, EVENT. PRODUCT, DATA, etc. The model uses BERT as based model, and trains on the Onenote5.0 (Weischedel et al., 2013), and finally achieves 81.18% accuracy in the test set.

[1] https://news.sina.com.cn/
[2] http://en.people.cn/
[3] https://github.com/ckiplab/ckip-transformers

Method	Example
Raw data	Tesla delivered approximately 140,000 electric vehicles worldwide in the third quarter of 2020.
MASK	[MASK] delivered [MASK] [MASK] [MASK] [MASK] worldwide in the third quarter of 2020.
PAD	[PAD] delivered [PAD] [PAD] worldwide in the third quarter of 2020.
Fine-grained	[MASK] delivered [PAD] electric vehicles worldwide in the third quarter of 2020.
Keyword	delivered ; in the third quarter of 2020

Table 4: Different methods of obsolete word replacement

4 Preliminary

4.1 Problem Defination

We randomly select 3,000 items from each news category in THUCnews(Sun et al., 2016), a total of 9 categories, and 27,000 items of data. Subsequently, we randomly selected 1,000 items for each category from our own datasets, for a total of 9,000 items. Two copies will be selected, one as the validation set and one as the test set. It is worth noting that during the training process, we only use the old datasets for training and do not add new data. This is the difference between our task and the normal news classification task. Specifically, we evaluate the models performance based on accuracy, precision, recall and Macro-F1, which computing the average of the F1 scores obtained by individual categories.

4.2 Training details

Specifically, we adopt pre-trained models in the HuggingFace Transformers toolkit[4](Wolf et al., 2020) through all of our works. Hyperparameters values of the training stage are listed in Table 5. We use a single RTX 3090 GPU for training. The best checkpoint of the model is searched during the validation stage. Specifically, we finetune all model parameters except pre-trained text embedding in this paper.

Hyperparameters values	
Number of epochs	5
Batch Size	16
Max Sentence Length	512
Optimizer	Adam (Kingma and Ba, 2014)
Learning rate	1e-5
Loss function	label smoothed cross-entropy (Szegedy et al., 2016)

Table 5: Hyperparameters values of training stage.

[4]https://github.com/huggingface/transformers

5 Experiments

In this section, we implement our experiment on supervised text classification built on common pretrianed model and fine-tuned with supervised softmax loss on labeled texts. We explore this problem from the following three perspectives.

5.1 Experimental Settings

Pre-trained Model Since different PLMs are suitable for different tasks, we fix other variables and only change the type of pre-training model for experimentation. We experiment on three common PLMs: `BERT-base-Chinese` (Devlin et al., 2019), which has 12 layers, 12 attention heads, 393M parameters, `Chinese-roberta-wwm-ext` (Liu et al., 2019), which has 12 layers, 12 attention heads, 393M parameters, `Chinese-xlnet-base` (Yang et al., 2019), which has 12 layers, 12 attention heads, 445M parameters.

Obsolete Word Replacement Following previous work, masked language modeling (MLM) (Taylor, 1953; Devlin et al., 2019), randomly masks some of the tokens from the input to learn an inner representation of language. We consider to cover up the outdated entity, focusing the study on the sentence structure and other important information. We first compare two replacement characters: `[MASK]`, which takes participating in the calculation when put the sentence into PLMs, and `[PAD]`, which means blank character, not having a hand in the calculation. Moreover, we adopt a fine-grained approach, dividing entities into three categories: entities with timeliness, entities that are not time-sensitive but affect classification, entities that are not time-sensitive and do not affect classification, taking the operations of masking, remaining, and padding of three types respectively. In addition, some keyword information would be ignored because it is impossible to classify the time-sensitive

Method	BERT				RoBERTa				XLNet			
	Pre.	Rec.	F1	Acc.	Pre.	Rec.	F1	Acc.	Pre.	Rec.	F1	Acc.
Baseline	82.70	76.17	75.24	76.17	82.27	78.06	76.76	78.06	**84.00**	**80.34**	**79.68**	**80.43**
MASK	82.72	78.27	77.30	78.27	82.35	78.46	76.90	78.46	82.26	79.04	77.26	79.04
PAD	82.94	78.46	77.19	78.46	81.33	77.91	76.27	77.91	83.26	79.42	78.26	79.42
Fine-Grained	81.42	78.29	76.83	78.29	81.94	78.39	77.15	78.39	81.88	77.78	76.38	77.78
MASK+KEY.	**83.38**	**79.53**	**78.69**	**79.53**	82.93	76.26	74.65	76.26	80.48	75.31	73.42	75.31
FG+KEY.	81.83	76.91	75.13	76.91	82.56	76.71	75.33	76.71	83.18	77.54	76.11	77.54

Table 6: Experimental results of baseline methods

characteristics from the recognized entities. We separate the data set into two copies, one to replace time-sensitive entities, one to extract keywords, and pass them through the same PLMs. By adjusting the weight of learning, we can learn the structure information of the sentence without ignoring keyword information. Different methods are shown in Table 4.

Datasets Distribution Furthermore, we want to explore whether is the distribution difference between different datasets that causes the problem. Apart from training on the old datasets and testing on the new datasets (old↔new), we design two other comparative experiments: training on the new datasets and testing on the old datasets (new↔old), training on the new datasets and testing on the new datasets (new↔new).

5.2 Results and Analysis

We first present the experimental results on the PLMs comparison and obsolete word replacement. The numbers are shown in Table 6. From the results, we can observe that XLNet (Yang et al., 2019) achieves the best performance 80.43%. Comparing with other two PLMs, XLNet combines BERT (Devlin et al., 2019) and Transformer-XL (Dai et al., 2019), which is more suitable for longer context. We believe that this model is more suitable for news text classification.

Then, we work on the influence of obsolete word replacement. The results are reported with the last five lines in Table 6. We have introduced five different strategies to eliminate the influence from the words' perspective. We discover that (1) Though the method, learning the sentences' structure without ignoring the keyword information, could make a slight improvement, there is still a considerable gap with 98.44% trained on the new datasets in the same setting. (2) It can be clearly seen that the effect of different replacements fluctuates greatly when main model is switched. We adopt two different strategies, PLMs and word encoder, as our approaches. However, the final improvement is very slight. We claim that (1) It is difficult for us to eliminate the influence from words' perspective. (2) There are still many issues remained to be solved in this problem.

Method	BERT	RoBERTa	XLNet
old↔new	75.24	76.76	79.68
new↔old	**97.59**	**97.72**	**97.56**
new↔new	98.44	99.03	98.89

Table 7: Comparison of news datasets distribution.

We then perform a further analysis on the different experimental settings, the result is shown in table 7. We surprisingly discover that both new↔old and new↔new achieve high performance. It certificates that we couldn't eliminate the influence from the perspective of data sample migration, since even if the training set and the testing set are exchanged, the problem of data migration should still exist. We believe that the main reason for this phenomenon is that the knowledge that did not appear in the finetune and pre-training stage appeared during the test, so how to eliminate this influence in the finetune stage has become the focus of our future research

6 Conclusion and Future Work

In this paper, we discover the impact of news timeliness on text classification. We investigate several ways to recognize and replace the outdated vocabularies. However, the results show that it is difficult to eliminate the influence of news timeliness from the words' perspective. Moreover, we propose a set of large-scale time-sensitive news datasets to facilitate the study of this problem. In future work, we can do this task on datasets of different time periods to explore whether such problems will occur in other tasks. We think this research is very meaningful under the pre-training paradigm.

References

Israel Abebe Azime and Nebil Mohammed. 2021. An amharic news text classification dataset. *arXiv preprint arXiv:2103.05639*.

David Cecchini and Li Na. 2018. Chinese news classification. In *2018 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 681–684. IEEE.

Yahui Chen. 2015. Convolutional neural network for sentence classification. Master's thesis, University of Waterloo.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*.

Omar Einea, Ashraf Elnagar, and Ridhwan Al Debsi. 2019. Sanad: Single-label arabic news articles dataset for automatic text categorization. *Data in brief*, 25:104076.

Olga Fuks. 2018. Classification of news dataset. *Standford University*.

Thorsten Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*, pages 137–142. Springer.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent convolutional neural networks for text classification. In *Twenty-ninth AAAI conference on artificial intelligence*.

Jialu Liu, Tianqi Liu, and Cong Yu. 2021. Newsembed: Modeling news through pre-trained documentrepresentations. *arXiv preprint arXiv:2106.00590*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Yuxian Meng, Wei Wu, Fei Wang, Xiaoya Li, Ping Nie, Fan Yin, Muyu Li, Qinghong Han, Xiaofei Sun, and Jiwei Li. 2019. Glyce: glyph-vectors for chinese character representations. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 2746–2757.

Fang Miao, Pu Zhang, Libiao Jin, and Hongda Wu. 2018. Chinese news text classification based on machine learning algorithm. In *2018 10th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, volume 2, pages 48–51. IEEE.

Kuncahyo Setyo Nugroho, Anantha Yullian Sukmadewa, and Novanto Yudistira. 2021. Large-scale news classification using bert language model: Spark nlp approach. In *6th International Conference on Sustainable Information Engineering and Technology 2021*, pages 240–246.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, pages 1–26.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Stephen Robertson. 2004. Understanding inverse document frequency: on theoretical arguments for idf. *Journal of documentation*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Maosong Sun, Jingyang Li, Zhipeng Guo, Z Yu, Y Zheng, X Si, and Z Liu. 2016. Thuctc: an efficient chinese text classifier. *GitHub Repository*.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Wilson L Taylor. 1953. "cloze procedure": A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433.

Canhui Wang, Min Zhang, Shaoping Ma, and Liyun Ru. 2008. Automatic online news issue construction in web environment. In *Proceedings of the 17th international conference on World Wide Web*, pages 457–466.

Zhen Wang, Xu Shan, and Jie Yang. 2021. N15news: A new dataset for multimodal news classification. *arXiv preprint arXiv:2108.13327*.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, et al. 2013. Ontonotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, PA*, 23.

Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Chuhan Wu, Fangzhao Wu, Yang Yu, Tao Qi, Yongfeng Huang, and Qi Liu. 2021. Newsbert: Distilling pre-trained language model for intelligent news application. *arXiv preprint arXiv:2102.04887*.

Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, and Ming Zhou. 2020a. MIND: A large-scale dataset for news recommendation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, et al. 2020b. Mind: A large-scale dataset for news recommendation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3597–3606.

Canwen Xu, Jiaxin Pei, Hongtao Wu, Yiyu Liu, and Chenliang Li. 2020. Matinf: A jointly labeled large-scale dataset for classification, question answering and summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3586–3596.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Xiang Zhang and Yann LeCun. 2017. Which encoding is the best for text classification in chinese, english, japanese and korean? *arXiv preprint arXiv:1708.02657*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28:649–657.

Jianming Zheng, Fei Cai, Honghui Chen, and Maarten de Rijke. 2020. Pre-train, interact, fine-tune: a novel interaction representation for text classification. *Information Processing & Management*, 57(6):102215.