Multi-Level Text Importance Classification Architecture Based on Deep Learning

Meizhen Huang[1], Jinshu Su[1], Zhong Liao[2], Shuhui Chen[1], Ziling Wei[1]
[1]College of Computer/National University of Defense Technology
[2]Hunan Provincial Key Laboratory of Media Fusion Content Aware and Security
Changsha, China
hmz_20@nudt.edu.cn, sjs@nudt.edu.cn, 18670099566@163.com, shchen@nudt.edu.cn, weiziling@nudt.edu.cn

ABSTRACT

In the era of information explosion, the Internet is full of spam and false information, making it more difficult for people to obtain effective information. Since text data is the main carrier for disseminating information and knowledge, we propose a multi-level text importance classification architecture based on deep learning to enable Internet users to quickly and accurately access text content of interest. Experiments demonstrate that the proposed architecture can achieve a good performance.

CCS CONCEPTS

• **Information systems** Information retrieval;

ACM Reference format:

Meizhen Huang, Jinshu Su, Zhong Liao, Shuhui Chen, Ziling Wei. 2022. Multi-Level Text Importance Classification Architecture Based on Deep Learning. In *6th Asia-Pacific Workshop on Networking (APNet 2022) (APNet '22), July 1–2, 2022, Fuzhou, China. ACM, New York, NY, USA, 2 pages.* https://doi.org/10.1145/3542637.3542640

1 INTRODUCTION

The huge amount of data in the Internet reduces the efficiency of people in finding and acquiring information. Text as the most widely distributed and data-rich information carrier in the Internet, the problem of information overload can be effectively solved by using text classification. The development of deep learning has led to a proliferation of research in this area, Kim et al. proposed the TextCNN[1] model, the Google team proposed the self-attention[2] mechanism, the Transformer-based pre-trained language model and the GNN-based model are the recent research hotspots. The existing models have shown their usefulness in text classification, but continuous improvement and exploration are needed in order to expect the models to "understand" text at the semantic level like humans.

To well obtain the important information for different users, we study how to realize the text importance classification. Different from traditional text classification that focuses on domain classification or sentiment classification, text importance classification is strongly related to the needs of users. The huge amount of data in the Internet puts a large burden on the model in terms of processing performance when realizing the text importance classification. To address the above challenge, we propose a multi-level text importance classification architecture in this paper. In addition, to well evaluate the performance of text importance classification model, we also propose an evaluation metric .

2 ARCHITECTURE DESIGN

The multi-level text importance classification architecture includes two layers: the coarse-grained domain classification layer and the fine-grained importance classification layer, as shown Fig. 1.

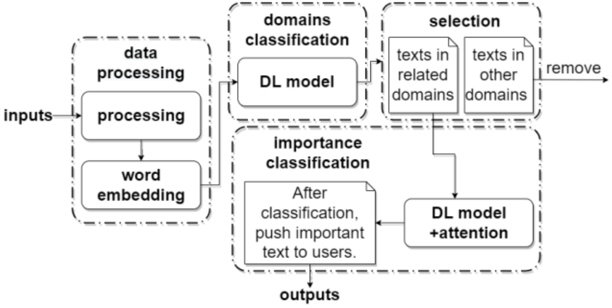

Figure 1: The proposed architecture

Coarse-grained domain classification layer: For different users, the interested text is not fully decided by the domain information. However, it should relate to the domain information. Thus, the function of this layer is to make a fast coarse-grained domain classification of the text so that a large number of texts in the irrelevant domain can be excluded first. Then, the texts in the domain related to user needs enter the fine-grained importance classification layer. By the above operation, it can reduce the amount of text processing in the fine-grained importance classification layer by removing a large number of irrelevant texts. We use the Fasttext[3] model, which has only a three-layer structure of input, hidden and output layers, thus enabling fast processing of large amounts of text.

Fine-grained importance classification layer: The function of this layer is to finely classify the importance of the remaining key texts. The importance of the text depends on the user requirements. In this case, the model needs to well understand the semantics of the text. We use a self-attention[2] model and a TextCNN[1] model that introduces attention for the purpose of classifying the text importance.

APNet '22, July 1–2, 2022, Fuzhou, China

ACM ISBN 978-1-4503-9748-3/22/07.
https://doi.org/10.1145/3542637.3542640

3 PERFOMANCE EVALUATION

3.1 Scenario Design

Before the experiment, we set up a user demand scenario as follows. We assume that the users focus on the information related to violent conflict events. The entities may be involved in this case are police, military, religion, racism, terrorists and etc. We set four tag values to measure the importance factor of the text content. "0" means the text content is not related to the set scenario. "1" means the entities in the text are related to the set scenario but the text content is not of interest to the user. "2" and "3" means the text content is consistent with the set scenario. "2" indicates the severity of violent conflict is low (e.g. personal injury or property damage), "3" indicates the severity of violent conflict is high (e.g. death). Therefore, these four score labels can be divided into user interest intervals and non-interest intervals.

3.2 Datasets

The current publicly available datasets are not applicable to the text importance classification. We annotated a small dataset named **20News Importance**.The datasets used are as follows.

20News Importance: Based on the 20 Newsgroups dataset, we selected five categories of texts that might be relevant to the user demand scenario for score labeling. The five categories of texts contains "talk.religion.misc", "soc.religion.christian", "talk.politics.mideast", "talk.politics.misc", "talk.politics.guns". The current dataset contains 499 articles.

AG News[1]: A subdataset of AG's corpus of news articles constructed by assembling titles and description fields of articles from the 4 largest classes ("World", "Sports", "Business", "Sci/Tech") of AG's Corpus. The AG News contains 30000 training and 1900 test samples per class.

20 Newsgroups[2]: This dataset is a collection of 18846 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups. Some of the newsgroups are related to each other, while others are highly unrelated (e.g. misc.forsale /soc.religion.christian).

3.3 Evaluation Metric

We define the actual score label as y and the score label predicted by the model as y_{pred}, the distance between the predicted value and the actual value is d:

$$d = |y - y_{pred}| \tag{1}$$

For the users, when y and y_{pred} belong to the same interval, the prediction result is positive. When y and y_{pred} belong to different intervals, then the prediction result is negative. Therefore, when $d = 1$, determine whether y and y_{pred} belong to the same interest intervals or non-interest interval, if not, then $d = d + 1$. We also give each test text a score f based on d:

$$f = \begin{cases} 100, \ d = 0 \\ 80, \ d = 1 \\ 0, \ d = 2 \\ -100, \ d = 3 \end{cases} \tag{2}$$

In this way, the final average score F can be obtained by:

$$F = \sum_{i=1}^{n} f_i \,/ N \tag{3}$$

N is the number of texts in the test dataset. F is used as a metric for the evaluation of the results of our experiments.

3.4 Experiment

In the coarse-grained domain classification layer, based on AG News and 20 Newsgroups, we conducted experiments with three simple models, Fasttext[3], TextCNN[1] and BiLSTM[4], respectively. Table 1 shows the experimental results of the three models. Fasttext takes tens of times faster than TextCNN and BiLSTM for each epoch. In addition, the accuracy of Fasttext is also acceptable. Thus, we take Fasttext as the model for this layer by taking into account the classification speed and accuracy.

In the fine-grained importance classification layer, we use the 20News Importance dataset to verify the text importance classification. User interest intervals are used as positive class samples and non-interest intervals are used as negative class samples to calculate precision, recall and F1 score. They are used to support the rationality of our proposed evaluation scheme. Table 2 shows the experimental results, which justify our proposed evaluation scheme.

Table 1: Coarse-grained classification layer evaluation results

	Max	Fasttext		TextCNN		BiLSTM	
	Length (words)	Time (s/Epoch)	ACC	Time (s/Epoch)	ACC	Time (s/Epoch)	ACC
20 News	200	2	77.75%	36	91.19%	236	88.22%
	500	3	76.34%	86	91.80%	590	88.75%
AG News	100	12	88.02%	123	89.70%	746	90.36%

Table 2: Fine-grained classification layer evaluation results

N	Accuracy	Precision	Recall	F1	F
39	0.615	1	0.667	0.800	82.051
49	0.571	0.800	0.667	0.727	79.592

4 SUMMARY AND FUTURE WORK

In this paper, we propose a multi-level structure to realize the text importance classification. By the proposed structure, it can effectively accelerate the processing procedure when dealing with large number of texts. Our future work mainly focuses on the annotation of the 20News Importance dataset and the optimization of the fine-grained importance classification layer.

REFERENCES

[1] Kim Y . Convolutional Neural Networks for Sentence Classification[J]. Eprint Arxiv, 2014.
[2] Vaswani A , Shazeer N , Parmar N , et al. Attention Is All You Need[J]. arXiv, 2017.
[3] Joulin A , Grave E , Bojanowski P , et al. Bag of Tricks for Efficient Text Classification[J]. 2017.
[4] Graves A , Jürgen Schmidhuber. Framewise phoneme classification with bidirectional LSTM and other neural network architectures[J]. Neural Networks, 2005, 18(5 - 6):602-610.

[1] http://groups.di.unipi.it/~gulli/AG_corpus_of_news_articles.html

[2] http://qwone.com/~jason/20Newsgroups/