CNewsTS - A Large-scale Chinese News Dataset with Hierarchical Topic Category and Summary

Quanzhi Li
LightSpeed Studios, Tencent
Bellevue, WA, USA
quanzhili@tencent.com

Yingchi Liu
Yohana
Palo Alto, CA, USA
yingchi.liu@yohana.com

Yang Chao
LightSpeed Studios, Tencent
Shenzhen, China
youngchao@tencent.com

ABSTRACT

In this paper, we present a large Chinese news article dataset with 4.4 million articles. These articles are obtained from different news channels and sources. They are labeled with multi-level topic categories, and some of them also have summaries. This is the first Chinese news dataset that has both hierarchical topic labels and article full texts. And it is also the largest Chinese news topic dataset. We describe the data collection, annotation and quality evaluation process. The basic statistics of the dataset, comparison with other datasets and benchmark experiments are also presented.

CCS CONCEPTS: • Computing methodologies~Artificial intelligence~Natural language processing • Information systems~Information systems applications~Data mining

KEYWORDS: Chinese news dataset, Hierarchical topic classification, News topic, News summary

ACM Reference format:
Quanzhi Li, Yingchi Liu, & Yang Chao. 2022. CNewsTS - A Large-scale Chinese News Dataset with Hierarchical Topic Category and Summary. In *Proceedings of the 31st ACM Int'l Conf. on Information and Knowledge Management (CIKM '22), Oct. 17–21, 2022, Atlanta, GA, USA.* ACM, New York, NY, USA. 6 pages. https://doi.org/10.1145/3511808.3557561

1 INTRODUCTION

Text classification is the process of categorizing texts (e.g., tweets and news articles) into organized groups. Typical text classification tasks include news categorization and sentiment prediction. Among different types of textual content, news contents are the most important information sources. A news topic classification system can help the users obtain information of interest in real-time, e.g., by discovering emerging news topics or recommending relevant news based on user interests [35, 63, 67].

For news topic classification related applications to work, a labeled news dataset is necessary. Unfortunately, existing Chinese news datasets either have only one level of topic category (just the top-level) or contain only article titles without full text. In contrast, there are English news datasets, such as [27, 46], that have full text articles with hierarchal topic categories. In this paper, we present a large Chinese news article dataset that has hierarchical topic information, including the 1st and 2nd level topic categories.

In addition to the hierarchical topic categories, some articles of this dataset also have summary. Text summarization is an important task in natural language processing, which requires the model to understand the document and generate a short text to summarize its main content. It is especially important for news domain, since nowadays most people read news from mobile devices with small screen and a short summary would help both the readers and news distributors [64, 65, 68]. Since this dataset has article topic categories and summaries, it will not only help the news classification or summarization task individually, but also provide a good resource for joint learning of the tasks.

We name this dataset as CNewsTS, since it is a **C**hinese **News** dataset with **T**opic category and **S**ummary. The main contributions of this paper are:

1. As far as we know, CNewsTS is the first Chinese news dataset that has both hierarchical topic labels and article full text. All these 4.4 million articles have a 1st level topic label, and this dataset is larger than existing Chinese news topic datasets. In addition to topic labels, half million articles also have summary.
2. Besides topic category and summary, each article also has a rich set of metadata, such as title, full text, channel (where we obtained the article), source (original publisher), source type (privately/state-owned, local/national), link, keywords and publication date. Users can use these metadata to do more news analysis, such as report bias analysis between privately and state-owned media.
3. Human evaluations of the quality of topics and summaries are conducted and presented in this paper.

This dataset could be used in various applications, such as classification for both short and long text, flat and hierarchical topic classification, summarization, keyword identification, title generation, topic detection, and news analysis. Because of its large size, it can also be used to pretrain or finetune language models [66]. The dataset and related information are available to the public at: *https://github.com/alievent/cnewsts/*

2 RELATED WORK

Text and news classification datasets. Based on their main target applications, text classification datasets can be categorized into

CIKM '22, October 17–21, 2022, Atlanta, GA, USA

ACM ISBN 978-1-4503-9236-5/22/10…$15.00 https://doi.org/10.1145/3511808.3557561

different groups, such as topic classification, news categorization and sentiment analysis [35]. The popular English news classification datasets are introduced below. The AG news dataset [58] is a collection of news articles collected from more than 2,000 news sources. The 20Newsgroups dataset [1] is a collection of newsgroup documents on 20 different topics. The Reuters-21578 news dataset [44] is one of the most widely used data collections for text categorization. We can consider news classification as one type of topic classification. There are many topic classification datasets: The DBpedia dataset [26] is a large-scale, multilingual knowledge base created from Wikipedia. PubMed [32] is for medical and biological scientific papers. The WOS dataset [25] is a collection of published papers from the Web of Science. The EUR-Lex dataset [33] includes different types of documents, and the most popular version of this dataset is based on different aspects of European Union law. MIND is a large-scale dataset for news recommendation [69]. There are also other types of text classification datasets: SQuAD [43], MS MARCO [38], SNLI dataset [2], Movie Review (MR) [39], The Stanford Sentiment Treebank (SST) dataset [48], and the MPQA dataset [6].

Several Chinese news topic datasets are available to the public: THUCNews [50], Sogou News [49], CNewsCorpus [4], NLPCC2017 [42], TNEWS [57], and Toutiao News [52]. However, these Chinese news topic datasets either have only one level of topic category (just the top-level) or contain only article titles but not the article full text. In contrast, there are English news datasets, such as RCV1 [24] and LDC2008T19 [46], that have both full text and the hierarchal topic categories. We compare our dataset to the Chinese news datasets mentioned above in Section 4.

News summarization datasets. Most news summarization datasets focus on English. Newsroom [15] is a large-scale news dataset crawled from 38 major news publication websites, with topics ranging from technology to sports. NYT [46] is a news summarization dataset created from New York Times Annotated Corpus. The CNN/DailyMail dataset [17] are modified by Nallapati et al. [37] and See et al. [47], and it is a very commonly used dataset for single-document summarization.

There are also several datasets on Chinese news summarization. LCSTS [19] and RASG [11] are two Chinese social media summarization datasets. The articles from TTNews [20] are all from only one news channel, Toutiao. CLTS [29] is a dataset extracted from the news site ThePaper. The articles of dataset CNewSum [54] are also from one channel, Toutiao. In contrast, the articles in our dataset are obtained from 14 different channels. We will compare our summarization dataset to the datasets mentioned above in Section 4. There are also several Chinese summarization datasets in other domains, not news [2, 21, 56].

3 DATA COLLECTION, ANNOTATION AND EVALUATION

Data collection. We collected data from 14 news channels from January 2020 to August 2021. They are either news aggregators (portals), such as Sina News, or big influential state-owned news media organizations, such as CCTV and Xinhua News. The data were obtained via their API services or their public websites. We pre-processed each article by extracting the text content from the HTML source and extracting metadata, such as publication datetime, title and link. The article full text is extracted by our inhouse HTML parsing tool. Unrelated content, such as ads and banners, are identified and not included in the full text. Each article has a topic field provided by either the article author or editors. Each article also has a source field, which is the publisher, who could be a state-owned news agency, a private news organization, or even a freelancer. Some articles also have a summary, a list of keywords and the location of the story. They are provided by the authors/editors.

Topic category annotation. After collecting and pre-processing the articles, we analyzed the topic labels extracted from them. Totally, there are 160 unique topic labels. Two annotators who are experienced news analysts classify them into 30 top level topics. The Classification and Code of News in Chinese schema (CCNC) published by the Chinese Standardization Administration was used as reference for this manual classification. Overall, the inter-agreement between the two annotators was 88.3%. For the ones they disagreed on, we intervened, and together we made a final decision after discussion.

Evaluation of topic labels provided by authors/editors. In order to assess the quality of the topic labels provided by the authors or editors, we did an evaluation by randomly taking 400 articles and asking two annotators to manually evaluate these labels. Overall, the inter-agreement between these two annotators was 96.5%. Both agreed that 96.5% of the 400 articles have correct labels, and 1% of them have wrong labels. There were 2.5% of the article labels they disagreed on. Most of the labels they disagreed on were due to the multi-label issue, meaning an article can be labeled with multiple topics. For example, an article talking about the business development of a technology company may be labeled as both "Business" and "Technology". The evaluation result shows that the quality of the topic labels provided by the authors/editors is high and reliable.

Evaluation of summary quality. We conducted an experiment to evaluate the quality of the summaries provided by the authors/editors. 200 articles were randomly selected, and we also used two unsupervised algorithms to automatically generate summaries for these 200 articles. They were mixed with the summaries provided by the authors/editors. Two annotators conducted the evaluation, using a 5-point scale and two criteria, informativeness and succinctness, following previous studies [24, 30]. Totally, there were 600 summaries to evaluate. Table 1 shows the average scores for these 200 articles. The result clearly shows that the summaries provided by the authors/editors are much better than the automatically generated ones, and the difference is statistically significant at $p<0.01$ level using *t-test* [45]. It also shows that the scores of the manual ones are quite high, 4.53 for informativeness and 4.68 for succinctness. The result demonstrates the high quality of the summaries in this dataset.

Table 1. The human evaluation result of summary quality.

Method	Informativeness	Succinctness
LEAD-3 [15]	4.12	4.25
TextRank [34]	3.78	3.81
By editors/authors	4.53	4.68

4 ARTICLE METADATA, STATISTICS AND COMPARISON WITH OTHER DATASETS

Article metadata: Table 2 presents the basic metadata for each article. Details about the *channel* field can be found in our GitHub site. The *topic assigned* field is determined by the topic category annotation process described in Section 3. For the 4.4 million articles, there are about 10,000 unique sources. Two experienced journalists annotated these sources to determine their *source types*. Most of them are quite easy to determine, since only these two dimensions are involved: privately owned vs. state-owned and local vs. national. For the hard ones, they looked up the registry data of the publisher and other resources to decide. Their inter-agreement is 95.7%. For the ones they disagreed on, we intervened, and together we made the final decision. The article *full text* is extracted from the corresponding HTML page, as described before. Other fields in Table 2 are directly extracted from the corresponding news channels.

We did some analysis on the dataset and the statistics over the whole dataset for articles, summaries and keywords can be found from the GitHub site.

Table 2. Metadata provided for each article.

Metadata	Description
title	article title
summary	summary, if there is one
original topic label	topic label provided by authors/editors
topic assigned	1st/2nd level topic. The 2nd level is optional
keywords	keywords, if provided
location	location of the news story, if provided
channel	channel name the article is obtained from
source	the publisher of the article
source type	privately or state-owned (government-owned), local or national media
publish date	publication date
body	full text body
link	link of the article

Comparison with other news topic datasets. Table 3 presents the comparison between our dataset and other Chinese news datasets: THUCNews [50], Sogou News [49], CNewsCorpus [4], NLPCC2017 [42], TNEWS [57], and Toutiao News [52]. From Table 3 we can see that our dataset is the largest one, and it is the only one that has both hierarchical topic labels and article full text. The Toutiao News dataset is also large and has hierarchical categories, but it only provides article title, not full text.

Comparison with other news summarization datasets. There are several datasets on Chinese news summarization. LCSTS [19] and RASG [11] are two Chinese social media summarization datasets. TTNews [20], CLTS [29] and CNewSum [54] contain news article and their summaries. Table 4 shows the comparison of these datasets with CNewsTS. This table shows that the articles of our dataset are from 14 channels, compared to just one channel that the other datasets get data from.

Table 3. Chinese news topic dataset comparison

Dataset	Size	Number of top -level category	Has hierarchical topics	Has full text
THUCNews	740,000	14	no	yes
Sogou News	2.9 million	5	no	yes
CNewsCorp	39,247	8	no	yes
NLPCC2017	12,000	18	no	yes
TNEWS	73,000	15	no	only title
Toutiao News	2.9 million	15	yes	only title
CNewsTS	4.4 million	30	yes	yes

Table 4. Chinese news summary dataset comparation

Dataset	Size	Data type	Data channel
LCSTS	2.4 million	social media	Weibo
RASG	863,826	social media	Weibo
TTNews	54,000	news article	Toutiao
CLTS	185,397	news article	ThePaper
CNewSum	304,307	news article	Toutiao
CNewsTS	0.5 million	news article	14 different channels

5 EXPERIMENTS

This section presents the benchmark experiments we conducted for the news topic categorization and summarization tasks. Their performance can serve as the baseline for future work.

Experimental settings. In the following experiments, all the models, except the pre-trained language model, e.g., BERT [7], which has their own text tokenizer, used Jieba [22], a popular Chinese tokenizer, to tokenize the Chinese text. For the BERT based models, we took the first 510 tokens of an article as input. All our neural network models were optimized using Adam optimizer with a 5e-5 learning rate, 10% warm-up proportion, and batch size of 32. For fine-tuning the pretrained language models, the dropout rate at the fully connected layers was 0.1. The models using pretrained language models are implemented based on Transformers [55].

5.1 Topic Classification

The topic classification experiment was done for the whole hierarchical categories, including the 1st and 2nd level categories. Half million articles were selected and split into train (70%), validation (10%) and test (20%) set.

There have been some existing studies on better output cost functions for hierarchical text classification [61, 62]. In this experiment, for the fastText and fine-turned BERT model, we use the cross entropy objective function to determine labels and adopt the simple but effective recursive regularization framework (RRF) proposed in [13, 41]. The idea is that if the two labels are parent and child in the hierarchy, it is assumed that the classification from these two labels to other labels are similar. It means the children label classifiers are close to the parent classifier. The three benchmark approaches tested in this experiment are:

- BoW + HSVM: HSVM is a hierarchical SVM [53] model.
- fastText + RRF: fastText with RRF framework.
- Fine-turned BERT + RRF: Fine-turned BERT-base with RRF framework.

Table 5. News topic classification result.

Method	Macro F1
BoW + HSVM	72.7
fastText +RRF	81.2
Fine-turned BERT +RRF	85.3

Table 6. Summarization experiment result.

Method	ROUGE-1	ROUGE-2	ROUGE-L
LEAD-3	31.04	18.77	26.86
NeuSum	32.23	19.36	28.41
BERT-abs	38.68	23.43	33.37

We use macro F1 as the evaluation metric, and the result is presented in Table 5. The result shows that the fine-turned BERT model performs the best, and the difference between it and the other two models are statistically significant at $p<0.01$ level using *t-test* [45]. The result also shows that it is still quite challenging for multi-level news topic classification. One reason is that there are many categories in the hierarchical classification task. Another reason is that we have more multi-label issues in the hierarchical classification case; we will discuss the multi-label issue more in Section 6.

5.2 Summarization

For this experiment, articles with summary were split into three sets (train/validation/test) using the 70/10/20 split ratio. The evaluation metrics used are ROUGE-1, ROUGE-2 and ROUGE-L [28], which are the most popular performance metrics for summarization tasks. There are two primary methods to generate summaries: extractive and abstractive. Three different types of approaches are tested in this experiment, and they belong to different methodologies (supervised vs. unsupervised, extractive vs. abstractive):

- LEAD-3 [15]: this is a simple but quite competitive unsupervised method, which just takes the first three sentences of the article as the summary. It is a common lower bound for news summarization dataset and has been used by previous studies [47, 37].
- NeuSum [59]: this is an extractive approach that jointly scores and selects sentences.
- BERT-abs [30]: this model is an abstractive method based on BERT-base-chinese.

The experiment result is presented in Table 6. The result shows that the abstractive model based on BERT has better result on Chinese news content, which is consistent with other studies [54].

6 LIMITATIONS AND CONCLUSION

Below we discuss some limitations of this dataset.

Multi-label issue. As mentioned before, some articles can have multiple topic labels. For example, an article talking about the financial issues of a healthcare company could have both Finance and Health labels, or an article about international politics could have both Politics and International topic labels. In our dataset, currently each article has only one topic path, i.e., 1st-2nd level with the last level optional. We are planning to add multi-label information to this dataset, using a semi-automatic process to reduce the workload.

One Article with multiple stories. When analyzing the articles, we found that there are cases that an article may have different news/stories covered. For example, an article may contain short summaries of different news about several companies, e.g., the 1st paragraph is about Google's new product announcement, the 2nd paragraph is a summary of new drug announcement from Merck, and the 3rd paragraph is about stock market update. This kind of news briefing articles pose challenges to both the data annotators and the NLP algorithms. We have tried to not include this type of articles in our dataset by analyzing the patterns of articles of different article sources. But there are still some in the dataset, although we estimated that it is less than 0.05%.

Category Imbalance. The distribution of articles among topic categories are imbalanced, with some categories having more than 10% and some others having around 1% of the total articles. This is a problem for any dataset covering a larger number of categories, unless it has only several popular categories. We are working on reducing the impact of this issue, by adding more articles in the minority topic categories.

In conclusion, this paper presents a new Chinese news article dataset with 4.4 million articles. This is the first Chinese news dataset that has both multi-level topic labels and article full text. In addition to topic categories, half million articles also have summary. Besides topic category and summary, each article also has a rich set of metadata, such as title, full text, channel, source and source type. Users can use these metadata to do more news analysis. To ensure the quality of the dataset, human evaluations were conducted to evaluate the quality of topic categories and summaries. We hope this dataset could be used in various related applications in the future.

REFERENCES

[1] 20Newsgroups, 2018. http://qwone.com/~jason/20Newsgroups/

[2] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," arXiv preprint arXiv:1508.05326, 2015.

[3] Ricardo Campos, Vítor Mangaravite, Arian Pasquali , Alípio Mário Jorge, Célia Nunes , and Adam Jatowt, 2018. YAKE! Collection-Independent Automatic Keyword Extractor, ECIR

[4] CNewsCorpus, 2017. Chinese news classification dataset. http://download.cnblogs.com/finallyliuyu/corpus.rar

[5] Cortes, Corinna; Vapnik, Vladimir N. (1995). "Support-vector networks" (PDF). Machine Learning. 20 (3): 273–297. CiteSeerX 10.1.1.15.9362. doi:10.1007/BF00994018.

[6] L. Deng and J. Wiebe, "Mpqa 3.0: An entity/event-level sentiment corpus," in Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: human language technologies, 2015

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. NAACL-HLT.

[8] B. Dolan, C. Quirk, and C. Brockett, "Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources," in Proceedings of the 20th international conference on Computational Linguistics, 2004

[9] Florescu, C. and Caragea, C. (2017) A new scheme for scoring phrases in unsupervised keyphrase extraction. In Proceedings of the Advances in Information Retrieval - 39th European Conference on IR Research, ECIR 2017

[10] Ygor Gallina, Florian Boudin, and Beatrice Daille. 2019. KPTimes: A Large-Scale Dataset for Keyphrase Generation on News Documents. In Proceedings of the 12th International Conference on Natural Language Generation. Association for Computational Linguistics

[11] Shen Gao, Xiuying Chen, Piji Li, Zhaochun Ren, Lidong Bing, Dongyan Zhao, and Rui Yan. Abstractive text summarization by incorporating reader comments. In Proc. of AAAI, 2019

[12] Shen Gao, Xiuying Chen, Piji Li, Zhangming Chan, Dongyan Zhao, and Rui Yan. How to write summaries with patterns? learning towards abstractive summarization through prototype editing. In Proc. of EMNLP 2019.

[13] Siddharth Gopal and Yiming Yang. 2013. Recursive regularization for large-scale classification with hierarchical and graphical dependencies. KDD.

[14] Siddharth Gopal and Yiming Yang. 2015. Hierarchical Bayesian inference and recursive regularization for large-scale classification. ACM Transactions on Knowledge Discovery from Data (TKDD)

[15] Max Grusky, Mor Naaman, Yoav Artzi1, 2018. NEWSROOM: A Dataset of 1.3 Million Summaries with Diverse Extractive Strategies, NAACL 2018.

[16] Hasan, K. S. and Ng, V. (2014) Automatic keyphrase extraction: A survey of the state of the art. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014

[17] Karl Moritz Hermann, Tom´as Kocisk´y, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. Proc. of NeurIPS 2015.

[18] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," arXiv preprint arXiv:1801.06146, 2018

[19] Baotian Hu, Qingcai Chen, and Fangze Zhu. LCSTS: A large scale Chinese short text summarization dataset. In Proc. of EMNLP, pages 1967–1972, 2015.

[20] Lifeng Hua, Xiaojun Wan, and Lei Li. Overview of the nlpcc 2017 shared task: single document summarization. In Proc. of NLPCC, pages 942–947. Springer, 2017.

[21] Kuan-Hao Huang, Chen Li, and Kai-Wei Chang. Generating sports news from live commentary: A Chinese dataset for sports game summarization. In Proc. Of AACL, pages 609–615, 2020

[22] Jieba, 2013, https://github.com/fxsjy/jieb

[23] Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, and T. Mikolov, "Fasttext. zip: Compressing text classification models," arXiv preprint arXiv:1612.03651, 2016

[24] Svetlana Kiritchenko and Saif Mohammad. 2017. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics.

[25] K. Kowsari, D. E. Brown, M. Heidarysafa, K. J. Meimandi, M. S. Gerber, and L. E. Barnes, "Hdltex: Hierarchical deep learning for text classification," in 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA).

[26] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. Van Kleef, S. Auer et al., "Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia," Semantic Web, 2015.

[27] David D Lewis, Yiming Yang, Tony G Rose, and Fan Li. 2004. RCV1: A new benchmark collection for text categorization research. Journal of Machine Learning Research 5, Apr (2004), 361–397.

[28] Chin-Yew Lin. Rouge: A package for automatic evaluation of summarie1. Text Summarization Branches Out, 2004

[29] Xiaojun Liu, Chuang Zhang, X. Chen, Yanan Cao, and Jinpeng Li. Clts: A new chinese long text summarization dataset. In Proc. of NLPCC, 2020.

[30] Yang Liu and Mirella Lapata, 2019, Text Summarization with Pretrained Encoders, EMNLP

[31] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov, 2019, RoBERTa: A Robustly Optimized BERT Pretraining Approach, https://arxiv.org/abs/1907.1169

[32] Z. Lu, "Pubmed and beyond: a survey of web tools for searching biomedical literature," Database, vol. 2011

[33] E. L. Mencia and J. Fürnkranz, "Efficient pairwise multilabel classification for large-scale problems in the legal domain," in Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, 2008, pp. 50–65

[34] Rada Mihalcea and Paul Tarau, 2004. TextRank: Bringing Order into Texts, EMNLP

[35] Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, Jianfeng Gao , 2021, Deep Learning Based Text Classification: A Comprehensive Review, https://arxiv.org/abs/2004.03705

[36] Mu, F., Yu, Z., Wang, L., Wang, Y., Yin, Q., Sun, Y., Liu, L., Ma, T., Tang, J., Zhou, X.: Keyphrase extraction with span-based feature representations. arXiv preprint arXiv:2002.05407 (2020)

[37] Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. Summarunner: A recurrent neural network based sequencemodel for extractive summarization of documents. In Satinder P. Singh and Shaul Markovitch, editors, Proc. of AAAI, 2017.

[38] T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, and L. Deng, "Ms marco: a human-generated machine reading comprehension dataset," 2016.

[39] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in Proceedings of the ACL conference on Empirical methods in natural language processing, 2002

[40] Eirini Papagiannopoulou, Grigorios Tsoumakas, 2019. A Review of Keyphrase Extraction, https://arxiv.org/abs/1905.05044

[41] Hao Peng, Jianxin Li, Yu He, Yaopeng Liu, Mengjiao Bao, Lihong Wang, Yangqiu Song, and Qiang Yang, 2018, Large-Scale Hierarchical Text Classification with Recursively Regularized Deep Graph-CNN, WWW

[42] Xipeng Qiu, Jingjing Gong, Xuanjing Huang, 2017, Overview of the NLPCC 2017 Shared Task: Chinese News Headline Categorization, NLPCC 2017

[43] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "Squad: 100,000+ questions for machine comprehension of text," arXiv preprint arXiv:1606.05250, 2016.

[44] Reuters-21578 Text Categorization Collection, 1999. http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html

[45] John A. Rice. 2006. Mathematical Statistics and Data Analysis (3rd ed.). Duxbury Advanced

[46] Evan Sandhaus. 2008. The New York Times Annotated Corpus LDC2008T19. In Linguistic Data Consortium. – Nytimes

[47] Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. In Proc. of ACL, 2017.

[48] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in Proceedings of the 2013 conference on empirical methods in natural language processing, 2013, pp. 1631–1642.

[49] C. Sun, X. Qiu, Y. Xu, and X. Huang, "Howto fine-tune bert for text classification?" in China National Conference on Chinese Computational Linguistics. Springer, 2019, pp. 194–206.

[50] Maosong Sun, Jingyang Li, Zhipeng Guo, Yu Zhao, Yabin Zheng, Xiance Si, Zhiyuan Liu. 2016. THUCTC: An Efficient Chinese Text Classifier. http://thuctc.thunlp.org/

[51] Si Sun, Zhenghao Liu, Chenyan Xiong, Zhiyuan Liu, and Jie Bao,Capturing Global Informativeness in Open Domain Keyphrase Extraction, https://arxiv.org/abs/2004.13639

[52] Toutiao, 2018. TouTiao Text Classification for News Titles (TNEWS) (CLUE Benchmark) Dataset, https://metatext.io/datasets/toutiao-text-classification-for-news-titles-(tnews)-(clue-benchmark

[53] Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, and Yasemin Altun. 2005. Large margin methods for structured and interdependent output variables. Journal of Machine Learning Research 6, Sep (2005)

[54] Danqing Wang, Jiaze Chen, Xianze Wu, Hao Zhou and Lei Li, 2021, CNewSum: A Large-scale Chinese News Summarization Dataset with Human-annotated Adequacy and Deducibility Level. https://arxiv.org/abs/2110.10874

[55] Thomas Wolf, Lysandre Debut, Victor Sanh, Ju et al. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing, https://arxiv.org/abs/1910.03771

[56] Xuefeng Xi, Zhou Pi, and Guodong Zhou. Global encoding for long chinese text summarization. ACM Trans. Asian Low-Resource Language Information Process., 19(6), 2020

[57] Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, et al., 2020. CLUE: A Chinese Language Understanding Evaluation Benchmark. https://arxiv.org/pdf/2004.05986.pdf

[58] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in Advances in neural information processing systems, 2015, pp. 649–657.

[59] Qingyu Zhou, Nan Yang, Furu Wei, Shaohan Huang, Ming Zhou, and Tiejun Zhao. Neural document summarization by jointly learning to score and select sentences. In Proc. of ACL, 2018.

[60] Bojanowski, Piotr and Grave, Edouard and Joulin, Armand and Mikolov, Tomas, 2017, Enriching Word Vectors with Subword Information, Transactions of the Association for Computational Linguistics, Volume 5.

[61] Eva Gibaja and Sebastian Ventura, 2014. Multi-label learning: a review of the state of the art and ongoing research. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, *4*(6), 411–444.

[62] Min-Ling Zhang and Zhi-Hua Zhou, 2013. A review on multi-label learning algorithms. IEEE Transactions on Knowledge and Data Engineering. https://doi.org/10.1109/TKDE.2013.39.

[63] Quanzhi Li, Armineh Nourbakhsh, Sameena Shah, Xiaomo Liu, 2017. Real-time novel event detection from social media, the IEEE 33rd International Conference on Data Engineering (ICDE 2017)

[64] Quanzhi Li, Qiong Zhang, 2020. Abstractive Event Summarization on Twitter, Proceedings of the Web Conference (WWW 2020)

[65] Quanzhi Li, Qiong Zhang, 2021. Twitter Event Summarization by Exploiting Semantic Terms and Graph Network, Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence

[66] Quanzhi Li, Sameena Shah, Xiaomo Liu, Armineh Nourbakhsh, 2017. Data sets: Word embeddings learned from tweets and general data, the 11th International AAAI Conference on Web and Social Media (ICWSM 2017)

[67] Quanzhi Li, Sameena Shah, Xiaomo Liu, Armineh Nourbakhsh, Rui Fang, 2015. Tweetsift: Tweet topic classification based on entity knowledge base and topic enhanced word embedding, in Proceedings of the 25th ACM CIKM

[68] Xiaomo Liu, Quanzhi Li, Sameena Shah, Robert Martin, John Duprey, 2017. Reuters tracer: Toward automated news production using large scale social media data, 2017 IEEE International Conference on Big Data (IEEE BigData 2017)

[69] Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, et al., 2020. MIND: A Large-scale Dataset for News Recommendation, ACL 2020