RESEARCH-ARTICLE

Data-Augmented and Retrieval-Augmented Context Enrichment in Chinese Media Bias Detection

LUYANG LIN, Chinese University of Hong Kong, Hong Kong, Hong Kong

JING LI, The Hong Kong Polytechnic University, Hong Kong, Hong Kong, Hong Kong

KAM FAI WONG, Chinese University of Hong Kong, Hong Kong, Hong Kong

Citation in BibTeX format

Data-Augmented and Retrieval-Augmented Context Enrichment in Chinese Media Bias Detection

LUYANG LIN, MoE Key Laboratory of High Confidence Software Technologies, The Chinese University of Hong Kong, Hong Kong, China

JING LI, Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China

KAM-FAI WONG, MoE Key Laboratory of High Confidence Software Technologies, The Chinese University of Hong Kong, Hong Kong, China

Warning: This article contains content that may be offensive or controversial.

With the increasing pursuit of objective reports, automatically understanding media bias has drawn more attention in recent research. However, most of previous work examines media bias from Western ideology, such as the left and right in the political spectrum, which is not applicable to Chinese outlets. Based on the previous lexical bias and informational bias structure, we refine it from the Chinese perspective and go one step further to craft data with seven fine-grained labels. To be specific, we first construct a dataset with Chinese news reports annotated by our newly designed system, and then conduct substantial experiments on it. However, the scale of the annotated data is not enough for the latest deep-learning technology, and the cost of human annotation in media bias, which needs a lot of professional knowledge, is too expensive. Thus, we explore some context enrichment methods to automatically improve these problems. In Data-Augmented Context Enrichment (DACE), we enlarge the training data; while in Retrieval-Augmented Context Enrichment (RACE), we improve information retrieval methods to select valuable information and integrate it into our models to better understand bias. Our results show that both methods outperform our baselines, while RACE methods are more efficient and have more potential.[1]

CCS Concepts: • **Computing methodologies → Natural language processing;**

Additional Key Words and Phrases: Chinese media bias detection, natural language processing, data augmentation, information retrieval

---

[1]The annotated corpus is released in https://github.com/lylin0/lin2023data

---

1 Introduction

Media bias is a common journalism phenomenon in news articles, where authors inject their personal viewpoints into the reports. Automatic detection of media bias has gained more attention in recent years due to its critical role in pursuing objective information dissemination [47, 51]. However, much of the existing research has been conducted within the context of Western ideology [25], such as the left and right in the political spectrum [1, 5, 12, 18, 21, 29].

While from the Chinese perspective, the problem is more complex. The most direct phenomenon is that language expression habits cannot be summarized according to the leaning of stances as Western countries do. In this situation, the political spectrum may not be the only reason for the formation of media bias [63], other factors, such as the capital market, also play a crucial role [28].

Therefore, we propose a new label system for Chinese media bias. Considering the limitations of most existing studies, which coarsely characterize media bias with either lexical [24, 36, 50, 62] or informational features [6, 8, 9, 43], we align with the concept proposed by Fan et al. [16], which consists both **lexical bias** and **informational bias**. Further, to define the bias more clearly and objectively, we craft seven fine-grained labels following the journalism theories [4] and discuss with domain experts in media and journalism.

To further illustrate how our formulation works in the Chinese perspective, even in some situations with distinct differences in political stances, Figure 1 shows two article excerpts about COVID-19, one (Article 1) from United Daily News (pro-CNP) and the other (Article 2) Liberty Times (pro-DPP).[2] Despite the opposite political stance, both of them exaggerate the outbreak in Wuhan through writing style. For content selection, they both omitted or framed some aspects in their report, whereas their opposite positions resulted in the fact that Article 1 weighs more on the government's voice compared to Article 2's focus on the individual's voice on social media.

For data collection, we gathered 300 Chinese articles reporting COVID-19 news from five media outlets in Hong Kong, Taiwan, and Singapore. Following the previous practice [16], these articles are formed into 100 triplets, reporting the same event and from a different source. In-house annotators and domain experts then manually label coarse-grained and fine-grained media bias on the token-level and sentence-level content based on our newly designed annotation system. Because one sentence may contain various methods, our multi-label assumption can be more comprehensive in understanding bias, compared with single-label applied in the existing practice [1, 40].

Based on the multi-label setup, we employ popular **Natural Language Processing** (**NLP**) baselines to predict both sentence-level and token-level media bias and the results show that the pre-trained **Bidirectional Encoder Representations from Transformers** (**BERT**) performs the best. Then, the joint training is conducted to analyze the interactions between token-level and sentence-level contexts in detecting media bias, while prior studies tend to discuss them separately [16, 38]. It enables better results than their separate training counterparts, which shows features within the sentence play an important role in lexical bias. As for informational bias, which can be benefited from encoding richer context (other articles from the same triplet), is more challenging and shows, that is, usually signaled by external context.

When conducting the qualitative analysis to examine the advances and challenges of existing NLP models in detecting bias, we observe that, due to the small scale of the data, overfitting would be a big problem when the latest and strongest models are applied to this task. In social science research, media bias is usually analyzed in quantitative measure [13], and the whole outlet is taken as a unit. Although smaller units, such as articles or sentences [1, 12, 18], are focused when exploring its automatic detection with various computational methods [6, 29], the scale of the media bias dataset is still far too small compared with other machine learning tasks. Especially

---

[2]**Chinese Nationalist Party** (**CNP**) and **Democratic Progressive Party** (**DPP**) are two major parties in Taiwan.

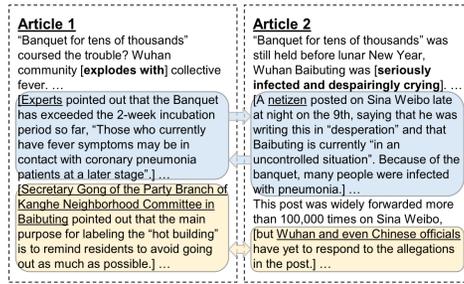| Article 1	Article 2
"Banquet for tens of thousands" coursed the trouble? Wuhan community [**explodes with**] collective fever. …	"Banquet for tens of thousands" was still held before lunar New Year, Wuhan Baibuting was [**seriously infected and despairingly crying**]. …
[Experts pointed out that the Banquet has exceeded the 2-week incubation period so far, "Those who currently have fever symptoms may be in contact with coronary pneumonia patients at a later stage".] …	[A netizen posted on Sina Weibo late at night on the 9th, saying that he was writing this in "desperation" and that Baibuting is currently "in an uncontrolled situation". Because of the banquet, many people were infected with pneumonia.] …
[Secretary Gong of the Party Branch of Kanghe Neighborhood Committee in Baibuting pointed out that the main purpose for labeling the "hot building" is to remind residents to avoid going out as much as possible.] …	This post was widely forwarded more than 100,000 times on Sina Weibo, [but Wuhan and even Chinese officials have yet to respond to the allegations in the post.] …

Fig. 1. Two COVID-19 news excerpts from media outlets with opposite political stances. The boldface [**tokens**] in brackets indicate writing bias in *exaggeration*. Text in content bias is marked with frames: blue indicates *omission* and yellow *framing*. Arrows point to complementary voices over the same topic and their sources are underlined for easy reading. (The original text in chinese is shown in Table 7.)

when high-quality data and labels are pursued, which require solid expertise, the cost is too high to annotate a large dataset by humans.

In further discussions, we explore whether the external corpus, which is large-scale but unlabeled, can contribute to the media bias detection task through some contextual enrichment methods. Inspired by the support of token-sentence joint training in lexical bias and the additional context in informational bias, we consider two main directions. One is **Data-Augmented Context Enrichment (DACE)**, which increases the size of the dataset by generating more training cases based on the token and sentence features. Another is **Retrieval-Augmented Context Enrichment (RACE)**, which improves information retrieval methods to select valuable information from a large outside knowledge source and feeds it to the model when a case is processed.

From the additional experiments, we notice that with the enrichment of the context, whatever method is used, the results outperform the baselines, which is also verified in the English dataset **Bias Annotation Spans on the Informational Level (BASIL)** [16]. In general, RACE methods get better results. When the context is considered during the retrieval process. the results are further improved. However, the performance of the model would be disturbed if the time sequence is ignored in the news-related task, where the reports are time-sensitive. We would consider this in future work.

In summary, to the best of our knowledge,

— *We design a novel annotation system to explore Chinese media bias, contributing the first media bias benchmark in a multi-label setting.*
— *We finetune the pre-trained models by the media bias dataset and comprehensively study the difference in the detection of lexical bias and informational bias.*
— *Inspired by the insights, we propose and verify that contextual enrichment methods, such as Data Augmentation and Information Retrieval, can benefit media bias detection tasks. Explore how external unlabeled data can contribute to such practical and small-scale tasks.*[3]

2 Media Bias Structure Design

Based on the **lexical bias** and **informational bias** bias structure proposed by Fan, Lisa, et al. [16], we go one step further to craft four fine-grained labels in lexical bias and three fine-grained labels in informational bias so that it could be clearer defined and better fitted to the Chinese context, which are listed below.

---

[3]We will make our annotated corpus publicly available upon publication.

Table 1. Our Media Bias Annotation Examples

Lexical Bias Annotation Examples
(a) ... They are [crazy for]$_{\text{Exaggeration}}$ air tickets, flying around the earth, hoping to "return to [the safest]$_{\text{Exaggeration}}$ place" ...
(b) ... some people even continue to go out to have dinner with friends. The "[Buddhism]$_{\text{Stereotype}}$" is surprising ...
(c) ... [some experts doubt the effectiveness of mass quarantine, pointing out that "the virus will actually spread"]$_{\text{Ventriloquism}}$
(d) ... and the situation [is bound to continue to deteriorate]$_{\text{Opinion}}$ ...

Informational Bias Annotation Examples
Annotation Set 1
Article 1 ... The Chinese song made many [foreign netizens sarcastically say in a "high-level black" way that this "North Korean boy" is too cute.] ... [According to "The Paper", the song has a detailed and in-depth description of the mobile cabin hospital, which helps patients fight the virus and relieve tension.]$_{\text{Framing (Positive)}}$ ...
Article 2 ... [Some netizens denounced the creator for "entertaining disasters"]. [The Beijing News online commentary even denounced this as "forcibly extolling disasters".]$_{\text{Framing (Negative)}}$ ... [Jiang Junrong said "A lot of places are "cleared" now. Don't we need to show a little bit of optimism? Don't we need to show a better mentality?"]$_{\text{Omission of Article 1}}$ ...
Article 3 ... [Some netizens even described the video as "seemingly back in the 1980s and 1990s".] ... [Composer Jiang Junrong said in response that "They don't know my state of life, and I don't make any calculations".]$_{\text{Omission of Article 1}}$ [The Beijing News commented on this, saying that turning it into the material of cheerful children's songs not only dilutes the disaster background, but also eliminates the seriousness of the fight against the epidemic.]$_{\text{Framing (Negative)}}$
Annotation Set 2
Article 1 ... "Shincheonji" chairman will hold a press conference to publicly respond to the epidemic for the first time.
Article 2 South Korean leader of "Shincheonji" kneels to apologize and thank the government for fighting the epidemic.
Article 3 Weird! The leader of "Shincheonji" kneels and [wears Park Geun-Hye's watch]$_{\text{Sensationalist}}$.

For simplicity, we do not show the whole annotation set for lexical bias examples. The informational bias annotations are also omitted for lexical bias examples (so as informational bias examples). We underline the source in each part of the articles for easier understanding. (The original text refers to Table 8.)

2.1 Lexical Bias

Although we note there are some prior studies for lexical bias [20, 44, 48], to the best of our knowledge, there does not exist a systematic study summarizing all sub-types of lexical bias for media bias. Building upon existing literature and also in consultation with domain experts, we design the following sub-types to measure the bias in writing styles.

Exaggeration. It is defined as using eye-catching wordings with strong emotional implications or describing things in a way, that is, more than it really is, to influence an audience [46]. For example, articles in Figure 1 use tokens "explode" and "despairingly" (in the 1st paragraph) to exaggerate the phenomenon, while other media (not in Figure 1) only uses "residents fever".

Stereotype. Sometimes a media source may favor or attack a particular group, or a person, and describe them with wordings reflecting their stereotype to them. We refer to this kind of bias as _Stereotype_, or we can call it labeling. It can also describe the events or people using words that may contain judgments [24, 42], e.g., Table 1(b) employs "Buddhism" to criticize the slack attitude to epidemic prevention.

Ventriloquism. We annotate text spans as _Ventriloquism_ when experts or witnesses are quoted in a way that intentionally voices the author's own opinion [23]. For instance, in Table 1(c), part of a sentence from the expert is quoted to emphasize the uselessness of mass quarantine, which may be just the author's opinion.

Opinion. Inspired by Anand [2], this sub-type covers cases like stories that focus not on what has occurred, but primarily on what might occur, or it is just the opinion of the author. We can also name it Speculative Content. In Table 1(d), "continue to deteriorate" is just the speculation of the author but not the fact.

2.2 Informational Bias

Informational bias, a.k.a., content bias in some studies, refers to the bias reflected by selecting or emphasizing part of the event. It can be indicated by comparing the report with the counterpart views from other media outlets [11, 16, 55]. Our design of informational bias sub-types is mostly inspired by Baker [4].

For certain informational bias sub-types, such as *Omission*, we rely on the concept of "stakeholders" to help identify the informational bias more objectively: the information from different sources is first sorted out into several stakeholders, and we then compare these stakeholders within or across articles to evaluate the existence of informational bias.

Omission. We use *Omission* to describe situations where some reports ignore information that tends to disprove their claims [4, Chapter 2]: if an article does not contain the content about stakeholders that appeared in other articles, then it has the informational bias of *Omission*. In our displayed cases, the missing descriptions of the composer's motivation for writing the song in Annotation Set 1 in Table 1 and the citizen's situation in Figure 1 are both omissions.

Framing. We say an article has the informational bias of *Framing* when multiple stakeholders are contained but not balanced or undue weight. Such imbalances can be raised from the following aspects: different positions [4, Chapter 4], different lengths of the description [4, Chapter 6], and different sentiment polarities of selected content. The case in Figure 1 reflects the bias on the different lengths of descriptions of the official response, while the Annotation Set 1 in Table 1 reflects the bias on the positive and negative sentiment of the reports.

Sensationalist. This informational bias sub-type refers to the event, that is, not closely relevant to the theme, but selected to excite the greatest number of readers [52]. For example, in the Annotation Set 2 of Table 1, the author focuses on the same style of watch in the report about the epidemic to draw readers' attention.

3 Dataset Construction

3.1 Data Collection

Different media outlets normally have different descriptions for the same event (as presented in Figure 1), providing us with an opportunity to study where bias raises. Motivated by this observation, in our paper, we mainly focus on studying bias from Chinese news related to the COVID-19 pandemic. We first collect a large collection of news articles from five media outlets over three different areas or countries, ranging from January 2020 to August 2020.[4] We then filter out articles that contain only objective information, such as COVID-19 case updates. Finally, 300 articles are selected following Fan, Lisa, et al. [16] to construct 100 sets, with each set containing three articles from different outlets but discussing the same event.

3.2 Human Annotation

Annotation Process. Given the different nature of lexical and informational bias, we annotate them separately, though the basic tool functions are shared. To illustrate our annotation tool, Figure 2 shows a sample interface for lexical bias while that for informational bias is put in Figure 7. We recruit in-house annotators from a journalism background and provide them with hands-on and specific training. All types of bias are explained through cases and they are required to exercise, discuss, and analyze the results to warm up the annotation. To further mitigate the bias of annotators, we detail our regulations to assist guideline understanding; e.g., four main source categories are

---

[4]Media outlets include Wen Wei Po, Apple Daily, United Daily News, Liberty Times, and Lianhe Zaobao.
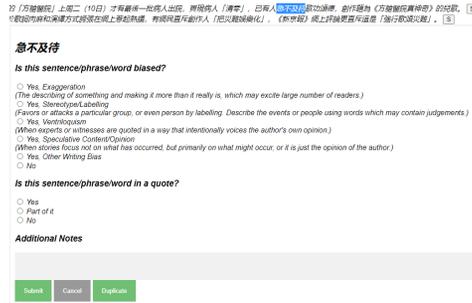
Fig. 2. The interface of lexical bias annotation tool. Except for the sentence-level annotation, the annotators can also annotate in spans for token-level annotation.

exampled (officers, citizens, experts, and media) to help annotators learn stakeholders' concepts for better labeling content bias. During the annotation, the constructed sets of articles through a web page are presented to annotators. Annotators should first go through all three articles in the set, and determine if a given sentence contains a certain sub-type of the bias.[5] Then, the annotators are asked to specify the text boundary that indicates bias. After that, media bias labels are assigned to two-level contexts, sentence-level, and token-level. In this way, our annotation not only indicates what media bias types are observed, but also how it is presented by certain text spans.

To further present labeling details, all the sub-types of lexical bias and informational bias that annotators can choose from (also referred to as **fine-grained** labels) are summarized in Table 3. Based on the existence of a certain bias sub-type, we also give each sentence a **coarse-grained** label, namely *containing a lexical bias*, *containing an informational bias*, and *no bias exists*. Both the coarse-grained and fine-grained annotations will be explored in our experiments in Section 4.

IAA. Each set is annotated by at least two in-house annotators. Cohen's Kappa [3] and F1 score is used following Fan, Lisa, et al. [16] and Toprak et al. [54] to measure the reliability between annotators in the sentence-level and token-level.

The IAA scores are calculated by Cohen's Kappa at both the sentence level and the token level. We mainly calculate in three categories, lexical bias, informational bias, and the combination of two, which is the overall result. As for the sub-category, in the combined category, we consider Biased-Grained with two labels (biased and non-biased), Coarse-Grained with three labels(lexical biased, informational biased, and non-biased), and Fine-Grained with all seven sub-types and non-biased labels. In the sub-type bias, for example, Exaggeration, we consider whether it is labeled as exaggeration or non-exaggeration and so on. The IAA scores are shown in Table 2.

We can observe that, the agreement scores between sentence level and token level do not vary much in informational bias, the boundary of bias text span is clear. However, in lexical bias, it is harder to specify the biased tokens. Also, some subtypes that require more subjective knowledge, such as Exaggeration, Ventriloquism, and Sensationalist, get lower agreement scores. In general, these are above the bar of good agreement indication [17], while also implying the challenges in media bias annotation, especially in finer granularity and in our public health-related data. It is because of the subjective nature of media bias and consistent with the previous findings that high IAAs are not usually observed in subjective and fine-grained NLP annotations [49]. Besides, fine-grained lexical bias shows relatively lower agreement than informational attributed to more categories involved.

---

[5]To exhaustively measure the media bias from language, text spans that are in quotation marks are also considered.

Table 2. Breakdown Statistics of **Inter-Annotator Agreement** (**IAA**) in
Sentence-Level and Token-Level of the Dataset

Type	Sub-Type	Sent-level	Token-level
Combined	Biased-Grained	0.671	0.613
	Coarse-Grained	0.617	0.595
	Fine-Grained	0.505	0.565
Lexical Bias	Biased-Grained	0.740	0.568
	Fine-Grained	0.603	0.534
	Exaggeration	0.471	0.322
	Stereotype	0.642	0.569
	Ventriloquism	0.225	0.144
	Opinion	0.625	0.535
Informational Bias	Biased-Grained	0.654	0.644
	Fine-Grained	0.628	0.613
	Omission	0.676	0.663
	Framing	0.485	0.516
	Sensationalist	0.223	0.251

In Biased-Grained, we only consider whether a sentence has been labeled as any
sub-type of bias, which means we calculate the agreement between two labels:
biased and non-biased. For Coarse-Grained and Fine-Grained, the definition can
refer to Section 3.2.

To further control the quality of our annotations, a domain expert is invited to manually check
the in-house annotations and decide the final labels (as the gold-stand), compared to which our
in-house annotators exhibit 0.89 F1 on average. It implies acceptable annotation quality with
hands-on training and in-house verification, consistent with previous practice to use expert anno-
tators [49] to improve the quality and the low IAA for similar media bias annotation tasks through
crowdsourcing [40].

3.3 Data Feature

Number of Bias Types per Sentence. In total, 5,012 sentences are used for our annotation task.
Among these, 59.8% of sentences contain at least one bias sub-type label. Sentences in our dataset
may have more than one annotated span, with each sentence or even each span possibly having
different bias sub-types: we observe that 1,313 sentences (26.2%) are such cases. It thus inspires us
to perform experiments in a multi-label framework. The breakdown statistics for each bias sub-type
are reported in Table 3.

Data Imbalance. From Table 3, we can observe that the numbers of sentences in each sub-type
are quite imbalanced and there is even a more than 20 times difference. One of the reasons is that
some sub-types, for example, *Sensationalist* seldom appear in health news, though widely applied
in other domains, e.g., entertainment. The data imbalance issue may concretely challenge models
to characterize certain media bias sub-types, although data augmentation may potentially mitigate
the problem, which will be discussed later.

Sentence Length and Position. We calculate the average length for sentences containing bias
sub-types. We observe that the sentences with informational bias on average contain 52.2 tokens,
which are longer compared to those with lexical bias (19.2 tokens). One possible explanation is that
the informational bias needs more length and context to present compared with lexical bias.

The position distribution of biased sentences is presented in the Figure 3. We observe that the
position distribution of lexical bias differs a lot from the informational bias. Compared with a

Table 3. Breakdown Statistics of Sentences for Different Bias Sub-Types

Type (Coarse)	Sub-Type (Fine-grained)	# Sent.
Contain Lexical Bias	Exaggeration	681
	Stereotype	690
	Ventriloquism	54
	Opinion	1,318
	Sent. with ≥1 lexical sub-type	2,022
Contain Informational Bias	Omission	1,569
	Framing	427
	Sensationalist	67
	Sent. with ≥1 informational sub-type	1,819

Each sentence has a coarse-grained label (see Section 3.2) and associated fine-grained bias sub-types if applicable.
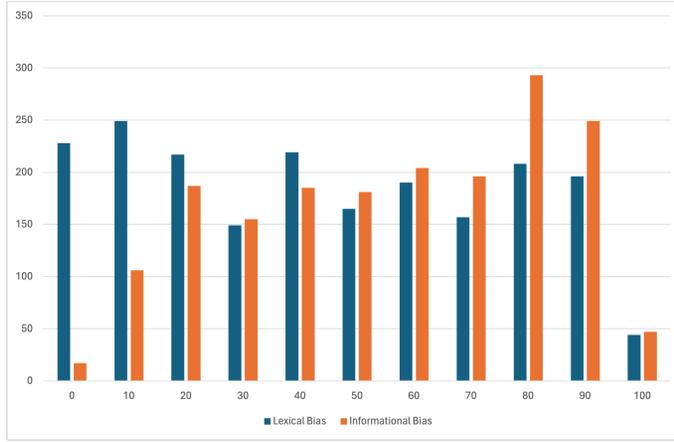

Fig. 3. The positions of biased sentences in the articles. x-axis: the position of the percentage in the article. 0 means the beginning of the article and 100 stands for the end of the article. y-axis: the number of biased sentences in the position range.

large amount of lexical bias that appears at the beginning of the article, more sentences with informational bias are in the second half of the article. We assume that lexically biased sentences are usually short and put at the beginning of the article to attract the attention of the readers so that the informational biased sentences which are more complicated can be read. It also aligns with our assumption of the length features.

4 Dataset Analysis

We first conduct some basic experiments on our dataset and analyze the results to test whether our dataset is feasible for the media bias detection task.

4.1 Basic Task Design

Sentence-level Classification. We first study whether it is feasible to develop automatic methods to measure the media bias from the text. To this end, we design a sentence-level multi-label bias classification task. Specifically, we consider the following settings (listed as four columns in Table 4).

Table 4. Results for Sentence-Level Classification, Measured by Hamming Loss (HL), Jaccard Index (JI), and Weighted F1 Metrics

Method	Coarse-Grained			Fine-Grained Combined			Fine-Grained Lexical			Fine-Grained Informational		
	HL (↓)	JI (↑)	F1 (↑)	HL (↓)	JI (↑)	F1 (↑)	HL (↓)	JI (↑)	F1 (↑)	HL (↓)	JI (↑)	F1 (↑)
Random	0.484	0.247	0.396	0.497	0.110	0.199	0.497	0.109	0.197	0.496	0.106	0.191
GPT3.5 (Zero-Shot)	0.328	0.045	0.085	0.135	0.046	0.087	0.131	0.120	0.214	0.162	0.036	0.070
GPT3.5 (Few-Shots)	0.334	0.123	0.220	0.128	0.096	0.175	0.136	0.140	0.245	0.150	0.072	0.134
SVM	0.450	0.200	0.333	0.355	0.122	0.218	0.393	0.093	0.170	0.303	0.168	0.288
CNN	0.299	0.360	0.529	0.116	0.200	0.333	0.107	0.314	0.478	0.142	0.232	0.377
RNN	0.287	0.195	0.326	0.124	0.135	0.238	0.124	0.203	0.338	__0.126__	0.189	0.317
LSTM	0.263	0.325	0.491	0.124	0.252	0.402	0.113	0.289	0.449	0.136	0.231	0.375
BERT	__0.259__	__0.381__	__0.552__	__0.115__	__0.342__	__0.510__	__0.101__	__0.390__	__0.561__	0.129	__0.310__	__0.474__
BERT+context	0.271	**0.396**	**0.568**	**0.112**	**0.352**	**0.521**	0.105	0.375	0.545	0.134	**0.321**	**0.486**
Joint Train	0.260	0.324	0.490	0.119	0.300	0.461	**0.095**	**0.421**	**0.592**	**0.126**	0.294	0.454

The best performance for models without additional information is underlined. The best performance of all is marked bold. ↑ indicates that a higher value is better and ↓ is the opposite, and the scale of HL, JI, and F1 are from 0 to 1. The design of tasks refers to Section 4.1.

In Coarse-Grained classification, we only consider whether a sentence has any type of lexical bias or informational bias, using coarse-grained labels discussed in Section 3.2. While in fine-grained classification, the model needs to specify four sub-types of lexical bias (denoted as Fine-Grained Lexical) or three sub-types of informational bias (henceforth Fine-Grained Informational). To further test the model capability, we combine all bias sub-types as a seven-label multi-label classification task (named as Fine-Grained Combined).

Token-level Sequence Tagging. Treating each sentence as a whole and predicting media bias labels is clearly insufficient for studying bias—we are more interested in identifying tokens that cause certain types of bias. Our annotated dataset supports this kind of experiment, as annotators are asked to select text spans that trigger bias during annotation (discussed in Section 3.2). We thus formulate our token-level media bias detection as a sequence tagging task, with the goal of marking the bias span tokens in each sentence. All annotated spans are first transformed into BIO scheme (with B tag denoting *begin*, I for *intermediate*, E for *end*, S for *single character* and O for *outside*) and then used for model training.

Joint Training. Then, we detect the bias by joint training of sentence-level and token-level classification tasks to examine the interactions of the whole sentence and local tokens. In the warm-up period, two classifiers are trained iteratively: after three epochs of training of sentence-level classification, the parameters in the classifier layer for the sentence are frozen, and then three epochs of training are conducted in the token-level classifier. The losses of two classification tasks are added together as the optimization target after the warm-up steps.

Context Feature Adding. Based on the human annotation process and previous work [55], we also want to learn the effects of additional context, and the experiment is conducted by adding sentences in the same sets as the context feature. As shown in Figure 4, we first encode the target sentence through BERT and at the same time, summarize the context sentences, which are from the same set as the target sentence. Due to the length limitation of BERT, we use the Longformer pre-train model to process the context content. In the next stage, after encoding by two separate **Long Short-Term Memory** (**LSTM**) [26] modules, the representation of the target sentence is concatenated with the context sentence representation and then classified to obtain sentence-level predictions.

Large Language Model (GPT). Recently, ChatGPT, a strong model that interacts in a conversational way, was released by OpenAI. It performs well in almost every task that machines have attempted to do in human current expectations. We also test our task on ChatGPT to find its
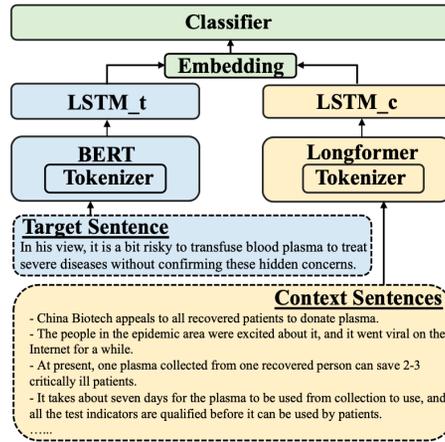
Fig. 4. The structure of our model with context feature added. The target sentence is encoded by BERT and an LSTM layer, context sentences are encoded by longformer and an LSTM layer. Their embeddings are concatenated as the input of the classifier for final results. In this case, the context sentences are all confident with the plasma treating method. They will be helpful for judging the target sentence, which doubts the feasibility of this method.

strengths and weaknesses, so that we can better plan our future work and cooperate with those strong large language models.

4.2 Experimental Setup

Our non-neural models include the **Support Vector Machine (SVM)** classifier in sentence-level classification and the **Conditional Random Field (CRF)** in token-level sequence tagging. Unigram and bigram features are used in SVM, while we only apply unigram features to CRF. Also in SVM, **One-vs-the-rest (OvR)** multi-class strategy is applied.

We train our dataset using a set of neural models, including **Convolutional Neural Network (CNN)** [37], **Recurrent Neural Network (RNN)** [15], and LSTM [26] models, 300-dimension pre-trained Chinese embedding provided by Li et al. [39] is applied in these neural models.

We also experiment with pre-trained models including BERT [14] with pre-trained tokenizer and pre-trained model, which is conducted through Hugging Face [59] and *Chinese* version is chosen.[6] Then we fine-tune BERT on our dataset by AdamW [41] optimizer. When the Longformer [7] model is applied, the basic setting is the same.

In GPT test experiments, we use the model *gpt-3.5-turbo-16k* through LangChain and give the descriptions of each type of bias. The prompts are shown in Appendix C. When conducting few-shots experiments, we randomly sample 16 cases from the training set.

We randomly divide our annotated 100 sets by 8:1:1 into training, validation, and test splits, with each containing 3,915 sentences, 608 sentences, and 489 sentences. We then select the model with the lowest validation loss used for evaluation. In our experiments, the learning rate is 5e-6, the batch size is 16, and the maximum length for each sentence is 128. In the context feature adding a task, we randomly selected eight sentences from the set instead of the whole set due to the limitation of our computing sources. We use Hamming loss, Jaccard index, and F1 metrics to measure our model

---

[6]*BertForSequenceClassification* is used for sentence-level classification and *BertForTokenClassification* is used for token-level classification.

Table 5. Experimental Results on Token-Level Bias Classification Task

Method	Lexical			Informational			Combined		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Random	0.745	0.201	0.298	0.500	0.197	0.279	0.478	0.199	0.277
CRF	0.728	0.853	0.786	0.434	0.529	0.440	0.432	0.452	0.408
RNN	0.815	0.847	0.825	0.591	0.602	0.589	0.609	0.605	0.600
LSTM	0.810	0.851	0.821	0.598	0.609	0.594	0.610	0.603	0.594
BERT	**0.860**	0.860	0.858	0.649	0.651	0.650	**0.675**	**0.675**	**0.674**
Joint Train	0.858	**0.866**	**0.861**	**0.673**	**0.658**	**0.661**	0.674	0.673	0.671

We use the same notations as in Table 4.

performance. The micro average method is used for the Jaccard index and F1 score in sentence-level classification tasks.

In all the experiments, we include a random baseline model, which randomly selects half of the sentences as containing a specific bias sub-type.

4.3 Basic Results

4.3.1 Sentence-level Results. We start the experimental discussions with the comparison results of sentence-level classification in Table 4 and the following observations can be drawn.

First, neural models in general outperform their non-neural counterparts. It indicates our task is challenging, where *n*-gram features may fail to signal media bias. Among neural models, BERT performs the best, especially in fine-grained scenarios, possibly due to its superior semantic understanding ability gained via large-scale pre-training and more complex structure.

Second, within fine-grained predictions, most models perform worse for informational bias than lexical. The possible reason is informational bias may require more knowledge to understand while some token-level indicators may sufficiently signal lexical bias. To probe into this point, we feed the BERT model with richer context (content from other articles in the same set), enabling it to compare different source's information.

The results of this BERT+context variant (BERT with context feature adding) are also listed in Table 4. It is observed that the learning of fine-grained content bias can indeed benefit from richer context, because models can compare the content selected by different authors and better capture semantics. This finding is consistent with that drawn by Fan et al. [16] for coarse-grained informational bias (as well as our coarse-grained results), and together inspire the future potential in exploring automatic context formation and learning to understand informational bias, which also discuss in Section 5.5.

4.3.2 Token-level Results. Then, we examine how models perform to understand token-level bias and report the results in Table 5. We observe that both non-neural and neural models present fine recall, whereas fine-tuned BERT still outperforms other baselines, indicating the helpfulness of language pre-training and complex structure for token-level bias understanding as well.

We also find the overall results for lexical bias are much better than informational bias. It is probably because informational bias tends to be reflected with richer context, rendering longer spans marked as its indicators (as discussed in Section 3.3) and lower chances to perfectly predict for all involved tokens.

4.3.3 Joint Training Results. We have discussed sentence-level and token-level results to learn that lexical bias and information bias get effects from different levels of context. To further examine their interactions, we jointly train a classifier to predict two-level media bias with multi-task

learning (henceforth Joint Train). The results are reported separately in the last row in Tables 4 and 5 for easier comparison purposes with other methods.

For sentence-level results, Joint Train provides more performance gain to fine-grained lexical bias detection, while informational bias benefits more from additional context (discussed in Section 4.3.1). It is because, as we speculated, the judgment basis of lexical and informational bias is quite different; the former may require more knowledge and the latter may be explicitly indicated by a few tokens. As a result, lexical bias prefers features that reflect internal semantics, such as tokens, while adding external information would better benefit informational bias learning.

On the token-level, the benefit of joint training is observed universally for both lexical and informational bias, suggesting a sentence-level view gained on comprehensive semantics helps capture token-level bias.

4.4 Analysis

4.4.1 Analysis of Fine-grained Classification. The above discussions mostly concern models' averaged fine-grained results. Here we further probe into each media bias sub-type and find model performs differently over varying fine-grained labels. A factor analysis is hence conducted over the output of Joint Train and BERT+context, respectively, champions the overall lexical and informational bias detection (Table 4). The following findings are drawn.

First, models' performances are largely affected by the available training data scales, whereas label imbalance obviously exhibits in our fine-grained annotations (Table 3). Take informational bias annotation as an example: roughly 75% of the sentences are labeled as *Omission*. Bert+Context thus achieves a 0.52 F1 in detecting *Omission*, much higher than the other two sub-types. These shed light on the importance of advancing robust training with imbalanced labels for fine-grained media bias prediction.

Second, the nature of labels also plays a crucial role. For lexical bias sub-types, Joint Train better predicts *Exaggeration* and *Stereotype* (both exhibiting 0.65 F1) compared to a 0.55 for *Opinion*, although the latter almost has twice training samples. It might be because most cases of the former two sub-types are explicitly indicated by certain tokens like "crazy" (Table 1(a)), "explode" (Figure 1), whereas expressions in *Opinion* are relatively more implicit and uncertain, sometimes requiring higher-level knowledge for understanding. It calls for future attention to injecting knowledge to understand media bias in implicit languages.

4.4.2 Analysis of Large Language Model. From Table 4, we observe that ChatGPT is strenuous in media bias detection. Although it is improved by few-shots experiments, the performance is still not as good as the fine-tuned model. To further explore the reason for poor performance, we experiment with the detection of only lexical bias and informational bias. We observe it has high precision but low recall scores. For example, in lexical bias, it got 0.846 in precision, but only 0.065 in the recall, which means the model is not sensitive to bias and misses most of the biased sentences.

We can also notice that there is a big difference in the performance of ChatGPT between Fine-Grained Lexical and Fine-Grained Informational bias detection tasks. It is harder for ChatGPT to tell informational bias. During our testing, we noticed that sometimes ChatGPT would request more context information instead of returning the answer directly. However, when giving the context, the limitation of input length also exists. Also, the model will omit some information when the text is too long. So, it is quite helpful to do further research in context enrichment to supply valuable information.

5 Context Enrichment Methods and Experiments

Inspired by the insights in our basic experiments Section 4.3, we explore some methods to enrich the context for our model to better understand bias. We consider two main directions of context

enrichment, one is DACE, which increases the size of the training set and another is RACE, which can supply more knowledge when training. In order to test the generality of our methods, we also conduct experiments on the English dataset BASIL [16]. In this section, we will introduce how we improve these methods for media bias detection tasks.

5.1 Data-Augmented Context Enrichment

We fine-tune the BERT with BASIL and our dataset as the baselines of the following several methods. In the selection of data augmentation methods, we choose some traditional supervised learning methods[60] that augmented the sentence at the token level. We also attempt a semi-supervised learning method that augments the data at the sentence level.

Synonym Replacement (SR) Randomly choose one word, that is, not a stop word and replace the word with its synonym. This operation is repeated n times.

Random Insertion (RI) Randomly choose one word, that is, not a stop word in the sentence, and find its synonym. The synonym word will be inserted into a random position in the sentence. This operation is repeated n times.

Random Swap (RS) Randomly choose two words in the sentence and swap their position. This operation is repeated n times.

Random Deletion (RD) Remove each word in the sentence with a random possibility p.

Easy Data Augmentation (EDA) EDA consists of the four methods we mentioned above, SR, RI, RS, and RD. The policies are selected according to the randomly produced possibilities.

Unsupervised Data Augmentation (UDA) Recently, it is common to apply the concept of consistency training to semi-supervised learning, which is to say, the model predictions should be invariant between the original unlabeled data and the unlabeled data with noise input. One of the important works is UDA [60], which adds noise by augmentation methods, such as back translation, **Term Frequency-Inverse Document Frequency** (**TF-IDF**) word replacement, and so on. To align with our other DACE methods, EDA is applied as the noise-adding module.

The loss function is constructed by two parts and balanced by a parameter λ, which is usually set as 1 [60]. One part is supervised cross-entropy loss of the labeled data (denoted as L_{sup}) and another is unsupervised consistency training loss, denoted as L_{unsup}, which is used to minimize a divergence metric between the two distributions, original data P_{ori}, and the data after augmented Q_{aug}. The equations are listed below.

$$L_{unsup} = D_{KL}(P_{ori}\|Q_{aug}), \tag{1}$$

$$L = L_{sup} + \lambda L_{unsup}. \tag{2}$$

5.2 Retrieval-Augmented Context Enrichment

In this part, the result of fine-tuned BERT is also regarded as the baseline of RACE methods, without any information from outside knowledge sources.

In this task, for every sentence in the dataset (denoted as a **target sentence**), some methods are used to retrieve related sentences from knowledge sources (denoted as **context sentences**) for bias determination. Then, the model of context feature adding in Section 4.1 is applied in this part. In our setting, the articles in the same set report the same event or focus on the same problem, which are important references for annotators to compare and determine the bias. So we regard the result when the context sentences are from the same set as the target sentence as the ground truth of retrieval tasks.

BM25 BM25[45] is a traditional method that calculates the relevance of a query and a document based on the probabilistic retrieval framework. For each target sentence, we regard it as a query and then calculate the BM25 scores of the sentences in knowledge sources. The top 10 sentences are retrieved as context sentences of the target sentence according to BM25 scores.

SimCSE SimCSE[19] is a simple contrastive learning framework by using "entailment" pairs as positives and "contradiction" pairs as negatives to optimize the distance between sentence embedding. The distance between the embedding of sentences with similar semantics is closer. We fine-tune the pre-trained SimCSE model supplied by Gao et al. [19] using our datasets. According to our setting of ground truth, the target sentence with the sentence in the same set is regarded as a positive pair, and the target sentence with a sentence, that is, not in the same set is regarded as a negative pair. In the retrieval step, we calculate the Cosine similarities between the target sentence embedding and the embedding of each sentence in the knowledge sources, then the top 10 are retrieved as context sentences.

Dense Passage Retrieval (DPR) The method DPR[32] is first proposed in the open-domain question-answering task. This model aims to rank the passage containing the answer top and put other passages in low ranking so that valuable information would be retrieved to guarantee the correct answer.

Inspired by this concept, we design our training on the DPR model based on our datasets. Each instance in the training data consists of a target sentence and several context sentences, denoted as $\{(t_i, c^+_{i,1}, ..., c^+_{i,n}, c^-_{i,1}, ..., c^-_{i,m})\}^p_{i=1}$. The n sentences from the same set that the target sentence is in are positive sentences, and the other m are negative sentences. Following the previous work, we calculate the similarity between the sentences using the embedding output from encoders, which are denoted as E_t and E_c. The model is optimized by **Binary Cross Entropy Loss** (**BCELoss**) as the Equation (4).

$$sim(t, c) = E_t(t)^T E_c(c), \tag{3}$$

$$L(t_i, c^+_{i,1}, ..., c^+_{i,n}, c^-_{i,1}, ..., c^-_{i,m}) =$$
$$-\sum_{p=1}^{n} \log \sigma(sim(t_i, c^+_{i,p})) - \sum_{q=1}^{m} \log \sigma(1 - sim(t_i, c^-_{i,q})). \tag{4}$$

Finally, the top-10 sentences are retrieved as context sentences according to the similarity scores.

Combining Titles Sometimes, the meaning of complex information cannot be correctly reflected in a single sentence. The content of the whole article that the sentence is in is also necessary to be considered. Usually, the title of a piece of news summarizes the main content. So, we conduct the experiments that feed the title as the sentence's context into the model, which saves a lot of memory compared with adding the whole report.

Based on the BM25 and SimCSE model, we concatenate a sentence with its report title for training and selection. As for DPR model, which is more flexible, we add another encoder for the title sentences. Three encoders in total are trained to get the embedding of target, context, and title sentences. Then the embedding of each sentence is concatenated with the embedding of its article title for further similarity calculation.

5.3 Experiments Setup

Data In the experiments, we use two main datasets, one is an English dataset BASIL, and another is our Chinese dataset. We also expand the two datasets with a large number of unlabeled reports online as the unsupervised dataset in UDA and knowledge sources in the information retrieval task. For BASIL, we use a large dataset consisting of 754,000 news articles [33]. For our Chinese

Table 6. Experimental results of Data Augmentation and Information Retrieval methods on Dataset BASIL and Ours

Method	BASIL						Our dataset											
	Lexical			Informational			Coarse-Grained			Fine-Grained Combined			Fine-Grained Lexical			Fine-Grained Informational		
	P ($\uparrow$)	R ($\uparrow$)	F1 ($\uparrow$)	P ($\uparrow$)	R ($\uparrow$)	F1 ($\uparrow$)	HL ($\downarrow$)	JI ($\uparrow$)	F1 ($\uparrow$)	HL ($\downarrow$)	JI ($\uparrow$)	F1 ($\uparrow$)	HL ($\downarrow$)	JI ($\uparrow$)	F1 ($\uparrow$)	HL ($\downarrow$)	JI ($\uparrow$)	F1 ($\uparrow$)
BERT	0.103	0.671	0.178	0.228	**0.729**	0.344	0.259	0.381	0.552	0.115	0.342	0.510	0.101	0.390	0.561	0.129	0.310	0.474
BERT+Context	0.129	0.589	0.209	0.314	0.624	0.415	0.271	0.396	0.568	0.112	0.352	0.521	0.105	0.375	0.545	0.134	0.321	0.486
Synonym Replace	0.111	0.627	0.185	0.283	0.529	0.342	0.265	0.428	0.599	0.115	0.311	0.474	0.097	0.363	0.533	0.143	0.329	0.496
Random Insertion	0.111	0.635	0.187	0.273	0.498	0.322	0.260	0.423	0.594	0.113	0.338	0.505	**0.094**	0.397	0.568	0.142	**0.343**	**0.511**
Random Swap	0.111	0.603	0.184	0.276	0.543	0.346	0.270	0.409	0.580	0.115	0.332	0.498	0.097	0.390	0.561	0.137	0.325	0.491
Random Deletion	0.110	0.654	0.188	0.269	0.539	0.331	0.267	0.409	0.580	0.116	0.302	0.464	0.100	0.373	0.543	0.138	0.341	0.508
EDA	0.116	0.649	0.194	0.285	0.525	0.345	0.288	0.406	0.578	**0.110**	0.349	0.517	0.098	0.382	0.552	0.152	0.331	0.498
UDA	0.133	**0.688**	0.222	0.260	0.651	0.371	0.260	**0.437**	**0.608**	0.118	0.343	0.511	0.096	0.408	0.580	0.156	0.321	0.486
BM25	0.134	0.546	0.214	0.302	0.543	0.385	0.268	0.423	0.594	0.114	0.341	0.508	0.102	0.372	0.543	0.128	0.311	0.475
SimCSE	0.146	0.616	0.235	0.316	0.615	0.414	0.274	0.415	0.586	0.112	0.350	0.518	0.101	**0.409**	**0.581**	**0.123**	0.322	0.487
DPR	0.139	0.588	0.223	0.315	0.621	0.415	**0.246**	0.430	0.602	0.116	0.333	0.499	0.101	0.388	0.559	**0.123**	0.314	0.478
BM25+title	0.122	0.419	0.188	0.290	0.552	0.374	0.260	0.424	0.596	0.116	0.328	0.494	0.113	0.363	0.533	0.131	0.273	0.429
SimCSE+title	0.137	0.602	0.222	0.300	0.642	0.404	0.264	0.425	0.597	0.122	0.361	0.530	0.102	0.375	0.545	**0.123**	0.314	0.478
DPR+title	**0.147**	0.635	**0.239**	**0.319**	0.616	**0.419**	**0.246**	0.430	0.602	**0.110**	**0.378**	**0.549**	0.098	0.398	0.570	**0.123**	0.316	0.480

For our dataset, the same notations as in Table 4 are used. For BASIL, Lexical and Informational stands for the lexical and informational bias classification at the sentence level.

dataset, we have about 120,000 unlabelled sentences of the reports that crawled from the COVID topic in several media outlets in the same period.

Setup Most of our media bias detection tasks are based on BERT [14]. The experiments are conducted through Hugging Face [59], and the pre-trained tokenizer and model are applied, according to the language of BASIL and our dataset, which are $BERT_{base-cased}$ and $BERT_{base-chinese}$. The synonym packages for EDA are also selected according to the languages.

In the experiments of dataset BASIL, the 100 triples are randomly divided by 8:1:1 into training, validation, and test splits. Then conduct lexical and informational bias detection tasks at the sentence level. However, we observe that the results are unstable and vary much in different data splits with different random seeds due to the small scale with extreme data imbalance in dataset BASIL. To help the model learn more bias features and make the result more robust, in the training set, we apply the traditional simple strategy [31], copying the sentences in the minor class to balance the dataset. In the data augmentation supervised learning experiments, we have enough data after the augmentation, so we sample sentences with different labels by the same scale from the augmented dataset instead of copying the sentences. Also, the final result is the average of the results with five different random seeds. Precision, Recall, and F1 score are selected to measure the performance in BASIL according to the previous work [16]. The experiments in this section are conducted at the sentence level, for the dataset BASIL, we detect the lexical bias and informational bias, for our dataset, four settings (Coarse-Grained, Fine-Grained Combined, Fine-Grained Lexical, and Fine-Grained Informational) are tested. Other settings are the same as the basic experiments in Section 4.2.

5.4 Results of DACE

In DACE, SR, RI, RS, RD, and EDA are supervised learning methods with augmentation in the token level, while UDA is a semi-supervised learning method that augments the data in the sentence level.

Token-Level Enrichment From Table 6, we observe that the policies of EDA in general enable BERT to perform better, although there are minor differences among each policy, indicating the overall helpfulness of data augmentation on our task. All four types of policies help to increase the scale of the training set, while the policy RI enriches the word corpus without content reduction, which shows better performance than other policies.

In conclusion, implementing the four policies together increases the robustness of the performance, and gained improvements in all tasks. This further confirms that EDA is suitable to be used
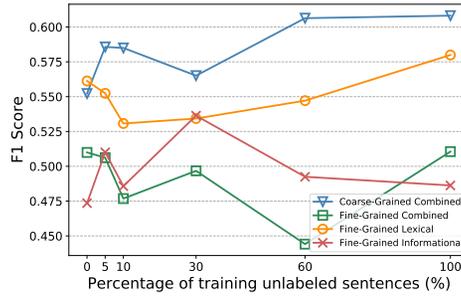
Fig. 5. Results of UDA with different scales of unlabeled data. x-axis: varying percentage of the extended dataset as unlabeled sentences when training UDA. y-axis: F1 score.

as a noise-adding approach in UDA, which can maintain the consistency of label distribution in media bias detection tasks.

Sentence-Level Enrichment From Table 6, we observe that UDA method further enhances the performance of media bias detection, especially in lexical bias detection. To further explore how UDA functions, we conduct experiments in several steps with the increase in the percentage of unlabeled sentences feeding to the model, and the results of each task are shown in Figure 5.

As the scale of training sentences increases, the performance of Coarse-Grained Combined grows continuously with small fluctuations. While the situations are quite different in Fine-Grained classification tasks. The F1 score of the Fine-Grained lexical bias task first rapidly drops then increases slowly, and maintains an upward trend. While there is a lot of fluctuation in the curve of the Fine-Grained informational bias task. The results indicate that enrichment in training corpus has a significant effect on lexical bias detection, however, the improvements in informational bias are unstable, which is also aligned with their natural features.

When go through the details of the predictions in the model, we observe the UDA's improvements in overfitting. It is more sensitive about the biased sentence and can better distinguish the trigger words. For example, most of the sentences with *Opinion* have the token "believe", while the word can also be used for other purposes. UDA can avoid labeling all the sentences with "believe" as biased sentences. Our experiments also verify that large-scale of unlabeled data can be used to enlarge the training set of some practical tasks with small-scale datasets and get improvements in the performance. Also, semi-supervised and even unsupervised learning has the potential to be further explored in the future.

5.5 Results of RACE

Information Retrieval Our RACE results are also shown in Table 6. From our ground truth model BERT+Context, adding external context sentences improves the results, which is aligned with the result of previous work [55]. Some information retrieval methods are conducted to select valuable information from external knowledge source. We can observe that the gap between traditional and neural network methods is not as large as other tasks. But the traditional method BM25 is weak in informational bias detection, deep learning models can make up and have more potential in this field. SimCSE and DPR have similar performance in informational bias performance. While the SimCSE model outperforms other models in lexical bias detection, because we fine-tune the SimCSE model based on a large pre-training language model, which already contains a lot of semantic information.

Combining Titles When the titles are also added as the context of the sentence, from Table 6, we notice that DPR+title gets best performance in general. Compared with the model only consider the single sentence itself, the combination of titles helps to retrieval more precise and valuable information. While the way to feeding titles is also important, concatenating the title and the sentence simply even sometimes hurts the performances. DPR model is more flexible, which can encode the titles and sentences separately, takes full advantage of the titles and further improves the results.

We also notice the shortage of our model in considering time series. For example, there is a case that the target sentence is,

"If the authorities still further restrict demonstrations and gatherings, they may have exceeded public health safety considerations."

We find the context sentences through model DPR+title, for example,

"Egypt opened the Rafah port connecting the Gaza Strip on the same day, allowing hundreds of Palestinians stranded in Egypt due to the epidemic to return to Palestine",

"From May 16th to 19th, the "stay-at-home order" will continue to be implemented nationwide",

"The spread of the epidemic in India is accelerating".

These sentences are similar to the target sentence in semantics, and can also provide information on epidemic situations of other countries as references. However, they will also bring confusion if the description is not for the same period. The original design of DPR is for the QA system, most of the queries and answers are about general knowledge which seldom changes as time goes on. While one of the significant differences between our task and the QA task is that reports in journalism are time-sensitive, which indicates us the future direction of improvement in the models.

5.6 Case Study

The adding of retrieved information provides more knowledge and improves the performance is easy to understand. In this part, to provide more insights into what can be learned by UDA, we conduct a case study on fine-grained lexical bias detection to explore the difference between basic BERT and UDA when detecting the bias after a larger scale of training data is supplied.

Because there is attention mechanism in both methods, we use the visualization tool BertViz [56] to show the attention relationship in each layer and head among the tokens. When we highlight the words according to the attention matrix, we can find that the basic BERT can also learn the pattern related to media bias in general, only minor difference appears in UDA, which improves the final performances.

Figure 6 shows the attention mechanism of the sentence "Hong Kong people who are in the UK are crazy for the air tickets to avoid quarantine". The ground truth shows that this sentence has lexical bias *Exaggeration*. Both of the methods correctly classify the bias at coarse-grained level, while in fine-grained, basic BERT wrongly classifies it as *Opinion*. From Figure 6, we observe that the allocation of attention in basic BERT is more average than that in UDA. As we know, in our dataset, informational bias is longer than lexical bias, and *Opinion* is usually longer than other bias in lexical bias. Due to the divergence in length of different bias types, it is easier for the model to give the result according to the length of the highlighted part, that is, why BERT labels it as *Opinion*. We notice that in the result of model UDA, the token "crazy", which is the trigger of bias *Exaggeration* in human judgment, got much more attention than other tokens. So, when the model cannot exactly find the important tokens may result in an incorrect result in fine-grained level, and UDA, which enrich data-augmented context and make up the gap.
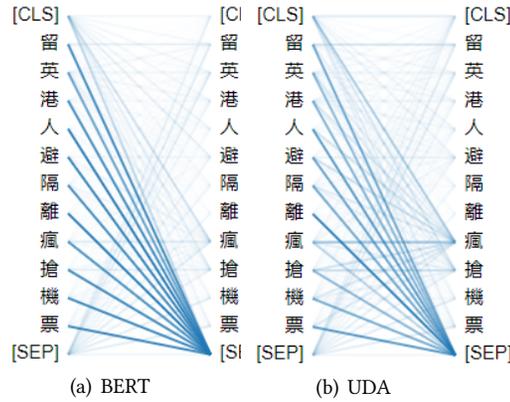
Fig. 6. The attention relationship presentation of model BERT and UDA. The darker color stands for larger value in attention parameters.

6 Related Work

6.1 Media Bias Detection

Our work is in line with prior research in media bias. Its automatic detection has been examined through substantial efforts with NLP methods. Most previous studies concern what linguistic features, such as tokens and phrases, will likely be used in news expressing dissenting political opinions [6, 62]. Later work [11, 16, 55] points out the limitation to focus only on lexical bias and calls for attention over information bias. They integrate some prior efforts in news framing [6, 9], aiming to learn news reporting perspectives, e.g., economy and society as prototype of informational bias, and propose that two types of bias should be considered together. However, the existing benchmark datasets are mostly labeled in a coarse granularity [40] or through political spectrum [16] according to the western ideology, which is not applicable to Chinese media bias detection. Meanwhile, recent studies have started to investigate bias in the Chinese context. Specifically, [27] studies bias in LLM-generated content across dimensions such as gender and race, rooted in Chinese culture; [53] examines Chinese public opinions on the use of large language models in education, highlighting the public concerns about content bias. However, both focus on content bias in large language models, whereas our work focuses on media bias, defined as slanted news coverage or internal bias reflected in news writing. We step further to design fine-grained labels and annotate them in both token and sentence level, and be the first to systematically analysis the media bias from the Chinese perspective. Also, limited attention has been paid to media bias in non-political news, whereas our paper about data in health field attempts to mitigate the gap.

6.2 Context Enrichment

Previous work points out that insufficient data in media bias detection tasks and more context is needed in detecting information bias [55], while we systematically propose two main directions that enrich the context to help the bias detection tasks.

Data-Augmented Context Enrichment Collecting a large and well-labeled dataset is expensive, so data augmentation approaches are widely applied in training deep learning models [10]. Preliminary work focuses on supervised learning tasks, most of which augmentation is at token-level. For example, synonym replacement [34], random deletion [30]. Our work follows Easy Data Augmentation [58], which consists of four basic data augmentation methods, and analyzes the

affection of each method. Later work attempts to augment data at the sentence level directly, for example, Yu et al. [64] proposes to enrich training data by paraphrasing. As the techniques push on, augmentation data can be used for semi-supervised learning for unlabeled data through consistency training [60]. We are the first to apply this concept in the media bias detection task.

Retrieval-Augmented Context Enrichment Information Retrieval is widely used as the retriever module in question-answering and fact-checking tasks, which need the support of large knowledge sources [22, 65]. Traditional methods, such as TF-IDF [35] and BM25 [57, 61], use sparse representations. Along with the success of deep learning, deep retriever models with dense representations have been developed. SimCSE [19] is a simple contrastive learning framework, which makes the distance of the sentences that have similar semantics closer through embedding training. DPR [32] trains different encoders to encode the question and document and then measures them by similarity functions. Based on the frameworks, we propose methods to retrieve more valuable information as context sentences of the target sentences, which largely improve the performance of media bias detection tasks.

7 Conclusion

Based on the media bias detection task, we refine the previous label structure and design a new annotation system from the Chinese perspective. We also contribute the first annotated Chinese media bias dataset in the public health field and its benchmark based on a multi-label setting. In a further step, we explore how context enrichment methods can be better applied to the practice problem with small-scale datasets. Through substantial experiments on it, we point out these crucial findings: (1) Training data scale may largely limit the model's capability to learn media bias, and contextual enrichment in both data augmentation and information retrieval is potentially helpful. (2) Retrieval-augmented methods work better than data-augmented methods, especially when more context information, such as the titles' feature, is added. However, there are also some details worth further exploring in future work.

Acknowledgments

References

[1] Dmitrii Aksenov, Peter Bourgonje, Karolina Zaczynska, Malte Ostendorff, Julian Moreno Schneider, and Georg Rehm. 2021. Fine-grained classification of political bias in German news: A data set and initial experiments. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*. 121–131.

[2] Bharat Anand, Rafael Di Tella, and Alexander Galetovic. 2007. Information or opinion? Media bias as product differentiation. *Journal of Economics & Management Strategy* 16, 3 (2007), 635–682.

[3] Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics* 34, 4 (2008), 555–596.

[4] Brent H. Baker, Tim Graham, and Steve Kaminsky. 1994. *How to Identify, Expose & Correct Liberal Media Bias*. Media Research Center Alexandria, VA.

[5] Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. 2020. We can detect your bias: Predicting the political ideology of news articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 4982–4991.

[6] Eric Baumer, Elisha Elovic, Ying Qin, Francesca Polletta, and Geri Gay. 2015. Testing and comparing computational approaches for identifying the language of framing in political news. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 1472–1482.

[7] Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. arXiv:2004.05150. Retrieved from https://arxiv.org/abs/2004.05150. (2020).

[8] Ceren Budak, Sharad Goel, and Justin M. Rao. 2016. Fair and balanced? Quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly* 80, S1 (2016), 250–271.

[9] Dallas Card, Amber Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. 438–444.

[10] Jiaao Chen, Derek Tam, Colin Raffel, Mohit Bansal, and Diyi Yang. 2023. An empirical survey of data augmentation for limited data learning in NLP. *Transactions of the Association for Computational Linguistics* 11 (2023), 191–211.

[11] Wei-Fan Chen, Khalid Al-Khatib, Benno Stein, and Henning Wachsmuth. 2020. Detecting media bias in news articles using gaussian bias distributions. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. 4290–4300.

[12] Wei-Fan Chen, Khalid Al-Khatib, Henning Wachsmuth, and Benno Stein. 2020. Analyzing political bias and unfairness in news articles at different levels of granularity. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*. 149–154.

[13] Tawnya J. Adkins Covert and Philo C. Wasburn. 2007. Measuring media bias: A content analysis of Time and Newsweek coverage of domestic social issues, 1975–2000. *Social Science Quarterly* 88, 3 (2007), 690–706.

[14] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 4171–4186.

[15] Jeffrey L. Elman. 1990. Finding structure in time. *Cognitive Science* 14, 2 (1990), 179–211.

[16] L. Fan, M. White, E. Sharma, R. Su, P. K. Choubey, R. Huang, and L. Wang. 2019. In plain sight: Media bias through the lens of factual reporting. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 6343–6349.

[17] Joseph L. Fleiss, Bruce Levin, and Myunghee Cho Paik. 2013. *Statistical Methods for Rates and Proportions*. john wiley & sons.

[18] Rama Rohit Reddy Gangula, Suma Reddy Duggenpudi, and Radhika Mamidi. 2019. Detecting political bias in news articles using headline attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. 77–84.

[19] T. Gao, X. Yao, and D. Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021. Association for Computational Linguistics (ACL)*. 6894–6910.

[20] Matthew Gentzkow and Jesse M. Shapiro. 2010. What drives media slant? Evidence from US daily newspapers. *Econometrica* 78, 1 (2010), 35–71.

[21] Tim Groeling. 2013. Media bias by the numbers: Challenges and opportunities in the empirical study of partisan news. *Annual Review of Political Science* 16 (2013), 129–151.

[22] Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics* 10 (2022), 178–206.

[23] Stuart Hall. 2016. 29 NOTES ON DECONSTRUCTING'THE POPULAR'. *People's History and Socialist Theory (Routledge Revivals)* (2016), 227.

[24] Felix Hamborg. 2020. Media bias, the social sciences, and NLP: Automating frame analyses to identify bias by word choice and labeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*. 79–87.

[25] Thomas Hanitzsch. 2019. Journalism studies still needs to fix Western bias. *Journalism* 20, 1 (2019), 214–217.

[26] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9, 8 (1997), 1735–1780.

[27] Y. Huang and D. Xiong. 2024. CBBQ: A chinese bias benchmark dataset curated with human-AI collaboration for large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING'24)*. 2917–2929.

[28] Koichi Iwabuchi. 2010. Globalization, east Asian media cultures and their publics. *Asian Journal of Communication* 20, 2 (2010), 197–212.

[29] Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. 2014. Political ideology detection using recursive neural networks. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1113–1122.

[30] Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 1681–1691.

[31] Justin M. Johnson and Taghi M. Khoshgoftaar. 2019. Survey on deep learning with class imbalance. *Journal of Big Data* 6, 1 (2019), 1–54.

[32] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020. Association for Computational Linguistics (ACL)*. 6769–6781.

[33] Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. 2019. Semeval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*. 829–839.

[34] Oleksandr Kolomiyets, Steven Bethard, and Marie-Francine Moens. 2011. Model-portability experiments for textual temporal analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Vol. 2. ACL; East Stroudsburg, PA, 271–276.

[35] B. Kratzwald and S. Feuerriegel. 2018. Adaptive document retrieval for deep question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 576–581.

[36] J. D. Krieger, T. Spinde, T. Ruas, J. Kulshrestha, and B. Gipp. 2022. A domain-adaptive pre-training approach for language bias detection in news. In *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries*. 1–7.

[37] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.

[38] Yuanyuan Lei, Ruihong Huang, Lu Wang, and Nick Beauchamp. 2022. Sentence-level media bias analysis informed by discourse structures. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 10040–10050.

[39] S. Li, Z. Zhao, R. Hu, W. Li, T. Liu, and X. Du. 2018. Analogical reasoning on Chinese morphological and semantic relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 138–143.

[40] Sora Lim, Adam Jatowt, Michael Färber, and Masatoshi Yoshikawa. 2020. Annotating and analyzing biased sentences in news articles using crowdsourcing. In *Proceedings of the 12th Language Resources and Evaluation Conference*. 1478–1484.

[41] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. arXiv:1711.05101. Retrieved from https://arxiv.org/abs/1711.05101. (2017).

[42] Anne Maass. 1999. Linguistic intergroup bias: Stereotype perpetuation through language. In *Proceedings of the Advances in Experimental Social Psychology*. Vol. 31. Elsevier, 79–121.

[43] Fred Morstatter, Liang Wu, Uraz Yavanoglu, Stephen R. Corman, and Huan Liu. 2018. Identifying framing bias in online news. *ACM Transactions on Social Computing* 1, 2 (2018), 1–18.

[44] Vincent Price, Lilach Nir, and Joseph N. Cappella. 2005. Framing public discussion of gay civil unions. *Public opinion quarterly* 69, 2 (2005), 179–212.

[45] Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval* 3, 4 (2009), 333–389.

[46] Arthur Schopenhauer. 2017. *Delphi Collected Works of Arthur Schopenhauer (Illustrated)*. Vol. 12. Delphi Classics. 109 pages.

[47] Michael Schudson. 2001. The objectivity norm in American journalism. *Journalism* 2, 2 (2001), 149–170.

[48] Jonathon P. Schuldt, Sara H. Konrath, and Norbert Schwarz. 2011. "Global warming" or "climate change"? Whether the planet is warming depends on question wording. *Public Opinion Quarterly* 75, 1 (2011), 115–124.

[49] Timo Spinde, Manuel Plank, Jan-David Krieger, Terry Ruas, Bela Gipp, and Akiko Aizawa. 2021. Neural media bias detection using distant supervision with BABE-bias annotations by experts. In *Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2021*. 1166–1177.

[50] Timo Spinde, Lada Rudnitckaia, Jelena Mitrović, Felix Hamborg, Michael Granitzer, Bela Gipp, and Karsten Donnay. 2021. Automated identification of bias inducing words in news articles using linguistic and context-oriented features. *Information Processing & Management* 58, 3 (2021), 102505.

[51] Richard Streckfuss. 1990. Objectivity in journalism: A search and a reassessment. *Journalism Quarterly* 67, 4 (1990), 973–983.

[52] Percy H. Tannenbaum and Mervin D. Lynch. 1960. Sensationalism: The concept and its measurement. *Journalism Quarterly* 37, 3 (1960), 381–392.

[53] Yao Tian, Chengwei Tong, Lik-Hang Lee, Reza Hadi Mogavi, Yong Liao, and Pengyuan Zhou. 2023. A perspective study on Chinese social media regarding LLM for education and beyond. arXiv:2306.04325. Retrieved from https://arxiv.org/abs/2306.04325. (2023).

[54] Cigdem Toprak, Niklas Jakob, and Iryna Gurevych. 2010. Sentence and expression level annotation of opinions in user-generated discourse. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. 575–584.

[55] Esther van den Berg and Katja Markert. 2020. Context in informational bias detection. In *Proceedings of the 28th International Conference on Computational Linguistics*. 6315–6326.

[56] J. Vig. 2019. A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 37–42.

[57] Z. Wang, P. Ng, X. Ma, R. Nallapati, and B. Xiang. 2019. Multi-passage BERT: A globally normalized BERT model for open-domain question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 5878–5882.

[58] J. Wei and K. Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 6382–6388.

[59] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 38–45.

[60] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems* 33 (2020), 6256–6268.

[61] Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2019. End-to-end open-domain question answering with bertserini. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*. 72–77.

[62] Tae Yano, Philip Resnik, and Noah A. Smith. 2010. Shedding (a thousand points of) light on biased language. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. 152–158.

[63] Jiafei Yin. 2008. Beyond the four theories of the press: A new model for the Asian & the world press. *Journalism & Communication Monographs* 10, 1 (2008), 3–62.

[64] A. W. Yu, D. Dohan, M. T. Luong, R. Zhao, K. Chen, M. Norouzi, and Q. V. Le. 2018. QANet: Combining local convolution with global self-attention for reading comprehension. In *International Conference on Learning Representations*.

[65] Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. 2021. Retrieving and reading: A comprehensive survey on open-domain question answering. arXiv:2101.00774. Retrieved from https://arxiv.org/abs/2101.00774. (2021).

Appendices

A   Original Text of Examples

Due to the length limitation of our paper, we show original Chinese text of Figure 1 in Table 7, and Table 1 in Table 8.

Table 7. The Original Text of Examples in Figure 1

Article 1 "Banquet for tens of thousands" coursed the trouble? Wuhan community [explodes with] collective fever....

萬家宴惹禍？ 武漢社區 [爆] 集體發熱。……

[Experts pointed out that the Banquet has exceeded the two-week incubation period so far, "Those who currently have fever symptoms may be in contact with coronary pneumonia patients at a later stage" .] ...

[專家表示，萬家宴於 1 月 18 日舉行，至今已經超過 2 周潛伏期，「目前有發熱症狀的，有可能是後期接觸了新冠肺炎的病人」。] ……

[Secretary Gong of the Party Branch of Kanghe Neighborhood Committee in Baibuting pointed out that the main purpose for labeling the "hot building" is to remind residents to avoid going out as much as possible.] ...

[百步亭社區康和居委會黨支部龔書記指出，發熱門棟主要是為提醒居民，盡可能避免出門。] ……

Article 2 "Banquet for tens of thousands" was still held before lunar New Year, Wuhan Baibuting was [seriously infected and despairingly crying]. …

年前仍擺萬家宴，武漢百步亭 [感染嚴重哭喊絕望]。……

[A netizen posted on Sina Weibo late at night on the 9th, saying that he was writing this in "desperation" and that Baibuting is currently "in an uncontrolled situation" . Because of the banquet, many people were infected with pneumonia.] ...

[網友9 日深夜在新浪微博發文，指自己是在「絕望中寫這些」，目前百步亭處於「無人管的境地」。因為舉辦「萬家宴」，導致很多人感染了武漢肺炎。] ……

This post was widely forwarded more than 100,000 times on Sina Weibo, [but Wuhan and even Chinese officials have yet to respond to the allegations in the post.] ...

這篇貼文發出後，在新浪微博被廣為轉發超過 10 萬次，[但武漢乃至於中國官方，目前尚未對貼文中指控的內容作出回應。] ... ...

Table 8. The Original Text of Examples in Table 7

Writing Bias Annotation Examples
(a) ... They are [crazy for]Exaggeration air tickets, flying around the earth, hoping to "return to [the safest]Exaggeration place" ...   ……他們 [瘋搶機票]，繞著地球飛，盼「回到 [最安全] 的地方」……
(b) ... some people even continue to go out to have dinner with friends. The "[Buddhism]Stereotype/Labeling" is surprising ...   ……甚至還有人繼續在外和朋友聚餐等，[佛系] 得讓人意外。
(c) ... [some experts doubt the effectiveness of mass quarantine, pointing out that "the virus will actually spread"]Ventriloquism   ……[有專家懷疑大規模隔離的成效，指「病毒實際上總會散播開去」]
(d) ... and the situation [is bound to continue to deteriorate]Speculative Content / Opinion …   ……情況 [想必會持續惡化]……

Content Bias Annotation Examples

Annotation Set 1

Article 1 ... The Chinese song made many [foreign netizens sarcastically say in a "high-level black" way that this "North Korean boy" is too cute.]…
中國歌曲讓不少 [外國網友以「高級黑」方式諷刺說，這「北韓男孩」太可愛了。

[According to "The Paper", the song has a detailed and in-depth description of the mobile cabin hospital, which helps patients fight the virus and relieve tension.]Framing / Undue Weight (Positive) …
據《澎湃新聞網》聲稱，該首歌曲對於方艙醫院有細緻、深入的具體刻畫，有助於患者對抗病毒、舒緩緊張

Article 2 ... [Some netizens denounced the creator for "entertaining disasters"]. [The Beijing News online commentary even denounced this as "forcibly extolling disasters".]Framing / Undue Weight (Negative) …
有網民直斥創作人「把災難娛樂化」]，[《新京報》網上評論更直斥這是「強行歌頌災難」。

[Jiang Junrong said "A lot of places are "cleared" now. Don't we need to show a little bit of optimism? Don't we need to show a better mentality?"]Omission of Article 1 …
蔣軍榮接受澎湃新聞訪問時表示，「現在好多地方都『清零』了。我們不需要拿出一點樂觀主義精神嗎？不需要拿出更好的心態來嗎？」

Article 3 ... [Some netizens even described the video as "seemingly back in the 1980s and 1990s".] ...
有網民甚至形容視頻 "好像回到了八九十年代"。

[Composer Jiang Junrong said in response that "They don't know my state of life, and I don't make any calculations".]Omission of Article 1
作曲家蔣軍榮回應時說：「他們不知道我的生命狀態，我也不會計較。"

[The Beijing News commented on this, saying that turning it into the material of cheerful children's songs not only dilutes the disaster background, but also eliminates the seriousness of the fight against the epidemic.]Framing / Undue Weight (Negative)
《新京報》對此發表評論文章認為，中國抗擊冠病疫情是一件嚴肅而沉重的事情，將方艙醫院化為歡快兒歌的素材，不僅淡化了抗疫的災難底色，也消解了抗疫的嚴肅性。

Annotation Set 2

Article 1 ... "Shincheonji" chairman will hold a press conference to publicly respond to the epidemic for the first time.
韓國「新天地」會長將開記者會，首次公開回應疫情。

Article 2 South Korean leader of "Shincheonji" kneels to apologize and thank the government for fighting the epidemic.
南韓新天地教主下跪道歉，又感謝政府抗疫。

Article 3 Weird! The leader of "Shincheonji" kneels and [wears Park Geun-Hye's watch]Sensationalism Content·
詭！新天地教主下跪，竟戴著朴槿惠手錶。

B   Annotation Interface

Interface of content bias annotation tool is shown in Figure 7.

Fig. 7. The interface of content bias annotation tool, which has the same functions with writing bias annotation tool. Several questions are set to assist the determination of annotators.

C   GPT Prompts

We show the prompt of Fine-Grained Combined bias detection task as the example:

Given the fine-grained class definitions of media bias as following (Exaggeration, Stereotype, Ventriloquism, Opinion, Omission, Framing, and Sensationalist):

Exaggeration(It is defined as using eye-catching wordings with strong emotional implications or describing things in a way, that is, more than it really is, to influence an audience.),

Stereotype(Sometimes a media source may favor or attack a particular group, or a person, and describe them with wordings reflecting their stereotype to them. We refer to this kind of bias as Stereotype, or we can call it labeling. It can also describe the events or people using words that may contain judgments.),

Ventriloquism(We annotate text spans as Ventriloquism when experts or witnesses are quoted in a way that intentionally voices the authors own opinion),

Opinion(This sub-type covers cases like stories that focus not on what has occurred, but primarily on what might occur, or it is just the opinion of the author. We can also name it Speculative Content.),

Omission(We use Omission to describe situations where some reports ignore information that tends to disprove their claims: if an article does not contain the content about stakeholders that appeared in other articles, then it has the informational bias of Omission.),

Framing(We say an article has the informational bias of Framing when multiple stakeholders are contained but not balanced or undue weight. Such imbalances can be raised from the following aspects: different positions, different lengths of the description, and different sentiment polarities of selected content.),

Sensationalist(This informational bias sub-type refers to the event, that is, not closely relevant to the theme, but selected to excite the greatest number of readers.)

——————————

Given the sentence: SENTENCE

Please answer the informational bias situation of this sentence.

Format: <Exaggeration (If Yes: 1, else: 0)>,<Stereotype (If Yes: 1, else: 0)>,<Ventriloquism (If Yes: 1, else: 0)>,<Opinion (If Yes: 1, else: 0)>,<Omission (If Yes: 1, else: 0)>,<Framing (If Yes: 1, else: 0)>,<Sensationalist (If Yes: 1, else: 0)>

Please only answer the number.