A Dataset for Analysing News Framing in Chinese Media

Owen Cook*, Yida Mu*, Xinye Yang*, Xingyi Song and Kalina Bontcheva

School of Computer Science, University of Sheffield
{oscook1, x.song, k.bontcheva}@sheffield.ac.uk

Abstract

Framing is an essential device in news reporting, allowing writers to influence public perceptions of current affairs. While automatic news framing detection datasets exist in various languages, none focus on news framing in the Chinese language, which presents unique challenges with complex character meanings and unique linguistic features. This study introduces the first Chinese News Framing dataset, to be used as either a stand-alone dataset or a supplementary resource to the SemEval-2023 task 3 dataset. We detail its creation and conduct baseline experiments to demonstrate the need for such a dataset and create benchmarks for future research, providing results obtained through fine-tuning XLM-RoBERTa-Base and using GPT-4o in the zero-shot setting. We find that GPT-4o performs significantly worse than fine-tuned XLM-RoBERTa across all languages. For the Chinese language, we obtain an F1-micro (the performance metric for SemEval task 3, subtask 2) score of 0.719 using only samples from our Chinese News Framing dataset and a score of 0.753 when we augment the SemEval dataset with Chinese news framing samples. With positive news frame detection results, this dataset is a valuable resource for detecting news frames in the Chinese language and is a useful supplement to the SemEval-2023 task 3 dataset.

Introduction

Framing is a fundamental process for understanding the world. It occurs when people both communicate and receive information (Entman 1993) and it shapes a person's conceptualisation of the world around them. While scholars do not agree on an exact definition (Hertog 2001; Van Dijk 2023), we refer to that of Gamson et al. (1992), who define a frame as "a central organising principle that holds together and gives coherence and meaning to a diverse array of symbols". A symbol may be anything from a small facial expression to a group of words that are commonly used to evoke particular emotions. The interpretation of symbols by a receiver is influenced by their existing knowledge and internal grouping of concepts, impacting their perception of information. The communicator also frames the symbols they project, allowing them to influence the receiver's interpretation of them.

In the news and media, framing is omnipresent; social media content creators, journalists, and politicians particularly influence how the public views and feels about current affairs. D'angelo (2018) refers to this as news framing (also known as media framing): "how journalists, their sources, and audiences work within conditions that shape the messages they construct as well as the ways they understand and interpret these messages".

Due to the vastness of information available online, there is a desire for automatic news frame detection. Automatic news frame detection refers to the use of computational methods to identify news frames. It has a variety of applications, such as understanding media bias (Morstatter et al. 2018), providing balanced framing of articles at the information retrieval stage (Reiter-Haas et al. 2024), automating large-scale content analysis (Kwak, An, and Ahn 2020; Alonso del Barrio and Gatica-Perez 2023), and detecting misinformation (Wang et al. 2024). Datasets available for the development of language models capable of automatically detecting news frames span a number of languages, yet news framing in the Chinese language is particularly underexplored.

This work introduces the Chinese News Framing dataset, facilitating research investigating the training of models capable of detecting news frames in Chinese. The dataset has been designed to also act as a complementary resource for the SemEval-2023 task 3 dataset (Piskorski et al. 2023), which does not include any Chinese data points. We provide baseline experiments highlighting the performance of the multilingual language model XLM-RoBERTa (Conneau 2019) on Chinese news framing samples, fine-tuned on the SemEval training set, our Chinese News Framing training set, and our augmented SemEval training set (a concatenation of both training sets). We also provide results obtained by GPT-4o in the zero-shot setting for all test samples in each language. We make our dataset and code available on Zenodo [1], and GitHub [2] with the CC-BY-NC-SA Licence.

*These authors contributed equally.

[1]Dataset: https://doi.org/10.5281/zenodo.14659362

[2]Code: https://github.com/GateNLP/chinese-news-framing

Related Work

News Frames

News framing can emphasise certain aspects of an issue, shaping public perception and interpretation (De Vreese 2005; Lecheler and De Vreese 2019) through the selective emphasis and organisation of specific elements within news stories.

By highlighting certain aspects of events while downplaying others, frames can profoundly influence how people think about issues, attribute responsibility, and evaluate potential solutions (Lecheler and de Vreese 2012). Frame analysis reveals how media organisations in different countries and languages present the same events through different cultural and ideological perspectives. Understanding these framing patterns also enables researchers to track how public discourse evolves over time on global issues such as climate change, public health, and international conflicts (Card et al. 2015b).

News Frame Detection

In an era of increasing digital news consumption, the ability to detect and analyse frames systematically has become essential for understanding media influence and bias across diverse platforms and contexts (Hamborg, Donnay, and Gipp 2019). To detect news frames, a number of approaches have been employed. These include topic modelling (DiMaggio, Nag, and Blei 2013), hierarchical topic modelling (Nguyen et al. 2015), cluster and sentiment analysis (Burscher, Vliegenthart, and Vreese 2016), and, more recently, neural network models, with transformer-based models now highly performant and widely used (Liu et al. 2019; Akyürek et al. 2020; Piskorski et al. 2023). One of the most commonly used models in the SemEval-2023 task 3 was the multilingual BERT-based model XLM-RoBERTa (Wu et al. 2023b; Jiang 2023; Liao, Lai, and Nakov 2023).

To achieve the automatic detection of news framing, Card et al. (2015a) developed the first large-scale dataset, the "Media Frames Corpus", and the corresponding annotation framework. The Media Frames Corpus consists of 20k English articles annotated into one or more frames from 15 categories introduced by Boydstun et al. (2014); the topics in this dataset include immigration, smoking, and same-sex marriage. Liu et al. (2019) introduced the Gun Violence Frame Corpus (GVFC), focusing solely on gun violence and the English language, using a different set of frames. Utilising the same frame definitions as Card et al. (2015b), Piskorski et al. (2023) introduced the first multilingual news framing dataset (SemEval-2023 task 3); the publicly available training and development sets contain samples in English, French, German, Italian, Polish, and Russian.

The lack of Chinese content in these existing datasets presents clear limitations for Chinese news framing analysis. The unique linguistic features of the Chinese language, such as the lack of explicit word boundaries, complex logographic character meanings and relationships, and a heavy reliance on context (Si et al. 2023; Gu et al. 2025), pose specific challenges in natural language processing. One such challenge is the tokenisation of Chinese text (Gu et al. 2025). The languages included in the SemEval-2023 task 3 dataset are primarily spoken in Western countries. This also highlights a cultural gap in the dataset, with languages influenced by Eastern culture being under-represented.

To the best of our knowledge, a Chinese dataset for automatic news framing detection has not yet been developed. To address this gap, we introduce the first Chinese News Framing dataset. This resource enables the analysis of framing patterns in Chinese media discourse across a variety of news sources and contributes to multilingual and cross-cultural news framing research.

Task Description

Given a news article, the task is to determine one or more frames applied in the article from a set of 14 generic framing dimensions (Card et al. 2015a; Piskorski et al. 2023): (1) Economic, (2) Capacity and Resources, (3) Morality, (4) Fairness and Equality, (5) Legality, Constitutionality and Jurisprudence, (6) Policy Prescription and Evaluation, (7) Crime and Punishment, (8) Security and Defence, (9) Health and Safety, (10) Quality of Life, (11) Cultural Identity, (12) Public Opinion, (13) Political, and (14) External Regulation and Reputation. We frame our task as a **multi-class, multi-label classification** problem at the news article level. Table 1 lists all the framing dimensions and their corresponding definitions.

Data

To annotate frames in Chinese articles, we use the pipeline adopted by Piskorski et al. (2023), which has been successfully applied to various languages. Specifically, our dataset development framework is divided into three distinct steps:

- *Data Collection.* To support our document-level annotation task, we begin by gathering a dataset of Chinese news articles T.
- *Data Sampling.* From T, we select a representative subset of articles, D, for annotation.
- *Data Annotation.* Lastly, we provide a detailed description of the annotation process applied to D.

Data Collection

News sources. Following Piskorski et al. (2023), we collected Chinese news articles from 14 different websites spanning 5 countries. We also selected these news outlets to ensure a balance of political biases, as determined by the Media Bias/Fact Check (MBFC) platform.[3] Table 2 presents the specifications of each news outlet according to MBFC.

Time and Topics. We collected Chinese news articles published between 2020 and the end of 2024. Our collection covers globally discussed events, including the COVID-19 vaccine, Israeli–Palestinian conflict, Russo–Ukrainian war, and US election. In total, we obtained approximately 300k news articles from 14 different news sites.

[3]https://mediabiasfactcheck.com/

Category	Definition
1: Economic	This type identifies parts of the articles referring to costs, benefits, or other financial implications.
2: Capacity and Resources	This type identifies parts of the articles referring to the availability of physical, human, or financial resources, and the capacity of current systems.
3: Morality	This type identifies parts of the articles referring to religious or ethical implications.
4: Fairness and Equality	This type identifies parts of the articles referring to the balance or distribution of rights, responsibilities, and resources.
5: Legality, Constitutionality and Jurisprudence	This type identifies parts of the articles referring to rights, freedoms, and authority of individuals, corporations, and government.
6: Policy Prescription and Evaluation	This type identifies parts of the articles referring to discussion of specific policies aimed at addressing problems.
7: Crime and Punishment	This type identifies parts of the articles referring to the effectiveness and the implications of laws and their enforcement.
8: Security and Defence	This type identifies parts of the articles referring to threats to welfare of the individual, community, or nation.
9: Health and Safety	This type identifies parts of the articles referring to health care, sanitation, and public safety
10: Quality of Life	This type identifies parts of the articles referring to threats and opportunities for the individual's wealth, happiness, and well-being.
11: Cultural Identity	This type identifies parts of the articles referring to traditions, customs, or values of a social group in relation to a policy issue.
12: Public Opinion	This type identifies parts of the articles referring to attitudes and opinions of the general public, including polling and demographics.
13: Political	This type identifies parts of the articles referring to considerations related to politics and politicians, including lobbying, elections, and attempts to sway voters.
14: External Regulation and Reputation	This type identifies parts of the articles referring to international reputation or foreign policy.

Table 1: Chinese news framing categories and definitions as used in Piskorski et al. (2023).

New Sources	Country	# of articles
BBC Chinese	UK	3k
Voice of America	USA	36k
Financial Times	UK	3k
Reuters	UK	23k
Deutsche Welle	Germany	24k
China Digital Times	USA	14k
The New York Times Chinese	USA	6k
Radio France Internationale	France	3k
Radio Free Asia	USA	11k
New Tang Dynasty	USA	79k
The Epoch Times	USA	100k
China Daily	China	3k
The Paper	China	2k
Xinhua News Agency	China	1k

Table 2: Chinese news outlets we collected data from. The "Country" column indicates the media outlet's base location.

Data Sampling

To ensure class balance in the final annotated dataset, we subsampled the collected news articles based on their topics. Specifically, given the news corpus $T = \{t_1, t_2, \ldots, t_n\}$, we first employed BERTopic (Grootendorst 2022) to assign a primary topic to each article (t_i). Next, we computed the cosine similarity between the BERTopic-generated topic representation and each framing category, then pre-assigned the article to the provisional framing category (f_j^*) based on the similarity between BERTTopic label and defined framing category. The BERTopic-generated topic representation and each framing category were encoded using the Sentence Transformer[4] (Reimers and Gurevych 2020).

$$\mathbf{z}_i = \text{SentenceTransformer}(\text{BERTopic}(t_i)) \quad (1)$$

$$\mathbf{f}_j = \text{SentenceTransformer}(f_j), \quad (2)$$

where z_i is the embedding of the BERTopic label for document t_i, and f_j is embedding of framing category j and $F = \{f_1, f_2, \ldots, f_k\}$. $\forall t_i \in T, f_j \in F$.

Framing Assignment. We assigned the most likely framing category f_j^* to each article t_i by maximising the cosine similarity between the topic embedding $\mathbf{z}_i$ and the framing category vectors $\mathbf{f}_j$:

$$f_j^* = \arg\max_{f_j} \frac{\mathbf{z}_i \cdot \mathbf{f}_j}{\|\mathbf{z}_i\| \|\mathbf{f}_j\|}. \quad (3)$$

Stratified Sampling. Based on the pre-assigned topics, we used a stratified method to sample a subset of 400 news articles (D), which is used for the training session and final annotation. Note that we chose a size of 400, which is comparable to the number of articles per language in Piskorski et al. (2023).

[4]https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

Figure 1: GATE Teamware user interface. Annotators can leave comments (see the bottom of the figure), which are used by the expert annotator during the data adjudication process (see *Data Adjudication*).

	Train set	Dev set	Test set	All
Statistics				
# of Samples	233	50	70	353
#avg. Frames	3.17	3.0	3.1	3.13
#Avg. Tokens	484	487	469	481
BBC Chinese	2	1	3	6
VOA	35	5	8	48
Reuters	21	8	7	36
NYT Chinese	3	1	1	5
DW	9	3	1	13
FT	3	2	1	6
RFA	3	1	1	5
RFI	33	6	10	49
CDT	2	1	1	4
Epoch Times	34	9	11	54
New Tang Dynasty	50	3	16	69
The Paper	11	3	4	18
Xinhua News	11	5	4	20
China Daily	14	2	4	20
Time				
2024	39	6	16	61
2023	54	11	12	77
2022	42	13	18	73
2021	47	14	8	69
2020	51	6	16	73
# of Samples per Category				
1: Economic	71	19	21	111
2: Capacity	70	15	17	102
3: Morality	22	6	7	35
4: Fairness	27	8	7	42
5: Legality	58	9	8	75
6: Policy	114	26	28	168
7: Crime	57	7	22	86
8: Security	75	15	23	113
9: Health	42	5	15	62
10: Life	41	10	10	61
11: Cultural	25	4	10	39
12: Public Opinion	33	4	11	48
13: Political	28	4	13	45
14: External	77	18	25	120

Table 3: Statistics of three subsets.

Data Annotation

Annotation Tool. We annotated Chinese news articles using an open-source data annotation tool, GATE Teamware[5] (Wilby et al. 2023), which has been employed in similar computational social science annotation tasks (Mu et al. 2023; Wu et al. 2023a; Cook et al. 2024).

Annotator Training. For annotation, we hired six native-Chinese-speaking undergraduate and postgraduate students from the University of Sheffield at a rate of £17 per hour. We first provided a two-hour training session for all participants, during which we: (i) introduced the task description (see *Task Description*), label definitions (see Table 1), and guidelines for using GATE Teamware (see Figure 1); and (ii) asked all participants to annotate 20 samples. Following this, the senior annotator released the gold-standard labels and explained their decisions for each article. The gold-standard labels for the annotator training set were created and validated by three expert annotators, all of whom are native Chinese speakers and experienced NLP researchers with expertise in media analysis.

All annotators were provided with an information sheet containing details about the task and signed consent forms in accordance with ethical guidelines.

Task Allocation. Following Piskorski et al. (2023), in the initial annotation stage, each news article was annotated by two different annotators. To distribute the samples among six annotators, such that each sample had two unique annotators, we employed a modified version of the EffiARA annotation framework (Cook et al. 2024). Using the EffiARA framework, we evenly distributed samples across annotators while maintaining the ability to assess inter- and intra-annotator agreement, as well as calculate a combined "annotator reliability factor" for each annotator.

Inter-annotator agreement can be assessed because each

[5]https://github.com/GateNLP/gate-teamware

annotator shares their samples evenly with four other annotators. This provides sufficient overlap to calculate pairwise inter-annotator agreement, visualised in Figure 2. To assess intra-annotator agreement, we assign 20 duplicate samples to each annotator. Here, we diverge slightly from Cook et al. (2024) by randomly sampling from the complete set of an annotator's annotations, rather than from the set of single-annotated samples; this is because our dataset contains only double-annotated samples.

In the annotation process, we used the EffiARA annotator reliability score to filter out an annotator with a significantly lower reliability score than all other annotators. The annotations removed based on this were re-annotated by a highly reliable annotator according the EffiARA reliability scores.

Annotator Agreement and Reliability. Each annotator's average inter-annotator agreement, intra-annotator agreement, and reliability factor were calculated as described in Cook et al. (2024). Figure 2 shows the annotator agreement within the dataset. Each node represents an annotator, the edges represent the pairwise inter-annotator agreement, and the intra-annotator agreement value is displayed next to each node. The agreement metric used in this study was the mean Krippendorff's alpha across each of the 14 news frames. For each pairwise agreement calculation, we computed the sum of agreement between two users per news frame and divide by the number of news frames. The average Krippendorff's alpha of all links (excluding annotations from annotator 5, the least reliable annotator, as their annotations did not contribute to the gold standard) is 0.465. While this is lower than the recommended level of 0.667, it is higher than that of Piskorski et al. (2023), who report an inter-annotator agreement of 0.342 in their dataset.

Note that there is an overlapping set of annotations between annotator 2 and annotator 5, as shown in Figure 2. This represents the complete set of annotations originally made by annotator 5, re-annotated by annotator 2.

Annotation Adjudication. Once all samples had been double-annotated, the senior annotator reviewed the annotations, resolving any identified conflicts as suggested in Piskorski et al. (2023). Any overlapping labels were automatically considered gold-standard, with the senior annotator using their discretion in cases of disagreement.

Chinese News Framing Dataset

The Chinese News Framing dataset consists of 353 Chinese articles from 14 different news sources, annotated with at least one news frame. Following the structure of the SemEval dataset (Piskorski et al. 2023), the Chinese Framing dataset is split into three subsets: training set (233), development set (50) and test set (70).

In Table 3, we present detailed descriptive statistics for the three subsets and the full dataset. For reference, we also provide descriptive statistics of our dataset alongside the SemEval news framing dataset (Piskorski et al. 2023) in Table 4.

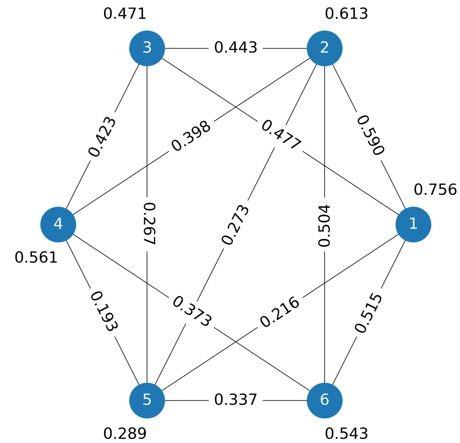

Figure 2: Inter- and intra-annotator agreement scores, with edges representing pairwise inter-annotator agreement between two annotators and the values next to nodes representing the intra-annotator agreement score, using a multi-label variant of Krippendorff's alpha. The reliability scores, using $\alpha = 0.5$ in the EffiARA reliability score calculation, are (from annotator 1 to 6): 1.309, 1.141, 0.945, 0.978, 0.576, 1.051.

Experimental Setup

In this section, we present experiments designed to evaluate the quality of our dataset and the value of its integration into the existing SemEval dataset.

The aims of our experiments are as follows:

(i) Provide baseline results for each language when models are fine-tuned in a multilingual setting, assessing the value of the Chinese News Framing dataset;

(ii) Provide baseline results for each language using a state-of-the-art decoder model (GPT-4o (OpenAI et al. 2024)) for comparison with fine-tuned XLM-RoBERTa models;

(iii) Understand whether our Chinese News Framing dataset facilitates the creation of monolingually fine-tuned models capable of classifying articles into a set of given news frames;

(iv) Understand the impact of augmenting the SemEval dataset with Chinese News Framing samples, using the classification performance of each language as the performance metric.

Augmenting the SemEval Dataset

As shown above, our dataset was developed following a similar methodology to that used to create the SemEval dataset (Piskorski et al. 2023). This suggests that our dataset can not only be used as a stand-alone resource for Chinese news framing analysis but also as a complementary addition to the SemEval dataset.

To assess the value of our dataset, we experiment with three training sets: the original SemEval dataset, our novel Chinese News Framing dataset, and the augmented SemEval dataset containing Chinese news framing samples. In these

Metric	Chinese	English	French	German	Italian	Polish	Russian	Georgian	Greek	Spanish
#docs	**353**	590	261	227	364	241	263	29	64	30
#avg. Frames	**3.1**	4.0	3.0	4.7	3.8	5.2	2.1	1.7	2.9	2.3
1: Economic	**111**	74	79	108	142	144	68	2	14	4
2: Capacity & Resources	**102**	56	62	104	120	88	34	4	10	44
3: Morality	**35**	231	62	39	62	63	31	2	5	7
4: Fairness & Equality	**42**	131	30	35	52	39	21	0	8	2
5: Legality & Jurisprudence	**75**	281	41	65	73	56	44	0	23	7
6: Policy Prescription	**168**	154	38	70	129	110	15	2	12	7
7: Crime & Punishment	**86**	274	22	44	57	57	51	3	11	4
8: Security & Defense	**113**	222	89	121	155	105	90	10	19	10
9: Health & Safety	**62**	86	60	107	97	144	37	4	8	3
10: Quality of Life	**61**	115	40	53	89	85	32	0	5	3
11: Cultural Identity	**39**	42	34	46	43	48	13	1	8	0
12: Public Opinion	**48**	68	34	50	58	74	22	4	10	3
13: Political	**45**	343	108	130	178	144	55	10	43	6
14: External Reputation	**120**	214	85	91	132	86	44	9	9	3

Table 4: Descriptive statistics of the Chinese Framing dataset and the individual languages from the SemEval dataset (Piskorski et al. 2023).

experiments, we treat the SemEval development set as the test set because the official SemEval test set is not publicly available. We consider all the available languages offered in the SemEval training and development sets: English, French, German, Italian, Polish, and Russian. Statistics for the augmented SemEval dataset are shown in Table 4.

Model

We conducted experiments using XLM-RoBERTa-base[6] (Conneau et al. 2020). XLM-RoBERTa was one of the most widely used transformer-based multilingual models among SemEval participants (Piskorski et al. 2023; Jiang 2023; Liao, Lai, and Nakov 2023).

Experimental Details

Text Pre-processing. To prepare the article text for classification, following Wu et al. (2023b), we conducted the following additional pre-processing steps:

- Add a newline character between the news title and body;
- Remove duplicate sentences occurring consecutively;
- Remove hyperlinks to websites and images;
- Remove strings detailing author biographies (such as names and affiliations).

Model Training. We fine-tuned XLM-RoBERTa-base for 100 epochs, with 10 warm-up epochs, a batch size of 8, and a max-sequence length of 512 tokens. We utilised the AdamW optimiser with a linear weight decay of 0.01. The loss function used in this multi-class, multi-label classification task was Binary Cross-Entropy with Logit Loss.

The learning rate was selected through initial experimentation on the Chinese News Framing development set, as our dataset offers train, development, and test splits. These initial experiments involved fine-tuning over only 30 epochs, maintaining a warm-up rate of 0.1. The learning rates tested were those used by Wu et al. (2023b) across their experiments: 1e-4, 5e-5, 3e-5, 2e-5, and 5e-6. The best-performing learning rate, measured by the F1-micro score on the Chinese News Framing development set, was selected for the final set of experiments.

All models are trained three times with the set of three seeds {555, 666, 777}, using the highest-performing learning rate obtained using the Chinese-only development set as described above. The key performance metric in this study, following the official performance metric of SemEval-2023 task 3 subtask 2 (Piskorski et al. 2023), is the F1-micro score; we report the mean and standard deviation across the three seeds for each language. The code and configuration files required to run all experiments are available at https://github.com/GateNLP/chinese-news-framing.

All experiments were run on an Nvidia RTX 4090 with 24GB of VRAM.

GPT-4o Experiments. We employed GPT-4o (OpenAI et al. 2024) in a zero-shot setting to generate frame labels in a comma-separated format. The temperature parameter was set to 0.0 for deterministic outputs. Following our supervised experiments, we conducted three runs per language and report the mean F1-micro score with standard deviation.

Results and Discussion

Table 5 displays the experimental results of GPT-4o in the zero-shot setting and the results of XLM-RoBERTa-base fine-tuned on the three different datasets: the SemEval training set, our Chinese News Framing training set, and the SemEval training set augmented with Chinese news framing samples.

Experimental Results

SemEval Baseline Performance. Our baseline results, using XLM-RoBERTa-Base fine-tuned on the SemEval training set and tested over three random seeds on the development set, closely align with the results obtained by Wu et al. (2023b), with the only results differing by more than an F1-micro score of over 0.01 being English, Polish, and Russian, which have differences of $+0.05$, -0.018, and $+0.035$;

[6]https://huggingface.co/FacebookAI/xlm-roberta-base

Training Data	Chinese	English	French	German	Italian	Polish	Russian
Zero-Shot GPT-4o	0.560±0.004	0.603±0.009	0.528±0.001	0.541±0.004	0.540±0.002	0.575±0.010	0.508±0.002
SemEval (Piskorski et al. 2023)	0.584±0.016	0.733±0.011	**0.589±0.017**	**0.643±0.014**	**0.599±0.011**	0.647±0.009	**0.584±0.014**
Chinese News Framing	0.719±0.012	0.570±0.013	0.433±0.011	0.515±0.013	0.546±0.003	0.516±0.016	0.434±0.024
SemEval + Chinese News Framing	**0.753±0.015**	**0.739±0.023**	0.578±0.007	0.639±0.023	0.592±0.004	**0.670±0.007**	0.542±0.008

Table 5: Experimental results for GPT-4o in the zero-shot setting and for XLM-RoBERTa-base fine-tuned with three different training sets, trained for 100 epochs, with a learning rate of 5e-5.

a positive result indicates that our model performs better. It is worth noting the difference in experimental setup. The experiments conducted by Wu et al. (2023b) involve a 3-fold cross-validation on the development set to create three different test sets, whereas we maintain the same test set (the SemEval development set) over three different random seeds.

While not trained on Chinese news framing samples, the model performs within 0.005 of the scores achieved for French and Russian. The F1-micro score achieved on the Chinese test set without explicit Chinese-language fine-tuning was 0.584.

Chinese News Framing Dataset Only. Training on the Chinese News Framing dataset, as expected, significantly increases the model's performance on Chinese news framing test samples. By training on only the Chinese News Framing data, the Chinese test score increases to 0.719; this is a higher performance than all languages except English in our SemEval baseline performance results. This positive performance on our dataset indicates that it is a valuable tool and benchmark for the Chinese news framing task.

Augmented SemEval Performance. After augmenting the SemEval dataset with our Chinese News Framing dataset, we observe a further increase in performance for the Chinese language, achieving an F1-micro score of 0.753, which is the highest score attained on any language in all of our experiments. While most languages show little change (less than 0.01 in F1-micro) in performance from our SemEval-only baseline experiments, the French news framing performance decreases by 0.011 and Russian by 0.042; Chinese classification performance increases by 0.171 and Polish by 0.023.

By augmenting the SemEval dataset with our Chinese News Framing dataset, we observe comparable classification performance in the majority of languages, with a significant increase in performance for the Chinese language. This indicates that our Chinese News Framing dataset serves as a valuable complementary addition to the SemEval dataset.

GPT-4o Performance. The experimental results show that GPT-4o achieves lower F1 scores than XLM-RoBERTa-Base (trained on the augmented SemEval set containing Chinese news framing samples) in all languages. This performance gap arises from the distinct prediction pattern of GPT-4o; despite achieving higher recall, it suffers from significantly lower precision. As a generative large language model, GPT-4o tends to assign more labels to texts, interpreting text-topic associations from an overly broad semantic perspective. This leads to the labelling of content that is only superficially related to target categories. Although this over-generalised approach improves topic discovery, it introduces numerous false positives, ultimately compromising classification accuracy.

Frames	P	R	F1	Support
1: Economic	0.83	0.90	0.86	21
2: Capacity & Resources	0.47	0.47	0.47	17
3: Morality	1.00	0.71	0.83	7
4: Fairness & Equality	0.71	0.71	0.83	7
5: Legality	0.26	0.62	0.37	8
6: Policy	0.69	0.71	0.70	28
7: Crime & Punishment	0.94	0.73	0.82	22
8: Security & Defence	0.82	0.78	0.80	23
9: Health & Safety	1.00	0.67	0.80	15
10: Quality of Life	0.83	0.50	0.62	10
11: Cultural Identity	0.88	0.70	0.78	10
12: Public Opinion	0.38	0.27	0.32	11
13: Political	1.00	0.85	0.92	13
14: External Regulation	0.84	0.84	0.84	25

Table 6: Classification report on the Chinese News Framing dataset, using XLM-RoBERTa-Base trained exclusively on our dataset. It shows the precision, recall, F1 score and the support (or number of samples) for each class.

Error Analysis

We also provide further error analysis on the performance of XLM-RoBERTa-Base trained only on our Chinese News Framing dataset. We analyse the performance of this model to highlight trends within our dataset alone, including potential strengths and weaknesses for the model based on our training set.

Classification Report. Table 6 shows the precision, recall, F1 score, and support for each individual class within the Chinese News Framing test set. We observe that the model has particular issues in both precision and recall for the frames (2) Capacity & Resources, (5) Legality, and (12) Public Opinion. When comparing this to the same metrics for the English language development samples in the SemEval dataset, we also observe that these frames are difficult for the model to classify. A key exception to this is in English and French, where (5) Legality is well classified with respect to the other frames for those languages. For English,

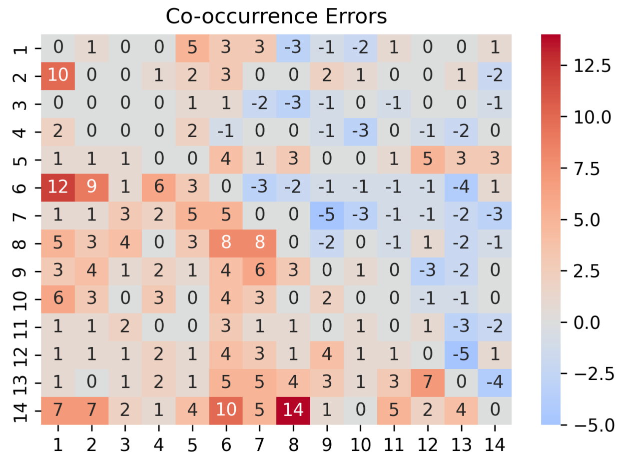

Figure 3: Co-occurrence Error Matrix on the Chinese News Framing dataset. Note that labels 1 to 14 denote the news frame orders as described in Task Description. Positive values represent over-prediction and negative values represent under-prediction of the co-occurrence of two classes.

this can be explained by the large set of samples containing the frame Legality & Jurisprudence. The frames that are most correctly identified are (1) Economic, (3) Morality, (4) Fairness & Equality, (13) Political, and (14) External Regulation. This does not align with any of the other languages that are trained only on the SemEval set and does not correspond to the number of samples supporting each class.

It is worth noting that with the Augmented SemEval set containing Chinese News Framing samples, similar trends are observed across languages. For the Chinese language, however, the performance of the model in identifying the frame (12) Public Opinion increases by an F1 score of 0.35. Through the analysis of each language's F1-macro performance, the Chinese language performs significantly better than other languages, achieving 0.74; the next best-performing language, based on F1-macro, is Polish with a score of 0.54. This indicates that our dataset is particularly effective for training models to successfully identify the majority of news frames as well as those with higher support (indicated by the high F1-micro). It is important to consider, however, that in downstream applications, such as balancing the framing within a user's news feed, frequent misclassifications could result in unintended bias or an imbalance in frame representation.

Co-occurrence Error Matrix. We also conducted co-occurrence error analysis, shown in Figure 3. The co-occurrence error matrix is used to understand how well the model is able to capture the relationships between pairs of classes in a multi-class, multi-label prediction task. As the co-occurrence matrix is symmetric, we display the number of true co-occurrences in the lower triangle and the difference in co-occurrences in the upper triangle; positive scores indicate that the model has over-predicted the co-occurrence of two news frames and negative scores indicate an under-prediction.

One class showing a number of over-predictions with other classes is (5) Legality. It is particularly overestimated to co-occur with (1) Economic, (6) Policy, (12) Public Opinion, and (13) Political. This is likely due to the poor precision identified for the (5) Legality news frame shown in Table 6.

The (13) Political news frame is shown to be under-predicted to co-occur with a number of news frames, including: (6) Policy, (11) Cultural Identity, (12) Public Opinion, and (14) External Reputation. As the (13) Political news frame is successfully identified by the model but not very well predicted with the correct co-occurring news frames, it indicates that the model does not capture the dependency between news frames. Politics often influences policy, cultural identity, public opinion, and external reputation, but this is not captured very well by the model. Capturing this relationship between news frames in the classification stage may be of interest in future research.

Conclusion

In this work, we have created and published the Chinese News Framing dataset, facilitating news framing experiments in the Chinese language and serving as a complementary asset to the SemEval dataset (Piskorski et al. 2023). We have demonstrated the need for Chinese news framing samples to be added to the SemEval set in order to effectively detect news frames in the Chinese language. We have also reported benchmark results, training on only our Chinese News Framing dataset and an augmented SemEval dataset containing Chinese news framing samples, achieving an F1-micro score of 0.753 on our Chinese News Framing test set; this score represents an improvement over training on only Chinese News Framing samples. Augmenting the SemEval dataset maintains similar F1-micro performance on the SemEval development set, while improving the news framing detection in the Chinese, English, and Polish languages.

While this work provides experiments to demonstrate this dataset's efficacy as an addition to the SemEval set, it also presents experiments achieving high classification performance as a stand-alone dataset. With the additional metadata available about individual annotations, this work allows for future research involving the development of advanced classification methods, utilising annotator identities (Cook et al. 2024).

Limitations

The articles within this dataset were collected between 2020 and 2024, offering news framing samples concerning a variety of topics. While this serves as a valuable resource, it does not allow for the analysis of news framing over time, allowing for future work to supplement this dataset with the annotation of articles from different time periods.

As the text is not directly made publicly available in the dataset, relying on the user retrieving content from a web link, there is a risk that some article texts may become unavailable after the release of the dataset.

Although we provided annotator guidelines and training, subjective interpretations of news frames are likely to be present in the dataset. We attempt to mitigate this by providing each individual annotation in our dataset, rather than only the aggregated "gold-standard" labels. This allows for further research investigating the uncertainty and subjectivity involved in this task.

Dataset Availability

To maximise the use of our dataset, we adhere to the FAIR principles (Wilkinson et al. 2016).

- *Findable.* Our dataset has been published to the Zenodo dataset sharing service: https://doi.org/10.5281/zenodo.14659362; our experimental code is also available at https://github.com/GateNLP/chinese-news-framing.
- *Accessible.* All data is publicly accessible through web links in the published dataset; we also provide a tool to obtain the news article text from the given link in our code repository.
- *Interoperable.* The CSV format is widely accepted and can be processed by a wide range of data processing tools.
- *Reusable.* The CSV format is widely accepted and can be processed by a wide range of data processing tools.

References

Akyürek, A. F.; Guo, L.; Elanwar, R.; Ishwar, P.; Betke, M.; and Wijaya, D. T. 2020. Multi-label and multilingual news framing analysis. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, 8614–8624.

Alonso del Barrio, D.; and Gatica-Perez, D. 2023. Framing the news: from human perception to large language model inferences. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, 627–635.

Boydstun, A. E.; Card, D.; Gross, J. H.; Resnik, P.; and Smith, N. A. 2014. Tracking the development of media frames within and across policy issues. In *APSA 2014 annual meeting paper*.

Burscher, B.; Vliegenthart, R.; and Vreese, C. H. d. 2016. Frames beyond words: Applying cluster and sentiment analysis to news coverage of the nuclear power issue. *Social Science Computer Review*, 34(5): 530–545.

Card, D.; Boydstun, A.; Gross, J. H.; Resnik, P.; and Smith, N. A. 2015a. 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.

Card, D.; Boydstun, A. E.; Gross, J. H.; Resnik, P.; and Smith, N. A. 2015b. The Media Frames Corpus: Annotations of Frames Across Issues. In Zong, C.; and Strube, M., eds., *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. Beijing, China: Association for Computational Linguistics.

Conneau, A. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, É.; Ott, M.; Zettlemoyer, L.; and Stoyanov, V. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8440–8451.

Cook, O.; Grimshaw, C.; Wu, B.; Dillon, S.; Hicks, J.; Jones, L.; Smith, T.; Szert, M.; and Song, X. 2024. Efficient Annotator Reliability Assessment and Sample Weighting for Knowledge-Based Misinformation Detection on Social Media. *arXiv preprint arXiv:2410.14515*.

De Vreese, C. H. 2005. News framing: Theory and typology. *Information design journal+ document design*, 13(1): 51–62.

DiMaggio, P.; Nag, M.; and Blei, D. 2013. Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of US government arts funding. *Poetics*, 41(6): 570–606.

D'angelo, P. 2018. Doing news framing analysis II. *Empirical and Theoretical Perspectives*.

Entman, R. M. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of communication*, 43(4): 51–58.

Gamson, W. A.; Croteau, D.; Hoynes, W.; and Sasson, T. 1992. Media images and the social construction of reality. *Annual review of sociology*, 18(1): 373–393.

Grootendorst, M. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Gu, Y.; Huang, Z.; Zeng, M.; Qiu, M.; and Park, J. 2025. Improving Automatic Grammatical Error Annotation for Chinese Through Linguistically-Informed Error Typology. In *Proceedings of the 31st International Conference on Computational Linguistics*, 2781–2798.

Gururangan, S.; Marasović, A.; Swayamdipta, S.; Lo, K.; Beltagy, I.; Downey, D.; and Smith, N. A. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8342–8360.

Hamborg, F.; Donnay, K.; and Gipp, B. 2019. Automated identification of media bias in news articles: an interdisciplinary literature review. *International Journal on Digital Libraries*, 20.

Hertog, J. 2001. A multiperspectival approach to framing analysis: A field guide. *Framing public life/Erlbaum*.

Jiang, Y. 2023. Team QUST at SemEval-2023 Task 3: A Comprehensive Study of Monolingual and Multilingual Approaches for Detecting Online News Genre, Framing and Persuasion Techniques. In Ojha, A. K.; Doğruöz, A. S.; Da San Martino, G.; Tayyar Madabushi, H.; Kumar, R.; and Sartori, E., eds., *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, 300–306. Toronto, Canada: Association for Computational Linguistics.

Kwak, H.; An, J.; and Ahn, Y.-Y. 2020. A systematic media frame analysis of 1.5 million new york times articles from 2000 to 2017. In *Proceedings of the 12th ACM Conference on Web Science*, 305–314.

Lecheler, S.; and de Vreese, C. H. 2012. News Framing and Public Opinion. *Journalism & Mass Communication Quarterly*, 89: 185 – 204.

Lecheler, S.; and De Vreese, C. H. 2019. *News framing effects: Theory and practice*. Taylor & Francis.

Liao, Q.; Lai, M.; and Nakov, P. 2023. MarsEclipse at SemEval-2023 Task 3: Multi-lingual and Multi-label Framing Detection with Contrastive Learning. In Ojha, A. K.; Doğruöz, A. S.; Da San Martino, G.; Tayyar Madabushi, H.; Kumar, R.; and Sartori, E., eds., *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, 83–87. Toronto, Canada: Association for Computational Linguistics.

Liu, S.; Guo, L.; Mays, K.; Betke, M.; and Wijaya, D. T. 2019. Detecting frames in news headlines and its application to analyzing news framing trends surrounding US gun violence. In *Proceedings of the 23rd conference on computational natural language learning (CoNLL)*, 504–514.

Morstatter, F.; Wu, L.; Yavanoglu, U.; Corman, S. R.; and Liu, H. 2018. Identifying framing bias in online news. *ACM Transactions on Social Computing*, 1(2): 1–18.

Mu, Y.; Jin, M.; Grimshaw, C.; Scarton, C.; Bontcheva, K.; and Song, X. 2023. Vaxxhesitancy: A dataset for studying hesitancy towards covid-19 vaccination on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, 1052–1062.

Nguyen, V.-A.; Boyd-Graber, J.; Resnik, P.; and Miler, K. 2015. Tea party in the house: A hierarchical ideal point topic model and its application to republican legislators in the 112th congress. 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)*, 1438–1448.

OpenAI; Achiam, J.; Adler, S.; Agarwal, S.; and et al. 2024. GPT-4 Technical Report. arXiv:2303.08774.

Piskorski, J.; Stefanovitch, N.; Da San Martino, G.; and Nakov, P. 2023. SemEval-2023 Task 3: Detecting the Category, the Framing, and the Persuasion Techniques in Online News in a Multi-lingual Setup. In Ojha, A. K.; Doğruöz, A. S.; Da San Martino, G.; Tayyar Madabushi, H.; Kumar, R.; and Sartori, E., eds., *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, 2343–2361. Toronto, Canada: Association for Computational Linguistics.

Reimers, N.; and Gurevych, I. 2020. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Reiter-Haas, M.; Klösch, B.; Hadler, M.; and Lex, E. 2024. FrameFinder: Explorative Multi-Perspective Framing Extraction from News Headlines. In *Proceedings of the 2024 Conference on Human Information Interaction and Retrieval*, 381–385.

Si, C.; Zhang, Z.; Chen, Y.; Qi, F.; Wang, X.; Liu, Z.; Wang, Y.; Liu, Q.; and Sun, M. 2023. Sub-character tokenization for Chinese pretrained language models. *Transactions of the Association for Computational Linguistics*, 11: 469–487.

Van Dijk, T. A. 2023. Analyzing frame analysis: A critical review of framing studies in social movement research. *Discourse Studies*, 25(2): 153–178.

Wang, G.; Frederick, R.; Haghighi, B. T.; Wong, B. W.; Rupar, V.; Li, W.; and Bai, Q. 2024. FramedTruth: A Frame-Based Model Utilising Large Language Models for Misinformation Detection. In *Asian Conference on Intelligent Information and Database Systems*, 135–146. Springer.

Wilby, D.; Karmakharm, T.; Roberts, I.; Song, X.; and Bontcheva, K. 2023. GATE Teamware 2: An open-source tool for collaborative document classification annotation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 145–151.

Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L. B.; Bourne, P. E.; et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3(1): 1–9.

Wu, B.; Li, Y.; Mu, Y.; Scarton, C.; Bontcheva, K.; and Song, X. 2023a. Don't waste a single annotation: improving single-label classifiers through soft labels. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 5347–5355.

Wu, B.; Razuvayevskaya, O.; Heppell, F.; Leite, J. A.; Scarton, C.; Bontcheva, K.; and Song, X. 2023b. SheffieldVeraAI at SemEval-2023 Task 3: Mono and Multilingual Approaches for News Genre, Topic and Persuasion Technique Classification. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, 1995–2008.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes.
   (e) Did you describe the limitations of your work? Yes, see Limitations.
   (f) Did you discuss any potential negative societal impacts of your work? Yes, see Ethics Statement.

(g) Did you discuss any potential misuse of your work? Yes, see Ethics Statement.

(h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, see Dataset Availability and Ethics Statement.

(i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes.

2. Additionally, if your study involves hypotheses testing...

(a) Did you clearly state the assumptions underlying all theoretical results? N/A.

(b) Have you provided justifications for all theoretical results? N/A.

(c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? N/A.

(d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? N/A.

(e) Did you address potential biases or limitations in your theoretical framework? N/A.

(f) Have you related your theoretical results to the existing literature in social science? N/A.

(g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? N/A.

3. Additionally, if you are including theoretical proofs...

(a) Did you state the full set of assumptions of all theoretical results? N/A

(b) Did you include complete proofs of all theoretical results? N/A

4. Additionally, if you ran machine learning experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes, see the footnote on page 1.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes, see Experimental Setup.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Yes, see Table 5.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes, see Experimental Details.

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes, see Results and Discussion.

(f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes, see Classification Report.

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...

(a) If your work uses existing assets, did you cite the creators? Yes.

(b) Did you mention the license of the assets? No, as all referenced datasets and assets were publicly available and used in accordance with their respective licenses. Appropriate citations have been included throughout to ensure compliance.

(c) Did you include any new assets in the supplemental material or as a URL? Yes, see the footnote on page 1.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes, see Ethics Statement.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, see Ethics Statement.

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? Yes, see Dataset Availability.

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? Yes, see the Zenodo link in Dataset Availability.

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? Yes, see the Zenodo link in Dataset Availability.

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? Yes, see Ethics Statement.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? Yes, see Data Annotation.

(d) Did you discuss how data is stored, shared, and deidentified? Yes, see Dataset Availability and Ethics Statement.

Ethics Statement

Our study has received ethical approval from the University of Sheffield. All participants provided informed consent by signing the consent form after reviewing the information sheet detailing the annotation task. Annotator identities remain anonymous and any news content itself will not be included in the published dataset.

The potential use of this dataset by malicious actors has been considered, potentially allowing the generation of maliciously framed content; this behaviour is already possible with existing Large Language Model technology (Gururangan et al. 2020). Proper use, however, has much more potential for good in the form of understanding media bias, providing balanced search results, and conducting large-scale content analysis in research.