Modeling Multi-level Context for Informational Bias Detection by Contrastive Learning and Sentential Graph Network

Anonymous Author(s)

ABSTRACT

Informational bias is widely present in news articles. It refers to providing one-sided, selective or suggestive information of specific aspects of certain entity to guide a specific interpretation, thereby biasing the reader's opinion. Sentence-level informational bias detection is a very challenging task in a way that such bias can only be revealed together with the context, examples include collecting information from various sources or analyzing the entire article in combination with the background. In this paper, we integrate three levels of context to detect the sentence-level informational bias in English news articles: adjacent sentences, whole article, and articles from other news outlets describing the same event. Our model, MultiCTX (Multi-level ConTeXt), uses contrastive learning and sentence graphs together with Graph Attention Network (GAT) to encode these three degrees of context at different stages by tactically composing contrastive triplets and constructing sentence graphs within events. Our experiments proved that contrastive learning together with sentence graphs effectively incorporates context in varying degrees and significantly outperforms the current SOTA model sentence-wise in informational bias detection.

CCS CONCEPTS

• **Information systems** → **Document representation**; • **Applied computing** → **Sociology**; • **Networks** → *Network design principles.*

KEYWORDS

bias detection, informational bias, media bias, news bias, graph neural networks, contrastive learning, contextual modeling

ACM Reference Format:

Anonymous Author(s). 2018. Modeling Multi-level Context for Informational Bias Detection by Contrastive Learning and Sentential Graph Network. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Informational bias broadly exists in news articles. As a sort of framing bias, it always frames a certain entity by specific aspects using narrow, speculative or indicative information to guide a particular interpretation, thus swaying readers' opinion.

For most of us, news articles, especially online news articles are the main source of information. Therefore, they play a central role in shaping individual and public opinions. However, news reports often show internal bias. The current research is often limited to the lexical bias. This form of bias rarely depends on the context of the sentence. It can be eliminated by deleting or replacing a small number of biased words. Contrarily, researchers Fan et al. [8] found that the informational bias is more common and more difficult to detect.

Different from other types of bias, the sentence-level informational bias detection largely depends on the context and this fact makes the task very challenging. A sentence alone can be expressed in a neutral manner, but it might be revealed as biased in consideration of the context. Take the second row in Table 1 as example: the sentence *"Mr. Mattis, a retired four-star Marine general, was rebuffed."* seems to be a very simple declarative sentence stating a fact. However, if we read the previous sentence *"Officials said Mr. Mattis went to the White House with his resignation letter already written, but nonetheless made a last attempt at persuading the president to reverse his decision about Syria, which Mr. Trump announced on Wednesday over the objections of his senior advisers."* (the first row in Table 1) , we will know that *'a retired four-star Marine general'* indicates a negative, even ironic tone towards Mr. Mattis and his last attempt. Therefore, sentence-level informational bias can only be revealed by collecting information from various sources and analyzing the entire article together with its background. Such subtleties of informational bias are more likely to affect unsuspecting readers, which indicates the necessity of research into new detection methods.

In this paper, we propose MultiCTX (Multi-level ConTeXt), a model composed of contrastive learning and sentence graph attention networks to encode three different levels of context: **1) Neighborhood-context**: adjacent sentences, i.e. sentences in the same article around the target sentence; **2) Article-context**: the whole article containing the target sentence; **3) Event-context**: articles from various news media reporting the same event. These three levels encompass the contextual information from the most local to the most global.

In order to make use of the context rather than be overwhelmed by the noise introduced, MultiCTX prioritizes contrastive learning which learns sentence embedings via discriminating among (*target*, *positive sample*, *negative sample*) triplets to distill the essence of the target sentence. The quality of the learned CSE (Contrastive Sentence Embedding) relies on that of triplets. Other than the traditional brute-force way to select triplets only based on their labels, MultiCTX further considers article-level information which creates higher-quality triplets. Such triplet formulation guarantees that our CSEs infuse the context and reflect sentences' inherent semantics instead of the shallow lexical features.

MultiCTX then builds a relational sentence graph using CSEs. Edges are connected between two sentences if they are logically

Woodstock '18, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

related in the same *neighborhood* or if they are continuous in entities or semantically similar within the same *event*. Finally we apply a Self-supervised Graph Attention Network (SSGAT) on our sentence graph to make the final informational bias prediction. The SSGAT structure encodes neighborhood-level and event-level context via edges, making it possible for textually distant but contextually close sentences to connect directly. The flexible graph structure extends beyond the sequential arrangement of traditional LSTMs, which also consider the surrounding context.

Although document graphs are not rare in NLP tasks, they are often short and built by token-wise dependency parsing. It may suffer from high complexity and considerable noise when applied on long texts which is our case with news articles. Our relational sentence graph uses sentence nodes and focuses on inter-sentence relationships. It recquires only minimal syntax parsing, takes on less noise and has better interpretability.

Few research studies sentence-level informational bias detection by infusing context. Fan et al. [8] first published a human-annotated dataset on this task, taking the context into account during annotation. However, sentences are still treated sentences individually in their model. van den Berg and Markert [17] did a primary research on incorporating different levels of context in the informational bias detection. However, they consider only one kind of context in each model. To our best knowledge, our model is the first to incorporate multi-level contextual information in sentence-level classification task.

In summary, we present the following contributions:

- We are the first to incorporate three different levels of context together in the sentence-level bias detection task.
- We propose a novel triplet formation for contrastive learning in bias detection. The methodology can be generalized for other tasks.
- We are the first to use a sentence graph to encode the textual context information in the bias detection task.
- Our model MultiCTX significantly outperforms the current state-of-the-art model by 2 percentage points F1 score. It indicates that contextual information effectively helps sentence-level informational bias detection and our model successfully infuses multi-level context.

2 METHODOLOGY

Figure 3 illustrates our model MultiCTX (Multi-level ConTeXt). First, we carefully construct triplets from the original dataset and then apply supervised contrastive learning on them to obtain sentence embeddings. Second, we build relational sentence graphs by joining sentence nodes according to their discourse relationships and semantic similarity. Finally, we apply a Self-supervised Graph Attention Network [13] to perform the bias detection as a node classification task. In essence, MultiCTX has two modules, Contrastive Learning Embedding (CSE) and Self-supervised Sentence Graph Attention Network (SSGAT). In order to investigate the role of the context and to imitate the way people learn from the news reports, we also apply a more reasonable and challenging cross-event data splitting.

2.1 Data splitting

First of all, let's think about the nature of the news reports and the way people learn about the world in real life. News articles always emerge almost simultaneously in large numbers along with a particular event, over which people reason based on their experience learnt from previous events. Moreover, people usually read an article as a whole instead of randomly picking up several sentences and they are unlikely to encounter a sentence from news events happened before. Additionally, people tend to collect information from more than one article to get a bigger picture of the new event. Therefore, in order to simulate the real human's learning process, different from the commonly-used data splitting which randomly distributes sentences to one of the three subsets (train/val/test), we use event-wise data splitting mentioned in van den Berg and Markert [17], Chen et al. [4]. We treat the articles reporting the same event as a unit and keeping sentences from the same event in the same subset. Part of the data is shown in Table 1 with clear 'adjacent sentences', 'article' and 'event' structure.

Furthermore, splitting by events is more reasonable and more demanding, in terms of model generalizability for identifying informational bias from unseen events. Experiments in van den Berg and Markert [17] and Chen et al. [4] also show that common models including BERT-based models all experienced a considerable performance drop when switching from random splitting to event-based splitting.

2.2 Sentence Embedding using Contrastive Learning

The idea of contrastive learning is that humans discriminate objects by "comparison", thus similar objects should be close to each other in the representation space, and different objects should be as far apart as possible. However, news sentences inherently have small differences in terms of pure text. Two sentences with opposite stances might be different in a few words, while two sentences expressing the same idea are likely to be formulated completely differently. To address the problem, we apply supervised contrastive learning with hard negatives described in Gao et al. [10]. The idea is to develop, from the original dataset, the triplets (x_i, x_i^+, x_i^-) each denotes target sentence, positive sample and negative sample respectively. Using the $\mathbf{h}_i, \mathbf{h}_{i^+}, \mathbf{h}_i^-$ representations of x_i, x_i^+, x_i^-, the objective function to minimize is InfoNCE Loss.

The difficulty is to mine the positive and negative samples for each target sentence from the original dataset. A good positive sample is supposed to capture the most essential features of the sentence, rather than being influenced by other factors, such as the writing styles of different news media. Therefore, the best positive sample is expected to be significantly different from the target sample in terms of sentence formation, while the best negative sample should be similar to the target sentence in terms of syntactic structure. In short, samples with different labels from the target sentence but with initial embedding in its vicinity are likely to be the most useful, providing significant gradient information during the training process.

Inspired by Baly et al. [1] which applies a triplet loss in training using news media in triplet selection, our final triplet follows article-based criteria and is composed of: x_i: target sentence; x_i^+: same label

Event	Source	Index	Sentence	Label
86	nyt	3	Officials said Mr. Mattis went to the White House with his resignation letter already written, but nonetheless made a last attempt at persuading the president to reverse his decision about Syria, which Mr. Trump announced on Wednesday over the objections of his senior advisers.	0
86	nyt	4	Mr. Mattis, a retired four-star Marine general, was rebuffed.	1
86	nyt	5	Returning to the Pentagon, he asked aides to print out 50 copies of his resignation letter and distribute them around the building.	0
11	fox	20	However, Democrats rejected the plan even before Trump announced it, and a Senate version of the plan failed to get the 60 votes needed on Thursday.	1
11	fox	21	A second bill, already passed by the Democrat-controlled House to re-open the government, also fell short.	0
2	hpo	10	There were roughly 520,000 arrests for unauthorized border crossings last year, which is about one-third of the 1.6 million arrests that happened in 2000.	0
2	hpo	11	Since 2014, a high proportion of those crossing have been Central American children and families seeking to make humanitarian claims such as asylum.	1

Table 1: BASIL dataset

and event with x_i, but from a different article; x_i^-: from the same article with x_i but with a different label. Figure 1 illustrates our triplet construction.

Thereby we essentially augment the original 7977-sentence corpus to a much larger dataset of around 300k triplets where sentences are no longer isolated but linked to two others. More importantly, triplets with the same target sentence provide altogether a 'context' to help its representational learning. This context integrates two advantages: 1) it provides article-level contextual information from the same article by the negative sample, and also event-level context from other articles of the same event; 2) it discourages the model to learn from superficial writing styles of the article or the news publisher. Our experiments also confirmed that this article-based triplet construction is better than either the news outlet-based triplet or the event-based one.

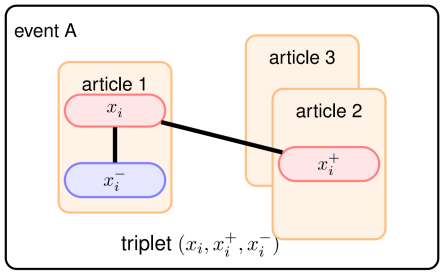


Figure 1: Triplet construction. Positive sample x_i^+ has same label (red) and event with target sentence x_i; negative sample x_i^- has different label (blue) but from the same article with x_i; note that three sentences must from same event.

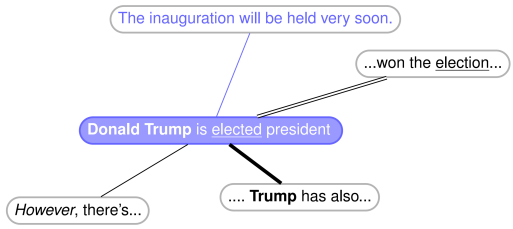
2.3 Relational Sentence Graph

Sentences are naturally suitable as nodes when encoding long documents, so we borrowed the idea from extractive text summarization from Christensen et al. [5] and Zhao et al. [18] to construct graphs. The graphs are formed by connecting the sentences in four different ways illustrated in 2(a):

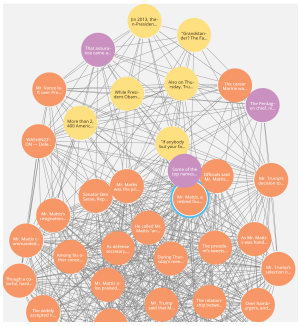
(1) Deverbal noun reference: when an action in verb form occurs in the current sentence, it is likely to be mentioned in the noun form in the following sentences. So we attach the current sentence with its downstream sentence when at least one semantically similar deverbal noun is found in the latter.
(2) Discourse marker: If the immediately subsequent sentence begins with a discourse marker (e.g., however, meanwhile, furthermore), the two sentences are linked.
(3) Entity continuation: we connect two sentences in the same event if they contain the same entity.
(4) Sentence similarity: sentence pairs in the same event with high cosine similarity are joined.

The four types of edge formation take different degrees of context into account: Type 1 and Type 2 consider only the subsequent sentences in the same article (neighborhood-context). In particular, Type 2 considers only the immediately following sentence. Type3 and Type 4 are not limited to adjacent sentences. Rather they consider the whole event (event-context). Note that edges occur only between in-event sentences, which is consistent with our event-based splitting.

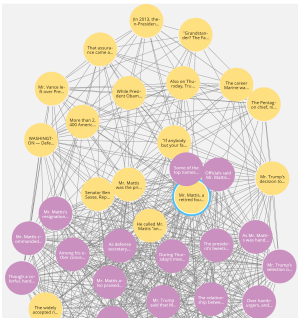
Figure 2(b) and 2(c) present the same subgraph taken directly from the real relational sentence graph in our study. We only present part of edges connected by the sentence *"Mr. Mattis, a retired four-star Marine general, was rebuffed.", NYT* described in Section 1, and the first sentence in Table 1 is effectively linked to it. Moreover, nodes in Figure 2(b) are colored according to news source (HPO/NYT/FOX) and we clearly see that relational sentence graph infuse information from different news media. In comparison of Figure 2(c), we found that most sentences from HPO and FOX (yellow and violet) related to target sentence are biased. Therefore event-context contained in articles of different news outlets effectively helps identify the biased sentences.



(a) Four types of edges in relational sentence graph. Underline: deverbal noun reference; *italic*: discourse maker; **bold**: entity continuation; colored: sentence similarity



(b) nodes are colored by news source



(c) nodes are colored by bias label

Figure 2: Relational sentence graph

Our graph composition is intended to mimic the way humans develop views: people acquire information through immediate context in article and reason by aggregating certain background knowledge from different news reports of the whole event.

2.4 Graph Attention Network

As one of the representative graph convolutional networks, Graph Attention Networks (GATs) introduces an attention mechanism to achieve better neighbor aggregation. By learning the weights of the neighbors, GAT can learn the representation of the target node by implementing a weighted aggregation of the neighbor node representations. However, it may suffer from graph noise introduced by incorrect node linking. In our study, we use Self-supervised Graph Attention Network Kim and Oh [13] which introduces, on top of the GAT, an edge presence prediction task and thus puts an emphasis on more on distinguishing misconnected neighbors.

The graph structure naturally places each sentence within its context, and as a result, different sentences are no longer isolated. The flexibility of the graph structure also allows it to move beyond the ordered arrangement of traditional LSTM. Therefore two sentences can be directly connected by edges, even if they are far apart in the original article or in different articles.

Note that our sentence graph doesn't contain edges between two events, therefore it assures no data leakage while training GAT on the whole graph.

3 EXPERIMENT AND RESULTS

We use BASIL (Bias Annotation Spans on the Informational Level) dataset proposed by Fan et al. [8] for the sentence-level informational bias detection task. We experiment with four baselines including the current state-of-the-art model and four variants of MultiCTX in order to fully demonstrate each module's utility. Our results suggest that MultiCTX greatly outperforms the current SOTA and effectively incorporates the contextual information in sentence-level informational bias detection.

3.1 Data

BASIL dataset provides sentence-by-sentence span-level annotation of informational bias for 300 online English news articles grouped in 100 triplets, each discussing on the same event from three news outlets. The articles are selected in order to make a fair coverage in terms of time and ideology: 1) From 2010 to 2019, 10 events are included each year in the dataset; 2) Fox News (FOX), New York Times (NYT) and Huffington Post (HPO), representative of conservative, neutral and liberal respectively in the US journalism, are chosen as three news sources.

As for the sentence-level informational bias detection task, we use the same data formulation in van den Berg and Markert [17]. In this sentence-wise binary classification task, a sentence is labeled as biased if at least one informational bias span occurs, and seven empty sentences are removed, resulting in a total of 7977 sentences with 1221 annotated bias.

Examples are shown in Table 1.

3.2 Set-up

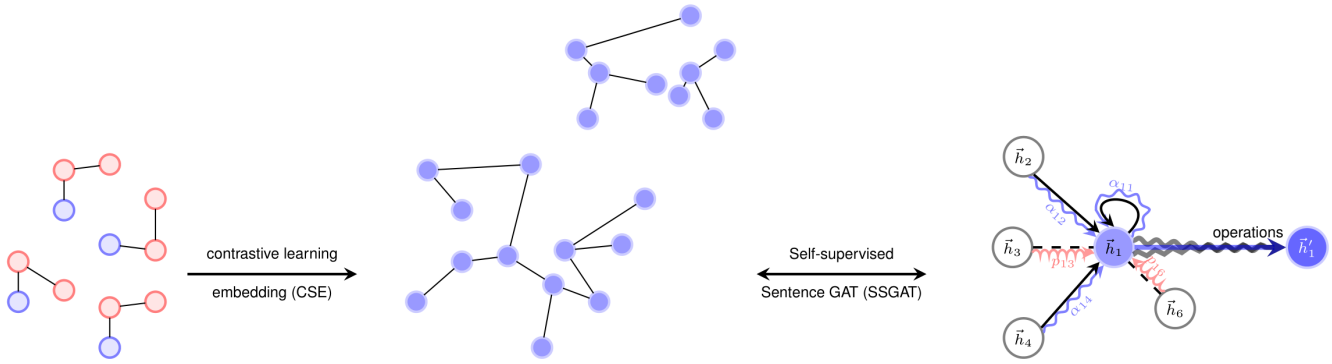
We use the same 10-fold cross-validation event-split in van den Berg and Markert [17] to facilitate the comparison.

Each fold has 80/10/10 non-overlapping events for train/val/test partition, and sentences from the same event never appear simultaneously in two different subsets within one fold. There are on average 6400/780/790 sentences in train/val/test set respectively. We use 5 different seeds for each method and the F1 score, precision and recall ('biased' is positive class) as the evaluation metrics. For each experiment, a mean value and standard deviation across 5 seeds will be reported if applicable.

We use the same hyper-parameters provided in [17] to reimplement BERT, RoBERTa and WinSSC baselines. However, for EvCIM, We cut the training epochs from 150 to 75 and increase the batch size from 32 to 64 due to the considerable time usage. For MultiCTX, We trained use a RoBERTa-based contrastive learning following the implementation in [10]. Due to unavoidable non-deterministic atomic operations in implementation of GAT, the result presented below may cannot be exactly reproduced, but we took an average on our experiments to reflect its range. All models are trained and evaluated on a GeForce GTX 1080 Ti GPU with 11G RAM and Intel(R) Xeon(R) CPU E5-2630 with 128G of RAM. Training details will be described in Appendix.

3.3 Baselines

There are few models in sentence-level informational bias detection. Fan et al. [8] has proposed BASIL dataset and corresponding BERT and RoBERTa benchmarks. Cohan et al. [6] has proposed several



(a) triplets carefully constructed using article-based rules (b) Relational sentence graph: we connect two sentences if they have discourse relationships or they are between semantically similar (c) Self-supervised GAT, where α_{ij} denotes attention between node i and j, p_{ij} denotes the probability of the presence of edge between node i and j

Figure 3: Our Model MultiCTX

models trying to incorporate context in different ways. We will take two of them, WinSSC and their best and also current SOTA model EvCIM, as our baselines. Few other works used BASIL dataset but with objectives other than sentence-level informational detection. Thus we have four baseline models:

- **BERT** [7] and **RoBERTa** [16]: we finetune the individual sentence informational bias detection task on BERT$_{base}$ and RoBERTa$_{base}$.
- **WinSSC** [17]
  WinSSC (windowed Sequential Sentence Classification) is a variant of SSC [6]. We include it as one of the baselines because SSC implements the very natural idea that comes to us when we think of using context: directly inputing sequences of consecutive sentences to BERT. SSC feeds the concatenation of sentences from a chunk of document to pretrained language models (PLMs),and then classifies each sentence using the embedding of the separator tokens `[SEP]` at its end. SSC makes non-overlapping chunks while WinSSC makes chunks by overlapping sentences at both ends, which eretains the contextual information for bookended sentences.
- **EvCIM** : PLM embeddings + BiLSTM
  EvCIM (Event Context-Inclusive Model) proposed by Cohan et al. [6] is the SOTA model on BASIL dataset and it also uses the contextual information. It takes the average of the last four layers of fine-tuned RoBERTa$_{base}$ as the sentence embedding, and then uses BiLSTM to encode each article from the same event as the target sentence. Finally it concatenates three article representations and the target sentence embedding to make the sentence-level prediction. Besides using the hyper-parameters from the original paper, we generate the result from a separate set of reasonable hyper parameters. We present below results both from the original paper and from our experiments.

3.4 Our Models

- **CSE: Contrastive Sentence Embedding**
  Classification by a logistic regression on sentence embeddings directly obtained from contrastive learning.
- **EvCIM w/ CSE**: CSE + BiLSTM
  Similar to EvCIM described in Section 3.3, we utilize BiLSTM-encoded context as well as the target sentence to perform the sentence-wise classification. However, instead of the average of the last four layers of fine-tuned RoBERTa$_{base}$ in EvCIM, we use CSE (Contrastive Sentence Embedding) in our study. Moreover, we also add news source embeddings before the final fully connected classification layer on top of BiLSTM-encoded in-event article embeddings.
  In the original paper [6], adding news source embeddings hurts EvCIM's performance, but because it is useful for EvCIM w/ CSE according to our experiments, we use this version here. This also indicates that CSE has better captured inherent properties of sentences compared to PLM embeddings. CSE can therefore well incorporate extra news media information rather than be disturbed by it.
- **MultiCTX w/o CSE**: PLM embeddings + SSGAT
  We use the original sentence embedding in EvCIM, which is the the average of the last four layers of fine-tuned RoBERTa$_{base}$ to build the relational sentence graph. We then apply Self-supervised GAT on the graph (SSGAT, Self-supervised Sentence GAT). In other words, we replace CSE in MultiCTX with EvCIM's sentence embedding.
- **MultiCTX**: our full model (CSE + SSGAT)
  MultiCTX first performs contrastive learning on carefully composed triplets to obtain CSE. It then builds relational sentence graph according to inter-sentence relationships. Finally, MultiCTX applies Self-supervised GAT above to get the final sentence informational bias prediction.

We can obtain the following observations and conclusions:

1. Encoding sequential sentences brutally by PLM may fail.

Here we use RoBERTa$_{base}$ as pretrained language model in WinSSC. However, it obtains worse result (F1=37.58) than the original RoBERTa$_{base}$ (F1=42.13). The result is similar as in van den Berg

Model*		Explanation	Precision	Recall	F1
baselines	BERT$_{base}$		40.44 ± 1.07**	31.65 ± 1.11	35.49 ± 0.67
	RoBERTa$_{base}$		44.588 ± 0.80	40.02 ± 2.22	42.13 ± 1.02
	WinSSC		41.47 ± 1.31	34.37 ± 0.57	37.58 ± 0.77
	EvCIM (our reproduction)	PLM embed. + BiLSTM	38.40 ± 0.64	48.53 ± 1.45	42.87 ± 0.69
	EvCIM (original paper)		39.72 ± 0.59	49.60 ± 1.20	44.10 ± 0.15
models	CSE		47.53***	40.13	43.51
	EvCIM w/ CSE	CSE+BiLSTM	48.53 ± 0.73	41.98 ± 0.36	45.01 ± 0.26
	MultiCTX w/o CSE	PLM embed.+ SSGAT	46.89 ± 0.71	42.88 ± 0.67	44.79 ± 0.63
	MultiCTX (full)	CSE+SSGAT	47.78 ± 0.94	44.50 ± 0.65	**46.08 ± 0.21****

* All results are implemented or reproduced by ourselves except for the second EvCIM record
** Mean value and standard deviation across 5 seeds are reported if applicable
*** CSE uses a linear regression therefore no randomness occurred, so the result is a deterministic value.
**** The best result on a single run obtained in our experiments is **F1=46.74**

Table 2: Results

and Markert [17]. There are two possible reasons: First, we take sentence chunks instead of individual sentences as input, and doing so may introduce data reduction. Second, BERT-based pretrained language models are not good at processing long text. They simply join neighboring sentences, which may introduce more noise and complexity rather than help integrate the context. Therefore, brute-force concatenation of sequential sentences can rarely make use of the contextual information, and it probably brings in more noise and reduces the data quantity.

2. Contrastive learning helps improve sentence embeddings. The results show that contrasive sentence embeddings (CSE) classified simply with a logistic regression (F1=43.51) beats our reproduction of EvCIM (F1=42.87); moreover, CSE combined with BiLSTM (EvCIM w/ CSE, F1=45.01) outperforms EvCIM even more in comparison with the declared F1=44.10 in its original paper [6].

Note that EvCIM uses the average of the last four layers of fine-tuned RoBERTa$_{base}$ as the sentence embedding. Therefore, our results prove that contrastive learning produces better sentence representations than BERT-based PLMs. this can be achieved via contextual information incorporation.

- BERT-based PLM tends to encode all sentences into a smaller spatial region, which results in a high similarity score for most of the sentence pairs, even for those that are semantically completely unrelated. Specifically, when the sentence embeddings are computed by averaging the word vectors, they are easily dominated by high-frequency words, making it difficult to reflect their original semantics.
- Instead of individual sentences, CSE considers for each target sentence a context built up by all its positive and negative counterparts in related triplets. Among them, negative samples provide an article-level context and positive samples provide an event-level context. With the goal of contrastive learning to "distill essence", it learns from its context and naturally suppresses such shallow high-frequency-words features, thus avoiding similar representations of semantically different sentences.

3. Sentence graph can effectively integrate context.

We can see that MultiCTX w/o CSE, i.e. PLM embed.+SSGAT (F1=44.79) outperforms EvCIM (F1=44.10 in the original paper and F1=42.87 from our reproduction). The two models both use averaged RoBERTa embedding as sentence embeddings. The former uses graph structure (SSGAT) while the latter use BiLSTMs to carry out classification.

The results prove that our sentence graph structure is better in encoding contextual information than sequential models such as BiLSTM. We will examine different levels of context, i.e., adjacent sentences, the article and the the event context in the ablation study in the next section.

4. Contrastive learning together with sentence graph achieves the best performance. Our full model MultiCTX achieves F1=46.08 in the sentence-level informational bias detection task, significantly outperforms the current State-of-the-Art model EvCIM [6] (F1=44.10 declared in original paper). Possible reasons are: 1) BiLSTMs are limited to the event context in EvCIM; 2) MultiCTX uses better sentence representations (CSE); 3) MultiCTX incorporates the context in varying degrees explicitly using graph structure and implicitly via contrastive learning.

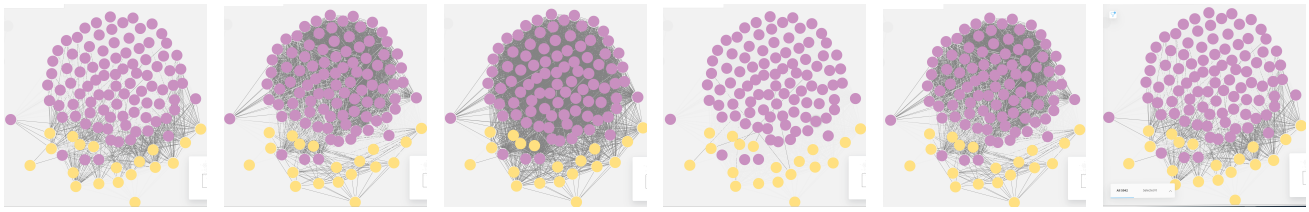
4 ABLATION ANALYSIS

We have proved that both CSE and SSGAT are essential for MultiCTX, and in this section, we will further explore roles of different inter-sentence relationships in our model. We keep CSEs fixed and modify our relational sentence graph by removing certain types of edges, and then report the results to see how each part contributes to MultiCTX in our informational bias detection task.

Edge types described in Section 2.3 can be briefly summarized in two categories: Type 1,2 and 3 are discourse relationships and Type 4 is semantic similarity. Besides, they can also be partitioned by level of context: Type 1 and 2 are neighborhood-level and article-level; Type 3 and 4 are event-level. we will focus on their utility in our ablation study.

Table 4 shows the ablation results. We can conclude that,

- **Context information transfers via edges in graph is better than via cells in BiLSTM.**



(a) Graph with edges of Type 1,2,3 (b) Graph with edges of Type 4 (c) Graph with edges of Type 3,4 (d) Graph with edges of Type 1,2 (e) Graph with edges of Type 1,2,4 (f) Graph with edges of Type 3

Figure 4: Ablation study on Relational sentence graph. Violet nodes are non-biased, Yellow nodes are biased.

Variant	MultiCTX	CSE + BiLSTM	Discourse relationship	Semantic similarity	Event-context	Neighborhood-context	w/o Entity continuation
Edge types	all=[1,2]*,3,4	No graph	[1,2],3	4	3,4	[1,2]	[1,2],4
Precision	47.78 ± 0.94	48.53 ± 0.73	47.43 ± 0.96	47.16 ± 0.27	47.07 ± 0.99	47.18 ± 1.08	47.56 ± 0.62
Recall	44.50 ± 0.65	41.98 ± 0.36	44.39 ± 0.84	43.47 ± 0.38	44.64 ± 0.37	44.01 ± 0.91	43.72 ± 0.76
F1	46.08 ± 0.21	45.01 ± 0.26	45.85 ± 0.35	45.24 ± 0.18	45.81 ± 0.42	45.53 ± 0.29	45.55 ± 0.34

* Type 1 together with Type 2 represents the Neighborhood-level context so we treat them as a whole in the ablation study.

Table 3: Ablation study on different types of edges in SSGAT. Mean and standard deviation across 5 seeds are reported.

All variants with graphs have better F1 score than the model without graph (CSE+BiLSTM, F1=45.01).

Figure 4 shows graphs with different edges in our ablation study using a subgraph of one event. Violet nodes are non-biased sentences, Yellow nodes are biased sentences.

- **Discourse relationship contributes more than the semantic similarity to SSGAT.** SSGAT with only discourse relationships (F1=45.85) still has close performance to the full MultiCTX, while SSGAT with only semantic similarity edges (F1=45.25) suffers a considerable decrease in its performance. Note that the semantic similarity is calculated based on CSEs, so SSGAT with only such edges didn't add much extra information but may introduce duplication. It can be explained by Figure 4(b): connections are mostly within non-based words while inter-communication between biased/non biased nodes are more frequent in Figure 4(a).
- **Event-level context is more important than the neighborhood-level context.** While they are both important according to our results, global event-level context contributes more than local neighborhood-level context. SSGAT with only adjacent sentences (Type 1,2) obtains F1=45.53 and with only Type 3,4 gets F1=45.81. The result is intuitive because edges of type 3,4 not only include adjacent sentences within article, but also extend to the whole event. We can also see the rare presence of edges of Type 1,2 in Figure 4(d) compared with the closely linked graph in Figure 4(c).
- **Adjacent sentences encoded by graph better interprets context information than brute-force PLMs.** In terms of neighborhood-level context, our SSGAT beats WinSSC by increasing massively the F1 score from F1=37.58 (Table 2) to F1=45.53.
- **Entity continuation is the most important edge type.** Among three ablation experiments removing respectively edges of Type 4 (F1=45.85), Type 1,2 (F1=45.81) and Type 3 (F1=45.55), the last one without Type 3 (entity continuation) suffers the largest performance drop. It suggests that entity continuation, or, coreference is the most important relation in our setting. We can clearly see that Type 3 edges are the main reason for inter-class communication from Figure 4(f).

5 SOCIAL IMPACT AND ETHICAL STATEMENT

This work aims to help people distinguish subtle internal informational bias presented in news reports thus prevent the readers from being directed to a specific point of view. In this way we encourage the readers to freely develop own opinions and to think independently and critically.

The news media has a great influence on individuals and the public's perception of the world, but news biases are widespread and the influence of biased news reports has been magnified by social media. Moreover, the recommendation systems make readers tend to focus only on news that is consistent with their established views and beliefs, resulting in the "echo chamber" effect. Over time, people will be trapped in such echo chamber and unable to contact or resist news that contradicts their views, and their internal prejudice will only be strengthened. Therefore, ideally the news reports should uphold the principle of objectiveness and neutrality, present the readers with a complete and impartial picture of the event. However, even the news articles might hardly be completely objective and neutral, we hope that our research could make various conflicting views to be presented fairly by unmasking their internal bias, serving as a reference tool for the readers to avoid being

Variant	MultiCTX	CSE + BiLSTM	Discourse relationship	Semantic similarity	Event-context	Neighborhood-context	w/o Entity continuation
Edge types	all=[1,2]*,3,4	No graph	[1,2],3	4	3,4	[1,2]	[1,2],4
Precision	47.78 ± 0.94	48.53 ± 0.73	47.43 ± 0.96	47.16 ± 0.27	47.07 ± 0.99	47.18 ± 1.08	47.56 ± 0.62
Recall	44.50 ± 0.65	41.98 ± 0.36	44.39 ± 0.84	43.47 ± 0.38	44.64 ± 0.37	44.01 ± 0.91	43.72 ± 0.76
F1	46.08 ± 0.21	45.01 ± 0.26	45.85 ± 0.35	45.24 ± 0.18	45.81 ± 0.42	45.53 ± 0.29	45.55 ± 0.34

* Type 1 together with Type 2 represents the Neighborhood-level context so we treat them as a whole in the ablation study.

Table 4: Ablation study on different types of edges in SSGAT. Mean and standard deviation across 5 seeds are reported.

misguided or getting trapped and to develop their own opinions. In this way we encourage the independent thinking and the critical thinking, maintain the communication of multiple viewpoints, and reduce the echo chamber effect.

The dataset used in this work is publicly available and is used under the data usage agreement. All news in examples are published by formal news agencies and can be found online. This study does not require IRB/ethical approval.

6 RELATED WORK

Media bias Detection. Datasets for media bias detection are limited and not standardized since it requires heavy workloads for humans to annotate manually the subtle bias with certain level of expertise. Besides, human annotators can suffer from implicit media bias and their judgments are subjective as well. However, there are still some sentence-level media bias datasets. **?**] proposed a large sentence-level corpus for political bias detection, **?**] proposed a sentence-level media bias dataset of 996 sentences from 46 news articles covering 4 topics, Huang et al. [11] further built a dataset consisting of more than 2000 sentences annoted with 43000 bias including subjectivity, hidden assumptions and representation tenddencies, **?**] also created a sentence-level media bias dataset with the emphasis on annotators' backgrounds, Fan et al. [8] presented the BASIL dataset used in our study.

Linguistic features-based techniques are initially utilized in media bias detection and they are still widely applied till today because they provide a strong descriptive and explanatory power. These techniques are systematically formulated in [**?**]. [**?**] explored linguistic patterns at word-level and article-level to analyze political bias in news articles; **?**] engineered various linguistic, lexical and syntactic features to detect media bias; Huang et al. [11] made used of syntactic structure to obtain generalized text embedding for news credibility check; and **?**] used a multi-task ordinal regression framework.

With the rise of deep learning, neural-based aproaches are broadly used in media bias detection. Iyyer et al. [12] used RNNs to aggregate the polarity of each word to predict political ideology on sentence-level. Gangula et al. [9] made use of headline attention to classify article bias. Li and Goldwasser [15] captured social information by Graph Convolutional Network to identify political bias in news articles. Fan et al. [8] used BERT and RoBERTa and van den Berg and Markert [17] used BiLSTMs as well as BERT-based models to detect sentence-level informational bias.

Contextual information in media bias detection. Contextual information is explored, though primarily, in media bias detection. Baly et al. [3] employed an adversarial news media adaptation using triplet loss; Kulkarni et al. [14] proposed an attention based model to capture views from news articles' title, content and link structure; Chen et al. [4] explored the impact of sentence-level bias to article-level bias; Li and Goldwasser [15] encoded social information using GCN; Baly et al. [2] made use of news media's cyber-features in news factuality prediction; Huang et al. [11] explored cross-media context by a news article graph.

Sentence-level informational bias is under-studied by only a few research and the methods described above are not applicable on this task. In order to infuse contextual information, we refer to extractive summarization Zhao et al. [19] and Christensen et al. [5] which used sentence graph to encode context.

7 CONCLUSION

Our work focus on incorporating different levels of context: neighborhood-level, article-level and event-level in sentence-level informational bias detection. We proposed MultiCTX, a model composed of contrastive learning and relational sentence graph attention network to encode such multi-level context at different stages.

Our model (F1=46.08) significantly outperforms the current state-of-the-art model (F1=44.10) by 20 percentage points. Therefore, we conclude that our model successfully learns from contextual information and that multi-level contextual information can effectively improves the identification of sentence-level informational bias.

REFERENCES

[1] 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)*. Association for Computational Linguistics, Online, 4982–4991. https://doi.org/10.18653/v1/2020.emnlp-main.404

[2] Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018. Predicting Factuality of Reporting and Bias of News Media Sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 3528–3539. https://doi.org/10.18653/v1/D18-1389

[3] Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. 2020. We Can Detect Your Bias: Predicting the Political Ideology of News Articles. arXiv:2010.05338 [cs.CL]

[4] 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*. Association for Computational Linguistics, Online, 4290–4300. https://doi.org/10.18653/v1/2020.findings-emnlp.383

[5] Janara Christensen, Mausam, Stephen Soderland, and Oren Etzioni. 2013. Towards Coherent Multi-Document Summarization. In *Proceedings of the 2013 Conference*

of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Atlanta, Georgia, 1163–1173. https://aclanthology.org/N13-1136

[6] Arman Cohan, Iz Beltagy, Daniel King, Bhavana Dalvi, and Dan Weld. 2019. Pretrained Language Models for Sequential Sentence Classification. 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)*. Association for Computational Linguistics, Hong Kong, China, 3693–3699. https://doi.org/10.18653/v1/D19-1383

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina 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)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[8] Lisa Fan, Marshall White, Eva Sharma, Ruisi Su, Prafulla Kumar Choubey, Ruihong Huang, and Lu 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)*. Association for Computational Linguistics, Hong Kong, China, 6343–6349. https://doi.org/10.18653/v1/D19-1664

[9] 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*. Association for Computational Linguistics, Florence, Italy, 77–84. https://doi.org/10.18653/v1/W19-4809

[10] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. arXiv:2104.08821 [cs.CL]

[11] Yen-Hao Huang, Ting-Wei Liu, Ssu-Rui Lee, Fernando Henrique Calderon Alvarado, and Yi-Shin Chen. 2020. Conquering Cross-Source Failure for News Credibility: Learning Generalizable Representations beyond Content Embedding. In *Proceedings of The Web Conference 2020* (Taipei, Taiwan) *(WWW '20)*. Association for Computing Machinery, New York, NY, USA, 774–784. https://doi.org/10.1145/3366423.3380158

[12] 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)*. Association for Computational Linguistics, Baltimore, Maryland, 1113–1122. https://doi.org/10.3115/v1/P14-1105

[13] Dongkwan Kim and Alice Oh. 2021. How to Find Your Friendly Neighborhood: Graph Attention Design with Self-Supervision. In *International Conference on Learning Representations*. https://openreview.net/forum?id=Wi5KUNlqWty

[14] Vivek Kulkarni, Junting Ye, Steve Skiena, and William Yang Wang. 2018. Multi-view Models for Political Ideology Detection of News Articles. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 3518–3527. https://doi.org/10.18653/v1/D18-1388

[15] Chang Li and Dan Goldwasser. 2019. Encoding Social Information with Graph Convolutional Networks forPolitical Perspective Detection in News Media. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 2594–2604. https://doi.org/10.18653/v1/P19-1247

[16] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692 [cs.CL]

[17] Esther van den Berg and Katja Markert. 2020. Context in Informational Bias Detection. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Barcelona, Spain (Online), 6315–6326. https://doi.org/10.18653/v1/2020.coling-main.556

[18] Jinming Zhao, Ming Liu, Longxiang Gao, Yuan Jin, Lan Du, He Zhao, He Zhang, and Gholamreza Haffari. 2020. *SummPip: Unsupervised Multi-Document Summarization with Sentence Graph Compression*. Association for Computing Machinery, New York, NY, USA, 1949–1952. https://doi.org/10.1145/3397271.3401327

[19] Jinming Zhao, Ming Liu, Longxiang Gao, Yuan Jin, Lan Du, He Zhao, He Zhang, and Gholamreza Haffari. 2020. *SummPip: Unsupervised Multi-Document Summarization with Sentence Graph Compression*. Association for Computing Machinery, New York, NY, USA, 1949–1952. https://doi.org/10.1145/3397271.3401327