Context in Informational Bias Detection

Esther van den Berg
Leibniz ScienceCampus
Heidelberg/Mannheim, Germany
berg@ids-mannheim.de

Katja Markert
Institute of Computational Linguistics
Heidelberg University, Germany
markert@cl.uni-heidelberg.de

Abstract

Informational bias is bias through sentences or clauses that convey tangential, speculative or background information that can sway readers' opinions towards entities. By nature, informational bias is context-dependent, but previous work on informational bias detection has not explored the role of context beyond the sentence. In this paper, we explore four kinds of context for informational bias in English news articles: neighboring sentences, the full article, articles on the same event, and articles from the same domain. We find that integrating article and event context improves classification performance over a very strong baseline. In addition, we perform the first error analysis of models on this task. We find that the best-performing context-inclusive model outperforms the baseline on longer sentences and sentences from politically more centrist articles.

1 Introduction

Informational bias is bias through sentences or clauses that convey tangential, speculative or background information that can sway readers' opinions towards entities (Fan et al., 2019). A natural place to look for informational bias is in news texts, where journalists use background information to place newsworthy events in a broader context. Examples of informational bias include: quotations of opinions from third parties about the target entity, allusions to what may have motivated the target entity to act as they did, and mentions of previous statements and actions of the same entity.

What separates informational bias from other kinds of bias is that it can be expressed in a completely factual and neutral way. While some instance of bias are recognisable outside of their context as introducing biased (e.g. quotations from third parties that contain opinions), others are mere statements of facts that do not raise suspicions of bias outside of their context. Consider example instance 1.3 in Table 1. This sentence contains no subjective language. Seen on its own, it is simply stating a fact. However, human annotators judged that it is a case informational bias (Fan et al., 2019). This is because this particular fact reflects positively on the target entity Mike Huckabee in the context of an announcement that he is running for president. Note that one can also imagine contexts where the implication is negative, for example the context of an article discussed the disconnect between olderl stape Republican candidates and a new generation of more progressive voters. The fact that instances of informational bias can be superficially neutral and context-dependent makes informational bias detection an exceptionally challenging task, which furthermore has a short research history and few available relevant resources.

Where previous work has performed informational bias detection on a token-level and on a sentence-level, we are the first to attempt to involve context beyond the sentence. We integrate four kinds of context: neighboring sentences (**direct textual context**), the full article (**article context**), articles on the same event (**event context**) and articles from the same domain (**domain context**).[1] Due to limited availability of data, existing methods of integrating context struggle to improve performance over a strong sentence-level-only baseline. However, our proposed models for leveraging article context (**ArtCIM**) and event context (**EvCIM**) do [SIGNIFICANTLY] improve over the baseline. Specifically, we find that

[1]Code and results available at: github.com/vdenberg/context-in-informational-bias-detection

Idx	Sentence	Inf	Src	ID
1.1	Former Arkansas Gov. Mike Huckabee announced Tuesday he is running for president [. . .].	0	FOX	53fox00
1.2	Mr. Huckabee opposes same-sex marriage, suggesting as recently as February that homosexuality is a lifestyle choice akin to drinking or swearing.	1	NYT	53nyt10
1.3	He was the longest-serving Arkansas governor, from 1996 to 2007.	1	HPO	53hpo15
2.1	"Trump says he wants to run the nation like he's run his business," Mr. Bloomberg said on Wednesday.	0	NYT	82nyt01
2.2	Bloomberg contrasted his business history with Trump's, saying "I've built a business, and I didn't start it with a million-dollar check from my father."	1	FOX	82fox05
2.3	"But Trump's business plan is a disaster in the making."	1	HPO	82hpo21
3.1	The states of Nebraska and Oklahoma filed a federal lawsuit in the U.S. Supreme Court Thursday [. . .]	0	HPO	21hpo00
3.2	"While Colorado reaps millions from the sale of pot, Nebraska taxpayers have to bear the cost."	1	FOX	21fox09
3.3	[Sheriff Adam Hayward of Deuel County, Neb.] has complained that marijuana arrests have strained his jail budget.	1	NYT	21nyt09

Table 1: Example instances from three stories in the BASIL corpus with their informational bias label (inf), news source (src) and BASIL ID.

the best-performing context-inclusive model (EvCIM) does better at handling long sentences, sentences from the politically more centrist news publisher, and sentences without subjective language.

1. The first systematic study of the impact of including context in informational bias detection.
2. The best-performing informational bias detection system for English thus far, tested on the BASIL corpus (Fan et al., 2019).
3. The first error analysis of informational bias detection systems, which identifies strengths and weaknesses of models with and without awareness of context.

2 Related work

Framing. Informational bias can be considered a type of framing with a focus on entities. The framing of entities has been studied for the construction of the English BASIL corpus of lexical and informational bias (Fan et al., 2019), which is the corpus the models proposed in this paper will be tested on. Fan et al. (2019) emphasize that, unlike more commonly studied kinds of bias, informational bias label assignments depend very heavily on context. The sentence-level and span-level BASIL annotations were thus provided by human annotators who saw sentences in their article context. The proposed computational models in this work only treat sentences in isolation. Framing of entities is also studied by Card et al. (2016), who examined how English-speaking news frames events through casts of characters, and van den Berg et al. (2020), who studied the effect of naming and titling on the perception of entities in English and German.

Most framing research focuses not on entities but on the framing of topics and events. The study of topic framing in news has a long history in social science (Entman, 1993; Berinsky and Kinder, 2006; Baumgartner et al., 2008; Gentzkow and Shapiro, 2010) and is beginning to attract attention from the natural language processing community (Tsur et al., 2015; Fulgoni et al., 2016; Field et al., 2018; Baumer et al., 2015). For topic framing research there exists the Media Frames Corpus of news annotated for the framing of same-sex marriage, smoking, and immigration (Card et al., 2015) and the Gun Violence Frame

Corpus (Liu et al., 2019a) annotated for framing in news on gun violence. Computational analysis and classification experiments have been done on framing in Russian news (Field et al., 2018), on detecting frames in English headlines (Liu et al., 2019a; Chen et al., 2018), and on detecting frames in a multi-label, multi-lingual setting (Akyürek et al., 2020).

Subjectivity. Related work also includes work on implicit sentiment through syntactic structures (Greene and Resnik, 2009) and partisan phrases (Yano et al., 2010), work on explicit stance and subjective language (Recasens et al., 2013; Pang et al., 2008; Wiebe et al., 2004; Hube and Fetahu, 2019), and work on the classification of documents or news outlets into leanings or ideologies (Iyyer et al., 2014). The difference between these various kinds of subjective language and informational bias lies in its in its exclusion of neutral and objective language. In framing research, any text that could lead an impartial third party to recognise a non-neutral viewpoint towards a topic or entity can be said to contain framing, even if it is objective and neutral intone.

Approaches. Sentence-level classification tasks have seen great increases in performance through the use of pre-trained language models (PLMs) such as BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019b). By further pre-training RoBERTa on domain-specific and task-specific datasets, Gururangan et al. (2020) made it possible to perform sentence-level classification using models that have been exposed to domain context. Several methods have been developed to allow PLM models to take larger sequences than sentences as their input (Pappagari et al., 2019; Adhikari et al., 2019). Of these, only one specifically performs sequential sentence classification (i.e. the task of providing labels for each of the sentences in the multi-sentence input) (Cohan et al., 2019). There exist non-PLM approaches to sequential sentence classification as well. These consist of hierarchical sequence encoders with a final CRF layer (Dernoncourt and Lee, 2017; Jin and Szolovits, 2018) and a bi-LSTM-based approach that contextualises Universal Sentence Encodings and also integrates information that is specific to the domain of movie plot synopses (Papalampidi et al., 2019). None of these techniques have previously been applied to informational bias detection.

3 Method

We experiment with different kinds of context to assess which one or which ones are helpful for informational bias detection. We define four types of context: direct textual context, article context, event context and domain context.

Direct textual context. Direct textual context consists of the directly neighbouring sentences around the target sentence. It may be helpful for disambiguating sentences with multiple possible interpretations, noticing patterns in the type of content preceding and following instances of informational bias, and noticing when a target sentence is part of a multi-sentence quote.

Article context. Article context consists of the full news article that the target sentence appears in. The article may be helpful for e.g. establishing the topic the target sentence relates to, the type of article the target sentence is from, and determining whether the target sentence is an outlier compared to the rest of the article.

Event context. Event context consists of all news articles that cover the same newsworthy event or topic as the article the target sentence appears in. The possible benefit of access to Event context is that it can help to notice when an article takes a unique angle to a topic or mentions information that is absent from other articles.

Domain context. Domain context consists of the domain of news articles that the target sentence is a part of. Domain refers to a group of text with shared lexical and structural properties, including topic, register and platform of publication. Following Gururangan et al. (2020), we consider the population of articles from which a corpus has been sampled a domain in its own right. Possible benefits of domain awareness when detecting informational bias include the ability to notice typical journalistic strategies for framing entities without attracting accusations of bias, the ability to distinguish between news outlets differing styles and ideologies, and increased awareness of domain-specific word connotations (e.g. “leading in the polls” or “declined to comment”).

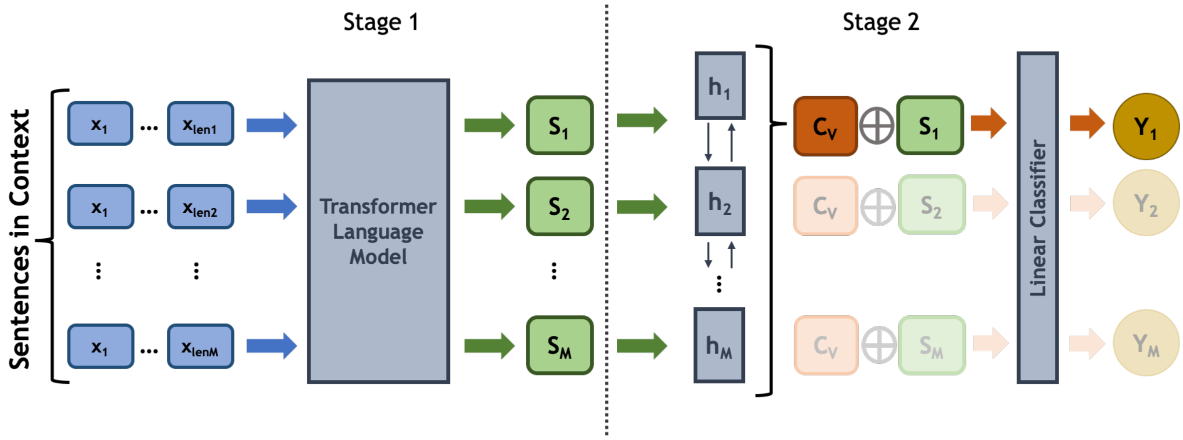

Figure 1: Context-Inclusive Model: Stage 1) Sentence embeddings are obtained by encoding sequences of words using a pre-trained language model. Stage 2) Sequences of sentences from a pre-determined context (article or coverage) are encoded by BiLSTMs (one BiLSTM per document, only one shown in diagram). The resulting context representation is concatenated and classified together with the target sentence representation to obtain a sentence-level prediction.

3.1 Data & Task

We use the currently only existing dataset with annotations of informational bias: the aforementioned BASIL dataset. The corpus contains 100 triples of news articles by Fox News (FOX), the New York Times (NYT) and the Huffington Post (HPO), each covering the same event. The dataset consists of 7,984 sentences of which 1,221 contain informational bias. During pre-processing, we remove 7 empty instances from the corpus with a sentence length of zero, leaving 7,977 instances. It provides span-level annotations which can be used for token classification or for binary sentence classification by converting them. To convert them, sentences are labeled as biased if it contains one or more spans with bias. The context-inclusion experiments in this work are only performed on sentence classification, since several of the proposed models are more suited for this task. For completeness, we also provide baseline results for token classification.

3.2 Approaches

Baseline. We compare our context-inclusive models to **BERT** (Devlin et al., 2018) and **RoBERTa** (Liu et al., 2019b) for sentence classification (binary in this case). These are both powerful transformer language models pre-trained on large amounts of data and proven to be effective when fine-tuned with low-resource task datasets. They take as input single sentences and output labels. They thus do not consider the context in which sentences appear in, but are optimised to make excellent use of any cues contained within the sentence.

Direct Textual Context. To involve direct textual context, we use a Windowed Sequential Sentence Classification method (**WinSSC**). Like Cohan et al. (2019)'s method of using pre-trained language models for sequential sentence classification, WinSSC takes multiple sentences as its input sequence, generates embeddings for the separator tokens in the sequence, and classifies these embeddings with a linear layer that outputs as many labels as there are sentences in the input sequence. Prior to embedding, the sequences are book-ended with the last sentence from the previous sequence, and the first sentence of the next sequence. These book-ends, which are ignored during evaluation, ensure that each sentence in the sequence has context at both ends, thus mitigating loss of information along the edges of sequences. This is important when segmenting news articles, as they tend to be long enough to require segmentation into several sections to avoid memory problems. We experiment with sections of 5 and 10 sentences to assess the effect of changing section sizes, and we compare our WinSSC method to the non-windowed SSC method from Cohan et al. (2019).

Article Context & Event Context. We integrate article context and coverage context by means of an

Task	Split	Model	Precision	Recall	F1-score
Token	Sentence	Fan et al. (2019)	25.56	14.78	18.71
		BERT	12.42 ± 1.31	28.31 ± 3.18	17.23 ± 1.61
		RoBERTa	36.1 ± 4.51	32.41 ± 2.92	34.03 ± 2.81
	Story	BERT	12.85 ± 1.06	22.12 ± 1.60	14.60 ± 0.91
		RoBERTa	32.44 ± 2.04	27.73 ± 1.54	29.86 ± 1.25
Sentence	Sentence	Fan et al. (2019)	43.87	42.91	43.27
		BERT	46.44 ± 2.51	33.0 ± 7.21	38.26 ± 5.29
		RoBERTa	47.55 ± 2.92	52.67 ± 6.41	49.89 ± 4.06
	Story	BERT	38.96 ± 5.55	35.79 ± 2.15	37.00 ± 2.11
		RoBERTa	45.42 ± 1.83	39.98 ± 1.51	42.48 ± 0.48

Table 2: Baseline performance of token and sentence classification of informational bias in isolated sentences divided across training and non-training sets by sentence or by story. Standard deviations are across 5 seeds for our models. Fan et al. (2019) report a minimum standard deviation of 3.36 and maximum of 12.44 for theirs.

Model	Prec.	Rec.	F1 with Std.
RoBERTa	46.26	40.11	42.14 ± 0.58
SSC-5	41.48	36.69	38.26 ± 1.09
SSC-10	43.58	34.18	37.50 ± 1.16
WinSSC-5	42.35	36.93	38.73 ± 0.71
WinSSC-10	43.03	34.53	37.18 ± 0.89

Table 3: Results of integrating direct textual context with a Sequential Sentence Classifier without (SSC (Cohan et al., 2019)) or with a window (WinSSC) and a maximum sequence length of 5 or 10.

Article Context-Inclusive Model (**ArtCIM**) and Event Context-Inclusive Model (**EvCIM**). Inspired by the Context Aware Model in Papalampidi et al. (2019), the Context-Inclusive Model uses a Bidirectional Long Short-Term Memory (BiLSTM; Hochreiter and Schmidhuber 1997) to encode news documents. In the case of ArtCIM, a single BiLSTM encodes the article. In the case of EvCIM, three BiLSTMs encode each document in the triple of Fox News, New York Times and Huffington Post articles on the same event. Sentence representations for the target sentence as well as for the input to the BiLSTMs are obtained by taking the average of the last four layers of fine-tuned base RoBERTa (Figure 1, Stage 1). We found this to be more effective than other kinds of pooling, and also more effective than sentence USE embeddings (Cer et al., 2018) or Sentence-Bert embeddings (Reimers and Gurevych, 2019). At the final stage, the encoding of the target sentence and the BiLSTM encodings of the context documents are concatenated and passed to a linear classifier (Figure 1, Stage 2). As such, classification is based both on the content of the target sentence, which the baseline capture very well, and the article or coverage context, which the baseline has no access to.

Domain Context. To integrate domain context, we apply domain-adapted of RoBERTa from Gururangan et al. (2020) that has been trained on news data (**DAPT**), a task-adapted version of RoBERTa that has been trained on the BASIL data (**TAPT**), or both (**DAPT-TAPT**).[2] Additionally, we experiment with including domain context by concatenating an embedding representing the source of an article (Fox News, New York Times or Huffington Post) as a feature in the CIM setting (at Stage 2 in Figure 1) (**ArtCIM*** and **EvCIM***)[3].

4 Results

4.1 Set-up

Previous work on the BASIL corpus has split data by dividing sentences across a training, development and test set (Fan et al., 2019). This type of split isolates target sentences from other sentences in the same article, and from other articles covering the same event. Distributing sentences across set types in this way is contrary to the goal of this work, which is to consider sentences within their context. In addition, distributing sentences from the same article across training and test data can be considered a type of leakage, as knowing of some sentences in an article that they are biased might help recognise similar sentences from the same article or another article on the same topic. Our setting, in which triples of articles are either in training or in test but not both, resembles a more realistic setting, where the hypothetical user of an information bias annotation system wants to identify bias in new articles on new events.

We use a split with sentences distributed across train and non-train sections - the **Sentence split** - only to report baseline results for the purpose of consistency and comparability to Fan et al. (2019). This split consists of 7,123 training instances, 408 development instances and 404 test instances. To test context-inclusive methods, we use a 10-fold cross-validation setting where *stories* (triples of articles) never appear in both a train and non-train section. Sizes of folds in this **Story split** vary slightly because of variation in the length of articles. Each consists of around 6,400 sentences designated for training, 780 for development and 790 for testing. All methods are tested 5 times with a different random seed. Further training details are provided in the appendix (Appendix A).

4.2 Baseline

To establish a baseline that classifies sentences in isolation we fine-tune transformer language models BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019b) on the BASIL corpus. For comparison with Fan et al. (2019), we fine-tune both to perform token classification and sentence classification and we compare the Sentence split to the Story split.

In line with the prediction that the Sentence split introduces leakage from test into training data, performance is consistently several f1-score points higher on the Sentence split than on the Story split (Table 2). The difference is largest for sentence classification with RoBERTa, where f1=49.89 on the Sentence split and 42.48 on the Story split.

In line with observations in Fan et al. (2019), we also observe that performance is lower for token classification than for sentence classification (f1=29.86 versus f1=42.48 (RoBERTa, Story split)).

We observe large improvements in performance of RoBERTa over BERT in all settings. Best performance on sentence classification was reported to be f1=43.27 on a Sentence split in Fan et al. (2019) by BERT. In our set-up using our seeds, BERT performance stands at f1=38.26 on the Sentence split, whereas RoBERTa's sentence classification performance on the Sentence split is 49.89. On the Story split the difference is also large: from f1=37 by BERT to f1=42.48 by RoBERTa.

4.3 Direct Textual Context

We experiment with integrating direct textual context by comparing two methods of sequential sentence classification (SSC and the novel WinSSC) to the best performing baseline sentence classifier. We find performance decreases when direct context is introduced in this manner. Increasing the length of the sequence from 5 to 10 harms performance of both the non-windowed SSC and WinSSC model (f1=38.19 to f1=38.22 and f1=38.67 to f1=37.44). It is likely that data sparsity is at fault here. When performing 10-fold cross-validation with the maximum sequence length set to 5, the number of sequences for training each iteration averages around 1654. With the maximum sequence length set to 10, this drops further to 856 sequences. This is likely too small a number of sequences for the models to generalize.

[2] The BASIL-adapted RoBERTa will be made available upon publication

[3] We experimented with combining AvCIM and EvCIM with domain-adapted and task-adapted RoBERTa, but combining the two lowered performance.

Model	Prec.	Rec.	F1 with Std.
RoBERTa	45.42 ± 1.83	39.98 ± 1.51	42.48 ± 0.48
ArtCIM			43.35 ± 0.38
ArtCIM*			42.76 ± 1.06
EvCIM	39.72 ± 0.59	49.6 ± 1.20	44.10 ± 0.15 †
EvCIM*	39.76 ± 1.50	46.88 ± 2.42	42.96 ± 0.34

Table 4: Results of article and event context with a Context-Inclusive Model with (*) or without news source as an added feature. A dagger indicates a significant improvement over the baseline.

Model	Prec.	Rec.	F1 with Std.
RoBERTa	45.42 ± 1.83	39.98 ± 1.51	42.48 ± 0.48
DAPT	46.87 ± 1.32	36.45 ± 1.27	40.97 ± 0.32
TAPT	46.49 ± 1.74	40.28 ± 1.91	43.12 ± 1.07
DAPT-TAPT	46.69 ± 1.50	37.41 ± 2.37	41.47 ± 1.00

Table 5: Results of domain-adapted, task-adapted and domain-and-task-adapted RoBERTa.

4.4 Article Context & Event Context

We use the Context-Inclusive Model to perform classification based on separate encodings of the target sentence and either just the news article the target sentence appears in (ArtCIM) or each member of the triple of coverage on the same event (EvCIM). As an additional coverage context experiment, we provide ArtCIM and EvCIM with a representation of the newspaper source (ArtCIM* and EvCIM*). Each of these models performs at least as well as the baseline. ArtCIM and EvCIM both obtain scores outside the range of variation of the baseline (Table 4). While RoBERTa achieves best precision, both CIM models achieve much higher recall and EvCIM achieves a significantly higher f1-score.

4.5 Domain Context

We test the performance of three adaptations of RoBERTa: domain-adapted to news (DAPT), task-adapted for informational bias detection on the BASIL corpus (TAPT) and domain-and-task-adapted (DAPT-TAPT). The DAPT and DAPT-TAPT model do not outperform the baseline, but the task-adapted model does (Table 5). These results echo the finding in Gururangan et al. (2020) that domain-adaptation and domain-and-task-adaptation are not as helpful in the news domain as in other domains. RoBERTa has likely seen a sufficient amount of news domain training data already (Liu et al., 2019b), making any further domain-specific pre-training marginally helpful.

5 Error Analysis and Discussion

To investigate whether the best-performing context-inclusive model (EvCIM) improves over the baseline by, in fact, leveraging context, we analyze dependence of performance on factors that we suspect influence the need for context when detecting informational bias. The factors that we consider are sentence length, the presence of quotes, the political leaning of the source article and the presence of subjective language.

Concretely, we expect that less context is needed and therefore *fewer* gains of EvCIM over the baseline can be expected for sentences with the following characteristics:

1. They are short, and more likely to contain a simple message that is easier to classify (e.g. 2.3 in Table 1).

2. They are a quote or contain quotes (e.g. 2.1 and 2.3 in Table 1). Quotes have been shown to introduce bias in news (Niculae et al., 2015) and informational bias in particular (Fan et al., 2019), as they maintain an air of neutrality on the part of the journalist.

Q	N	Bias	RoBERTa	EvCIM
0-90	2018	13.58%	44.87	44.50
91-137	2008	15.94%	40.76	43.57
138-192	1965	17.66%	42.81	45.30
193-647	1986	14.10%	41.65	42.79
All	7977	15.31%	42.48	44.10

Table 6: Performance (F1) by sentence length (quartiles) of RoBERTa and CIM with event context.

In Quote	N	Bias	RoBERTa	EvCIM
Yes	1291	19.67%	47.38	49.78
No	6686	14.46%	41.08	42.54
All	7977	15.31%	42.48	44.10

Table 7: Recall of bias inside or outside of a quote of RoBERTa and EvCIM (biased instances only).

3. They are from an article with a non-centrist political leaning (e.g. 2.3 and 3.2 in Table 1).

4. They contain subjective language (e.g. 3.3 in Table 1).

Consequently, we predict *larger* gains in performance by EvCIM on long sentences, sentences that are not quotes, sentences from centrist articles and sentences without subjective language.

5.1 Sentence Length

We examine whether EvCIM outperforms the baseline on longer sentences by partitioning data into bins corresponding to quartiles of sentence length. We then compare the f1-score computed on predictions by RoBERTa and EvCIM. We observe that, as suspected, short sentences are relatively easy to classify for the baseline (f1=44.87) and EvCIM does not outperform it (f1=44.50). On longer sentences, baseline performance drops and EvCIM makes gains, most notably on the third quartile (from f1=42.81 to f1=45.30) (Table 6).

5.2 Quotes

Quoting patterns have been shown to introduce bias in news (Niculae et al., 2015) and informational bias in particular (Fan et al., 2019; Niculae et al., 2015), by introducing opinions through a third party proxy. We predict that neural approaches notice this relationship and rely on it to some degree to make their predictions.

The BASIL corpus contains annotations that specify for each instance of informational bias whether it is part of a quote or not. We can therefore analyze differences in recall of informational bias inside and outside of quotes (Table 7). Table 7 shows that both models have considerably better recall of informational bias in quotes. We predicted that the baseline would have an easier time with quotes, and that the gains of EvCIM with respect to the baseline would be higher on non-quotes. The findings do not confirm this: EvCIM does not outperform the baseline by a higher margin on non-quotes than on quotes.

5.3 Political Leaning of the Source

To examine whether bias is easier to detect in newspapers and articles with a more pronounced political leaning, we first compare performance on the different news publishers represented in the BASIL corpus. According to Budak et al. (2016), FOX is strongly right-leaning, NYT slightly left-leaning, and HPO strongly left-leaning, meaning NYT is the most centrist of the three. We also compare performance on subsets of data labeled as right-leaning, centrist or left-leaning according to human annotators who read the articles.

We observe that both models perform much better on Fox News articles compared to the other publishers (Table 8). Fan et al. (2019) have stated that there are differences in the polarity and target of biased

Src	N	%Bias	RoBERTa	EvCIM
FOX	2633	15.65%	45.85	47.31
NYT	3048	14.93%	40.57	43.01
HPO	2296	15.42%	40.90	41.82
All	7977	15.31%	42.48	44.10

Table 8: Performance (F1) by publisher of RoBERTa and EvCIM.

Lean	N	%Bias	RoBERTa	EvCIM
Right	2010	15.82%	43.46	43.92
Center	3660	14.07%	42.55	45.40
Left	2307	16.82%	41.61	42.58
All	7977	15.31%	42.48	44.10

Table 9: Performance (F1) by article leaning (left, center, right) of RoBERTa and EvCIM.

Lex	N	Bias	RoBERTa	EvCIM
Yes	448	9.82%	27.48	25.30
No	7529	15.63%	43.77	45.79
All	7977	15.31%	42.48	44.10

Table 10: Performance (F1) on items with and without lexical bias of RoBERTa and CIM with event context.

Subj	N	Bias	RoBERTa	EvCIM
Yes	2415	18.92%	45.80	47.17
No	5562	13.74%	40.31	42.08
All	7977	15.31%	42.48	44.10

Table 11: Performance (F1) on items with and without subjectivity clues of RoBERTa and CIM with event context.

sentences in the three news sources included in the BASIL corpus. The RoBERTa and EvCIM systems may be capitalizing on these and other differences to make better predictions for Fox News articles.

We predict that more gains are made by EvCIM over the baseline on centrist articles. This suspicion is confirmed. When looking at news publishers, EvCIM outperforms the baseline more on the centrist news source the New York Times (from f1=40.57 to f1=43.01) than on the right-leaning (f1=45.85 to f1=47.31) and left-leaning source (from f1=40.90 to f1=41.81). When comparing the source of sentences to the human-annotated label for political leaning assigned to articles, baseline performance follows the same trend as for source difference: right-leaning articles are classified better than centrist and left-leaning articles. EvCIM performs best on centrist articles, and its performance exceeds baseline performance for centrist articles (from f1=42.55 to f1=45.40) than for right-leaning (from f1=43.46 to f1=43.92) and left-leaning ones (from f1=41.61 to f1=42.58), confirming that EvCIM's gains over the baseline are higher on sentences from more neutral articles.

5.4 Subjective language

We suspect that EvCIM will make larger gains on sentences without sentiment-bearing language. The BASIL corpus contains annotations of lexical bias, i.e. bias through word choice, that can be used to investigate the impact on performance of the presence of such language. According to the BASIL annotation protocol, sentences contain lexical bias if the annotator found their opinion to be swayed by the choice of words. According to the annotators, this was the case in only 448 sentences, and informational bias was less likely to occur in sentences with lexical bias (9.82%) than sentences without lexical bias (15.63%) (Table 10). For comparison, we also compute the number of sentences that contain at least one strongly subjective subjectivity clue from the MPQA Subjectivity Lexicon (Wilson et al., 2005). We find that this number is higher: 2415 instances, and informational bias was *more* likely to occur in sentences with subjectivity (18.92%) than sentences without subjectivity (13.74%) (Table 11). We suspect that the latter numbers are a more realistic assessment of the amount of subjective language in the BASIL corpus. When comparing performance on instance with and without subjective language using the subjectivity lexicon, we find that both models perform better on items with subjective language and that the EvCIM gain over the baseline is higher for instances without subjective language (from 40.31 to 42.08) than with subjective language (from 45.80 to 47.17).

6 Conclusion

We explore the impact of including four kinds of context on informational bias detection. We integrate direct textual context, article context, context from other articles on the same event (coverage context), and domain context into sentence classification methods and test performance on the BASIL corpus of informational bias. We find that direct textual context, article context and domain context are difficult to integrate in a way that boosts performance beyond the strong RoBERTa baseline. Our proposed Context-Inclusive Model, however, outperforms RoBERTa in two settings, most noticeably when using coverage context.

Error analysis of RoBERTa and CIM with event context (EvCIM) shows that EvCim performs better than the baseline on longer sentences as well as on sentences from politically centrist news sources (New York Times) and centrist articles. Furthermore, both models perform better on BASIL's Fox News instances than its New York Times or Huffington Post instances. Additionally, both models are more likely to predict that bias is present if the target sentences contains quotation marks.

Future work could explore domain-adaptation to unlabeled data from the same population of articles that the BASIL corpus was drawn from. Given the differences in performance on Fox News Articles compared to other sources, domain-adaptation to specific sources is also a promising avenue. In addition, future datasets may need to ensure a balance of sources that represent different layers and sections of society. Future work could also extend context-inclusion experiments to token classification.

References

Ashutosh Adhikari, Achyudh Ram, Raphael Tang, and Jimmy Lin. 2019. Docbert: Bert for document classification. *arXiv preprint arXiv:1904.08398*.

Afra Feyza Akyürek, Lei Guo, Randa Elanwar, Prakash Ishwar, Margrit Betke, and Derry Tanti Wijaya. 2020. Multi-label and multilingual news framing analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8614–8624.

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

Frank R Baumgartner, Suzanna L De Boef, and Amber E Boydstun. 2008. *The decline of the death penalty and the discovery of innocence*. Cambridge University Press.

Adam J Berinsky and Donald R Kinder. 2006. Making sense of issues through media frames: Understanding the Kosovo crisis. *The Journal of Politics*, 68(3):640–656.

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

Dallas Card, Amber Boydstun, Justin H Gross, Philip Resnik, and Noah A Smith. 2015. The Media Frames Rorpus: 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)*, pages 438–444.

Dallas Card, Justin H Gross, Amber Boydstun, and Noah A Smith. 2016. Analyzing framing through the casts of characters in the news. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1410–1420.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Wei-Fan Chen, Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2018. Learning to flip the bias of news headlines. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 79–88.

Arman Cohan, Iz Beltagy, Daniel King, Bhavana Dalvi, and Daniel S Weld. 2019. Pretrained language models for sequential sentence classification. *arXiv preprint arXiv:1909.04054*.

Franck Dernoncourt and Ji Young Lee. 2017. Pubmed 200k rct: a dataset for sequential sentence classification in medical abstracts. *arXiv preprint arXiv:1710.06071*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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

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)*, pages 6344–6350, Hong Kong, China, November. Association for Computational Linguistics.

Anjalie Field, Doron Kliger, Shuly Wintner, Jennifer Pan, Dan Jurafsky, and Yulia Tsvetkov. 2018. Framing and agenda-setting in russian news: a computational analysis of intricate political strategies. *arXiv preprint arXiv:1808.09386*.

Dean Fulgoni, Jordan Carpenter, Lyle Ungar, and Daniel Preotiuc-Pietro. 2016. An empirical exploration of moral foundations theory in partisan news sources. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 3730–3736, may.

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

Stephan Greene and Philip Resnik. 2009. More than words: Syntactic packaging and implicit sentiment. In *Proceedings of human language technologies: The 2009 annual conference of the north american chapter of the association for computational linguistics*, pages 503–511.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Christoph Hube and Besnik Fetahu. 2019. Neural based statement classification for biased language. In *Proceedings of the twelfth ACM international conference on web search and data mining*, pages 195–203.

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)*, volume 1, pages 1113–1122.

Di Jin and Peter Szolovits. 2018. Hierarchical neural networks for sequential sentence classification in medical scientific abstracts. *arXiv preprint arXiv:1808.06161*.

Siyi Liu, Lei Guo, Kate Mays, Margrit Betke, and Derry Tanti Wijaya. 2019a. 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)*, pages 504–514.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Vlad Niculae, Caroline Suen, Justine Zhang, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2015. Quotus: The structure of political media coverage as revealed by quoting patterns. In *Proceedings of the 24th International Conference on World Wide Web*, pages 798–808.

Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135.

Pinelopi Papalampidi, Frank Keller, and Mirella Lapata. 2019. Movie plot analysis via turning point identification. *arXiv preprint arXiv:1908.10328*.

Raghavendra Pappagari, Piotr Zelasko, Jesús Villalba, Yishay Carmiel, and Najim Dehak. 2019. Hierarchical transformers for long document classification. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 838–844. IEEE.

Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1650–1659.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *arXiv preprint arXiv:1908.10084*.

Oren Tsur, Dan Calacci, and David Lazer. 2015. A frame of mind: Using statistical models for detection of framing and agenda setting campaigns. 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)*, volume 1, pages 1629–1638. ACL.

Esther van den Berg, Katharina Korfhage, Josef Ruppenhofer, Michael Wiegand, and Katja Markert. 2020. Doctor who? Framing through names and titles in german. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4924–4932, Marseille, France, May. European Language Resources Association.

Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. 2004. Learning subjective language. *Computational linguistics*, 30(3):277–308.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, pages 347–354.

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