What to Fact-Check: Guiding Check-Worthy Information Detection in News Articles through Argumentative Discourse Structure

Tariq Alhindi[‡] **Brennan Xavier McManus**[‡] **Smaranda Muresan**[†‡]
[‡]Department of Computer Science, Columbia University
[†]Data Science Institute, Columbia University
tariq@cs.columbia.edu, {bm2530, smara}@columbia.edu

Abstract

Most existing methods for automatic fact-checking start with a precompiled list of claims to verify. We investigate the understudied problem of determining what statements in news articles are worthy to fact-check. We annotate the argument structure of 95 news articles in the climate change domain that are fact-checked by climate scientists at climatefeedback.org. We release the first multi-layer annotated corpus for both argumentative discourse structure (argument components and relations) and for fact-checked statements in news articles. We discuss the connection between argument structure and check-worthy statements and develop several baseline models for detecting check-worthy statements in the climate change domain. Our preliminary results show that using information about argumentative discourse structure shows slight but statistically significant improvement over a baseline of local discourse structure.

1 Introduction

The proliferation of misinformation in online portals is increasing at a scale that calls for the automation of the slow and labor-intensive manual fact-checking process (Vosoughi et al., 2018). The need for automation is even bigger in highly controversial topics such as climate change. An end-to-end automatic fact-checking system needs to accomplish three main tasks: 1) find claims that are worth fact-checking, 2) retrieve relevant evidence from credible sources, and 3) determine the veracity of that claim given the retrieved evidence. Most previous attempts at automating fact-checking focus on the latter two steps by comparing a manually prepared list of claims against automatically- or manually-retrieved evidences from (trusted) sources such as Wikipedia or news articles from credible publishers (Thorne et al., 2018; Ferreira and Vlachos, 2016; Pomerleau and Rao, 2017). However, less attention is given to automatically compiling a list of check-worthy statements that can then be inspected and fact-checked by a human fact-checker (or by a fact-checking system). A small number of previous studies developed datasets and models for identifying check-worthy statements in political news and debates (Hassan et al., 2017; Jaradat et al., 2018; Arslan et al., 2020).

We look at the problem of deciding what sentences to fact-check in news articles and in particular in the climate change domain. We hypothesize that selecting segments for fact-checking in news articles, particularly for controversial topics, is related to the overall argumentative structure of the article, more specifically to the argument component type (e.g., claim, premise) and to the incoming and outgoing argumentative relations (e.g., support, attack) from or to the argument components. By looking at some of the fact-checked articles, we notice that the segments selected for fact-checking by climate scientists sometimes contain a claim, a premise, or a combination of both a claim and a premise. When we look at the context around the fact-checked segments, we notice patterns related to the argumentative structure. For example, human fact-checkers tend to fact-check a claim when it is not supported by an evidence (premise) or only supported by another claim, and fact-check a premise when it is used to support an claim (e.g., to challenge the relevance of that evidence in support for the claim). Not all fact-checked segments are chosen on a basis related to the argumentative structure as we show in our analysis, however, having annotations of both fact-checked segments and argument component types allow us to understand and model this relation. Figure 1 shows an excerpt from one article in our dataset with its argument and fact-checked segments annotations.

Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 380–391
July 29–31, 2021.

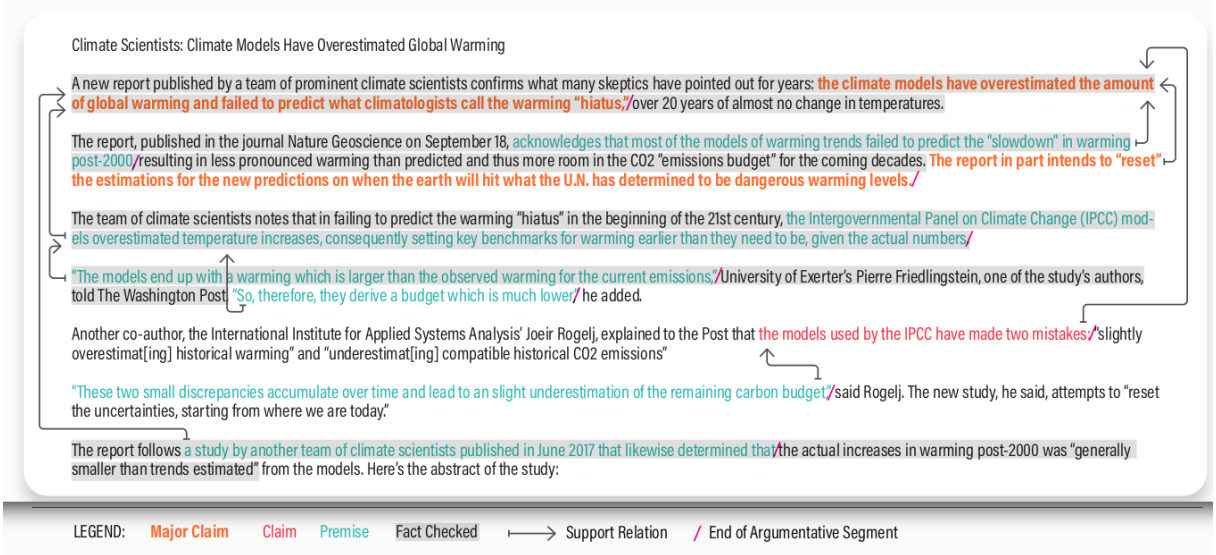

Figure 1: Fact-Checked Segments and Argument Components and Relations in one Article

Our contributions in this paper are as follows[1]:

1. We introduce a new dataset of 95 climate change news articles with annotations of fact-checked segments (Section 3.1).

2. We annotate the argumentative discourse structure of these 95 articles (Section 3.2), thus introducing the first multi-layer annotated corpus both for argumentative discourse structure and check-worthy statements that allows us to deepen our understating of the connection between the two (Section 4).

3. We show that a BERT model (Devlin et al., 2019) that incorporates information about argumentative discourse structure provides a slight but statistically significant improvement over a BERT model that uses just local discourse context (Sections 5 and 6).

2 Related Work

Previous work on fact-checking has focused on different steps of the fact-checking pipeline (Thorne and Vlachos, 2018; Graves, 2018), the majority of which is work on predicting the veracity of claims either by comparing them against evidence from Wikipedia (Thorne et al., 2018), trusted news outlets (Ferreira and Vlachos, 2016; Pomerleau and Rao, 2017), discussion forums (Joty et al., 2018), or debate websites (Chen et al., 2019), or by analyzing the linguistic properties of false and true claim (Pérez-Rosas et al., 2018; Rashkin et al., 2017) in addition to the speaker's history (Wang, 2017; Alhindi et al., 2018). Other work focuses on estimating the credibility of sources by using an external list of bias per publisher (Baly et al., 2018) or by modeling conflicting reports on a claim from different sources (Zhang et al., 2019). However, all of these methods either report bias at the publisher level or start with a list of claims to fact-check.

Previous work on detecting check-worthy claims focus on text from the political domain. The two main existing systems for check-worthy claim detection are ClaimBuster (Hassan et al., 2017) and ClaimRank (Jaradat et al., 2018). ClaimBuster is trained on sentences from political debates and uses sentence level features such as TF-IDF weights and sentiment. ClaimRank extends this to Arabic (in addition to English) and uses a richer feature set that includes the context. Other more recent work include datasets that are bigger in size and across longer time spans (Arslan et al., 2020) or in other languages such as Dutch (Berendt et al., 2020). Covering multiple domains (political speeches, tweets, Wikipedia) and task formulations (check-worthiness, rumor detection, and citation detection), Wright and Augenstein (2020) use positive unlabelled learning (Bekker and Davis, 2020) to perform a comparison of datasets across domains where the notion of check-worthiness vary greatly.

Over the past three years, the CLEF check-that lab introduced tasks for detecting check-worthy political claims from debates and social media (Nakov et al., 2018; Elsayed et al., 2019; Barrón-

[1]The annotated dataset, guidelines, and code are available here: https://github.com/Tariq60/whatToFactcheck

Cedeño et al., 2020), where the best teams in the 2019 task (Hansen et al., 2019) uses syntactic features and word embeddings in an LSTM model. More recently on the same datasets, Kartal et al. (2020) introduce a logistic regression model using BERT-based features, presence of comparative and superlative adjectives, augmented with data from controversial topics. Finally, Meng et al. (2020) use adversarial training on transformer neural network models for detecting check-worthy statements. However, all of these models are trained on political text from debates, speeches and tweets, or lists of claims previously checked by various fact-checking agencies such as `FactCheck.org`. We on the other hand work on a dataset from a different genre: *news articles*, and from a different domain: *climate change*, and investigate the question whether argumentative discourse structure helps in detecting check-worthy statements.

Argument mining is a field concerned with finding argument structure in text from argument components (claim, premises) to relations (support, attack) as covered extensively by Lawrence and Reed (2020). Several argumentation corpora are available on texts from multiple genres such as student essays (Stab and Gurevych, 2014), and social-media threads (Hidey et al., 2017), which have been used in applications such as writing assistance (Zhang and Litman, 2016) and essay scoring (Somasundaran et al., 2016). Freeman (2000) has argued that statements have different types which affects the type of evidence they need or lack thereof. This was empirically explored by works that attempted to identify the appropriate type of support for statements in user comments (Park and Cardie, 2014) and controversial topics in the social media (Addawood and Bashir, 2016). In this work, we provide a resource and a model that aims to deepen our understanding of the relations between argumentative discourse structure and check-worthiness.

3 Multi-Layer Annotated Corpus

We describe below the dataset, its fact-checked segment annotation by climate scientists, and our argumentative discourse structures annotation on the same dataset.

3.1 Fact-Checked Segments Annotation

We introduce a new dataset of 95 climate change news articles fact-checked at the sentence-level by climate scientists at the `climatefeedback.org` website. The articles are from 40 publishers mainly in the U.S., UK and Australia (e.g., *The New York Times, The Guardian, The Washington Post, The Wall Street Journal, The Australian, The Telegraph, Forbes, USA today, Breitbart,* and *Mashable*).[2] Each article is fact-checked by 3 to 5 climate scientists that evaluate scientific reasoning, add relevant information missed by the article and check for: factual accuracy, scientific understanding, logical reasoning, precision/clarity, sources quality, and fairness/objectivity[3]. The articles are given an article-level credibility assessment from very low to very high by the fact-checkers in addition to the segment-level annotation. Table 1 shows the number of articles in each of the nine degrees of credibility for news articles. The annotations of fact-checked segments vary in length from a fragment of a sentence to multiple sentences. We thus map this to binary labels at the sentence-level: fact-checked sentences or non-fact-checked sentences. Each sentence is labeled as 'fact-checked' if it was fact-checked, or it has a fact-checked fragment, or it is part of multi-sentence fact-checked segment. We use NLTK sentence segmenter (Loper and Bird, 2002) to split both the original articles and the fact-checked segments into a list of sentences.

Credibility	Count	Credibility	Count
very-low	23	high	21
very-low/low	7	high/very-high	8
low	10	very-high	18
neutral	7	mixed	1

Table 1: Number of articles per credibility level

There are a total of 134 articles that are fact-checked by `climatefeedback.org` at the time of crawling this data (*May 2020*). However, we only include articles that have segment-level annotations and thus the final dataset has a total of 95 articles. We split the dataset to 68 articles in the training set (4,353 sentences in total, 824 are fact-checked), 7 articles in the development set (249 sentences in total, 55 are fact-checked), and 20 articles in the test set (970 sentences in total, 220 are fact-checked). We consider article credibility, publisher, and the ratio of fact-checked sentences when doing the split to make sure all data splits have articles from a diverse set of credibility levels, publishers

[2] We collect the articles from LexisNexis, which licenses the use of data for research purposes.

[3] https://climatefeedback.org/process/

and styles. The ratio of fact-checked sentences in all three splits is around 20-25% of total number of sentences in the data.

3.2 Argumentative Discourse Structure Annotation

We also annotate the argumentative discourse structure of the 95 fact-checked articles. Our annotation scheme is a slight modification of the one introduced by Stab and Gurevych (2017). It has three types of argument components: Major-Claim, Claim, and Premise. Each consist of a single proposition. Major-Claims are propositions that express the main stance the author takes about the text's main issue. Claims are stances relating to the text's main issue that can support or undermine a major claim, or another claim. Finally, Premises are propositions which express reasons to believe a given claim. Our scheme uses four types of relations: Support, Attack, Restate, and Joint. Relations are directed connections between components, such that each component may have no more than one outgoing relation. Besides the classical Support and Attack relations, we introduce a Restate relation that indicates that two components of the same type (such as two claims) are the same (e.g., the author introduces a Main Claim and then restate it at the end of the article). Finally, a Joint relation, which occurs only between two adjacent Premises, indicates that the two should be taken as a single argumentative unit. They are distinct propositions, but neither can be considered argumentative without the other.

Our annotation study consisted of six annotators, all undergraduate students. We recruited potential annotators from the departments of Linguistics, English, and Comparative Literature, trained them on a sample of articles, then assigned each a 32-article batch. The articles were distributed such that each batch had three annotators. We used the Brat web server as our annotation tool.[4]

We create gold annotations for each article by synthesizing all three of its annotators' contributions. The text span for each gold component consists of the minimum common span of all overlapping components from the three annotations. We use majority voting to decide the label of the new gold component, with the label that occurs most often in the overlapping individual annotations being chosen as the gold label. In cases with a three-way tie between unlabelled, Premise, and Claim or Major-Claim, we determine highest quality annotator of that span, where annotator quality is an ordinal ranking of all annotators in the study in descending order of their average pairwise agreement across all articles, and use the label the highest quality annotator provided. Once the gold argument components are created, we generate gold relations. First, we collect all outgoing relations from the individual annotators' components associated with a given gold argument component. We then remove any relations which begin or end at a component which was not included in the creation of a gold component. Then, for each gold argument component, we determine the gold relation by, in order of priority: adherence to guidelines, annotator quality, and the frequency with which the given relation type appears in our corpus. Adherence is a binary True or False depending on whether the proposed relation is consistent with our annotation schemes, such that an adherent relation is chosen when possible. To assess the quality of the resulting gold annotations, an expert meta-annotator then examined 18 of the resulting 95 annotated articles, and recorded any instances in which they disagreed with the gold annotation. This comparison resulted in an agreement with the gold annotations 85.3% of the time.

We calculate inter-annotator agreement using two versions of dkpro-statistic's open-source [5] implementation of Krippendorff's alpha, which measures on a scale from -1 to 0 to 1 from inverse agreement, to agreement only by chance, to perfect agreement (Bär et al., 2013; Krippendorff, 2011). When using the coding version, which uses only the labels assigned to each component, we find an overall inter-annotator agreement of .4368, with category agreements of .1745 for Premises, .2175 for Claims, and .3782 for Major-Claims. Using the unitizing version, which takes into account both the label of each argument component and the span each annotator selected, we find an overall agreement of .2763, with agreements of .2803 for Premises, .2463 for Claims, and .4312 for Major-Claims. We also use the unitizing version to calculate each annotator's average pairwise overall agreement for the purpose of assessing annotator quality, finding a range from .1776 to .4641.

The dataset comes from multiple publishers and countries, and includes numerous types of articles

[4] brat.nlplab.org

[5] dkpro.github.io/dkpro-statistics

Best Annotator		Gold Annotations	
AC Type	Frequency	AC Type	Frequency
Claim	110	Claim	91
Premise	100	Premise	76
Premise Premise	40	Major-Claim	22
Claim Claim	26	Premise Premise	20
Claim Premise	25	Claim Premise	17
Major-Claim	21	Claim Claim	12
Premise Claim	13	Premise Claim	9
Premise Premise Premise	10	Premise Claim Claim	4
Claim Claim Claim	8	Premise Premise Claim	4
Premise Claim Premise	7	Claim Premise Claim	4

Table 2: The most frequent argument component (AC) types of fact-checked segments.

AC Type	Total Rel.	Relation Type	Frequency
Claim	1	$\xrightarrow{\text{sup}}$ Claim	18
	1	$\xrightarrow{\text{sup}}$ Major-Claim	13
Premise	1	$\xrightarrow{\text{sup}}$ Claim	79
	2	$\xrightarrow{\text{att}}$ Claim, $\xleftarrow{\text{sup/oth}}$ Premise	9
Major	$\geq$5	$\xleftarrow{\text{sup}}$ Claim (all)	13
Claim	1	$\xrightarrow{\text{oth}}$ Major-Claim	3

Table 3: Relation types counts for best annotator

AC Type	Total Rel.	Relation Type	Frequency
Claim	1	$\xrightarrow{\text{sup}}$ Claim	12
	1	$\xrightarrow{\text{sup}}$ Major-Claim	11
Premise	1	$\xrightarrow{\text{sup}}$ Claim	54
	1	$\xrightarrow{\text{sup}}$ Premise	4
Major	$\geq$4	$\xleftarrow{\text{sup}}$ Claim (all)	10
Claim	1	$\xrightarrow{\text{oth}}$ Major-Claim	2

Table 4: Relation types count in gold

such as editorials, op-eds, news analysis and news reporting. This increases the complexity of the annotation task which could explain the low Krippendorff's alpha scores for inter-annotator agreement.

4 Analysis of Argumentation in Fact-Checked Segments

To further understand the relation between argumentative discourse structure and fact-checked segments, we analyze the argument components types and relations of the fact-checked segments in the training data. To see the effect of our strategy in selecting gold argumentative spans and relations on the overlap with fact-checked segments, we do our analysis using the annotations of the best annotator for each article (overall highest in pairwise agreement with other annotators), and the gold annotations. We look at the original fact-checked segments before they are split to sentences as described in Section 3.1. This results in 589 fact-checked segments that mostly consist of multiple sentences (splitting them to sentences increases the number to 824 fact-checked sentences).

Argument Component Types. We first look at the best annotator's coding. Out of the 589 fact-checked segments, 430 map to argument components in the articles. Out of argumentative fact-checked segments, 53% consist of a single argument component: 95 are Claims, 82 are Premises and 17 are Major-Claims, while the remaining consist of two (25%), three (10%), or four or more argument components (12%). Table 2 shows the most frequent argument component types of the fact-checked segments.

When we use the gold annotations, the number of annotated segments in most articles decreases due to only including segments that are annotated by two or more annotators. This reduces the argumentative fact-checked segments from 430 to 307 out of the 589 total fact-checked segments. This reductions cascades to the frequency of argument component types (Table 2) and relations counts (Table 4) in fact-checked segments.

Argumentative Relations. When we look at the relations from and to argument components that are fact-checked (as annotated by the best annotator), we notice that a Premise is fact-checked when it has one relation (mostly an outgoing sup-

port relation) and a Claim is fact-checked when it has many relations (up to four) with mixed directions (incoming, outgoing) and types (support, attack). This essentially maps to fact-checking a Premise when it is used as a supportive evidence and fact-checking a Claim when it is central to the overall argument of the article. Also, Claims and Major-Claims are fact-checked when they are only supported by other Claims (which could signal that the author is not providing an evidence, thus showcasing an "*evading the burden of proof*" fallacy). The most frequent relation counts of fact-checked segments are shown in Table 4.

The general patterns found in the annotations of the best annotator still hold for the gold annotations. The only exception in the gold annotations is that a Major-Claim is fact-checked more often that segments consisting of two Premises or two Claims, which is mainly due to the smaller count of argument component (and relations) in the gold annotations. More detailed counts are shown in Appendix B.

5 Experimental Setup

We use the climate scientists' decision to fact-check a sentence as our gold labels for check-worthiness. In order to understand the capability of machine learning models to decide whether a sentence should be fact-checked, we introduce an experimental setup as follows. In line with previous work, we formulate this problem in two ways: a) **sentence classification task**, i.e. determining whether a given sentence should be fact-checked or not, and b) **sentence ranking** by check-worthiness. For the sentence classification task, we use Macro F1 scores as our evaluation metric, while for ranking we use Mean Average Precision (MAP). We experiment with fine-tuning BERT (Devlin et al., 2019) using the transformers library by huggingface (Wolf et al., 2020) with and without argumentation-based selection of context as described below.

Baselines. We fine-tune BERT for 3 epochs (*bert-base-uncased*, max sequence length 256, batch size 16, learning rate 2e-5) using three different inputs to establish a baseline for this task. The first baseline is fine-tuning using only the target sentence for classification as the input (SENT). The other two configurations utilize the capability of BERT to handle two inputs. Therefore, we experiment with passing the target sentence with its previous sentence as input (PREV+SENT) and with its next sentence (SENT+NEXT). These two configurations essentially provide local **discourse context** following the natural order of sentences in the article.

Argumentation Context. One simple way to test our hypothesis on the relation between argumentation and check-worthiness is by selecting a context for the target sentence using the argumentative discourse structure. We refer to such context as the argumentation context in our discussion. If the target sentence is argumentative, we look at its outgoing and incoming argumentative relations. If the sentence has an incoming relation, then the source of that relation is passed as the first input of BERT and the target sentence is passed as the second input. If the relation is outgoing from the target sentence, then the target sentence is passed as the first input and the target of the relation is passed as the second. As a single sentence could consist of more than one argument component, which in turn could have many relations, this creates many pairs for the target sentence.

We explore three configurations for using the argument structure to select context. First, we keep all pairs for each target sentences, thus increasing the number of instances in the data and maintaining the same gold label for each repeated target sentence in the training data that is matched with a different argumentation context. We denote such configuration as AC(ALL) in our discussion. The final label during inference time can be determined in two ways: via majority label of predictions for each target sentence, and via favoring the minority class, i.e., if one prediction is to fact-check then we consider that as the final label.

Second, we select some of the argumentation context by keeping the most frequent relations in fact-checked segments seen in training as discussed in Section 4. If the target sentence has a Claim or Major-Claim, then we only keep incoming support relations from other Claims or Major-Claims. However, if the target sentence has a Premise, we keep outgoing relations to Claims or Major-Claims. We also limit the total number by either 3 (AC(3)) or 1 (AC(1)) selecting at random if the remaining relations exceed the limit. In case the target sentence is not argumentative, we revert to the discourse context by selecting the previous sentence.

Third, we experiment with prepending argument component type of the target sentence and its context to the input text (e.g., if the sentence has a

Group	Model Input	Not-Checked	Fact-Checked	Macro F1	MAP
Baselines	SENT	0.83	0.23	0.53	0.296
	PREV+SENT	0.83	0.29	0.56	**0.387**
	SENT+NEXT	0.83	0.27	0.55	0.296
Argument Context (Text only)	SENT+AC(1)	0.84	**0.33**	**0.58**	0.366
	SENT+AC(3)^{v1}	0.82	0.31	0.57	0.299
	SENT+AC(3)^{v2}	0.82	0.32	0.57	0.299
	SENT+AC(ALL)^{v1}	0.83	0.26	0.54	0.318
	SENT+AC(ALL)^{v2}	0.81	0.30	0.56	0.318
Argument Context (Text+Type)	SENT+AC(1)+T	0.83	0.29	0.56	0.359
	SENT+AC(3)+T^{v1}	0.84	0.27	0.57	0.305
	SENT+AC(3)+T^{v2}	**0.85**	0.29	0.57	0.305
	SENT+AC(ALL)+T^{v1}	0.82	0.32	0.57	0.281
	SENT+AC(ALL)+T^{v2}	0.82	0.31	0.57	0.281

Table 5: Results on the Development Set. Per-class F1, Macro F1 for sentence classification, and MAP for sentence ranking. ^{v1}Majority prediction to determine the final label. ^{v2}Final prediction is to Fact-Check if at least one prediction for the target sentence is as such. ^{v1,v2}Voting strategies do not affect MAP as we take the average of the prediction probabilities for each target sentence.

Input	NC	FC	F1	MAP
SENT	**0.85**	0.28	0.56	0.398
PREV+SENT	0.82	0.29	0.56	0.384
SENT+NEXT	0.84	0.26	0.55	0.385
SENT+AC(1)	0.83	0.30	0.57	0.413
SENT+AC(1)+T	0.84	**0.33**	**0.59**†	**0.420**†

Table 6: Per-class F1, Macro F1 and MAP on the Test Set. †significant over the baseline (PREV+SENT)

claim, the input will be "_CLAIM_" followed by the sentence; for non-argumentative sentences we use "_NONE_"). We denote experiments with such configurations with the letter (T).

6 Results and Discussion

We show the results of our experiments in Table 5 for the development set and Table 6 for the test set. We can see in the baseline experiments in both tables that PREV+SENT condition is better than SENT+NEXT condition both in terms of Macro F1 score and the Fact-Checked class F1 score (FC_{class} F1). Looking at the results on the dev set, we can see that the argument context of SENT+AC(1) has the highest FC_{class} F1 of 0.33, which is ***4 points*** above PREV+SENT and ***6 points*** above SENT+NEXT. It also has the highest Macro F1 of 0.58, which is ***2 points*** above PREV+SENT and ***3 points*** above SENT+NEXT. This indicates that providing a context based on argument relations that could be either before or after and not necessarily adjacent to the target segment is more informative for check-worthiness than providing local discourse context of the previous or next sentence. The same holds for the test set where the best argument context of SENT+AC(1)+T has the best FC_{class} F1 of 0.33 (***4 points*** above PREV+SENT and ***7 points*** above SENT+NEXT), best Macro F1 of 0.59 (***2 points*** above PREV+SENT and ***3 points*** above SENT+NEXT), and best MAP of 0.420 (***2 points*** above SENT, which is the highest baseline with MAP score). The test set SENT+AC(1)+T Macro F1 and MAP results are ***statistically significant*** over all three baselines SENT, PREV+SENT, and SENT+NEXT.

However, providing more than one sentence does not improve the results in the AC(3) and AC(ALL) experiments as shown in Table 5, regardless whether the final prediction at inference time is decided via majority voting or favoring the FC class. Therefore, we only run AC(1) and AC(1)+T experiments on the test set. It is worth noting that adding the argumentative type to the target sentence and its context has the highest results on the test set but not on the development set. This could be due to the small size of the development set of 249 sentences from 7 articles, which could have lead to high variability from the general trend in the data. The sentence type information has also the highest MAP score for the sentence ranking task. The ranking is done based on the prediction probability of the model for all sentences in an ar-

ticle. The MAP value is computed by taking the mean of all average precision scores on all articles in one data split. This is a simplified version of the classification task where the model does not need to have correct prediction for every single sentence in the article as long as it highly ranks most of the fact-checked sentences in an article.

Argumentative Segments. In order to have a better understanding of the true potential of the argumentative discourse context for this task, we look at the accuracy of predictions on the argumentative segments of the articles. All non-argumentative segments have no incoming or outgoing argumentative relations. Therefore, there is no way of providing an argumentative discourse context for them so they are matched with their previous sentence as mentioned earlier. Thus, the reported results on all AC conditions is on a mix of pairs where some sentences have an argumentation context while other have a discourse context. Out of the 249 sentences in the dev set, 133 are argumentative of which 37 are Fact-Checked. If we look at the model performance on this subset of the dev set, we see scores of 0.31 FC_{class} F1 and 0.53 Macro F1 for PREV+SENT, while having scores of 0.41 FC_{class} F1 and 0.60 macro F1 for SENT+AC(1). A gain of 10 F1 points in the FC_{class} on the argumentative subset of the dev set compared with 4 points difference in FC_{class} F1 on the whole set shown in Table 5. The same observation holds for the test set that includes 485 argumentative sentences (out of 970) of which 123 sentences are Fact-Checked. The results on this subset are 0.33 FC_{class} and 0.55 macro F1 for PREV+SENT, and 0.38 FC_{class} and 0.61 macro F1 for SENT+AC(1)+T. This is again a wider margin of 5 F1 points on FC_{class} compared to the 4 points difference in FC_{class} F1 reported in Table 6 on the whole test set. These numbers show that using argumentation context for determining check-worthiness of sentences in an article is more clearly beneficial on the argumentative segments of the article. We leave further experimentation and modeling for future work that includes complimenting this approach with other linguistic information to determine check-worthiness of the non-argumentative parts of the articles.

Error Analysis. We closely examine a few examples where the argumentative discourse context helped the model in making a correct prediction. One fact-checked "Major-Claim" saying: *"Updated data from NASA satellite instruments reveal the Earth's polar ice caps have not receded at all since the satellite instruments began measuring the ice caps in 1979."* was the first sentence in the article so it was paired with title in the PREV+SENT model that did not make a correct prediction. However, the AC(1)+T paired it with another "Major-Claim" (*The updated data contradict one of the most frequently asserted global warming claims* ...) that comes 3 sentences later in the article and has a support relation to the target sentence. Another example is the "Major-Claim" (*The brutal weather has been supercharged by human-induced climate change*) supported by a "Claim" (*Climate models for three decades have predicted exactly what the world is seeing this summer*). Both of these examples have been correctly predicted by the AC(1)+T model, which indicates the benefit of providing both argument component type and its argumentation context to determine its check-worthiness, especially for "Major-Claims". On the other hand, AC(1)+T makes several wrong predictions to fact-check sentences from the Not-Checked class, which were predicted correctly by SENT and PREV+SENT models. This happens in cases where both the target and context sentences are Claim/Major-Claim, which indicates that such relations are providing a strong signal to fact-check. However, the climate scientist might have decided that those sentences were not check-worthy due to their own knowledge in the field rather than reasons related to the argumentation structure.

7 Conclusion

We introduced a corpus of news articles with multi-layer annotations of check-worthiness and argumentative discourse structure to further our understanding of the relation between argumentation and fact-checking. We approached the task of determining what sentences to fact-check in a news articles formulated as a sentence classification task and as a sentence ranking task. We showed that providing an argumentative discourse context along with the target sentence when fine-tuning BERT improves over baselines of the target sentence alone or with its local discourse context, especially on the argumentative part of the articles.

In future work, we want to compare using the gold annotations of argument structure with predicted argument components and relations by training another model that generate argumentation fea-

tures to be used for the main task as done in previous work (Alhindi et al., 2020). Also, we want to explore the use of other linguistic features tested in previous work and other variations of argumentation context and features such as counts of relations for the target argumentative segment. BERT is pretrained on the next sentence prediction task, which makes an out-of-order argumentation context to be further away from the distribution of the pretraining data. To remedy this, we plan to adaptively pretrain BERT on more argumentation context extracted from multiple argumentation corpora. Finally, we want to study the relation of check-worthiness to intrinsic clause types such as facts and testimony, and to argument fallacies not related to the argument structure.

Acknowledgements

The first author is supported by the KACST Graduate Studies Scholarship. This research is based upon work supported in part by the National Science Foundation (award #1847853). The views and conclusions herein are those of the authors and should not be interpreted as necessarily representing official policies, expressed or implied of NSF or the U.S. Government. We thank the anonymous reviewers for constructive feedback.

References

Aseel Addawood and Masooda Bashir. 2016. "what is your evidence?" a study of controversial topics on social media. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 1–11.

Tariq Alhindi, Smaranda Muresan, and Daniel Preotiuc-Pietro. 2020. Fact vs. opinion: the role of argumentation features in news classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6139–6149, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 85–90, Brussels, Belgium. Association for Computational Linguistics.

Fatma Arslan, Naeemul Hassan, Chengkai Li, and Mark Tremayne. 2020. A benchmark dataset of check-worthy factual claims. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 821–829.

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*, pages 3528–3539.

Daniel Bär, Torsten Zesch, and Iryna Gurevych. 2013. Dkpro similarity: An open source framework for text similarity. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 121–126.

Alberto Barrón-Cedeño, Tamer Elsayed, Preslav Nakov, Giovanni Da San Martino, Maram Hasanain, Reem Suwaileh, and Fatima Haouari. 2020. Checkthat! at clef 2020: Enabling the automatic identification and verification of claims in social media. In *European Conference on Information Retrieval*, pages 499–507. Springer.

Jessa Bekker and Jesse Davis. 2020. Learning from positive and unlabeled data: a survey. *Mach. Learn.*, 109(4):719–760.

Bettina Berendt, Peter Burger, Rafael Hautekiet, Jan Jagers, Alexander Pleijter, and Peter Van Aelst. 2020. Factrank: Developing automated claim detection for dutch-language fact-checkers.

Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. 2019. Seeing things from a different angle: Discovering diverse perspectives about claims. 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)*, pages 542–557.

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)*, pages 4171–4186.

Tamer Elsayed, Preslav Nakov, Alberto Barrón-Cedeno, Maram Hasanain, Reem Suwaileh, Giovanni Da San Martino, and Pepa Atanasova. 2019. Overview of the clef-2019 checkthat! lab: automatic identification and verification of claims. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 301–321. Springer.

William Ferreira and Andreas Vlachos. 2016. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 1163–1168.

James B Freeman. 2000. What types of statements are there? *Argumentation*, 14(2):135–157.

D Graves. 2018. Understanding the promise and limits of automated fact-checking.

Casper Hansen, Christian Hansen, Jakob Grue Simonsen, and Christina Lioma. 2019. Neural weakly supervised fact check-worthiness detection with contrastive sampling-based ranking loss. In *CLEF*.

Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anil Kumar Nayak, et al. 2017. Claimbuster: the first-ever end-to-end fact-checking system. *Proceedings of the VLDB Endowment*, 10(12):1945–1948.

Christopher Hidey, Elena Musi, Alyssa Hwang, Smaranda Muresan, and Kathleen McKeown. 2017. Analyzing the semantic types of claims and premises in an online persuasive forum. In *Proceedings of the 4th Workshop on Argument Mining*, pages 11–21.

Israa Jaradat, Pepa Gencheva, Alberto Barrón-Cedeño, Lluís Màrquez, and Preslav Nakov. 2018. ClaimRank: Detecting check-worthy claims in Arabic and English. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 26–30, New Orleans, Louisiana. Association for Computational Linguistics.

Shafiq Joty, Lluís Màrquez, and Preslav Nakov. 2018. Joint multitask learning for community question answering using task-specific embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4196–4207.

Yavuz Selim Kartal, Busra Guvenen, and Mucahid Kutlu. 2020. Too many claims to fact-check: Prioritizing political claims based on check-worthiness. *arXiv preprint arXiv:2004.08166*.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

John Lawrence and Chris Reed. 2020. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Edward Loper and Steven Bird. 2002. Nltk: the natural language toolkit. *Proceedings of the Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistic*.

Kevin Meng, Damian Jimenez, Fatma Arslan, Jacob Daniel Devasier, Daniel Obembe, and Chengkai Li. 2020. Gradient-based adversarial training on transformer networks for detecting check-worthy factual claims. *arXiv preprint arXiv:2002.07725*.

Preslav Nakov, Alberto Barrón-Cedeno, Tamer Elsayed, Reem Suwaileh, Lluís Màrquez, Wajdi Zaghouani, Pepa Atanasova, Spas Kyuchukov, and Giovanni Da San Martino. 2018. Overview of the clef-2018 checkthat! lab on automatic identification and verification of political claims. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 372–387. Springer.

Joonsuk Park and Claire Cardie. 2014. Identifying appropriate support for propositions in online user comments. In *Proceedings of the first workshop on argumentation mining*, pages 29–38.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401.

D. Pomerleau and D. Rao. 2017. Fake news challenge. http://www.fakenewschallenge.org/. (Accessed on 12/06/2019).

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2931–2937.

Swapna Somasundaran, Brian Riordan, Binod Gyawali, and Su-Youn Yoon. 2016. Evaluating argumentative and narrative essays using graphs. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1568–1578.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

James Thorne and Andreas Vlachos. 2018. Automated fact checking: Task formulations, methods and future directions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3346–3359, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426.

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

Dustin Wright and Isabelle Augenstein. 2020. Fact check-worthiness detection as positive unlabelled learning. *arXiv preprint arXiv:2003.02736*.

Fan Zhang and Diane Litman. 2016. Using context to predict the purpose of argumentative writing revisions. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1424–1430.

Yi Zhang, Zachary Ives, and Dan Roth. 2019. Evidence-based trustworthiness. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 413–423.

A Experiment Reproducibility

As the main objective of the paper is not optimizing for the best hyperparameters for our task but rather introduce the resource and develop some baseline models, we do not experiment for many hyperparameters and stick to the ones recommended by (Devlin et al., 2019) as mentioned in Section 5. We train 3 times for the baseline conditions SENT, PREV+SENT, and SENT+NEXT and take the average of those runs. After seeing stability in the numbers across the three runs, we only train once for the remaining conditions.

B Relation Counts

Here we list more detailed tables of the most frequent types of relations of fact-checked segments. Table 8 is a detailed version of Table 3, and Table 7 is a detailed version of Table 4. Both of Tables 3 and 4 are discussed in Section 4.

AC Type	Total Rel.	Relation Type	Frequency
Claim	1	$\xrightarrow{\text{sup}}$ Claim	12
	1	$\xrightarrow{\text{sup}}$ Major-Claim	11
	2	$\xrightarrow{\text{sup}}$ Major-Claim, $\xleftarrow{\text{sup}}$ Premise	10
	0	–	8
	1	$\xleftarrow{\text{sup}}$ Premise	3
	3	$\xrightarrow{\text{sup}}$ Major-Claim, $\xleftarrow{\text{sup}}$ Premise (2)	3
	1	$\xrightarrow{\text{att}}$ Claim	3
Premise	1	$\xrightarrow{\text{sup}}$ Claim	54
	1	$\xrightarrow{\text{sup}}$ Premise	4
	0	–	4
	1	$\xrightarrow{\text{sup}}$ Major-Claim	4
Major Claim	$\geq$4	$\xleftarrow{\text{sup}}$ Claim (all)	10
	1	$\xrightarrow{\text{oth}}$ Major-Claim	2

Table 7: Relation types count in gold

AC Type	Total Rel.	Relation Type	Frequency
Claim	1	$\xrightarrow{\text{sup}}$ Claim	18
	1	$\xrightarrow{\text{sup}}$ Major-Claim	13
	2	$\xrightarrow{\text{sup}}$ Claim, $\xleftarrow{\text{sup}}$ Premise	8
	2	$\xrightarrow{\text{sup}}$ Major-Claim , $\xleftarrow{\text{sup}}$ Premise	8
	2	$\xrightarrow{\text{sup}}$ Major-Claim, $\xleftarrow{\text{sup}}$ Claim	6
	4	$\xrightarrow{\text{sup}}$ Major-Claim , $\xleftarrow{\text{sup}}$ Premise (3)	5
	3	$\xrightarrow{\text{sup}}$ Major-Claim, $\xleftarrow{\text{sup}}$ Premise (2)	4
	0	–	4
Premise	1	$\xrightarrow{\text{sup}}$ Claim	79
	2	$\xrightarrow{\text{att}}$ Claim, $\xleftarrow{\text{sup/oth}}$ Premise	9
	1	$\xrightarrow{\text{sup}}$ Major-Claim	4
	1	$\xrightarrow{\text{sup}}$ Premise	3
Major Claim	≥ 5	$\xleftarrow{\text{sup}}$ Claim (all)	13
	1	$\xrightarrow{\text{oth}}$ Major-Claim	3

Table 8: Relation types counts for best annotator