# Argumentatively Phony? Detecting Misinformation via Argument Mining

MUHENG YAN, YU-RU LIN, and DIANE J. LITMAN, University of Pittsburgh, USA

As the Internet media grows, fact-checking news articles online becomes increasingly difficult as it requires a vast amount of background knowledge. Recent studies proposed the concept of *reason-checking*, which focuses on analyzing the argumentative reasoning style of texts to identify low-quality news articles. While argument mining techniques are leveraged in automatic systems analyzing the quality of formally written texts such as essays, both its efficiency on news-editorial texts and its benefit in fake news detection are under-investigated. To this end, we analyze the performance of argument mining algorithms on fake news and explore how argumentation knowledge will help computational systems to identify fake news.

**ACM Reference Format:**
Muheng Yan, Yu-Ru Lin, and Diane J. Litman. 2018. Argumentatively Phony? Detecting Misinformation via Argument Mining. *ACM Trans. Graph.* 37, 4, Article 111 (August 2018), 6 pages. https://doi.org/10.1145/1122445.1122456

## 1 INTRODUCTION

The growing size of Internet media in part gives rise to the creation of fake news due to the lack of information gatekeepers in traditional journalism [13]. Fact-checking, a procedure in which professional human workers examine the veracity of events reported in the online news articles, has become a significant approach of fighting the online fake news and misinformation [28]. In recent years, an abundance of studies in computer science tries to automate this procedure as a fake news classification task [18; 19; 21; 22]. All these approaches, whether automated or not, relies heavily on the vast knowledge base of facts and the semantics in the news [7], which can be potentially biased and time-variant – for example, at the beginning of the COVID-19 pandemic, the U.S. CDC claimed "wearing face-masks is not recommended", which later became invalid around the summer of 2020. Furthermore, fake news can be misleading even with facts if they use incorrect reasoning in making arguments [1]. As suggested by Visser et al. [27], besides *fact-checking*, *reason-checking* can be an alternative approach to distinguish fake news online. In this study, we aim to leverage the advances in *Argument Mining* [12] to provide knowledge on reasoning, in fake news detection tasks.

[1]For example, a reasoning error ignorance that reads "I have yet to hear a reasonable argument against quitting my job and moving to the wild. Therefore it must be the right choice to make."

Authors' address: Muheng Yan, muheng.yan@pitt.edu; Yu-Ru Lin, yurulin@pitt.edu; Diane J. Litman, yurulin@pitt.edu, University of Pittsburgh, Pittsburgh, Pennsylvania, USA, 15260.


0730-0301/2018/8-ART111 $15.00
https://doi.org/10.1145/1122445.1122456

One obstacle in mining arguments from fake news is the lack of annotated data for supervised learning. Identifying the boundary and types of argumentation discourse units (ADUs) is the fundamental task in argument mining pipelines [12]. This task is shown to be sensitive to the context, and semantic of the corpus [1]. To this end, fake news is shown to be different from credible ones in context (for instance, they may have conspiracy theories) [29], and lexical features [23; 29]. For this reason, to accurately model the argumentation in fake news, it is essential to validate the existing argument mining method's efficacy on fake news corpus.

It has been shown that the general discourse knowledge of texts can improve the fake news classification task. By extracting the discourse structure measured by sentence similarity, Karimi et al. improved the performance of a fake news classification task by 2% [10], and established the state of the art performance on that corpus. Unlike the existing works that treat each sentence similarly, we aim to further specify their roles in argumentation. Following the theory presented by Al Khatib [3], we specify the role of elementary ADUs, validate the mined units, and explore how the argument styles formed by those units discriminate fake news articles.

This is the first study evaluating the viability of argument mining instruments on fake news corpus to our best knowledge. The contribution of this study is two-fold. Firstly, we address the research gap by validating the argument mining method's performance on a unique type of text corpus – fake news. Secondly, we test how the argument mining techniques would boost the detection of fake news on the Internet. Despite the argument mining models are domain-sensitive, and the fake news texts differ from credible texts in argumentation (ref Section 5), we find that the argument mining models trained on credible news article corpus comprehend well on fake news corpus. The inferred argumentation knowledge from news articles can boost the performance of fake news detection from the state-of-the-art linguistic model, BERT [8]. These findings together provide valuable empirical results for future fake news detection applications.

## 2 RELATED WORKS

**Argument Mining in News Articles.** Argument mining is a natural language processing topic that commonly characterized as 1) identifying argumentative ADUs, at different granularity levels such as sentence or clause, and 2) discovering relations between ADUs [12]. While techniques of argument mining have been widely deployed in a variety of forms of texts such as essays [1], online discourse [24], and political debates [14], arguments in news-editorial texts is under-investigated. Lippi et al. first considered news articles as one source of texts for argumentation discourse analysis [15]; however, the discussion is limited to ten news articles in this study. In 2016, Al Khatib et al. [3] proposed the first and the only annotated news-editorial corpus of argumentation. This study defines the basic concepts and theory for argumentation in new articles

and annotates clauses in 300 news with their argumentative roles. On top of this corpus, Ajjour et al. examined whether the ADU detection models developed for essays can be transferred to news editorial corpus and concluded that the model's architectures are transferable while the training of models is sensitive to the corpus types [1]. Furthermore, Al Khatib et al. [3] characterized the argumentation strategies of news articles by their argumentation flow – the ordered sequence of ADUs. In this paper, we leverage the annotated news-editorial corpus and validate if models trained on this corpus are generalizable on fake news.

**Fake News Detection with Argumentation Knowledge.** As argument mining is under-investigated in news texts, there have not been any studies discuss how argumentation knowledge would interact with the detection of fake news. Visser et al. [27] proposed the idea of *reason-checking* – using the reasoning structures of news to help humans to judge the reliability of online news articles – and demonstrated the effectiveness of it in a BBC teenager education project. Nevertheless, they do not discuss how computational systems can benefit from it. While there lack studies linking argumentation discourse analysis to fake news detection, we may find hints from studies leveraging general discourse analysis in classifying fake news. Karimi and Tang [10] presented the earliest study combining discourse analysis and fake news detection. They modeled the discourse structure between sentences by similarity measurements and used the structures to form a latent representation for fake news classification. Bonet-Jover et al. [5] segment articles into functional discourse units (such as headlines or conclusions) and use lexical features in units to predict the veracity of the articles. The argumentative roles of ADUs in fake news texts are still under-investigated in all prior studies.

## 3 METHOD

This section presents the theoretical grounds of argumentation in fake news, the method of argument component detection, and methods of fake news detection with argument information.

### 3.1 Argumentation in Fake News Articles

We follow the argumentation theory for editorial texts developed by Al Khatib et al. [3]. In this theory, the granularity of ADU is set to clause level. Each ADU (clause) can have one of the six distinct functional roles in forming arguments in news articles: *Assumption* (AS), *Anecdote* (AN), *Testimony* (TS), *Statistics* (ST), *Common-grounds* (CG), and *Other* (OT). Within this taxonomy, the AS proposition that represents opinions or judgments of the article authors can receive supports from the evidence provided by the AN, TS, or ST propositions. We include their published dataset – the *Webis-16 Editorial dataset* [3], as a benchmark dataset to train our machine learning algorithms in the argument component detection task (ref Sec 3.2). This dataset contains 300 articles from three different sources [2], and is pre-split in training, validating, and testing sets.

As this study aims to unfold the relationship between argumentation and fake news detection, we collect a set of articles from non-credible sources that spread on social media and another set of articles from credible sources for comparison. During the COVID-19 pandemic in 2020, there has been an abundance of online articles shared on the Internet providing credible or fake information about the disease and healthcare, which facilitates the body of our targeted corpus. Leveraging a Twitter collection tracked during the pandemic [6] (from Jan.2020 to May.2020), we identify Twitter-shared news by filtering the URLs in this tweet collection. We adopted a list of flagged problematic and trustworthy news sources as a proxy of the article's credibility. For the problematic sources, we first include a combination of reported lists by Grinberg et al. [9] [3]. Furthermore, we extend the problematic list by appending two other lists: "conspiracy and pseudoscience sources" and "questionable sources" curated by the Media Bias and Fact Check (MBFC) organization. We also construct the credible sources list by including the domains considered safe in the same study [9]. We filter the Twitter-shred articles by the problematic and trustworthy lists. This results in 87340 articles, with 62551 of them are from credible domains.

As the argumentation modeling is demonstrated to be topic-sensitive [1], we use LDA modeling to cluster the articles by their topics. In total, 15 topics are detected based on the elbow rule on the topic coherence. These topics are further manually merged into four major clusters by the semantic of their top-ranked keywords. The four clusters are COVID's impact on life (57270 articles), Authorities' Response (5436), Travel and Lockdown (10331), and Science/Medicare of COVID-19 (14303). All the articles participate in the argumentation modeling. However, to balance the labels and topics in the fake news detection task, we down-sample the articles based on the 2 (credible/fake) by 4 (topic clusters) grids (ref Sec 3.3). 1600 articles are randomly sampled in each grid, and in total 15200 articles participate in the fake news detection task.

### 3.2 Argument Component Detection

**Task Definition and Model.** Typically, the detection and classification of argument components from free texts are formed as a sequence tagging task [12]. Each token in the texts is labeled with the Beginning-Inside-Outside (BIO) scheme [4] [17], and the computational models are trained to predict the labels for each token. Combined with ADU types, we create 13 labels for each token to represent their roles in argumentation, including *beginning* (B) or *inside* (I) of each one of the six ADU types, plus *outside* (O) any ADUs. There has been no method published for this task. Hence, we use the latest language model that has shown its strong performance in many NLP tasks, BERT [8], as our baseline for this task [5]. In specific, we leverage the pre-trained BERT encoding layers, add two dense layers for the sequence tagging task, and fine-tune the model on the *Webis-16 Editorial dataset*.

**Evaluation.** The argument component detection model is evaluated in both the *Webis-16 Editorial dataset* and the *fake news dataset*. For the *Webis-16 Editorial dataset* evaluation, we use the default training, validating, and testing split presented by Al Khatib et al. [3]. The BERT model for argument component detection is first trained on the training split of *Webis-16 Editorial dataset* (while the

[2]https://webis.de/data/webis-editorials-16.html

[3]The domains labeled as "black", "red", and "orange" are included in the problematic list, as they are defined as deceptive and with little regards of the truth

[4]In this labeling scheme, the first token of a text span is labeled as "B", the rest tokens within the span are labeled as "I", and the outside-span tokens are labeled as "O"

[5]BERT is proved to be the most effective model in a similar task where the ADUs are characterized by "claim" and "premise" instead of the fine-grained labels [25].

hyper-parameters are searched based on the validating split) and tested directly on the test split.

To facilitate the validity of the model on the *fake news dataset*, we manually annotate a portion of the dataset for evaluation purposes. We randomly sample 10 articles from each one of the 2 (credibility label) by 4 (topic cluster label) grids to create a collection of 80 articles for annotation. Similar to the method used in the *Webis-16 Editorial dataset* annotation, the 80 articles are first segmented into clauses. The annotator (first author) first learns about the annotation scheme presented by Al Khatib et al. [3], and tests his understanding of the scheme by annotating nine (three from each domain in the dataset) articles from the *Webis-16 Editorial dataset*. Within 2016 clauses for annotation from the nine articles, the annotator achieved an agreement of $k = 0.677$ in Fleiss' Kappa (among two annotators, the published dataset is considered an annotator in the calculation) on this 2016 clause with six levels. This is comparable to the reported Kappa $k = 0.560$ (calculated among three annotators, on all 14313 clauses in the dataset with six levels) in the original annotation task. After the evaluation of understanding on the annotation scheme, this annotator then annotates the sampled 80 articles from the *fake news dataset*. With the annotated labels, we test the validity of the argument component detection model and report the performance.

### 3.3 Fake News Detection

The BERT has been shown to excel in various tasks, and is reported effective in fake news detection [11]. Using BERT as the baseline, we present two methods of utilizing the argumentation information in fake news detection. Let there be a text sequence consist of tokens $\mathbf{W} = \{w_1, w_2, ...w_n\}$ where $n$ represents the length of the text. The bert first embed $\mathbf{W}$ into a sequence of embeddings $\mathbf{E} = \{\mathbf{e}_1, \mathbf{e}_2, ...\mathbf{e}_n\}$, and encode the embeddings with six layers of transformer encoder (details in Appendix B). We use $bert(\cdot)$ to refer to the encoder layers. The output of the BERT encoder $\mathbf{h}^b = bert(\mathbf{E}), \mathbf{h}^b \in \mathbb{R}^{768}$ is then input to the dense layers to calculate the output labels $\hat{y} = dense(\mathbf{h}), \hat{y} \in \mathbb{R}^1$.

We use two methods to encode the argument component tags in this prediction task. Let there be a tag sequence $\mathbf{T}$ with the same length of the text $\mathbf{W}$. Each tag $\mathbf{t} \in \mathbb{R}^{13}$ is an one-hot vector of 13 dimensions, each corresponds to one possible argument tags (ref Sec 3.2). In method (a) *argument-as-embedding*, we create an *argument embedding layer* to encode the tag sequence as $\mathbf{a}_i$. The argument tag embedding is summed with the original embedding of BERT before input to the encoder layers $\hat{\mathbf{e}}_i = \mathbf{e}_i + \mathbf{a}_i$. The prediction of the method (a) is made upon the merged embedding vectors.

In method (b) *argument-lstm*, we create a Bi-LSTM [20] as the encoder for the tag sequences. This encoder yields a sequence of latent vectors $\mathbf{k}_i$ corresponding to each tag $\mathbf{t}_i$. Following the encoder layers, we weight the latent vectors with an attention layer $attn(\cdot)$ [4], the output which is $\mathbf{h}^t = \sum a_i \mathbf{k}_i$, where $a_i = attn(\mathbf{k}_i)$. The $\mathbf{h}^t$ is concatenated with the output of the BERT encoder, $\mathbf{h}^b$, before input to the dense layers. Besides the argumentation-enriched models, we also include a *baseline model without BERT components*, that only contains the LSTM encoder for argument tags as described in the *argument-lstm*. This baseline helps to illustrate the contribution of argumentation information in the fake news classification task.

## 4 RESULT

### 4.1 Evaluation of Argument Component Detection

We fine-tune the BERT model on the *Webis-16 Editorial dataset* training set and evaluate the model on both the *Webis-16 Editorial dataset* testing set and the annotated *fake news dataset*.

**Argument Component Detection Baseline on *Webis-16 Editorial dataset*.** The last column of Table 1 shows the evaluation result of the BERT sequence tagging model on the testing set of the *Webis-16 Editorial dataset*. All metrics in the table are calculated on the token-level prediction. As mentioned in Section 3.2, there have not been methods proposed for this task. Thus the BERT model sets a baseline for this 13-way prediction task with $F1 = 0.489$.

**Evaluation on *fake news dataset*.** Similar to the evaluation on *Webis-16 Editorial dataset*, we evaluate the model on *fake news dataset* on the token labeling level. The 80 annotated articles contain in total 103168 tokens. The *anecdote*, *assumption*, and *testimony* labels are the major labels in this corpus, while no tokens belong to the *common-ground* and *other* labels. We report the F1 score for each label separately in Table 1. The extended performance table with precision and recall scores and the confusion matrix for this evaluation can be found in Appendix A. Generally, the model's performance on the "I" tags is better compared to the "B" tags. The macro-averaged F1 score for all labels is 0.476. The performance of the model on the *fake news dataset* is comparable to the baseline performance on the *Webis-16 Editorial dataset* test set, where the overall F1 score is 0.467. The performance of the model on the fake news is also comparable to its performance on the credible articles within *fake news dataset*, with F1 scores of 0.476 and 0.448, respectively. While 13-way classification is a hard task, the model achieved fairly good performance on most labels besides B-CG and I-CG. Due to their rarity, the model does not give any prediction on these two labels, and yields zero for all metrics, which greatly affects the macro-averaged performance. Without these two tags, the overall F1 reaches 0.589 on *fake news dataset* for all articles.

### 4.2 Fake News Detection with Argumentation Tags

As illustrated in Sec 3.1, we experiment on the 15200 sampled articles evenly distributed on the 2 (credible/fake) by 4 (topic clusters) grids. Articles of each topic cluster are split into training (80% of the full size), validating (10%), and testing (10%). To exclude the influence from the topics of the articles, we build fake news detection models per topic cluster, in addition to the models for articles from all topic clusters. For each model architecture, we train five instances specifically on each one of the topic-cluster subset, plus the full dataset. Each trained model is then evaluated with test sets corresponding to each topic cluster.

Table 2 reports the accuracy from the hold-out experiment of the fake news detection task for a) the baseline BERT model with binary classification layers (without argumentation components), b) the argument baseline model with only the LSTM encoding the tags, c) the *argument-as-embedding* model, and d) the *argument-lstm* model. As can be seen from the table, the argumentation baseline achieved 0.670 - 0.700 accuracy in within-topic cases. Without further processing on the mined argumentation knowledge, merely encode the argumentation tag sequence can improve the fake news detection greatly from the naive baseline of random guessing. Overall,

| | | B-AS | I-AS | B-AN | I-AN | B-ST | I-ST | B-TS | I-TS | B-CG | I-CG | O | Macro |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Webis-16 Editorial dataset* | F1 | 0.656 | 0.675 | 0.552 | 0.639 | 0.581 | 0.670 | 0.218 | 0.663 | 0.000 | 0.000 | 0.724 | 0.489 |
| *fake news dataset* **All** Articles | F1 | 0.534 | 0.695 | 0.594 | 0.692 | 0.594 | 0.680 | 0.151 | 0.556 | 0.000 | 0.000 | 0.803 | 0.476 |
| *fake news dataset* **Credible** Articles | F1 | 0.131 | 0.687 | 0.614 | 0.718 | 0.581 | 0.697 | 0.110 | 0.570 | 0.000 | 0.000 | 0.821 | 0.448 |
| *fake news dataset* **Fake** Articles | F1 | 0.575 | 0.694 | 0.588 | 0.684 | 0.577 | 0.650 | 0.139 | 0.545 | 0.000 | 0.000 | 0.798 | 0.477 |

Table 1. **Argument Extraction Performance on *Webis-16 Editorial dataset* and *fake news dataset*.** The table reports the F1 scores for the prediction of each tags in both *Webis-16 Editorial dataset* and *fake news dataset*, followed by the macro-averaged overall performance. To confirm the generalizability of the model on fake news articles, we report the metrics on particularly the fake articles and credible articles separately. The prediction on "common-ground" label (CG) are zero as the trained model does not yield any of these labels in the prediction.

| Train On | Test On | Arg-LSTM (no BERT) | BERT | Arg-emb | Arg-LSTM |
|---|---|---|---|---|---|
| Impact | Impact | 0.700 | 0.892 | 0.879 | **0.918** |
| | Travel | 0.629 | **0.884** | 0.863 | 0.871 |
| | Medicare | 0.674 | 0.892 | 0.855 | **0.895** |
| | Authority | 0.626 | 0.866 | 0.882 | **0.903** |
| Travel | Impact | 0.629 | 0.876 | 0.876 | **0.884** |
| | Travel | 0.676 | **0.897** | 0.895 | 0.892 |
| | Medicare | 0.626 | 0.861 | 0.861 | **0.868** |
| | Authority | 0.553 | **0.903** | 0.887 | 0.892 |
| Medicare | Impact | 0.668 | **0.863** | **0.863** | 0.855 |
| | Travel | 0.624 | 0.847 | **0.863** | 0.858 |
| | Medicare | 0.697 | 0.884 | **0.897** | 0.887 |
| | Authority | 0.632 | **0.884** | **0.884** | 0.879 |
| Authority | Impact | 0.684 | 0.871 | 0.874 | **0.882** |
| | Travel | 0.639 | 0.855 | **0.874** | 0.816 |
| | Medicare | 0.637 | 0.868 | **0.874** | 0.850 |
| | Authority | 0.668 | 0.908 | **0.934** | 0.921 |
| All | All | 0.670 | 0.918 | 0.897 | **0.927** |
| | Impact | 0.700 | 0.916 | 0.882 | **0.921** |
| | Travel | 0.645 | 0.897 | 0.882 | **0.926** |
| | Medicare | 0.674 | 0.918 | 0.905 | **0.921** |
| | Authority | 0.663 | 0.937 | 0.921 | **0.942** |

Table 2. **Fake News Detection Performance.** The table contains the accuracy scores for all ablation conditions. Each ablation condition is with balanced positive and negative label frequencies. In the top part of the tabel (besides the "All" conditions), the bold numbers are the best per training dataset, and the underlined numbers are the best per test dataset.

besides rare cases, all models perform better in the same-topic training and testing scenarios, with higher inference accuracy scores. In many cases, either the *argument-as-embedding* or the *argument-lstm* models outperform the BERT baselines, suggesting the extra information introduced by the argumentation tags of the texts boosts the fake news detection task. More importantly, the argumentation enriched model *argument-lstm* trained on **all** articles achieved the overall best accuracy in all ablation conditions, compared to the other model architectures. While boosting same-topic training and testing performances, the argumentation-enriched models often show lower performance compared to BERT baselines in cross-topic scenarios. For instance, the *argument-lstm* model has lower accuracy scores (Acc.) in "authority-travel" ($Acc. = 0.816$ compared to BERT $Acc. = 0.855$) and "authority-medicare" ($Acc. = 0.850$ compared to BERT $Acc. = 0.868$) while having better performance in "authority-authority" ($Acc. = 0.921$ improved from $Acc. = 0.908$). Comparing the two alternatives of the argumentation-enriched models, the *argument-lstm* shows better generalization performance while the *argument-as-embedding* shows better same-topic inference performance. One possible reason is that the *argument-lstm* contains relatively fewer parameters than the *argument-as-embedding* [6], and so that the former one is less likely to overfit to topic-specific patterns.

## 5 ERROR ANALYSIS

To better understand the advantage of the argumentation-enriched model in fake news detection, we make a qualitative analysis on the advantaged cases (argumentation model made correct predictions while baseline BERT did not) and the disadvantaged cases. There are 49 advantaged cases (within which 9 with the label "credible") and 35 disadvantaged cases (27 with the label "credible"). From the perspective of fighting against fake news, correctly identifying possible fake news is more important than correctly recognizing credible news, as the failure in the former one leads to much harmful consequences than the failure in the latter one. The argumentation-enriched model has a larger advantage margin in this situation as it correctly identified 40 more fake news while failed to detect eight others, compared to the BERT baseline.

We further discuss what argumentation signals the model leverages in identifying fake news. We extract the frequency of uni-gram and bi-gram from their ADU sequence (the detected ADUs are smoothed as described in Appendix C). For instance, an article that first introduce an *assumption* and then support it with *statistics* and *testimony* will yield uni-grams of *AS*, *ST*, *TS* and bi-grams of *AS-ST*, *ST-TS*. These frequencies reflect the *argumentation strategy* of articles organizing ADUs, as illustrated by [2].

From 693 correctly predicted **fake** articles and 716 **credible** articles, we find that the fake articles has significantly less TS ($p < 0.001$ from Mann-Whitney U test), and less ST ($p < 0.001$). While **not** having less assumptions, the fake articles **do** have less *AS-TS* ($p < 0.01$), *AS-AN* ($p < 0.05$), and *AS-ST* ($p < 0.001$). Likewise, the fake articles also have less *AN-AS* ($p < 0.05$), *ST-AS* ($p < 0.001$), and *TS-AS* ($p < 0.001$). The observations suggest that credible news tends to give evidence (AN, ST, or TS) directly next to assumptions, while fake news does not. The argumentation-enhanced model gives good separation for correctly samples on these characteristics. On the other hand, on those wrongly predicted samples, none of the above significance holds. It demonstrates that, from the perspective of argumentation strategy, these samples do not have clear boundaries among the characteristics, and consequently, the argumentation-enriched model would fail.

## 6 DISCUSSION

This paper studies the role of argument mining in fake news detection. We evaluate the efficiency of existing argument detection algorithms on fake news texts. Our analysis reveals that the argumentation knowledge of texts can improve the accuracy of fake news identification, combined with typical language models. Our results add valuable insights for studies on the fake news detection topic. Besides, this study also contributes to argument mining by showing the generalizability of language models on the fake news domain, enabling future research on leveraging argument mining in detecting, analyzing, and counteracting fake news on the Internet. Resonating results from prior studies [1], our results demonstrate the topic-sensitivity of argument mining models.

We list two limitations of this study as the future study directions. Firstly, we establish the baseline performance of argumentation detection for fake news with the BERT model. This baseline can be

[6] to model the argumentation tags, the *argument-as-embedding* uses $768 \times 13$ parameters, while the *argument-lstm* uses $13 \times 64 \times 2$ parameters in the Bi-lstm with hidden size of 64, and $13 \times 2 \times 1$ in the attention layers.

further improved, and the downstream task of fake news detection can benefit from it. Secondly, as suggested by other studies [16], the BERT model can take advantage of specific linguistic cues in making judgments. Thus, its roles in both argument component detection and fake news detection need specification in future studies.

## 7 ACKNOWLEDGEMENT

The authors would like to acknowledge the support from NSF #1739413, #2027713, and AFOSR awards. Any opinions, findings, and conclusions or recommendations expressed in this material do not necessarily reflect the views of the funding sources.

## REFERENCES

[1] Ajjour, Y., Chen, W.-F., Kiesel, J., Wachsmuth, H., and Stein, B. Unit segmentation of argumentative texts. In *Proceedings of the 4th Workshop on Argument Mining* (2017), pp. 118–128.

[2] Al Khatib, K., Wachsmuth, H., Hagen, M., and Stein, B. Patterns of argumentation strategies across topics. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (2017), pp. 1351–1357.

[3] Al Khatib, K., Wachsmuth, H., Kiesel, J., Hagen, M., and Stein, B. A news editorial corpus for mining argumentation strategies. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (2016), pp. 3433–3443.

[4] Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[5] Bonet-Jover, A., Piad-Morffis, A., Saquete, E., Martínez-Barco, P., and García-Cumbreras, M. Á. Exploiting discourse structure of traditional digital media to enhance automatic fake news detection. *Expert Systems with Applications 169* (2021), 114340.

[6] Chen, E., Lerman, K., and Ferrara, E. Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. *JMIR Public Health and Surveillance 6*, 2 (2020), e19273.

[7] Conroy, N. K., Rubin, V. L., and Chen, Y. Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology 52*, 1 (2015), 1–4.

[8] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[9] Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., and Lazer, D. Fake news on twitter during the 2016 us presidential election. *Science 363*, 6425 (2019), 374–378.

[10] Karimi, H., and Tang, J. Learning hierarchical discourse-level structure for fake news detection. *arXiv preprint arXiv:1903.07389* (2019).

[11] Kula, S., Choraś, M., and Kozik, R. Application of the bert-based architecture in fake news detection. In *Conference on Complex, Intelligent, and Software Intensive Systems* (2020), Springer, pp. 239–249.

[12] Lawrence, J., and Reed, C. Argument mining: A survey. *Computational Linguistics 45*, 4 (2020), 765–818.

[13] Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., et al. The science of fake news. *Science 359*, 6380 (2018), 1094–1096.

[14] Lippi, M., and Torroni, P. Argument mining from speech: Detecting claims in political debates. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2016), vol. 30.

[15] Lippi, M., and Torroni, P. Margot: A web server for argumentation mining. *Expert Systems with Applications 65* (2016), 292–303.

[16] Niven, T., and Kao, H.-Y. Probing neural network comprehension of natural language arguments. *arXiv preprint arXiv:1907.07355* (2019).

[17] Ramshaw, L. A., and Marcus, M. P. Text chunking using transformation-based learning. In *Natural language processing using very large corpora*. Springer, 1999, pp. 157–176.

[18] Reis, J. C., Correia, A., Murai, F., Veloso, A., and Benevenuto, F. Explainable machine learning for fake news detection. In *Proceedings of the 10th ACM conference on web science* (2019), pp. 17–26.

[19] Reis, J. C., Correia, A., Murai, F., Veloso, A., and Benevenuto, F. Supervised learning for fake news detection. *IEEE Intelligent Systems 34*, 2 (2019), 76–81.

[20] Salehinejad, H., Sankar, S., Barfett, J., Colak, E., and Valaee, S. Recent advances in recurrent neural networks. *arXiv preprint arXiv:1801.01078* (2017).

[21] Shu, K., Cui, L., Wang, S., Lee, D., and Liu, H. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (2019), pp. 395–405.

[22] Shu, K., Sliva, A., Wang, S., Tang, J., and Liu, H. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter 19*, 1 (2017), 22–36.

[23] Sitaula, N., Mohan, C. K., Grygiel, J., Zhou, X., and Zafarani, R. Credibility-based fake news detection. In *Disinformation, Misinformation, and Fake News in Social Media*. Springer, 2020, pp. 163–182.

[24] Swanson, R., Ecker, B., and Walker, M. Argument mining: Extracting arguments from online dialogue. In *Proceedings of the 16th annual meeting of the special interest group on discourse and dialogue* (2015), pp. 217–226.

[25] Trautmann, D., Daxenberger, J., Stab, C., Schütze, H., and Gurevych, I. Fine-grained argument unit recognition and classification. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2020), vol. 34, pp. 9048–9056.

[26] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. *arXiv preprint arXiv:1706.03762* (2017).

[27] Visser, J., Lawrence, J., and Reed, C. Reason-checking fake news. *Communications of the ACM 63*, 11 (2020), 38–40.

[28] Wang, W. Y. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648* (2017).

[29] Zhou, X., and Zafarani, R. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR) 53*, 5 (2020), 1–40.

## A DETAILS OF ARGUMENT DETECTION ON *FAKE NEWS DATASET*

Table 3 is the extended performance table with precision and recall metrics. Table 4 reports the confusion matrix of the BERT tagging model's performance on the ADU detection task. As can be seen from the confusion matrix, the most common errors made by the model are the confusion between AS, AN, and TS. The boundaries between these three components are hard to distinguish for even humans. In news articles, there exist a number of sentences following the styles "Someone says something", or "Something according to someone". Depending on the emphasis from the semantic, these sentences can be:

1) AS: where the author use someone else's word to propose his/her own opinion;

2) AN: where the purpose of the sentence is to describe the fact that someone makes speech;

3) TS: where the purpose of the sentence is to emphasize the "something" as evidence.

It is not surprising that the machine algorithms make most of the errors between these three types of components; however, future studies may focus on improving this accuracy as it is fundamental to many topics in argument mining in news articles.

## B BERT MODEL BASICS

Let there be a text sequence consist of tokens $\mathbf{W} = \{w_1, w_2, ...w_n\}$ where $n$ represents the length of the text. The BERT model first embeds each token in the sequence into vectors $\mathbf{e}_i \in \mathbb{R}^{768}$. The embedded vectors are consists of two components: $\mathbf{e}_i = \mathbf{s}_i + \mathbf{l}_i$, where $\mathbf{s}_i$ is the semantic embedding of $w_i$ calculated through an embedding layer $\mathbf{s}_i = emb(w_i)$, and $\mathbf{l}_i$ is he location embedding [7]. The embedded sequence $\mathbf{E} = \{\mathbf{e}_1, \mathbf{e}_2, ...\mathbf{e}_n\}$ is then encoded by six layers of the transformer layers [26].

## C SMOOTHING ADUS

We adopt a smoothing algorithm to compensate for "grammatical" errors (such as ADUs leading by "I" tags or ADUs with mixed type

[7] Refer the original BERT paper [8] for detailed information about the location embedding

| | | B-AS | I-AS | B-AN | I-AN | B-ST | I-ST | B-TS | I-TS | B-CG | I-CG | O | Macro |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Webis-16 Editorial dataset* | F1 | 0.656 | 0.675 | 0.552 | 0.639 | 0.581 | 0.670 | 0.218 | 0.663 | 0.000 | 0.000 | 0.724 | 0.489 |
| | Precision | 0.587 | 0.702 | 0.496 | 0.594 | 0.608 | 0.716 | 0.165 | 0.675 | 0.000 | 0.029 | 0.791 | 0.488 |
| | Recall | 0.744 | 0.650 | 0.623 | 0.691 | 0.556 | 0.629 | 0.323 | 0.651 | 0.000 | 0.000 | 0.666 | 0.503 |
| *fake news dataset* All Articles | F1 | 0.534 | 0.695 | 0.594 | 0.692 | 0.594 | 0.680 | 0.151 | 0.556 | 0.000 | 0.000 | 0.803 | 0.476 |
| | Precision | 0.458 | 0.656 | 0.679 | 0.762 | 0.791 | 0.612 | 0.319 | 0.512 | 0.000 | 0.000 | 0.836 | 0.511 |
| | Recall | 0.630 | 0.736 | 0.525 | 0.630 | 0.457 | 0.720 | 0.086 | 0.594 | 0.000 | 0.000 | 0.769 | 0.468 |
| *fake news dataset* Credible Articles | F1 | 0.131 | 0.687 | 0.614 | 0.718 | 0.581 | 0.697 | 0.110 | 0.570 | 0.000 | 0.000 | 0.821 | 0.448 |
| | Precision | 0.073 | 0.613 | 0.708 | 0.806 | 0.871 | 0.672 | 0.288 | 0.548 | 0.000 | 0.000 | 0.821 | 0.491 |
| | Recall | 0.675 | 0.783 | 0.542 | 0.648 | 0.435 | 0.724 | 0.068 | 0.593 | 0.000 | 0.000 | 0.822 | 0.481 |
| *fake news dataset* Fake Articles | F1 | 0.575 | 0.694 | 0.588 | 0.684 | 0.577 | 0.650 | 0.139 | 0.545 | 0.000 | 0.000 | 0.798 | 0.477 |
| | Precision | 0.533 | 0.661 | 0.674 | 0.755 | 0.770 | 0.595 | 0.319 | 0.503 | 0.000 | 0.000 | 0.839 | 0.514 |
| | Recall | 0.624 | 0.730 | 0.521 | 0.625 | 0.461 | 0.718 | 0.089 | 0.595 | 0.000 | 0.000 | 0.760 | 0.466 |

Table 3. **Argument Extraction Performance on *Webis-16 Editorial dataset* and *fake news dataset*.** The full version with precision and recalls.

| | | True | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | B-AS | I-AS | B-AN | I-AN | B-ST | I-ST | B-TS | I-TS | B-CG | I-CG | O | Sum |
| Predicted | B-AS | **1009** | 129 | 205 | 37 | 24 | 6 | 491 | 156 | 0 | 0 | 147 | 2204 |
| | I-AS | 219 | **20086** | 12 | 2944 | 9 | 507 | 58 | 6149 | 0 | 0 | 658 | 30642 |
| | B-AN | 96 | 5 | **831** | 110 | 51 | 6 | 57 | 19 | 0 | 0 | 48 | 1223 |
| | I-AN | 45 | 4083 | 206 | **22546** | 3 | 570 | 20 | 1366 | 0 | 0 | 735 | 29574 |
| | B-ST | 6 | 0 | 15 | 0 | **102** | 3 | 3 | 0 | 0 | 0 | 0 | 129 |
| | I-ST | 15 | 480 | 6 | 882 | 22 | **3052** | 6 | 519 | 0 | 0 | 3 | 4985 |
| | B-TS | 15 | 3 | 158 | 19 | 0 | 0 | **103** | 13 | 0 | 0 | 12 | 323 |
| | I-TS | 34 | 1622 | 70 | 9143 | 3 | 82 | 302 | **12562** | 0 | 0 | 725 | 24543 |
| | B-CG | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **0** | 0 | 0 | 0 |
| | I-CG | 6 | 144 | 0 | 12 | 0 | 0 | 3 | 81 | 0 | **0** | 6 | 252 |
| | O | 156 | 732 | 80 | 117 | 9 | 14 | 148 | 267 | 0 | 0 | **7770** | 9293 |
| | Sum | 1601 | 27284 | 1583 | 35810 | 223 | 4240 | 1191 | 21132 | 0 | 0 | 10104 | 103168 |

Table 4. **Confusion Matrix of the ADU detection by BERT.** The rows represent the true labels of the tokens while the columns stand for the model's prediction. The bold diagonal numbers are the correct predictions.

tags) output by the BERT. We first segment ADUs by "B" and "O" tags, and use the majority type within each segment as the ADU type. For segments with the first half tokens of one type and the second half tokens of another (frequency of the two types of tokens within 0.4 to 0.6), we further split the segments in two. For instance, the sequence "O I-AS I-AS I-AN I-AN O" will be smoothed as two ADUs of AS and AN.