Are You Trying to Convince Me or Are You Trying to Deceive Me? Using Argumentation Types to Identify Deceptive News

Ricardo Muñoz Sánchez **Emilie Francis** **Anna Lindahl**
Språkbanken Text, University of Gothenburg, Sweden
{ricardo.munoz.sanchez,emilie.francis,anna.lindahl}@gu.se

Abstract

The way we relay factual information and the way we present deceptive information as truth differs from the perspective of argumentation. In this paper, we explore whether these differences can be exploited to detect deceptive political news in English. We do this by training a model to detect different kinds of argumentation in online news text. We use sentence embeddings extracted from an argumentation type classification model as features for a deceptive news classifier. This deceptive news classification model leverages the sequence of argumentation types within an article to determine whether it is credible or deceptive. Our approach outperforms other state-of-the-art models while having lower variance. Finally, we use the output of our argumentation model to analyze the differences between credible and deceptive news based on the distribution of argumentation types across the articles. Results of this analysis indicate that credible political news presents statements supported by a variety of argumentation types, while deceptive news relies on anecdotes and testimonial.

1 Introduction

The spread of disinformation has taken a toll on public trust in news media (Lee, 2024). The effects of this are reflected in social and political unrest, such as the 2016, 2020, and 2024 United States presidential elections (see Allcott and Gentzkow, 2017; Benkler et al., 2020; Arısoy Gedik, 2025, respectively) and the COVID-19 pandemic (Rocha et al., 2021). There is a widespread perception that journalists have not only failed to shield the public from disinformation, but have also contributed to its spread by aligning themselves with bad actors (Harrington et al., 2024). On top of that, there is a belief that news media prioritizes profit over veracity, treating it as some sort of advertisement (Amazeen and Wojdynski, 2019).

Although disinformation takes many forms, we focus on deception based on the definitions of news media watchdog organizations, such as Media Bias/Fact Check.[1] These are often determined based on political bias and on the amount of false information, be it of the articles themselves or of the outlets that publish them. We focus on political news, as it has become a loci of public concerns over the role that news media plays in global politics and the influence disinformation has in it (Benkler et al., 2020; Harrington et al., 2024).

Political persuasion and disinformation are closely related (Gil de Zúñiga et al., 2025). We assume that credible news aims to inform, while deceptive news attempts to persuade readers in favour of a certain viewpoint. We hypothesize that this will be reflected in the argumentation within the articles themselves. Our research questions are as follows:

RQ1: Can argumentation features be used to detect deceptive news?

RQ2: What insights can we acquire by comparing argumentation types in credible and deceptive news?

We implement a two-step approach to test this. We start by training a BERT (Devlin et al., 2019) model to identify argumentation types in English news articles. We extract argumentation features from this model and feed them to a Bi-LSTM (Hochreiter and Schmidhuber, 1997) to identify deceptive news. We go into more detail of our architecture and related design choices in Section 3.

We report the results of our experiments in Section 6. Our approach outperforms other models from the literature, having less variance compared to the other non-deterministic methods used in our

[1] https://mediabiasfactcheck.com/methodology/

Proceedings of the The 9th Workshop on Online Abuse and Harms (WOAH), pages 355–372
August 1, 2025 ©2025 Association for Computational Linguistics

experiments. We also show that feature-based models can outperform simple transformer baselines.

We do an analysis of the argumentation types between credible and deceptive news in Section 7. We show that deceptive news tends to present more anecdotes and testimonies, while credible news tends to have more assumptions supported through evidence.

2 Related Work

2.1 Misinformation and Deception Detection

Misinformation detection is a task that has arisen in order to combat the influence of false or misleading information. Oshikawa et al. (2020) note that, even though it is often framed as a binary veracity classification, it has also been framed using scales of truth (Rashkin et al., 2017; Wang, 2017) or through political bias (Potthast et al., 2018).

Research from psychology has shown that liars attempt to relieve the cognitive burden of deception by distancing themselves from their false statements (Newman et al., 2003). Similar effects have been reported when looking at "trolls" on social media (Addawood et al., 2019). However, it is important to note that veracity can be complicated to establish, which can lead to issues such as sampling biases (Zhou et al., 2021).

Ruffo et al. (2023) note that a lot of terminology in this area tends to have fuzzy or ambiguous definitions. They argue that terms such as "fake news" are often ill-defined, even in an academic setting. They mention that this blurs the lines between misinformation detection and similar tasks, such as automated fact-checking, propaganda, and hyper-partisan bias detection.

Although automated fact-checking is a distinct task that has been applied to various types of media,[2] it is also used in knowledge-based approaches to detect misinformation. An example of this is Kumar et al. (2025), who used factual statements to form knowledge graphs to provide models with updated contextually relevant information for fact-checking.

Several other approaches have used features within the text, such as syntactic (Huang et al., 2020) or discourse features (Karimi and Tang, 2019) One such approach by Ghanem et al. (2021) modelled emotional shifts throughout an article and employed the information as features for fake news detection. Oshikawa et al. (2020) note that the most commonly used content features tend to be bag-of-words features, frequency of punctuation, and psycholinguistic features from LIWC.[3]

Another common way to tackle misinformation detection uses metadata, such as social media interaction or web traffic. An example of this is a study by Baly et al. (2018), which establishes a link between news article reliability and publisher credibility by checking for the existence of a Wikipedia page or Twitter account.

Credibility, partisanship, and misinformation have also been investigated in prediction and detection tasks. Rather than explicit fact-checking, Potthast et al. (2018) argue that stylistic differences in partisan news are sufficient to detect disinformation. Potthast et al. (2018) and Baly et al. (2019) noted that hyper-partisan news articles across the political spectrum are more similar to each other in terms of style than to more balanced news.

Furthermore, the political orientation of a reader can affect how believable or factual a piece of information is perceived to be. Landreville and and (2019) note that if the political orientation of a news outlet aligns with that of its reader, it is considered to be more reliable. This is the case even if said statements are opinions instead of facts. On the other hand, Morris et al. (2020) point out that news readers in the United States tend to consider a news outlet more trustworthy if it is critical of the opposing political party. Even though this effect is present across the whole political spectrum, they note that it is stronger in conservative readers. Both of these studies point out that these effects increase the likelihood of believing disinformation as long as it aligns with our political values or is critical to those perceived to be opposing.

2.2 Argumentation Mining of News

Argumentation mining is a subfield of NLP that studies argumentation, ranging from identifying argumentative passages to analyzing argumentative structures and reasoning (Stede and Schneider, 2019; Lawrence and Reed, 2019). Argumentation mining of news media has generally focused on annotation of editorials and opinion pieces. Rocha et al. (2022) created a dataset of opinion articles in Portuguese annotated with argumentative discourse

[2] See Thorne and Vlachos, 2018 for an overview of the task up to 2018 or the yearly FEVER Workshop, organized since 2018: https://fever.ai/

[3] Linguistic Inquiry Word Count, originally introduced by Pennebaker and Francis (1999).

units,[4] argumentative components, and relations. Another corpus created by Habernal and Gurevych (2017) annotated user comments on news articles, discussion forums, and blog posts related to controversial issues in education. Similarly, Goudas et al. (2014) collected documents in Greek from social media (including news articles) and annotated them to identify sentences containing argumentation and whether they are claims or premises.

Several studies have bridged misinformation and argument mining. Rhetorical structure theory (RST) has been used to detect deceptive content (Vargas et al., 2022), while stance detection has close ties with argumentation mining (Weinzierl and Harabagiu, 2024; Saha et al., 2024) and has often been studied alongside news credibility (e.g.Kotonya and Toni, 2019 and the Fake News Challenge[5] shared task).

In this study, we use the Webis-16 dataset (Al-Khatib et al., 2016). It consists of news editorials annotated with argumentation types and information for the argumentative role they play. The paper that introduced the dataset used it to investigate patterns in argumentation strategies across various news topics. It has also been used by Ajjour et al. (2017) to identify argumentative segments in written news media.

2.3 Arguments and Persuasion in News and Politics

In a study of news editorials, El Baff et al. (2018) classified articles as challenging or reinforcing. Challenging editorials make the reader rethink their prior stance, while reinforcing editorials strengthen their prior stance. They show in a later paper (El Baff et al., 2020) that persuasive reinforcing editorials often start and end with negative tone. They also observe that persuasive articles often start with an engaging hook and fortify arguments with a 'punchy' closing. On the other hand, they note that ineffective articles tend to feel inauthentic and have positive tone in the article body.

Yu et al. (2021) focuses on the emotional aspect of news articles. They show that persuasive articles leverage the reader's emotions by using loaded language and logical fallacies, such as straw-man arguments and ad-hominem attacks.

Political speech in online news media often takes the form of advertisement, mimicking the style and format of the platform on which it appears (Amazeen and Wojdynski, 2019). Nelson et al. (2021) show that readers are not very successful at identifying this type of advertising. They also note that, unlike commercial ads, regulations guiding truth in advertising are typically not applied to political content (Nelson et al., 2021). Given the impact that news media has on society, this makes for a powerful political tool (Konieczny, 2023).

[4] Argumentative units are categorized according to the role they play in argumentation.

[5] http://www.fakenewschallenge.org/

Type	Explanation
Anecdote	Provides evidence through examples or personal experiences.
Assumption	Assumptions that need support to be accepted by the reader.
Testimony	Provides evidence by quoting a figure of authority.
Other	Establishes shared knowledge, presents statistics, or does not add to the argument.

Table 1: Argument types and their definitions.

3 Our Approach

As we want to analyze whether the argumentative structure of an article can be used to identify deception, we perform a two-step process inspired by Alhindi et al. (2021). This allows us to determine whether our model learns from argumentation in the text and provides us with information about the types of argumentation in news articles, which we analyze in Section 7. We use argumentation types instead of argumentation roles (such as premise and conclusion), as Al-Khatib et al. (2016) note that the latter encode the strategy an author uses to persuade readers.

We split the articles into an ordered set of sentences and assign them an argumentation type with BERT. We use four argumentation types: *anecdote*, *assumption*, *testimony*, and *other*. Table 1 explains the different argumentation types, while Section 4.1 details why these specific ones were chosen. We use the final transformer layer from this model to generate sentence embeddings, which are fed to a Bi-LSTM model to classify news articles as credible or deceptive. The architecture of this process is represented in Figure 1.

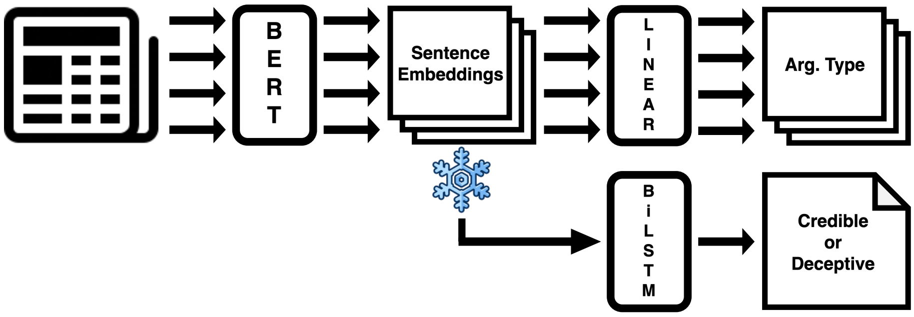


Figure 1: Diagram of our proposed approach and its different components. The argumentation type classifier (on top) assigns an argumentation type to each sentence. The deceptive news detection model (on the bottom) uses sentence embeddings to determine whether an article is credible or deceptive. These embeddings are taken from the frozen argumentation type model, represented in the diagram by a snowflake.

While decoder-only models[6] have been shown to work well with argumentation (El Baff et al., 2024), they have give mixed results in disinformation detection tasks (Hu et al., 2024; Su et al., 2024). We do not use them for this study to avoid the introduction of artifacts in any part of our pipeline. We use BERT over similar but larger models to avoid overfitting as our argumentation type dataset is quite small. Exploratory experiments revealed that BERT struggles with the least represented argument type (see Section 4.1). This issue is likely to be more prominent in larger models.

van Dijk (1989) and Yarlott et al. (2018) note that ordering is important for the argumentative role of text in written news media. We chose a Bi-LSTM for the deceptive news detection task, as these models will intrinsically take into account the ordering of the argumentation types.

We explain each model and how they are implemented in more detail throughout the rest of this section. The specific hyper-parameters used for our experiments can be found in Appendix A.

Argumentation Type Classifier: We fine-tune a BERT model[7] on the argumentation type dataset. This model is shown individual sentences and must assign an argumentation type to each of them. We use the output of the [CLS] token from the final transformer layer for classification. As the latter BERT layers typically learn task-specific features (Rogers et al., 2020), we expect the final layer to encode argumentation-related features for the whole sentence. This model is then frozen for the rest of the experiments to prevent its weights from changing later on, thus making sure that it retains its knowledge about argumentation types intact.

Deceptive News Classifier: Given a news article, we split it into sentences. These sentences are passed through the now-frozen argumentation type classifier. We use the output of the final transformer layer corresponding to the [CLS] token as a sentence embedding. These embeddings are then fed to a Bi-LSTM model to determine whether the article is credible or deceptive.

4 Datasets

In this section, we describe the different datasets used in our two tasks. For the argumentation type classification task, we use the Webis-Editorials-16 dataset (Section 4.1). For the deceptive news detection task, we use two datasets: one with article-level annotations (Section 4.2) and another with source-level annotations (Section 4.3).

All three datasets contain news articles in English collected prior to 2020. Although the landscape of deceptive news and misinformation is likely to have changed since these articles were originally published, these datasets are still valuable as they only contain human-generated news. Machine-generated mis- and disinformation is very different from that generated by humans (Tewari et al., 2021) and detecting it is another task in and of itself (Beigi et al., 2024). Therefore, we choose established datasets from before content produced by generative language models flooded the web.

4.1 Webis-16 Dataset

The Webis-Editorials-16 dataset (or Webis-16 for short) was originally introduced by Al-Khatib et al. (2016). It consists of news editorials in English from three established news sources. One hundred editorials were selected for each of the three

[6] Such as OpenAI's GPT line of models.

[7] https://huggingface.co/google-bert/bert-base-uncased

Class	PolitiFact	FakeNews-2018
Credible	131	8,117
Deceptive	242	14,962
Credible	372	23,079

Table 2: Number of articles for each class after having filtered the datasets for length.

publishers. The included texts were originally published between December 2014 and January 2015 and were selected such that they would have a length of at least 250 word and had at least five comments. This dataset does not to distinguish between true and false statements, which is beneficial for our task as it reduces the risk of introducing artifacts into the deceptive news classification task.

Each token in the text was assigned one of eight labels. Six of these labels correspond to argumentation types, namely *common ground*, *assumption*, *testimony*, *statistics*, *anecdote*, and *other*. The *continuation* label means that a token has the same argumentation type as the next argumentation type label that appears, thus forming spans of argumentative units. Some tokens, such as punctuation, are labelled as *non-argumentative* as they do not form part of an argumentative unit, regardless of surrounding tokens.

It is important to note that argumentative units do not necessarily correspond to sentences. A sentence may contain multiple clauses, each its own argumentative unit. It is also possible for argumentative units to span two or more sentences. This poses a problem for our task. Although token-level classification is useful for studying argumentative units in the context of argumentation types, the difference in granularity can harm our downstream task as it is document-level classification.

Because of this, we cast the argumentation type labels so that each sentence has one and only one. We do this in the following way: (i) *continuation* labels take on the same label as the next argumentation type label; (ii) sentences with more than one argumentation type label are discarded; and (iii) if all tokens in a sentence not labelled *non-argumentative* share an argumentation type label, the whole sentence gets that label.

One issue that arose during exploratory analysis was that models performed very well on the majority class, but very poorly on under-represented classes. This was still the case when applying early stopping and keeping the best performing checkpoint. A model that severely under-performs on one or more of the classes will not allow for good analysis of the data. To get around this issue, we collapsed some of the minority classes together. The labels for *assumption*, *anecdote*, and *testimony* were preserved, while *common ground* and *statistics* were grouped into *other*. This resulted in better performance of the deceptive news classifier and allowed us to conduct a more accurate analysis of argumentation in the articles. Appendix B goes into more detail on how the number of labels was chosen.

4.2 PolitiFact

FakeNewsNet, originally introduced by Shu et al. (2020), contains the PolitiFact and GossipCop datasets. They have article-level annotations obtained from their name-sake fact-checking websites.[8] The labels are binary and represent verifiable truth. As our analysis focuses on political news, only the PolitiFact dataset is used in our approach. It originally contained 948 articles accessible through links provided by the authors to preserve copyright. Unfortunately, many articles are no longer retrievable due to broken links.

Article length has been shown to be a strong indicator of deceptive news (Levi et al., 2019). We filter the dataset to ensure both credible and deceptive articles are within an range of 100 to 800 tokens. This helps make sure the model learns from argumentative structure rather than length. Motivation for these bounds can be found in Appendix C. The final number of articles after filtering can be found in Table 2.

4.3 FakeNews-2018 Dataset

The FakeNews-2018 dataset, originally introduced by Francis (2018), contains over 81,000 political news articles in English collected from various sources from the U.S., Canada, and the U.K. published between 2013 and 2017. Articles are labelled as credible or deceptive based on the source, according to the factuality and credibility scores from Media Bias/Fact Check,[9] AllSides,[10] and Ad-Fontes Media[11] to categorize sources as credible or deceptive.

[8] https://www.politifact.com/ and http://www.gossipcop.com/, now defunct.
[9] https://mediabiasfactcheck.com/
[10] https://www.allsides.com/
[11] https://adfontesmedia.com/

Some sources labelled as deceptive in the dataset are described as satire. Even though satirical news differ from non-satirical news (Horne and Adali, 2017), research has shown it is challenging to distinguish satire from deceptive news (Horne and Adali, 2017; Rubin et al., 2015). While satirical news are meant to be entertainment, disinformation outlets often present themselves as satire to protect themselves from legal consequences (Golbeck et al., 2018). Even when this is not the case, satirical news has the potential to mislead readers through its mimicry of actual news (Francis, 2024).

5 Baselines

We compare our argumentation type classifier against a random classifier and a majority class baseline. This is done to ensure the model is actually learning from the data and not simply assigning labels arbitrarily. We focus on both general performance and performance on the lowest scoring label.

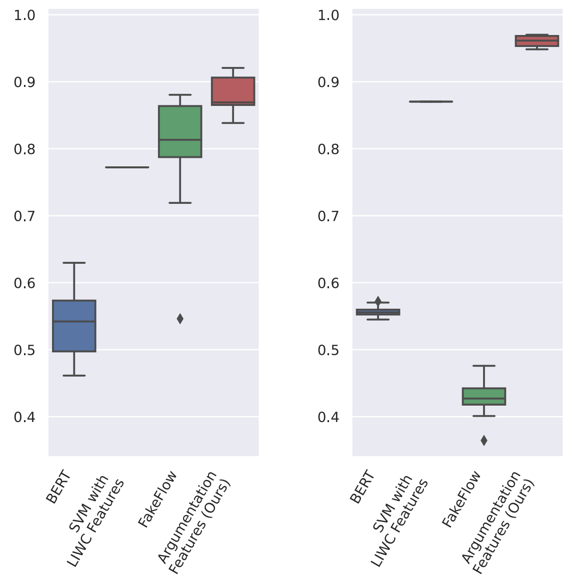
We compare our approach to three models: a BERT classifier, an SVM using LIWC[12] features, and FakeFlow (Ghanem et al., 2021). We choose BERT as it has been shown to perform well for a variety of tasks and is simple to implement. Classical machine learning models using LIWC features have been used successfully for deceptive news detection in the past (e.g. Che et al., 2018; Pérez-Rosas et al., 2018). We follow the implementation of Horne and Adali (2017), using an SVM classifier and the same feature selection process. The final model used for comparison is FakeFlow, which uses a CNN (Kim, 2014) to model article topics and a Bi-GRU (Cho et al., 2014) to model emotions in the text.

6 Results and Discussion

Throughout this section we present the quantitative results of both the argumentation type and deceptive news classification tasks. Appendix D contains additional tables with more detailed results from our experiments.

Each experiment was run multiple times in order to assess not only the performance of the models, but also their variance across runs. Only the random seed was changed across runs, all other hyperparameters remained the same. We performed

[12]Linguistic Inquiry Word Count, originally introduced by Pennebaker and Francis (1999).



(a) Weighted F1 scores for the PolitiFact dataset.

(b) Weighted F1 scores for the FakeNews-2018 dataset.

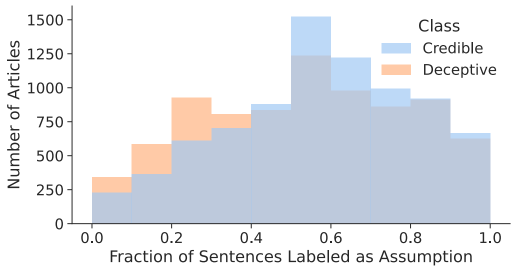
Figure 2: Weighted F1 scores for the deceptive news classification task. Our model (in red) outperforms the other models on average.

five runs for each argumentation type classification model. The deceptive news models were also trained an additional five times for each, resulting in a total of 25 runs. This allows us to assess the variance of the deceptive news model not only in terms of the training process but also of the representations it was fed.

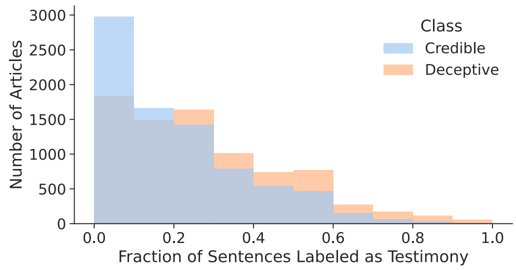
Argumentation Type Classifier: Our classifier achieves an average weighted F1 score of 0.84, which is significantly higher than those of the random and majority baselines (0.56 and 0.61, respectively). It is important to verify the F1 score of the worst-performing label at this step. This is used in analyzing the argumentation types of news articles (see Section 7) and is the motivation for collapsing some of the labels into a single one (see Section 4.1 and Appendix B). The lowest F1 score observed is 0.47 for the *other* class, which is a large improvement over 0.04 for the same class using the random baseline.

Deceptive News Classifier: As shown in Figure 2, using argumentation features outperforms the other models. There is a noticeable improvement over BERT and the SVM with LIWC features for both datasets.

We notice different patterns when comparing our model with FakeFlow. For the PolitiFact dataset, both models show an overlap in performance. How-



(a) Distribution of the *assumption* label



(b) Distribution of the *testimony* label

Figure 3: Distributions for the ratios of the argumentation types *assumption* and *testimony* in the FakeNews-2018 dataset. Deceptive news tends to make less assumptions and presents more testimonies.

ever, the three lower quartiles for FakeFlow are lower than the higher three ones for the model with argumentation features, meaning that the latter performs better on average. When looking at the FakeNews-2018 dataset, we notice that FakeFlow performs much worse than on any of the other models. Previous research has shown that features of deceptive news can be topic dependent, which may explain why some models under-perform on specific deception detection datasets (Francis, 2024).

Overall, our approach showed a lower variance in its performance when compared to the other statistical models (i.e. BERT and FakeFlow). This indicates it is more stable and less prone to the effects of randomness, such as the chosen random seed.

7 How Do People Argue in Deceptive News?

Given an argumentation type label and an article, we look at the fraction of sentences within the article with that label. We then compare how these values are distributed in each deceptive news dataset (see Figure 3). To ensure balanced sample sizes for the analysis, we under-sample the most represented class for each dataset.

We use a two-tailed Kolmogorov–Smirnov test (Hodges, 1958) to determine whether the distributions for credible and deceptive news are different[13] and, if so, how much they differ. It is important to note that all four distributions are related to each other as the ratios for a given article must sum up to one.[14] Thus, we must apply a Bonferroni correction for $n = 4$. That means that we need a p-value of 0.0125 instead of 0.05 to be able to reject the null hypothesis that the distributions for the credible and deceptive news articles are the same.

[13]Given that we are dealing with distributions of ratios, we can safely assume that they are not normally distributed.

[14]This is because each of the sentences in an article must have one and only one of the four argumentation type labels.

Label	PolitiFact	FakeNews-2018
Anecdote	***0.23***	0.02
Assumption	0.10	***0.10***
Testimony	***0.24***	***0.17***
Other	***0.27***	***0.15***

Table 3: Values of the Kolmogorov–Smirnov test, denoting the largest difference in the cumulative distribution functions. Statistically significant results are highlighted. Due to the Bonferroni correction, we need a p-value of 0.0125 to reject the null hypothesis.

The Kolmogorov–Smirnov statistic, shown in Table 3, tells us the largest difference between the two cumulative distribution functions. Excluding *anecdote* in the FakeNews-2018 dataset and *assumption* in PolitiFact, the results are statistically significant and show a large difference.

We will go over the differences between the distributions of the four labels, focusing on the FakeNews-2018 dataset as we consider that these distributions can give us potentially interesting insights.

When looking at the distribution of the *anecdote* argumentation type, we notice that anecdotes appear more often in articles labelled *fake* in the PolitiFact dataset. Usage of anecdotes may be a strategy used by deceptive news outlets to strengthen arguments in lieu of factual evidence. Previous literature has also noted that more persuasive articles use logical fallacies, such as arguments from anecdote, which leverage readers' emotions (Yu et al., 2021). Meanwhile the Komogorov-Smirnov statistic shows that the difference on the FakeNews-2018 dataset is small and not statistically significant.

In general, the label *assumption* is the most evenly distributed across the articles, regardless of the dataset or whether they are deceptive or not. As we can see in Figure 3, assumptions are less represented in deceptive news in the FakeNews-2018 dataset.

The label *assumption* appears more often in articles, regardless of whether they are deceptive or not. In contrast, the other labels tend to represent a small proportion of the sentences of an article. This does not mean that there are no differences between credible and deceptive news, as assumptions are less represented in deceptive news in the FakeNews-2018 dataset (as shown in Figure 3). Gelfert (2018) notes that the modern wave of disinformation stems partly from conspiracy theories. Conspiracy theorists avoid making explicit assumptions to avoid accountability for their claims, using the excuse of "just asking questions" (Egelhofer and Lecheler, 2019). As mentioned previously, the difference between the distribution for credible and deceptive news is not statistically significant for the PolitiFact dataset, likely due to the small size of the dataset.

The *testimony* label is represented more in deceptive news for both datasets. Figure 3 shows this for the FakeNews-2018 dataset. This may be related to the use of news as a medium for political advertising (Nelson et al., 2021). Studies have shown that testimonials positively impact consumer bias and that consumers identify more strongly with testimonials from individuals they consider peers (Shimp et al., 2007; Appiah, 2007). It has also been observed that partisan loyalty has an effect on believability, as readers are more likely to report information from sources that share their political affiliation as factual (Morris et al., 2020; Landreville and and, 2019). On the other hand, this could also be due to deceptive news using fallacious strategies such as appealing to authority (Yu et al., 2021).

The *other* label is represented the least in both datasets and both types of news, but appears more in credible news articles than in deceptive ones. It is important to note that the *other* label contains the *statistics* and *ground-truth* labels from the original Webis-16 dataset (as noted in Section 4.1). This suggests that credible news substantiates claims more often than deceptive.

As mentioned previously in this section, Figure 3 shows the distributions for the labels *assumption* and *testimony* in the FakeNews-2018 dataset. The histograms comparing the distributions for all the argumentation type labels can be found in Appendix E.

8 Conclusions

Factuality in news media is closely related to similar phenomena, such as partisan bias, propaganda, and satire (Ruffo et al., 2023). The rapid spread of deceptive news and misinformation has been linked to instability in the global political climate, as well as erosion of trust in news media (Lee, 2024). (Gelfert, 2018) and (Harrington et al., 2024) argue that it is important to study these complex phenomena in order to mitigate the risks and consequences they engender.

In this paper we hypothesized that argumentation in credible and deceptive political news articles would differ as a reflection of their role as informers or vectors for ideology. We proposed an approach exploiting argumentation types of sentences within an article to detect deceptive news. On average, our approach outperformed three models from the existing literature, namely BERT, an SVM with LIWC features, and FakeFlow. It also shows a lower variance than the non-deterministic baselines.

Some interesting patterns appear when analyzing the distributions of argumentation types. We found that deceptive articles tend to use more testimonies and, for one of the datasets, more anecdotes. Although credible news tend to have more assumptions, they appear to support them with evidence or by establishing shared knowledge. This matches previous findings from the literature that point out that deceptive news uses logical fallacies, such as overusing anecdotes or by appealing to authority (Yu et al., 2021).

It is important to note that the work we present in this paper is not any sort of "truth detector". Our model was trained and tested to be used in news articles and should only be used for that kind of media. The datasets have binary truth annotations and were curated with that purpose in mind. This means that things living in the in-between of truth and falsehood might potentially be misrepresented. Moreover, there are different kinds of mis- and disinformation (such as propaganda or hyper-partisan news) that are not explicitly studied in the present paper to better isolate features pertaining deceptive news.

The results of this study show that stylistic features, such as argumentation type, can improve classification performance and enrich our understanding of complex phenomena such as deceptive news and misinformation. Not only that, but they can also help develop systems that are both more interpretable and perform as well as other classification systems, if not better.

It is also important to note that we focus on the style of the text rather than on its content. One of our assumptions is that outlets publishing deceptive content online do so knowingly. This ignores the possibility that people who write deceptive news articles legitimately believe what they are writing. It also ignores propaganda in news media that is often regarded as trustworthy, be it backed by the State and/or by for-profit organizations.

Limitations

A possible limitation of our work could be the scope of the data. To the best of our knowledge, the Webis-16 dataset is one of the most thoroughly annotated news media datasets for argumentation types. However, the editorials it contains come from only three publishers. Despite this, we achieve good results in our downstream application. It is also important to note that Lindahl (2024) argues that it can be complicated to annotate argumentation in text due to ambiguity or multiple plausible interpretations.

Moreover, the annotations of this dataset do not take veracity into account. This makes it so that we can properly model argumentation on its own, without introducing biases in the deceptive news classification task. It is not possible to do joint training for the whole pipeline for that reason.

In a similar vein, the data we use for the deceptive news detection task comes predominantly from English outlets in the United States, Canada, and the United Kingdom. Furthermore, previous studies show that features of deceptive news can vary depending on news topic Francis (2024). Therefore our results might not generalize well to other languages, cultural contexts, or topics.

Regardless of these limitations, we consider our results to be useful in showcasing how other areas of NLP can give us a deeper insight into how deceptive news works. We encourage people using our methodology in different linguistic or cultural contexts to verify that is is an appropriate approach before doing any sort of implementation.

Ethical Considerations

The study of automatic detection of disinformation can be a complicated task. There is always the risk of the models being misused due to maliciousness, lack of information, or misinterpreting the purpose of the model.

An example of the first case could be a government or company looking to censor news articles that show them in an unfavourable light. Even though some of the assumptions we made in this paper might not hold true in this case, models that classify news articles could potentially be repurposed for other tasks.

Another issue could be blindly trusting the outputs of the model. Given that our model statistically selects the class that an article is most likely to belong to, there is always the risk of it being wrong. Because of this, it is important to always keep a human-in-the-loop approach when using these kinds of models.

People may also mistakenly use these kinds of models as a "truth detector" with other kinds of media. We have discussed this issue in the Limitations Section.

On top of that, there are the issues of where we get the data from and how it is annotated. Even though the datasets we used obtain their annotations from independent fact-checking organizations, there is always the risk of conflicts of interest or unstated agendas.

Even though we take steps to mitigate these issues, we are aware that some of them might still linger, especially those regarding possible misuse of the model.

References

Aseel Addawood, Adam Badawy, Kristina Lerman, and Emilio Ferrara. 2019. Linguistic cues to deception: Identifying political trolls on social media. *Proceedings of the International AAAI Conference on Web and Social Media*, 13(01):15–25.

Yamen Ajjour, Wei-Fan Chen, Johannes Kiesel, Henning Wachsmuth, and Benno Stein. 2017. Unit segmentation of argumentative texts. In *Proceedings of the 4th Workshop on Argument Mining*, pages 118–128, Copenhagen, Denmark. Association for Computational Linguistics.

Khalid Al-Khatib, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein. 2016. A news editorial corpus for mining argumentation strategies. In *Proceedings of COLING 2016, the*

26th International Conference on Computational Linguistics: Technical Papers, pages 3433–3443, Osaka, Japan. The COLING 2016 Organizing Committee.

Tariq Alhindi, Brennan McManus, and Smaranda Muresan. 2021. What to fact-check: Guiding check-worthy information detection in news articles through argumentative discourse structure. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 380–391, Singapore and Online. Association for Computational Linguistics.

Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–236.

Michelle A. Amazeen and Bartosz W. Wojdynski. 2019. Reducing native advertising deception: Revisiting the antecedents and consequences of persuasion knowledge in digital news contexts. *Mass Communication and Society*, 22(2):222–247.

Osei Appiah. 2007. The effectiveness of "typical-user" testimonial advertisements on black and white browsers' evaluations of products on commercial websites: Do they really work? *Journal of Advertising Research*, 47:14–27.

Cansu Arısoy Gedik. 2025. The role of ai-driven content, smart technologies, and disinformation in the 2024 u.s. presidential elections. *UPA Strategic Affairs*, 6(1):177–202.

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, Brussels, Belgium. Association for Computational Linguistics.

Ramy Baly, Georgi Karadzhov, Abdelrhman Saleh, James Glass, and Preslav Nakov. 2019. Multi-task ordinal regression for jointly predicting the trustworthiness and the leading political ideology of news media. 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 2109–2116, Minneapolis, Minnesota. Association for Computational Linguistics.

Alimohammad Beigi, Zhen Tan, Nivedh Mudiam, Canyu Chen, Kai Shu, and Huan Liu. 2024. Model attribution in llm-generated disinformation: A domain generalization approach with supervised contrastive learning. In *2024 IEEE 11th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10.

Yochai Benkler, Casey Tilton, Bruce Etling, Hal Roberts, Justin Clark, Robert Faris, Jonas Kaiser, and Carolyn Schmitt. 2020. Mail-in voter fraud: Anatomy of a disinformation campaign.

Xunru Che, Danaë Metaxa-Kakavouli, and Jeffrey T. Hancock. 2018. Fake news in the news: An analysis of partisan coverage of the fake news phenomenon. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '18 Companion, page 289–292, New York, NY, USA. Association for Computing Machinery.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Jana Laura Egelhofer and Sophie Lecheler. 2019. Fake news as a two-dimensional phenomenon: A framework and research agenda. *Annals of the International Communication Association*, 43(2):97–116.

Roxanne El Baff, Khalid Al Khatib, Milad Alshomary, Kai Konen, Benno Stein, and Henning Wachsmuth. 2024. Improving argument effectiveness across ideologies using instruction-tuned large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4604–4622, Miami, Florida, USA. Association for Computational Linguistics.

Roxanne El Baff, Henning Wachsmuth, Khalid Al-Khatib, and Benno Stein. 2018. Challenge or empower: Revisiting argumentation quality in a news editorial corpus. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 454–464, Brussels, Belgium. Association for Computational Linguistics.

Roxanne El Baff, Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2020. Analyzing the Persuasive Effect of Style in News Editorial Argumentation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3154–3160, Online. Association for Computational Linguistics.

Emilie Francis. 2018. MisInfoWars: A linguistic analysis of deceptive and credible news. *Master Thesis. Simon Fraser University*.

Emilie Francis. 2024. Variation between credible and non-credible news across topics. In *The First International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security*, pages 86–96.

Axel Gelfert. 2018. Fake news: A definition. *Informal Logic*, 38(1):84–117.

Bilal Ghanem, Simone Paolo Ponzetto, Paolo Rosso, and Francisco Rangel. 2021. FakeFlow: Fake news detection by modeling the flow of affective information. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 679–689, Online. Association for Computational Linguistics.

Homero Gil de Zúñiga, Pablo González-González, and Manuel Goyanes. 2025. Pathways to political persuasion: Linking online, social media, and fake news with political attitude change through political discussion. *American Behavioral Scientist*, 69(2):240–261.

Jennifer Golbeck, Matthew Mauriello, Brooke Auxier, Keval H. Bhanushali, Christopher Bonk, Mohamed Amine Bouzaghrane, Cody Buntain, Riya Chanduka, Paul Cheakalos, Jennine B. Everett, Waleed Falak, Carl Gieringer, Jack Graney, Kelly M. Hoffman, Lindsay Huth, Zhenya Ma, Mayanka Jha, Misbah Khan, Varsha Kori, Elo Lewis, George Mirano, William T. Mohn IV, Sean Mussenden, Tammie M. Nelson, Sean Mcwillie, Akshat Pant, Priya Shetye, Rusha Shrestha, Alexandra Steinheimer, Aditya Subramanian, and Gina Visnansky. 2018. Fake news vs satire: A dataset and analysis. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '18, page 17–21, New York, USA. Association for Computing Machinery.

Theodosis Goudas, Christos Louizos, Georgios Petasis, and Vangelis Karkaletsis. 2014. Argument extraction from news, blogs, and social media. In Aristidis Likas, Konstantinos Blekas, and Dimitris Kalles, editors, *Artificial Intelligence: Methods and Applications*, volume 8445, pages 287–299. Springer International Publishing. Series Title: Lecture Notes in Computer Science.

Ivan Habernal and Iryna Gurevych. 2017. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179.

Stephen Harrington, Axel Bruns, Phoebe Matich, Daniel Angus, Edward Hurcombe, and Nadia Jude. 2024. 'big lies': understanding the role of political actors and mainstream journalists in the spread of disinformation. *Media International Australia*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780. Conference Name: Neural Computation.

J. L. Hodges. 1958. The significance probability of the smirnov two-sample test. *Arkiv för Matematik*, 3(5):469–486.

Benjamin D. Horne and Sibel Adali. 2017. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In *Eleventh International AAAI Conference on Web and Social Media*.

Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. 2024. Bad actor, good advisor: exploring the role of large language models in fake news detection. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press.

Yen-Hao Huang, Ting-Wei Liu, Ssu-Rui Lee, Fernando Henrique Calderon Alvarado, and Yi-Shin Chen. 2020. Conquering cross-source failure for news credibility: Learning generalizable representations beyond content embedding. In *Proceedings of The Web Conference 2020*, pages 774–784. Association for Computing Machinery.

Hamid Karimi and Jiliang Tang. 2019. Learning hierarchical discourse-level structure for fake news detection. 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 3432–3442, Minneapolis, Minnesota. Association for Computational Linguistics.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2019. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations*.

Marcin Konieczny. 2023. Ignorance, disinformation, manipulation and hate speech as effective tools of political power. *Policija i sigurnost*, 32(2):123–134.

Neema Kotonya and Francesca Toni. 2019. Gradual argumentation evaluation for stance aggregation in automated fake news detection. In *Proceedings of the 6th Workshop on Argument Mining*, pages 156–166, Florence, Italy. Association for Computational Linguistics.

Anuj Kumar, Pardeep Kumar, Abhishek Yadav, Satyadev Ahlawat, and Yamuna Prasad. 2025. KG-FakeNet: A knowledge graph-enhanced model for fake news detection. In *Proceedings of the Workshop on Generative AI and Knowledge Graphs (GenAIK)*, pages 109–122, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Kristen D. Landreville and Cassie Niles and. 2019. "and that's a fact!": The roles of political ideology, psrs, and perceived source credibility in estimating factual content in partisan news. *Journal of Broadcasting & Electronic Media*, 63(2):177–194.

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Francis L. F. Lee. 2024. Disinformation perceptions and media trust: The moderating roles of political trust and values. *International Journal of Communication*, 18:23.

Or Levi, Pedram Hosseini, Mona Diab, and David Broniatowski. 2019. Identifying nuances in fake news vs. satire: Using semantic and linguistic cues. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 31–35, Hong Kong, China. Association for Computational Linguistics.

Anna Lindahl. 2024. Disagreement in argumentation annotation. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 56–66, Torino, Italia. ELRA and ICCL.

David S. Morris, Jonathan S. Morris, and Peter L. Francia and. 2020. A fake news inoculation? fact checkers, partisan identification, and the power of misinformation. *Politics, Groups, and Identities*, 8(5):986–1005.

Michelle R. Nelson, Chang Dae Ham, and Eric Haley. 2021. What do we know about political advertising? not much! political persuasion knowledge and advertising skepticism in the united states. *Journal of Current Issues & Research in Advertising*, 42(4):329–353.

Matthew L. Newman, James W. Pennebaker, Diane S. Berry, and Jane M. Richards. 2003. Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin*, 29(5):665–675.

Ray Oshikawa, Jing Qian, and William Yang Wang. 2020. A survey on natural language processing for fake news detection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6086–6093, Marseille, France. European Language Resources Association.

James Pennebaker and M Francis. 1999. Linguistic inquiry and word count: LIWC, 1999. Erlbaum Publishers.

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, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 231–240, Melbourne, Australia. Association for Computational Linguistics.

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, Copenhagen, Denmark. Association for Computational Linguistics.

Gil Rocha, Luís Trigo, Henrique Lopes Cardoso, Rui Sousa-Silva, Paula Carvalho, Bruno Martins, and Miguel Won. 2022. Annotating arguments in a corpus of opinion articles. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1890–1899, Marseille, France. European Language Resources Association.

Yasmim Mendes Rocha, Gabriel Acácio de Moura, Gabriel Alves Desidério, Carlos Henrique de Oliveira, Francisco Dantas Lourenço, and Larissa Deadame de Figueiredo Nicolete. 2021. The impact of fake news on social media and its influence on health during the COVID-19 pandemic: a systematic review. *Journal of Public Health*.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Victoria Rubin, Nadia Conroy, and Yimin Chen. 2015. Towards news verification: Deception detection methods for news discourse. In *Proceedings of the Hawaii International Conference on System Sciences (HICSS48)*, pages 5–8.

Giancarlo Ruffo, Alfonso Semeraro, Anastasia Giachanou, and Paolo Rosso. 2023. Studying fake news spreading, polarisation dynamics, and manipulation by bots: A tale of networks and language. *Computer Science Review*, 47:100531. Publisher: Elsevier.

Rudra Ranajee Saha, Laks V. S. Lakshmanan, and Raymond T. Ng. 2024. Stance detection with explanations. *Computational Linguistics*, 50(1):193–235.

Terence A. Shimp, Stacy L. Wood, and Laura Smarandescu. 2007. Self-generated advertisements: Testimonials and the perils of consumer exaggeration. *Journal of Advertising Research*, 47:453–461.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. FakeNewsNet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big Data*, 8(3):171–188.

Manfred Stede and Jodi Schneider. 2019. *Argumentation Mining*, 1st edition. Number 40 in Synthesis lectures on human language technologies. Morgan & Claypool Publishers.

Jinyan Su, Claire Cardie, and Preslav Nakov. 2024. Adapting fake news detection to the era of large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1473–1490, Mexico City, Mexico. Association for Computational Linguistics.

Shubhra Tewari, Renos Zabounidis, Ammina Kothari, Reynold Bailey, and Cecilia Ovesdotter Alm. 2021. Perceptions of human and machine-generated articles. *Digital Threats*, 2(2).

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.

Teun A. van Dijk. 1989. Dimensions of discourse. In *Handbook of Discourse Analysis*, 3. print edition, pages 104–112. Academic Press.

Francielle Vargas, Jonas D'Alessandro, Zohar Rabinovich, Fabrício Benevenuto, and Thiago Pardo. 2022. Rhetorical structure approach for online deception detection: A survey. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5906–5915, Marseille, France. European Language Resources Association.

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, Vancouver, Canada. Association for Computational Linguistics.

Maxwell Weinzierl and Sanda Harabagiu. 2024. Discovering and articulating frames of communication from social media using chain-of-thought reasoning. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1617–1631, St. Julian's, Malta. Association for Computational Linguistics.

W. Victor Yarlott, Cristina Cornelio, Tian Gao, and Mark Finlayson. 2018. Identifying the discourse function of news article paragraphs. In *Proceedings of the Workshop Events and Stories in the News 2018*, pages 25–33, Santa Fe, New Mexico, U.S.A. Association for Computational Linguistics.

Seunghak Yu, Giovanni Da San Martino, Mitra Mohtarami, James Glass, and Preslav Nakov. 2021. Interpretable propaganda detection in news articles. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1597–1605, Held Online. INCOMA Ltd.

Xiang Zhou, Heba Elfardy, Christos Christodoulopoulos, Thomas Butler, and Mohit Bansal. 2021. Hidden biases in unreliable news detection datasets. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2482–2492, Online. Association for Computational Linguistics.

A Hyperparameters of the Models

In this appendix we present the hyperparameters and other implementation details from our models.

A.1 Argumentation Type Classifier

The argumentation type classifier we used was implemented using the HuggingFace[15] package for python[16] using a PyTorch[17] backend.

We used the model `bert-base-uncased` from the Transformers package. For this, we used the class `AutoModelForSequenceClassification`.

The hyperparameters used were the default ones except for the following ones:

- Evaluation strategy: steps
- Evaluation steps: 100
- Evaluation delay: 1
- Number of training epochs: 3
- Load best model at the end: True
- Per device training batch size: 8

A.2 Deceptive News Classifier

The deceptive news classifier was implemented in PyTorch using the Adam (Kingma and Ba, 2019) optimizer. We used a single Bi-LSTM layer followed by a linear layer. The last hidden states from each direction were concatenated and then fed to the linear layer for classification.

The hyperparameters we used were the following:

- Learning rate: 1e-4
- LSTM hidden dimension: 64
- Batch size: 32
- Dropout: 0.5
- Max number of epochs: 2000
- Early stopping at n steps: 15

[15] https://huggingface.co/
[16] https://www.python.org/
[17] https://pytorch.org/

B Number of Labels of the Argumentation Dataset

During our preliminary exploration of the argumentation type classifier that the least represented class was getting misclassified in all of our experiments. Thus we decided to explore the possibility of collapsing some of the least represented labels into a single one.

We took into account the macro and weighted scores of the model, as well as the F1 score of the least represented class. An important criterion when selecting the number of labels was to keep as many labels as possible. This is particularly important as we want both our deceptive news classifier to learn the most out of the argumentative structure of the articles. Moreover, we want to be able to look at the argumentation types in the articles to get further insights.

As we can see from Figure 4, the less labels we keep, the better the performance of the model. This was to be expected given that the more labels there are available, the less likely a model is to get a correct result if it is choosing randomly.

When looking a the validation scores (see Figure 5) for the deceptive news classification task, we realize that the models that kept just four labels model does slightly better than the others in average. However, it is important to note that the boxplots for all groups overlap.

We decided to keep four labels as opposed to two or three as we believe that it would help with when qualitative analysis from Section 7, while keeping more labels would mean that there is a risk that neither the argumentation nor the deceptive news classifiers would work as well as they would otherwise.

C Analyzing the Length of News Articles

While looking through the datasets during our preliminary exploration we noticed that the length of the articles varied greatly between credible and deceptive ones. The distributions of the lengths of articles can be seen in Figure 6. This length is seen in terms of tokens according to the sentence tokenizer from NLTK.[18]

We decided to only maintain articles up to a certain length for two reasons. The first one is that we want to focus on the argumentation types within an article as a way of identifying whether

[18] https://www.nltk.org/api/nltk.tokenize.sent_tokenize.html

(a) Macro F1 scores

(b) Weighted F1 scores

(c) Lowest F1 score among the classes

Figure 4: Performance on the validation split when comparing different numbers of labels for the argumentation type dataset. Unexpectedly, the fewer labels we keep, the better the performance of the model, excluding outliers.

it is deceptive or not. One way to ensure this is controlling for variables that are not relevant to our hypothesis but that a model might pick up and learn spurious correlations from, such as the length of an article. The other reason is that the length of an article impacts how its discourse units interlock (van Dijk, 1989; Yarlott et al., 2018), meaning that argumentation will differ from shorter to longer texts.

We decided to maintain articles from 100 to 800 for the PolitiFact dataset those from 100 to 500 for the FakeNews-2018 dataset as this is where the summary statistics for both distributions start to converge.

D Detailed Results for the Classification Tasks

This appendix contains tables presenting the numerical results from our models. It is meant to complement the plots and values reported in Section 6, as well as the analyses contained within.

The results from the argumentation type classification task are reported in Table 4. The results for the deceptive news classification task are reported in Tables 5 and 6 for the Politifact and FakeNews-2018 datasets, respectively.

E Argumentation Types in Deceptive News Articles

Here we present the histograms comparing the distributions of the ratio of argumentation type labels of the articles between credible and deceptive news. The analysis of how these distributions vary can be found in Section 7.

There is a plot for each argumentation type label and for each dataset. We have grouped them by argumentation type in order to more easily allow comparisons across datasets. Figure 7 contains the histograms for the *anecdote* label, Figure 8 those for *assumption*, Figure 9 those for *testimony*, and Figure 10 those for *other*.

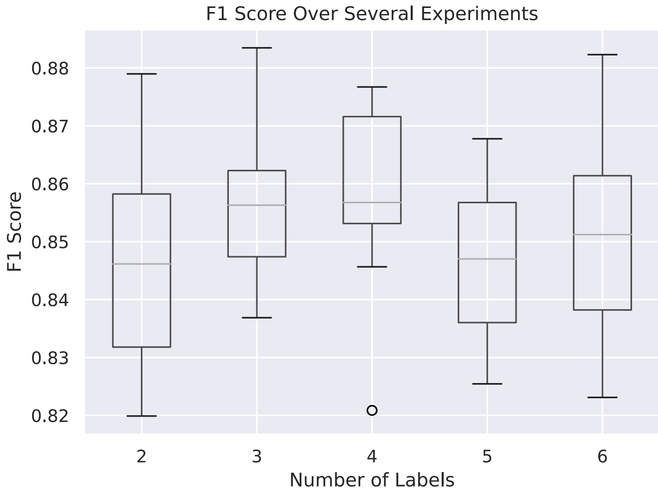
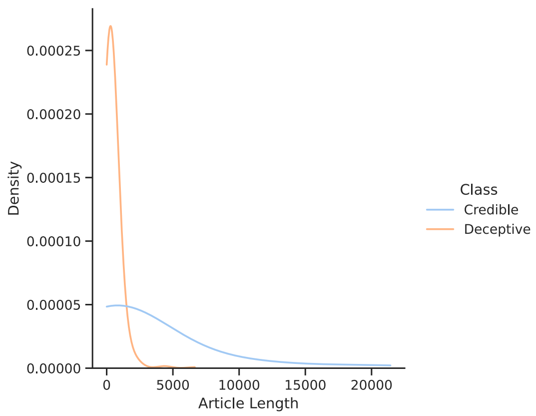


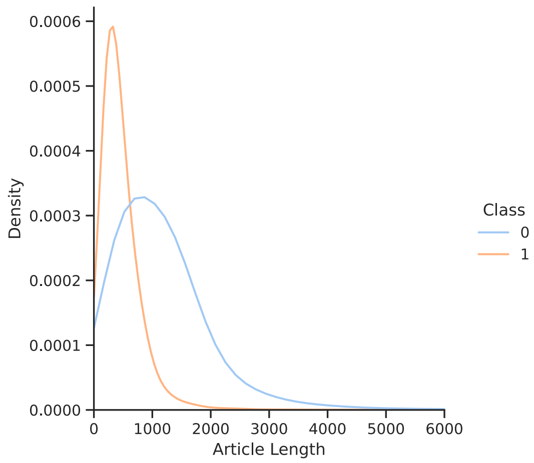
Figure 5: Boxplot from the F1-score in the validation set of the PolitiFact dataset. While the boxes overlap across all groups, we see that the one with four labels performs slightly better than the others.

	F1 Macro	F1 Weighted	Accuracy	Min F1 Score
Majority Baseline	0.21	0.605	0.722	0
Random Baseline	0.249 ± 0.006	0.556 ± 0.008	0.558 ± 0.01	0.038 ± 0.01
BERT	$\mathbf{0.69 \pm 0.003}$	$\mathbf{0.842 \pm 0.001}$	$\mathbf{0.844 \pm 0.002}$	**0.465**

Table 4: Results from our argumentation type classification task. We report the average accuracy and both the macro and weighted F1 scores across 5 runs, as well as the standard deviation. We also report the F1 score for the minimum class to ensure the model works reasonably well across all labels.



(a) Distribution of the lengths of the articles in the PolitiFact dataset.



(b) Distribution of the lengths of the articles in the FakeNews-18 dataset.

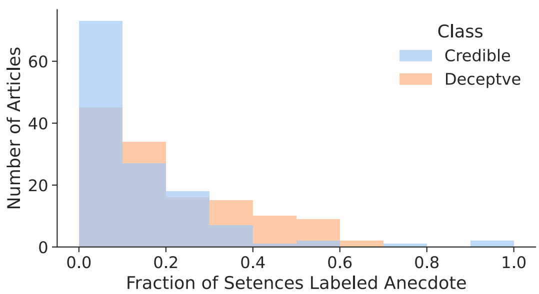
Figure 6: Lengths of the articles in both datasets for the deceptive news detection task. As we can see, there is a strong tendency for deceptive news to be shorter.

	F1 Macro	F1 Weighted	Accuracy
BERT	0.486 ± 0.054	0.537 ± 0.047	0.540 ± 0.046
SVM with LIWC Features	0.747	0.772	0.773
FakeFlow	0.780 ± 0.094	0.806 ± 0.075	0.814 ± 0.056
Argumentation Features (ours)	**0.868** ± **0.027**	**0.880** ± **0.024**	**0.881** ± **0.024**

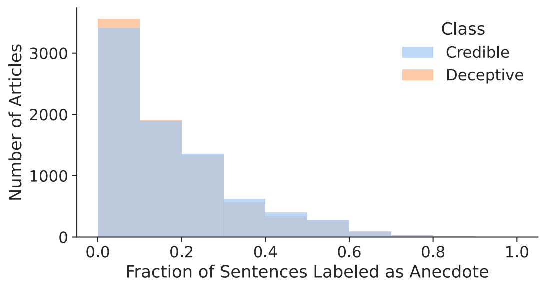
Table 5: Results from the deceptive news classification task on the PolitiFact dataset. We report the average accuracy, both the average F1 macro and weighted scores across 25 runs, and the standard deviation. Our approach (in bold) outperforms all the baselines we compared to. Of note is that the standard deviation of our model is also smaller than that of the other probabilistic models we are comparing with.

	F1 Macro	F1 Weighted	Accuracy
BERT	0.496 ± 0.007	0.557 ± 0.006	0.570 ± 0.014
SVM with LIWC Features	0.856	0.870	0.873
FakeFlow	0.407 ± 0.021	0.427 ± 0.030	0.415 ± 0.027
Argumentation Features (ours)	**0.957** ± **0.009**	**0.961** ± **0.008**	**0.961** ± **0.008**

Table 6: Results from the deceptive news classification task on the FakeNews-2018 dataset. We report the average accuracy and both the average F1 macro and weighted scores across 25 runs, as well as the standard deviation. Our approach (bold) outperforms all the baselines we compare it to. The standard deviation of our model is also smaller than that of the other models we compare it with.

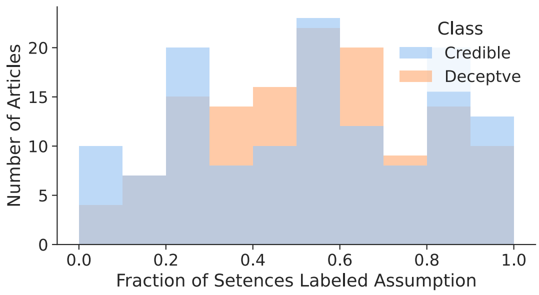


(a) Distribution for *anecdote* in the PolitiFact dataset.

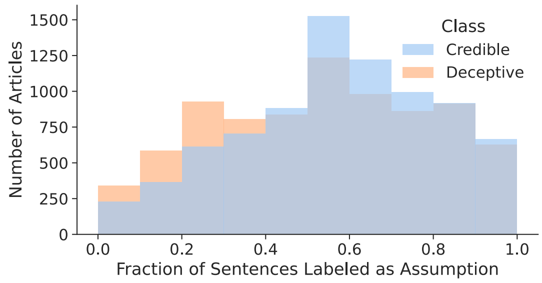


(b) Distribution for *anecdote* in the FakeNews-2018 dataset.

Figure 7: Histograms showing the distribution of the ratio of sentences labelled *anecdote* for both credible and deceptive news. Anecdotes are more represented on deceptive articles on the PolitiFact dataset, while they appear at roughly the same rate the FakeNews-2018 dataset.

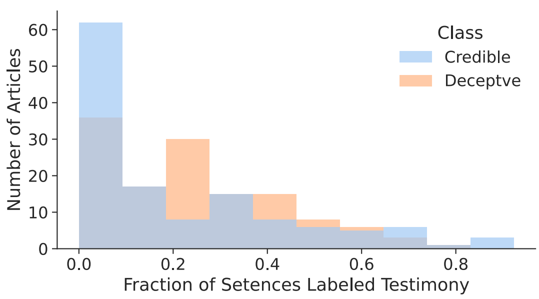


(a) Distribution for *assumption* in the PolitiFact dataset.

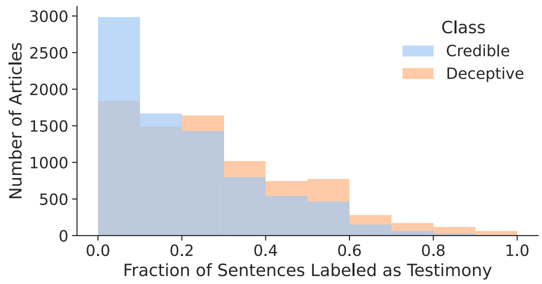


(b) Distribution for *assumption* in the FakeNews-2018 dataset.

Figure 8: Histograms showing the distribution of the ratio of sentences labelled *assumption* for both credible and deceptive news. Assumptions appear less often on deceptive articles on the FakeNews-2018 dataset. There difference for the distributions of the PolitiFact dataset is not statistically significant, meaning that we cannot rule out random chance as the reason behind this.

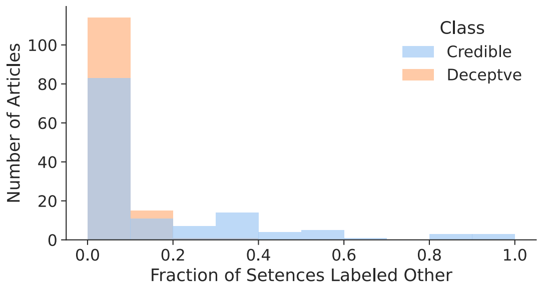


(a) Distribution for *testimony* in the PolitiFact dataset.

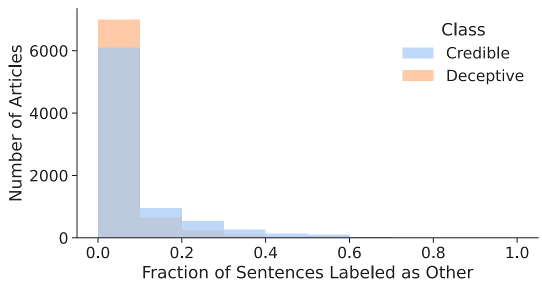


(b) Distribution for *testimony* in the FakeNews-2018 dataset.

Figure 9: Histograms showing the distribution of the ratio of sentences labelled *testimony* for both credible and deceptive news. Testimonies appear more often on deceptive articles, regardless of the dataset.



(a) Distribution for *other* in the PolitiFact dataset.



(b) Distribution for *other* in the FakeNews-2018 dataset.

Figure 10: Histograms showing the distribution of the ratio of sentences labelled *other* for both credible and deceptive news. This label appears more often in credible articles, regardless of the dataset. This label includes the labels *statistics* and *common-ground* from the Webis-16 dataset, as noted in Section 4.1.