Argument & Computation 1 (2024) 1–6
IOS Press

When automated fact-checking meets argumentation: unveiling fake news through argumentative evidence

Xiaoou Wang [a,*], Elena Cabrio [a] and Serena Villata [a]
[a] *Université Côte d'Azur, CNRS, Inria, I3S, France*
E-mails: xiaoou.wang@univ-cotedazur.fr, elena.cabrio@univ-cotedazur.fr, serena.villata@univ-cotedazur.fr

Abstract. The need for automated fact-checking has become urgent with the rise of misleading content on social media. Recently, Fake News Classification (FNC) has evolved to incorporate justifications provided by fact-checkers to explain their decisions. In this work, we argue that an argumentative representation of fact-checkers' justifications can improve the precision and explainability of FNC systems. To address this challenging task, we present LIARArg, a novel linguistic resource composed of 2,832 news and their justifications. LIARArg extends the 6-label FNC dataset LIAR-PLUS with argumentation structures, leading to the first FNC dataset annotated with argument components (claim and premise) and fine-grained relations (attack, support, partial support and partial attack). To integrate argumentation in FNC, we propose a novel joint learning method combining, for the first time, Argument Mining and FNC which outperforms state-of-the-art approaches, especially for news with intermediate truthfulness labels. Besides, our experimental setting demonstrates that fine-grained relations allow an extra performance boost. We also show that the argumentative representation of human justifications can be exploited in a Chain-of-Thought manner both in prompts and model output, paving a promising avenue for research in explainable fact-checking. Finally, our fully automated pipeline shows that integrating argumentation into FNC is not only feasible but also effective.

Keywords: Argumentation, Fake News Classification, Argumentat Mining, Disinformation, Social Media

1. Introduction

Social media are the main platforms for online social interaction and transmission of information. While these platforms have democratized access to information and facilitated worldwide communication, they also increased the circulation of online disinformation which refers to all forms of false, inaccurate, or misleading information designed, presented and promoted to intentionally cause public harm or for profit [17]. Fact-checking, i.e., the claims assessment task, is extremely complex and time-consuming to scale with the quantity and speed at which misleading information circulates.

Currently, automated fact-checking methods mainly rely on various pieces of evidence retrieved from the Web to assess the truthfulness of news claims [26]. The use of Large Language Models (LLMs) has significantly improved the task of Fake News Classification (FNC), and most works focused on experimenting with various neural network architectures and training strategies [5, 29] or how to better combine different pieces of evidence [37]. However, successfully recognizing online disinformation depends not only on understanding whether factual statements are true, but also on interpreting and

*Corresponding author. E-mail: xiaoou.wang@univ-cotedazur.fr.

critically assessing the reasoning and the arguments provided in support of the raised conclusions [75]. In this paper, we tackle this challenging issue and we answer the following research question: *how to better leverage evidence to improve the performance of automated FNC systems?*

Argument Mining (AM) [13, 33, 34] aims at identifying argumentative spans in natural language texts, classifying each argument component into major types such as *claims* and *premises*, and finally, predicting the relations among the classified components. Notable applications of Argument Mining include automated student persuasive essay scoring, where either coarse-grained argumentation features, such as the number of claims and premises, are added [25, 44], or an additional dimension of argumentation quality is introduced [78]. AM also enhances the quality of automatic debate systems [1, 7] and supports collective decision-making processes [63]. In the medical field, AM aids in detecting the effects of interventions (e.g., improved, increased) by incorporating argument features [39] and facilitates the analysis of scientific papers in biomedical text mining [35]. In legal domains, it helps justify decisions [6] and improves legal text summarization [21, 87]. Two particularly related areas to FNC are argument-based sentiment analysis [36] and stance detection [67], where identifying a user stance with respect to a certain topic is critical. Combining AM with Fake News Classification to better understand the relationship between a news claim and evidence is a natural step. This integration benefits FNC systems in two key ways: first, it mirrors the work of professional fact-checkers, who analyze claims and premises supporting or refuting the news claim; second, it enhances the explainability of automated FNC systems by allowing argument components to justify the final predicted label. More precisely, our research question breaks down into the following sub-questions: *i)* Does an argumentative representation for news claims and human justification profitably impact the task of FNC? and *ii)* How to leverage this representation to improve the performance of automated FNC systems?

The main contribution of this paper is threefold:

(1) We build a novel annotated linguistic resource called LIARArg, which extends LIAR-PLUS [4]. To the best of our knowledge, LIARArg is the first dataset integrating AM to FNC. It contains 2,832 news claims and justifications provided by professional fact-checkers enhanced with argument components as well as fine-grained argument relations.
(2) After establishing strong FNC baselines by combining the strength of recent knowledge-enhanced approaches [38, 86], we propose a novel joint-learning architecture to train AM and FNC together to transfer the argument knowledge contained in the annotation to FNC, as well as a novel Chain-of-Thought (COT) [85] based framework to explicitly inject the argument structure into the prompts. These novel approaches outperform SOTA approaches in FNC on the same dataset, especially for news with intermediate truthfulness labels.
(3) Through our extensive experimental setting, we demonstrate the crucial contribution of argument relations (particularly fine-grained relations, which allow an extra performance boost on this 6-label dataset) to assist the FNC task, highlighting therefore a promising research direction to tackle the problem of half-truths classification [22].

The remainder of this paper is structured as follows: Section 2 provides an overview of existing literature on Fake News Classification. In Section 3, we present the novel LIARArg dataset, including sampling methods, filtering criteria, annotation process, and agreement computation. Section 4 presents the experimental setup, baselines and introduces the two architectures we propose. Additionally, we introduce a fully automated, argumentation-enhanced FNC pipeline. Section 5 presents the key outcomes and insights derived from our experiments. Conclusions end the paper with a discussion about the main contributions and limitations of our approach, highlighting directions for future research.

2. Related Work

The automated fact-checking process generally involves 4 stages [26]: *i)* claim detection to identify or rank the claims to verify; *ii)* evidence retrieval to find sources supporting or refuting a claim; *iii)* claim verification or FNC to assign truthfulness labels to claims, and finally, *iv)* justification production which explains the verdict. This section focuses on works related to the FNC phase, which is the focus of this work.

Early studies on FNC have been conducted using only news claims as input, in a binary classification fashion [76]. FNC has then considerably evolved. Firstly, the input has been enhanced with meta-data such as the speaker's background and the news context [82]. The evidence has also been incorporated into the input in the form of text [56], knowledge graphs [66] or tabular data [15]. Secondly, several datasets have employed finer-grained classification schemes, e.g., including the addition of an extra "lack of information" label in FEVER [69], or more importantly, some supplementary labels to represent degrees of truthfulness such as in LIAR [82] and FakeCovid [65]. Lastly, in terms of modeling strategies, early studies used stylistic features or bag-of-words representation of news claims and meta-data without employing external evidence [57, 82]. Some studies use both claims and evidence as input, and frame FNC as a Natural Language Inference (NLI) problem. In this case, the evidence is used as premise to refute or support the hypothesis represented by a news claim. When multiple pieces of evidence are available, a weighted aggregation [62] is often required to take into account the reliability of the evidence. Another line of research considers the evidence as reliable by default [70]. Early studies used classical methods in Natural Language Inference such as the decomposable attention model of Parikh et al. [47] which scored best on the Stanford NLI corpus [10]. The advent of large pre-trained language models (LLM) such as BERT [18] has garnered attention also in the FNC community [5, 8, 58] and led to significant improvements in FNC systems performances [52]. Most recently, systems based on knowledge-enhanced LLMs such as KnowBert [53] and KEPLER [84] have further improved the classification performance [86]. Rather than using LLMs where the knowledge is injected during the training phase, the SOTA approach in FNC of [38] constructs an entity graph from the news text by aligning entities and their corresponding first-order neighbors in Wikidata [77]. The graph is then fed into a graph attention network [73] to produce a knowledge-enhanced graph representation. Finally, the graph representation is concatenated with a BERT-based textual representation provided as input to a linear classifier. In this paper, we leverage the strength of these two knowledge injection methods to set up a strong baseline.

Other approaches aim to enhance the FNC module by linking it with an evidence retrieval module. These approaches have been mostly tested on the FEVER dataset, grounding on the general idea of jointly training the evidence retrieval module and the claim verification module (i.e., FNC). For instance, Yin and Roth [89] proposed CNNs and attentive convolutions to extract sentence representations of the claim and evidence so that the two tasks are trained jointly. Hidey and Diab [27] trained the two tasks jointly by using pointer networks [74] for the sentence selection subtask, and an Multi-Layer Perceptron based architecture for FNC. Finally, Niet et al. [45] used semantic relatedness scores and ontological WordNet features to compare claims and evidence so that evidence retrieval and FNC can be trained together. With respect to these approaches, it is important to highlight that in our work we assume that the right evidence has already been retrieved. Essentially, our module enhances the stage following evidence collection, demonstrating how to effectively utilize this kind of information. We do so by making explicit the inherent argumentative structure within a specific piece of evidence. This assumption ensures that the evidence used is trustworthy and sufficient, thereby averting scenarios where

FNC systems appear to improve under false pretenses, namely instances where they correctly predict the veracity of a claim based on incorrect or irrelevant evidence, a typical case of being right for wrong reasons. The disentanglement of the evidence retrieval process from the evidence modeling process permits us to concentrate on how argumentation-based models can support FNC systems without being confounded by evidence quality concerns, thereby reducing the risk of false improvements.

Finally, as for the LIAR-PLUS dataset [4], several studies have been conducted to improve the FNC performance. Among the approaches employing both claims and justifications as inputs, Sadeghi et al. [60] proposed a BERT-based NLI model which outperformed the baselines of Alhindi et al. [4]. However, the full version of the justification texts has been used instead of the simple justifications (composed of some sentences) provided in the original dataset. Mehta et al. [40] proposed a triple BERT network to encode separately news claims, metadata and justifications. However, this approach does not outperform the ngram-based model of Alhindi et al. [4], i.e., 0.70 F1 for binary classification and 0.37 F1 for six-way classification. This highlights the need for improved methods, and particularly, for a more effective representation of justifications on this challenging dataset.

3. The LIARArg dataset

This section describes the LIARArg dataset which extends the LIAR-PLUS dataset [4] with argumentative labels (i.e., components and relations). To the best of our knowledge, this is the first dataset for FNC annotated with argumentative components and relations[1].

3.1. Data collection and filtering

LIARArg is built on LIAR-PLUS [4] which is itself an extension of the LIAR dataset [82] consisting of 12,836 news claims taken from POLITIFACT[2] and labeled with truthfulness, subject, context, speaker, state, party, and prior history. We choose LIAR because this dataset is particularly challenging, with six truthfulness labels: Pants-On-Fire, False, Mostly-False, Half-true, Mostly-True and True. The increased number of labels in comparison to datasets like FNC-1 [54] (4 labels) and Check-COVID [81] (2 labels) is warranted by the nature of the claims evaluated on POLITIFACT. As shown in Figure 1, claims typically take the form of "X says Y". The assessment of truthfulness focuses not on the fact that X made the statement, but rather on the accuracy of Y, considering additional contextual factors.

The extreme categories, such as "Pants-on-fire", "True", and "False", often include factually verifiable statements, for example:

- "A photograph of 21-year-old Hillary Clinton featured a Confederate battle flag in the background." (Pants-on-fire)
- "When undocumented children are picked up at the border and told to appear later in court... 90 percent do not then show up." (False)
- "New Hampshire has the third-highest property tax in the country." (True)

The intermediate labels, however, involve claims that require a more nuanced assessment of plausibility, particularly in cases where:

[1] LIARArg and the source code of our experiences will be made available upon paper acceptance.
[2] https://www.politifact.com/

Statement:"Says Rick Scott cut education to pay for even more tax breaks for big, powerful, well-connected corporations."
Speaker: Florida Democratic Party
Context: TV Ad
Label: half-true
Extracted Justification: A TV ad by the Florida Democratic Party says Scott "cut education to pay for even more tax breaks for big, powerful, well-connected corporations." However, the ad exaggerates when it focuses attention on tax breaks for "big, powerful, well-connected corporations." Some such companies benefited, but so did many other types of businesses. And the question of whether the tax cuts and the education cuts had any causal relationship is murkier than the ad lets on.

Fig. 1. Excerpt from the LIAR-PLUS dataset [4] where each news claim (statement) is paired with an automatically extracted justification provided by fact-checkers.

(1) *Causal relationships are examined:* For instance, whether a statistic can be attributed to a politician or department's policy, as in "The Texas Department of Transportation misplaced a billion dollars." (Mostly-true)
(2) *Temporal elements are involved:* For example, whether a politician has shifted their stance on an issue over time, as seen in "McCain supported George Bush's policies 95 percent of the time." (Half-true)
(3) *Speculative claims about the future are made:* As in "It is a fact that it costs more to run the schools in August." (Barely-true)
(4) *Positions on controversial issues are considered:* As in "Military spending cuts, known as the sequester, were President Barack Obama's idea." (Half-true)

As suggested by Uscinski et al. [72], the definition of truth or fact can vary depending on the nature of the claims being evaluated. This layered approach to truthfulness is specific to POLITIFACT and contrasts with datasets like Check-COVID, which focus more on factual accuracy.

Alhindi et al. [4] extended this dataset by automatically extracting for each claim a summary that has a headline "our ruling" or "summing up", which serves as justification provided by professional fact-checkers. When no summary exists, the last five sentences in the fact-checking article were extracted. Figure 1 shows an instance of the LIAR-PLUS dataset. Note that, in LIAR-PLUS, verdict phrases, such as "it is true" or "this is misleading", have been filtered to minimize label leakage.

To extend this dataset with the argumentative layer, we first randomly sampled an equal number of news claims for each label. Then we annotate each claim and justification with the following information: argument components (claim and premise), and fine-grained argumentative relations among the identified components (i.e., support, partial support, attack and partial attack). .

Besides the additional argumentation-based annotation layer, the filtering of invalid items is also a substantial contribution of our work: the justifications in LIAR-PLUS are very noisy, whether they are summaries or the last five sentences. Furthermore, on many occasions, these justifications are not sufficient to support the truthfulness of the news claim. This quality issue can affect the reliability of our approach. We therefore also annotated the quality of each justification as good, insufficient or incomprehensible. Lastly, we annotated certain news claims as incomplete because sometimes the news comprises

only two or three words, and has no meaningful truthfulness. In total, we annotated 3,934 texts of which 1,005 justifications are insufficient, 12 are incomprehensible and 85 statements are incomplete. The final dataset includes therefore 2,832 statements with a valid justification.

3.2. Annotation scheme

Typically, the annotation process in Argument Mining can be divided into three key subtasks: identifying argumentative components and their boundaries, recognizing types of argumentative discourse units (ADUs), and annotating the relationships between arguments. For the boundaries of argumentative components, clauses are typically the units. Discourse markers such as "however" are included in ADUs, as well as prepositional phrases such as "according to X". As for types of ADUs and relations, early studies have focused on annotating thesis and conclusion statements in student essays [12] and premises and conclusions in legal texts [46]. Peldszus and Stede's works on the microtext scheme [49, 51] draw inspiration from Freeman's theory of argumentation's macro-structure [24] and introduce a tree-based annotation framework that includes two key argumentative roles: the proponent, who presents and defends claims, and the opponent, who critically questions them. Their schema features fine-grained argument relations, differentiating between simple support (where a single premise suffices) and linked support (where multiple premises must be considered together), as well as rebuttal (where a statement is deemed invalid) and undercutting (where a statement is irrelevant to supporting or refuting another). The persuasive essay scheme annotates, besides claim/premise trees, a central component of student essays named major claim and only distinguishes two types of relations: support and attack. This scheme differentiates between only two types of relations: support and attack. Depending on the specific nature of the texts being annotated and the desired level of granularity, the roles of ADUs (claim or premise) can be omitted, as seen in [30] for scientific texts, or more nuanced relations can be incorporated [59] using Walton Argumentation Schemes [79].

We decided to keep the information of argument roles and annotate two kinds of argument components: claim and premise[3]. The news text is especially well-suited for this dichotomy because each text contains generally some claims which represent statements denoting opinions or standpoints, and premises which contain statements that can be verified to some extent, including typically some quotes from original documents or concrete statistics. Unlike persuasive essays, where identifying the major claim throughout the text is essential, in our case, it is unnecessary to annotate the major claim, as it is always the first claim presented to the annotator. Note that the claim-hood and premise-hood are not intrinsic features of a statement but determined also by the relationships between statements. For example, the same statement *the unemployment rate is the highest in 45 years* can be annotated as a claim when it is used as a news claim to be verified, but it can also be annotated as a premise when used in a justification, as evidence to support or refute another claim.

Due to the challenging nature of the LIARArg dataset, which includes 6 fine-grained truthfulness labels, we decided to annotate 4 types of argument relations, i.e., support, attack, partial attack, and partial support. Our decision to annotate more than two relations while excluding finer distinctions, such as simple support, linked support, or Walton Schemes, represents a compromise. While greater granularity would add more information, it would also lead to an increase in annotation time, without the certainty to include enough instances for each class to allow the model to learn on them in order to achieve good classification results for this particular corpus. Most crucially, argument relations in LIARArg often

[3]The detailed annotation guidelines can be found at https://anr-attention.github.io/gd.pdf

follow a one-to-one pattern, with complex structures like linked support being relatively uncommon. That said, incorporating a finer-grained annotation layer to corpora with lengthy and intricate arguments could open new avenues for training more sophisticated models, a highly promising direction for future research. The definitions of these four types of relations are provided below. An argument component supports another when it validates this component, and it attacks another when it contradicts the proposition of the target component. Partial support is used when an argument component validates certain aspects of another component but diverges in some other aspects. Partial attack is used when the source argument component is not in full contradiction, but it weakens the target component. Example (1) shows an instance of support and partial attack[4]. In Example (1), $Premise_2$ supports $\mathbf{Claim}_1$ while $Premise_3$ partially attacks the same claim. Examples like (1) are typically labeled using intermediate labels such as half-true.

(1) **[Hillary Clinton supported NAFTA and permanent China trade]**$_1$.**[Pennsylvania lost thousands of jobs]**$_2$.
{Another study by EPI concluded Pennsylvania lost another 44,173 jobs between 1993 and 2004 due to NAFTA}$_1$. *{It is true that Clinton has in the past verbally supported NAFTA and permanent trade with China}*$_2$. *{Yet it is also true that she has spoken forcefully about the need to reform NAFTA and to much more stringently enforce trade agreements with China}*$_3$.

Example (2) shows an instance of attack and partial support where $Premise_1$ attacks $\mathbf{Claim}_1$, and $Premise_2$ partially supports $\mathbf{Claim}_1$.

(2) **[Iran might not be a superpower, but the threat the government of Iran poses is anything but "tiny" as Obama says]**$_1$.
One could argue whether it's wise to meet with leaders of rogue nations. One could also debate whether Obama wrongly downplayed the threat posed by Iran. *{But Obama never said the threat from Iran was "tiny" or "insignificant,"}*$_1$, *{only that the threat was tiny in comparison to the threat once posed by the Soviet Union}*$_2$.

3.3. Annotation process

Two annotators with a background in computational linguistics carried out the first annotation phase. A data sample of 150 texts (i.e., 25 texts per label) was first annotated, followed by a first reconciliation phase to examine the main sources of disagreement, which concerned especially the annotation of partial support and partial attack. Then a second sample of 150 texts was annotated to investigate the effect of the reconciliation phase. After the second round of reconciliation, a last sample of 90 texts was annotated to make sure that the reconciliation led to consistent results. The Inter-Annotator Agreement (IAA) was calculated for each sample, as shown in Table 1. For argument components, IAA is computed for text spans which are identified as argumentative (claim or premise). For relations, all pairs of text spans which are annotated as linked are considered when computing IAA scores. Annotation is considered as agreed when both the relation label and the assigned target components are the same. We achieved a

[4]In the examples, claims are in bold and marked by brackets, and premises are denoted in italics in braces. Note that the first line is always the news statement, and the following text is the justification provided by the fact-checker.

Fleiss' kappa[5] of 0.73 (substantial agreement) for component annotation, and 0.61 (moderate agreement) for relation annotation[6].

Table 1
Inter-annotator agreement (Fleiss' kappa) for argument component and relation annotation.

Sample	Argument component	Argument Relation
150 texts	0.72	0.48
150 texts	0.71	0.59
90 texts	0.73 (substantial)	0.61 (moderate)

Figures 2 and 3 present the confusion matrices for the annotation of argument components and relations on the final 90 texts. 98.6% of the relations contain the same components. The results show that claims and premises are generally well distinguished. The primary disagreements in argument relations, as described at the beginning of this section, occur between support and partial support, and particularly between attack and partial attack.

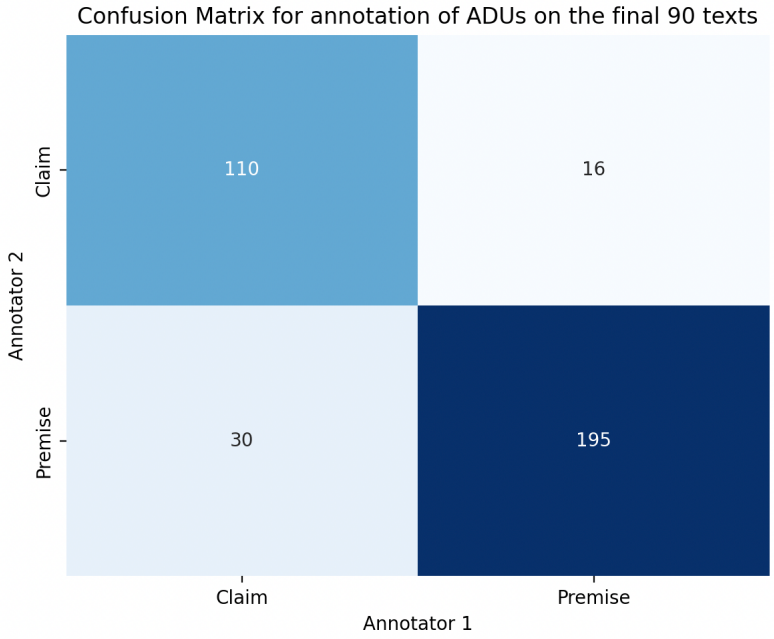

Fig. 2. Confusion matrix for two annotators' annotation of argument components on the final 90 texts.

It is important to note that argument relation annotation is a challenging task. The corpus ComArg [9], where 3 annotators annotate 2,249 comment-argument pairs with attack/support relations, achieves a Fleiss's kappa of 0.49. The same Fleiss's kappa score is calculated for 3 annotators on 30 Randomized

[5]Although only two annotators are involved, Fleiss' kappa is calculated instead of Cohen's kappa to ensure a certain degree of comparability with other works in the literature where often more than two annotators are involved.

[6]We refer to the scale introduced by Landis and Koch in [32] for interpretation of Kappa statistics where a value above 0.61 is considered to indicate substantial agreement. Since the original score for relation annotation is 0.606, we consider it as a moderate agreement.

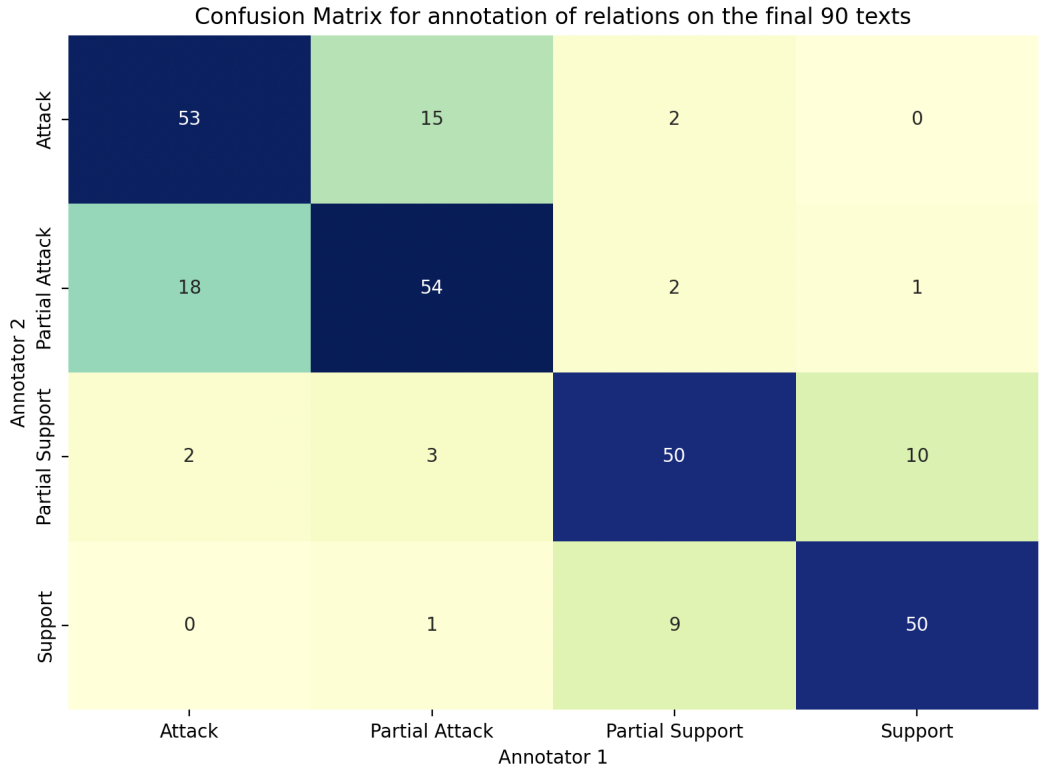

Fig. 3. Confusion matrix for two annotators' annotation of argument relations on the final 90 texts.

Controlled Trial abstracts and is equal to 0.62 (for 2 types of relations). On MicroText [49], a Fleiss's kappa of 0.58 is observed for distinguishing attack and support relations. Given these results, and considering that LiarArg involves four instead of two types of relations, we consider a Fleiss's kappa of 0.61 acceptable in the current work. We obtained a final score of 0.98 for the justification quality, and 0.96 for claim completeness in the IAA assessment, indicating almost complete agreement. This level of agreement is expected as most justifications which are insufficient are the results of automatic extraction of the last five sentences in the fact-checking articles when no human-written summaries are accessible. As for incomplete claims, they are typically truncated sentences consisting of two or three words, making them easy to identify, as discussed in Section 3.1. The annotation task was then completed by one of the two annotators.

Table 2 reports on the number of annotated items per label with the average number of claims and premises for each category. The final dataset is relatively balanced, and the number of claims and premises is similar across the labels.

4. Experimental setting

In this section, we first outline the general settings of our experiments. Following that, we introduce the baselines and describe the architectures we propose.

Table 2
Average number of claims and premises across the labels.

Label	Items	Claims	Premises
True	443	1.2	2.3
Mostly-True	529	1.4	2.5
Half-True	522	1.5	2.6
Barely-True	476	1.5	2.7
False	471	1.4	2.5
Pants-on-fire	391	1.4	2.4
Total	2,832	1.4	2.5

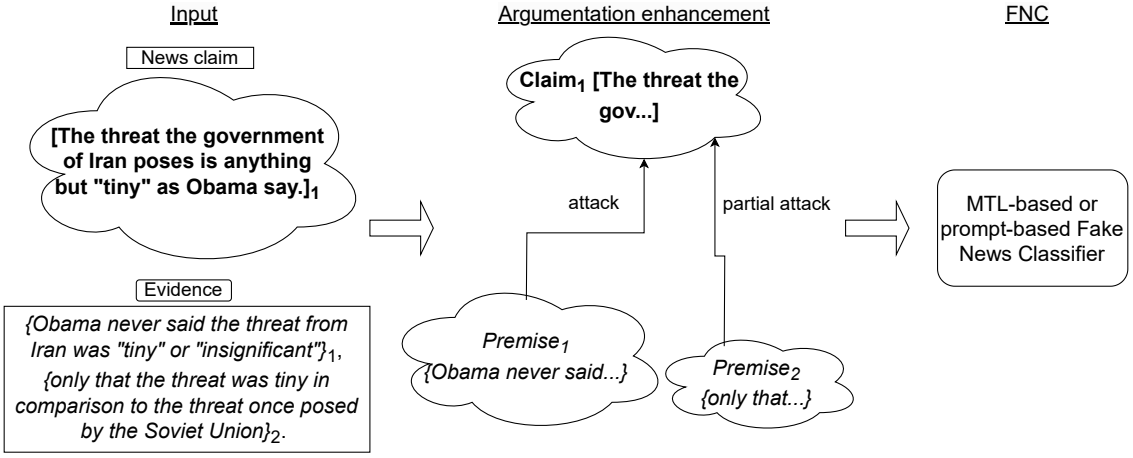

Fig. 4. The experimental pipeline.The input consists of a textual claim and a textual justification. The justification is then enriched with the identification of the specific argument components and relations identified in the text. Finally, the classifiers employ Multi-Task Learning or prompting to leverage the argument structure contained in the paired text to improve the FNC task performance.

As mentioned in Section 2, our work centers on the phase immediately following the evidence retrieval process. The whole pipeline of our experiment is illustrated in Fig. 4. A news claim paired with an evidence piece forms the primary input. Subsequently, the claim-evidence pair is enriched with argument components and relations. Finally, the classifiers employ Multi-Task Learning or prompting to leverage the argument structure contained in the paired text to improve the FNC task performance.

4.1. General approach

Our approach for fake news classification is based on two frameworks: Multi-Task Learning to implicitly integrate argumentation to FNC, and Chain-of-Thought to explicitly inject the argumentative information in the input.

Multi-Task Learning or joint learning has been extensively studied and successfully applied in various scenarios in Machine Learning [90] and Natural Language Processing [91]. Multi-Task Learning aims to leverage useful information shared across multiple related tasks to improve the generalization performance on all tasks [14]. A main task is defined with respect to some auxiliary tasks so that the knowledge learned in auxiliary tasks can help the main task, and, at the same time, prevent overfitting. Multi-Task

Learning is typically implemented with either hard or soft parameter sharing of hidden layers. The former is applied by sharing the hidden layers between all tasks while keeping several task-specific output layers. In the latter case, each task has its own model with its own parameters, and the distance between the parameters of different models is regularized so that they are encouraged to be similar across models through metrics such as L2 distance [20].

Chain-Of-Thought (COT) is a prompting technique [85] demonstrating a significant performance improvement on a range of arithmetic, commonsense, and symbolic reasoning tasks. The idea is to explicitly decompose the reasoning process such as calculation procedure during the prompting process. For example, instead of asking "What is the sum of 14 and 18?", and providing "32" as the answer, COT breaks down the procedure step by step: "4 + 8 = 12. Write down the 2 and carry over the 1. 1 + 1 + the carried over 1 = 3. The answer is 32.". Chain-Of-Thought has been shown to elicit reasoning in LLMs as GPT-3 [11], and it is particularly relevant to our scenario since argumentative information can be directly injected into the prompting in a similar manner.

4.2. Baselines

Since LIARArg is a subset of LIAR-PLUS, we used the best-performing model on LIAR-PLUS [4] as a simple baseline. Each claim and justification are concatenated, and unigram features of the concatenated text are fed into a Logistic Regression model (**LG**).

Given the recent advancements in FNC driven by the incorporation of knowledge into LLMs, we establish two other strong baselines by drawing insights from the approaches of Whitehouse et al. [86] and Ma et al. [38]. First, we concatenate each claim and justification by inserting [SEP] between the two. A special token [CLS] is then added to the beginning of each sentence pair, from which the final embedding of the input is extracted. The baseline **KB** uses KnowBert [53] as back-end, as it was the best-performing model in [86]. **KGB** uses the same setting as Ma et al. [38], meaning that we construct the entity graph of each concatenated text based on Wikidata, and we extract the graph embedding using graph attention networks [73]. However, we use KnowBert as the textual feature extractor (best-performing knowledge-enhanced LLM for FNC) instead of the BERT base model used in [38]. A softmax layer is applied to the final embedding (textual embedding for **KB**, and concatenation of graph and textual embeddings for **KGB**) to get the logits for each label. The loss function is CrossEntropy. We call these baselines "single task models" **(ST)** to distinguish them from LG and models trained using more than one task.

4.3. Proposed architectures

Multi-Task Learning (MTL). To fully explore the potential of the justification texts, we adopt the hard parameter sharing and define FNC as the main task. Argument component classification (CC) and relation classification (RC) are considered as auxiliary tasks. Figure 5 illustrates our neural architecture for MTL-based FNC. Each instance of LIARArg is processed into 3 components: concatenated news claim and justification, the argument components (i.e., 3 labels with unknown as an extra component), and argument relation pairs (i.e., 5 labels with neutral as an extra relation[7]). Knowledge-enhanced embeddings based on the concatenated text are produced in the same way as for ST classifiers, namely **KB** and **KGB**. The embeddings of each argument component and relation pair are produced by KnowBert with their own softmax layer to produce logits. All the models for the three tasks apply Adam optimizer,

[7]The additional labels are intended solely for training purposes and do not require extra annotation. Neutral relations include all non-linked identified ADUs pairs, while unknown components encompass all non-ADUs.

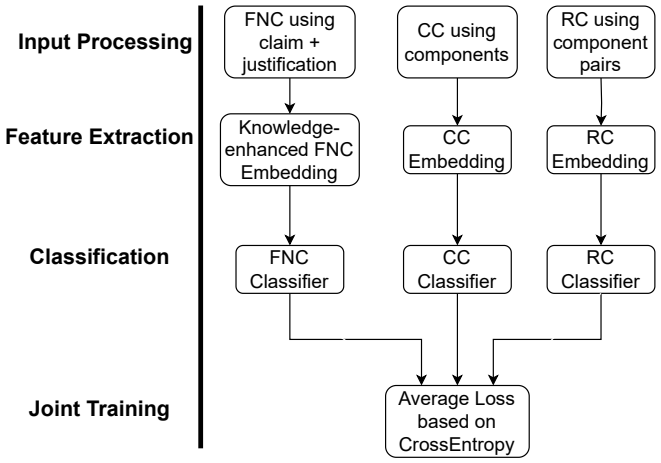

Fig. 5. Model architecture for MTL-based FNC. The embedding of the concatenated claim and justification is produced by KnowBert (KB) or KnowBert augmented with graph embeddings. The embeddings of each argument component and relation pair are produced by KnowBert. Each embedding is then fed into a softmax layer to produce the logits for each label.

learning rate 3e-5, dropout 0.1, warmingup ratio 0.1, batch size 16, CrossEntropy as loss function and run 5 epochs. The loss is produced per classifier i.e., FNC ($loss_{fnc}$), component classification ($loss_{cc}$), and relation classification ($loss_{rc}$). We experimented with different combinations of the loss functions (i.e., manual or learnable scaling of each loss), and found that the best results are obtained when the loss is the average of all losses. The final loss is then back-propagated to update the parameters of all classifiers.

Chain-Of-Thought (COT). We use the text-davinci-003 variant of GPT-3 and the publicly available GPT-3 API to make inferences[8]. All the prompts start with "The task in question is to assess the truthfulness of a news claim based on a justification text by outputting one of the following six labels: True, Mostly True, Half True, Mostly False, False, Pants on Fire. The definitions of these labels are as follows:". We used the definitions provided by POLITIFACT[9].

To assess the impact of argumentation on FNC, we set up 3 prompting strategies. All the strategies are based on a few-shot setting with 1, 10 and 20 examples for each label followed by a new instance, the difference residing in how the examples and the new instance are presented. To mitigate potential biases in the model arising from specific examples, we construct 10 random samples for the 3 different example sizes used in the few-shot setting. The final evaluation is calculated as the average across 10 trials for each example size, thereby providing a robust assessment of the model's performance.

In the following, we illustrate these three strategies with a news item containing two claims, as it is the most complex case in LIARArg[10]. The output of COT-based approaches is in `json` format to facilitate the parsing process. The three prompting strategies are as follows. To summarize, the rationale is not used at all in STP, used in examples and new instances in COTP, used in examples but not new instances on COTPS.

- **Standard prompting (STP)** where the claim and the justification are concatenated in the same way *both in the examples and in the new instance*.

[8] https://platform.openai.com/docs/api-reference/making-requests

[9] https://www.politifact.com/article/2018/feb/12/principles-truth-o-meter-politifacts-methodology-i/

[10] The connecting phrases are marked in italics.

- ∗ Example: *The news claims that* “Hillary said that I can’t sign money. That’s illegal.”. *According to the justification*, although defacing dollars is illegal, she could have signed that buck... *Based on this justification*, the truthfulness label of this news should be barely true.
- ∗ New instance: *The news claims that... According to the justification*, ...
  Based on this justification, assess the truthfulness of this news claim by selecting from one of the six labels and output the following format: {“label": “..."}
- ∗ Output example: {“label": “True"}

- **COT-like prompting (COTP)** in which the justification text is automatically converted into an argumentative text *both in the examples and in the new instance*. It is worth noting that the standard output of Chain-of-Thought contains also the rationale used by the model to produce the label, while our approach only outputs the label.

  - ∗ Example: *The news claims that* “Hillary said that I can’t sign money. That’s illegal.”. There are two claims in this news claim. Claim 1: Hillary said that I can’t sign money and Claim 2: Hillary said that that’s illegal. *According to the justification*, “Although defacing dollars is illegal” is a premise supporting Claim 2, “she could have signed that buck without fear of prosecution“ is a premise attacking Claim 1. *Based on this justification*, the truthfulness of this news should be half true.
  - ∗ New instance: *The news claims that... According to the justification*, ...is a premise supporting, ...is a premise attacking...
    Based on this justification, assess the truthfulness of this news claim by selecting from one of the six labels and output the following format: {“label": “..."}
  - ∗ Output example: {“label": “True"}

- **COT prompting from scratch (COTPS)** where the justification is automatically converted into an argumentative text *only in the examples*. For the new instance, the model is only provided with the justification in its original form. This means that in COTPS no annotated data is needed for new instances (differently from COTP). Note that both the label and the rational are produced as output, the latter being a particularly relevant feature for explainable AI.

  - ∗ Example: the same as in COTP.
  - ∗ New instance: *The news claims that... According to the justification*, although defacing dollars is illegal, she could have signed that buck...
    Based on this justification, please output the truthfulness of this news claim by selecting from the six labels mentioned above. Produce also the rationale for your output by imitating the reasoning process given in the examples. Put the label and the rationale in the following format: {“label": “...", “rationale": “..."}
  - ∗ Output example: {“label": “True", “rationale": “There are 2 claims in... According to..., X is a premise attacking Claim1... "} Refer to Ex. (3) for a full-fledged example.

In the subsequent sections, we append 1, 10 and 20 to STP, COTP and COTPS to denote the different number of examples used in the few-shot setting.

4.4. Fully automated FNC pipeline

A key challenge in the presented approach, with the exception of the COTPS approach, is represented by the required extensive annotation effort. To assess the feasibility of a fully automated FNC pipeline

enhanced through the argumentation module, we adapt the SOTA automatic AM parser of Morio et al. [43] to our task. Morio et al. [43]'s parser is trained on various types of data including student essays [68], argumentative microtexts [50] and scientific articles [2]. We retrain the parser with the same cross-corpora multi-learning approach as in [43] but we add LIARArg as an extra corpus. This fully automated pipeline (Argument Mining parser + FNC classifier) allows us to replicate the methodology discussed above on a significantly larger scale – namely 8,902 texts compared to the 2,832 texts in LIARArg – without requiring further human annotations.

Since the parser of Mario et al. [43] has been adapted to the type of data close to LIARArg, it is important to further assess the effectiveness of this pipeline on out-of-domain data. For this purpose, we conduct an additional evaluation campaign on FNC-1 [54] and Check-COVID [81]. The FNC-1 dataset is a well-known benchmark for FNC challenge derived from the Emergent Dataset [23], containing 75385 labeled headline and article pairs across more than 20 topics. Check-COVID is a benchmark of 1504 claims about COVID-19 where each news claim is paired with evidence from scientific journal articles. We choose these two datasets because both provide claim-evidence pairs that can be used as input to our automated FNC pipeline (see Fig. 4). FNC-1 is framed as a stance detection task with 4 labels: agree, disagree, discuss and unrelated, while Check-Covid is a binary classification task with Refute and Support as labels. We employ the state-of-the-art models in the literature for each dataset as baselines: the augmentation-based ensemble learning approach for FNC-1 [61] and the dual RoBERTa-based model for Check-COVID [81] where two RoBERTa models are fine-tuned to first select relevant sentences in evidence then to classify the claim-sentences pair.

Considering the large size of these three datasets, we change from GPT-3 to Mistral 7B [28] as back-end for the automated pipeline. Mistral 7B is a recent generative Large Language Model leveraging grouped-query attention [3] for faster inference and sliding window attention [16] to handle longer sequences more efficiently. It outperformed Llama2 13B [71] on a wide range of benchmarks, and Llama2 34B on mathematics and code generation. Its low consumption of hardware resources makes it particularly suitable for large-scale experiments. The official checkpoint on HuggingFace, Mistral-7B-v0.1[11], is used with all the default parameters unchanged.

4.5. Evaluation setup

We perform 10-fold cross-validation using by StratifiedKFold of Scikit-learn [48], splitting the LIARArg dataset into 80:10:10 proportions for training, validation, and test sets. Stratified sampling is used to maintain consistent label distribution across all subsets, ensuring that the training, validation, and test sets reflect the overall dataset's label proportions. For the automated pipeline, we split the rest of the LIAR-PLUS dataset into 80:10:10 with 10 crossfolds. For the two out-of-domain datasets, we adhere to the default data splits as outlined in their respective studies. Specifically, for FNC-1, the split comprises 49,972 instances for training set, and 25,414 for test test. The Check-COVID dataset follows a distribution of 70% for training, 15% for validation, and 15% for testing. Due to the presence of data imbalance in FNC-1, the macro-averaged F1 score is used as metric as in the baseline [81].

Ablation studies are conducted on the MTL setting to demonstrate the impact of argumentative features. We append **+CC**, **+RC** and **+CCRC** to different settings according to whether component classification, relation classification or both, have been used as auxiliary tasks. To assess the impact of fine-grained relations in FNC, we conduct another ablation study by re-running the same model training experiments while merging the partial attack to attack and the partial support to support.

[11] https://huggingface.co/docs/transformers/en/model_doc/mistral

5. Results

In this section, we present the results of the MTL-based and COT-based approaches presented above, as well as those of the fully automated FNC pipeline.

5.1. Results of MTL-based FNC

Table 3

Results of MTL-based FNC models in F1 score compared with LG and single-task based models. Bold texts indicate statistically significant improvements. Note that the combination of KnowBert and graph embeddings provides better results (KGB *vs.* KB). Also, adding relation classification as an auxiliary task further improves the performance (columns with +RC). For statistical significance, ST is compared with LG and +RC is compared with ST, respectively. Additionally, the two settings with relation classification as subtasks (+RC) have also been compared against each other.

Split \ Model	LG	KB				KGB			
		ST	+CC	+RC	+CCRC	ST	+CC	+RC	+CCRC
Binary Valid	0.58	**0.64**	0.63	**0.69**	0.66	**0.67**	0.67	**0.73**	0.72
Binary Test	0.60	**0.65**	0.64	**0.70**	0.66	**0.68**	0.68	**0.74**	0.72
6-way Valid	0.30	**0.33**	0.31	**0.38**	0.31	**0.35**	0.33	**0.41**	0.39
6-way Test	0.31	**0.34**	0.31	**0.39**	0.32	**0.36**	0.32	**0.42**	0.38

Table 3 reports the results of the MTL approach on LIARArg depending on the number and type of the tasks involved. The two-sided Wilcoxon signed rank test has been performed across the 10 crossfolds and statistically significant improvements have been highlighted in bold. First of all, it can be seen that the LLM-based baselines significantly surpass the LG approach of [4], confirming the strong improvement driven by the use of LLMs in the FNC task. Secondly, it is worth noting that combining the graph representation of the entities contained in texts and the KnowBert embedding of the texts themselves produces even better results (KB *vs.* KGB). The combination of these two knowledge-injected methods provides therefore a strong baseline for our study. Thirdly, adding the argument component classification as an auxiliary task reduces the performance. While we could conclude that adding component information does not aid FNC, we underline that the increasing complexity of the training setup might have counteracted the benefits of the additional information provided by components. Finally, relation classification, despite the strong baselines provided by the knowledge-enhanced representation of news and justifications, clearly improves both binary and 6-way FNC, showing that the information contained in argument relations is crucial for more efficient FNC.

5.2. Results of COT-based FNC

Table 4 reports the results obtained with the prompting techniques discussed in Section 4, compared to the best MTL setting. 20 examples per label have been used.

It can be seen that although the prompting approach produces less good results than our best joint learning approach, injecting argumentative knowledge into the prompts can significantly improve the classification results, as shown by the large gap observed between STP and COT-based methods. Indeed, by using 20 examples from each label, COTP achieves an F1-score of only 0.04 under our best model in MTL setting. It could be contented that COTP necessitates annotated new instances which are impractical in real-world settings. Nonetheless, COTPS, which does not require new instances to be annotated as mentioned in Section 4.3, also benefits significantly through argumentative enhancement,

Table 4

Results of prompt-based FNC using 20 examples compared to the best results achieved in the MTL setting. Bold texts indicate statistically significant improvements. Although the prompting approach produces less good results than our best joint learning approach (KGB+RC), injecting argumentative knowledge into the prompts can significantly improve the classification results, as shown by the large gap observed between STP and COT-based methods. For statistical significance, COTP20 and COTPS20 have been compared against STP20.

Model	KGB+RC	STP20	COTP20	COTPS20
Binary valid	0.73	0.58	**0.72**	**0.68**
Binary test	0.74	0.57	**0.73**	**0.70**
6-way valid	0.41	0.27	**0.39**	**0.35**
6-way test	0.42	0.28	**0.38**	**0.36**

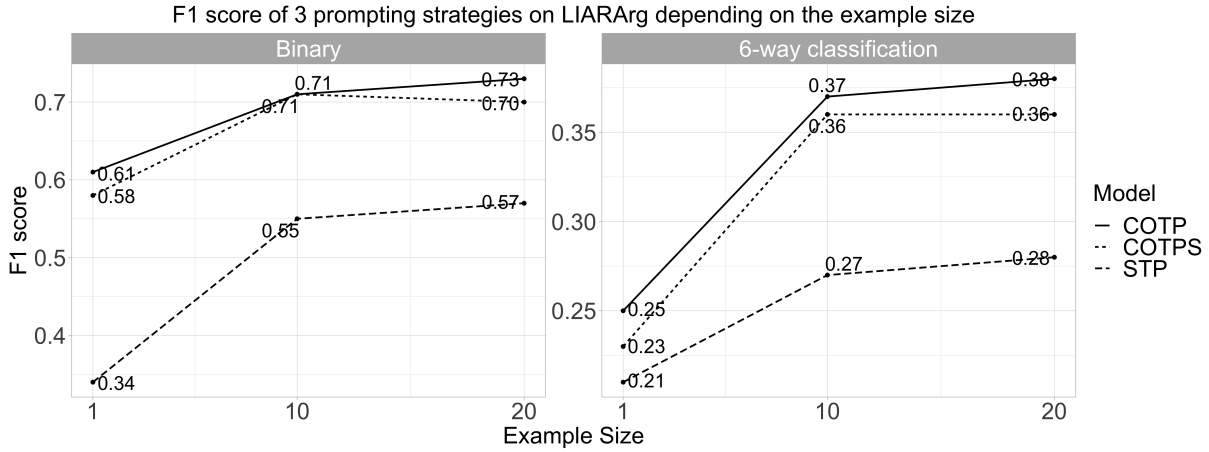

Fig. 6. The impact of example size on the performance of COT-based models. Improvements are statistically significant for all the models when the number of examples increases from 1 to 10, while the improvement from 10 to 20 is marginal.

producing an F1-score of 0.06 under the best MTL model for 6-way classification. This indicates that using 20 annotations per label in our few-shot setting leads to results comparable to the best joint learning model on LIARArg when argument features are introduced explicitly (COTP). However, without these features, the performance falls short of its potential (COTPS). Furthermore, Figure 6 shows the effect of example size on the performance of COT-based models. We conduct the Wilcoxon signed-rank test on the F1 scores of 10 runs for each model and example size. Significant improvements can be observed for all the models when the number of examples increases from 1 to 10, while the improvement from 10 to 20 is marginal and not statistically significant. This suggests that 10 examples would be sufficient to produce a model with a performance close to our best MTL model. These findings show that COT combined with annotated argumentative information can significantly assist the task of FNC with a very small number of examples, reducing the amount of annotated data needed to train fact-checking systems.

It is important to highlight that, in addition to yielding results comparable to those of COTP, COTPS produces rationals in argumentative form, which can then be used to improve the transparency of automated fact-checking [41][12]. In particular, some rationales produced by COTPS are more concise than human-written justifications, as shown in Example (3) (labeled as False). This provides a promising

[12]For readers' reference, it is important to highlight that, in our experiments, the inclusion or exclusion of rationales in the output of COTPS did not affect its performance.

alternative to existing explanatory methods, such as highlighting the salient tokens [19] or extracting sentences [64].

(3) **Claim:** Iran might not be a superpower, but the threat the government of Iran poses is anything but “tiny” as Obama says.
Human-written Justification: One could argue whether it’s wise to meet with leaders of rogue nations. One could also debate whether Obama wrongly downplayed the threat posed by Iran. But Obama never said the threat from Iran was “tiny" or “insignificant,", only that the threat was tiny in comparison to the threat once posed by the Soviet Union.
COTPS: The news claims... According to the justification, “Obama never said the threat from Iran was ‘tiny’ or ‘insignificant’" is a premise attacking the news claim. “only that the threat was tiny in comparison to the threat once posed by the Soviet Union” is a premise partially supporting the news claim.

5.3. Ablation study on argument relations

Table 5
F1 scores of the ablation study on the impact of fine-grained relations in FNC. The results show that without fine-grained relations (-F), the performance of all the models drops significantly. Comparisons of statistical significance have been performed between each experimental setting and its corresponding -F variant.

Model	Binary valid	Binary test	6-way valid	6-way test
KGB+RC	0.73	0.74	0.41	0.42
KGB+RC-F	**0.67**	**0.68**	**0.29**	**0.33**
COTP1	0.61	0.61	0.26	0.25
COTP1-F	**0.57**	**0.58**	**0.23**	**0.23**
COTP20	0.72	0.73	0.39	0.38
COTP20-F	**0.61**	**0.60**	**0.25**	**0.23**
COTPS1	0.59	0.58	0.24	0.23
COTPS1-F	**0.53**	**0.54**	**0.20**	**0.21**
COTPS20	0.68	0.70	0.35	0.36
COTPS20-F	**0.56**	**0.57**	**0.24**	**0.24**

The 4-class relation annotation aims to investigate the role of fine-grained relations in multi-class classification. Table 5 reports the F1-scores of our best-performing models with or without merging fine-grained relations, where “-F” indicates that we merge “partial support” with “support” and “partial attack” with “attack”. We also re-run COT-based models using only one example from each label to investigate the interaction between the number of examples and the obtained F1 score. It can be seen that without fine-grained argument relations, the performance of all the models drops significantly. It can also be observed that fine-grained relations form a synergy with the number of examples fed into COT-based models. For example, when fine-grained relations are provided in the prompt, F1 score of COTP1 increases from 0.25 to 0.38 compared to COTP20 for the 6-way classification, while this size effect is much smaller in -F settings (0.23 in COTP1-F *vs.* 0.24 in COTP20-F). The same observation applies to COTPS models. This highlights the importance of fine-grained relations for few-shot learning, suggesting that adding more examples will not significantly increase the classification performance unless relevant information, i.e., fine-grained argument relations, is conveyed.

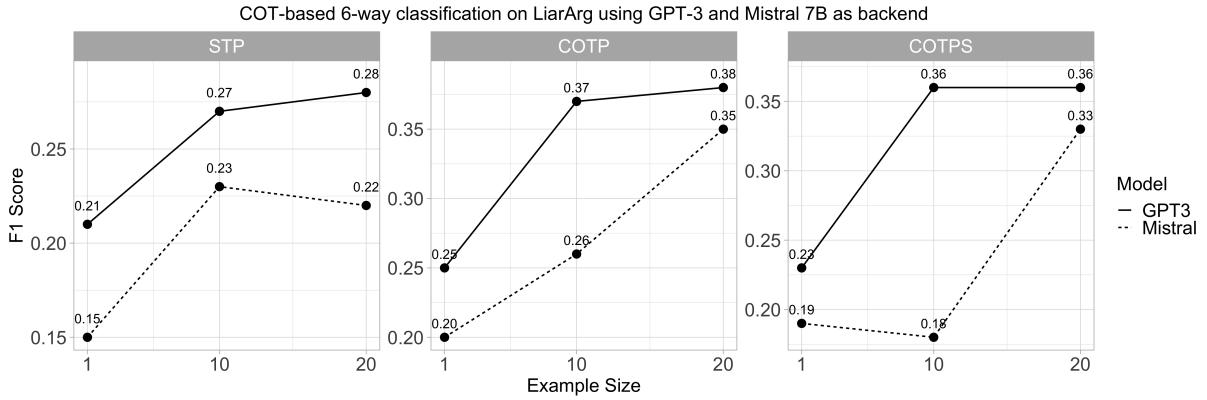

Fig. 7. The performance of Mistral 7B compared to GPT-3 in the 6-way classification on LIARArg. Note the comparable performance between Mistral 7B and GPT-3 when the number of examples is 20 and when argument features are provided.

Table 6

Results of MTL-based and prompt-based FNC models in F1 score compared with LG and single-task based models on automatically parsed LIAR [82] using an adapted version of Morio et al. [43]'s model. Knowledge-enhanced LLMs (KB and KGB) remain strong baselines compared to the previous baseline (LG). The addition of relation prediction significantly improves the performance of FNC models. ST has been compared against LG, +RC against ST and finally, COTP and COTPS against STP.

Split \ Model	LG	KB				KGB				STP	COTP	COTPS
		ST	+CC	+RC	+CCRC	ST	+CC	+RC	+CCRC			
Binary Valid	0.65	**0.71**	0.70	**0.76**	0.72	**0.73**	0.73	**0.77**	0.72	0.62	**0.73**	**0.70**
Binary Test	0.64	**0.72**	0.71	**0.77**	0.72	**0.74**	0.71	**0.78**	0.72	0.61	**0.74**	**0.69**
6-way Valid	0.33	**0.39**	0.41	**0.43**	0.40	**0.41**	0.39	**0.45**	0.40	0.28	**0.40**	**0.36**
6-way Test	0.32	**0.40**	0.40	**0.43**	0.39	**0.43**	0.40	**0.44**	0.37	0.29	**0.40**	**0.37**

5.4. Results for the automated FNC pipeline

To evaluate the efficiency performance of Mistral 7B compared to GPT-3, we first run the same COT-based experiments on LIARArg using Mistral 7B. Figure 7 shows the results of both LLMs in the 6-way classification. It can be observed that Mistral 7B generally falls short of GPT-3's performance in the simple STP setting regardless of the number of examples employed. However, in cases where argument features are integrated, Mistral 7B shows performance on par with GPT-3 when using 20 examples. This suggests Mistral 7B as a relevant alternative to GPT-3 in our setting when the number of examples is large. For this reason, we use 20 examples for the automated pipeline. As described in Section 4.3, 10 samples of 20 examples have been used to provide a robust evaluation.

Table 6 reports the results obtained by the fully automated pipeline (i.e., the automatic AM parser paired with the MTL-based or COT-based FNC models) on LIAR-PLUS from which LIARArg has been removed. To the best of our knowledge, our framework is the first one that integrates AM and FC in a pipeline fashion using an MTL approach and COT-based methods. It can be observed that all the variants of our pipeline outperform the previous baseline produced by LG. Knowledge-enhanced LLMs remain strong baselines. The same pattern is observed as in Section 5.1: notably, jointly training argument relation prediction significantly enhances the performance of FNC models. For prompt-based methods, COTP and COTPS outperform STP by a large margin. It is important to note that, for 6-way

Table 7

Results of MTL-based and prompt-based FNC models in F1 score compared with SOTA [61] and single-task based models on automatically parsed FNC-1 [55] using an adapted version of Morio et al. [43]'s model. On out-of-domain data, the best MTL-based pipeline (KGB+RC) and the best prompt-based pipeline (COTPS) achieve results nearly matching the state-of-the-art. +RC has been compared against ST while COTP and COTPS against STP.

Model / Split	SOTA	KB				KGB				STP	COTP	COTPS
		ST	+CC	+RC	+CCRC	ST	+CC	+RC	+CCRC			
4-way Test	0.90	0.82	0.81	0.85	0.85	0.84	0.83	**0.89**	0.89	0.78	**0.86**	**0.83**

Table 8

Results of MTL-based and prompt-based FNC models in F1 score compared with SOTA [81] and single-task based models on automatically parsed Check-COVID [55] using an adapted version of Morio et al. [43]'s model. On domain-specific data (Covid), the best MTL-based pipeline (KGB+RC) and the best prompt-based pipeline (COTPS) achieve results close to state-of-the-art performance. +RC has been compared against ST while COTP and COTPS against STP.

Model / Split	SOTA	KB				KGB				STP	COTP	COTPS
		ST	+CC	+RC	+CCRC	ST	+CC	+RC	+CCRC			
Binary Test	0.72	0.61	0.60	0.67	0.64	0.63	0.62	**0.69**	0.69	0.52	**0.66**	0.62

classification, it is widely recognized that LIAR-PLUS is a challenging dataset, with most works still struggling to achieve F1 scores higher than 0.35. For instance, Koloski et al. [31] cited a SOTA F1 of 0.37, Yang et al. [88] achieved 0.29, and the most recent work by Wang et al. [80] achieved 0.31. The best result to our knowledge is from Sadeghi et al. [60], who achieved 0.41 using full-length justifications. The best F1 score reported in our work (0.44 by KGB+RC) is therefore a significant improvement on LIAR-PLUS.

Table 7 and Table 8 report the results of the automated pipeline on out-of-domain data, namely on the FNC-1 and Check-COVID datasets. It can be seen that the best MTL-based automated pipeline (KGB+RC), using an AM parser trained on data non-specific to the test data, achieves results nearly matching the state-of-the-art models for both datasets (0.89 *vs.* 0.90 for FNC-1 and 0.69 *vs.* 0.72 in the case of Check-COVID), demonstrating the robustness of our approach on out-of-domain data and, more specifically, confirming the relevance of argument relations in the task of FNC. It is important to note that the current best models for FNC-1 and Check-COVID both use specific domain knowledge to enhance the original input, while our automated pipeline is domain-agnostic. Regarding the pipeline employing prompt-based models, although there is a larger gap between the highest F1 score and the previous baseline (0.86 *vs.* 0.90 for FNC-1, and 0.66 *vs.* 0.72 for Check-COVID), it is essential to underline that STP is outperformed by a large margin compared to COTP and COTPS, highlighting the valuable improvement brought by the integration of argumentation to the Fake News Classification task. These outcomes demonstrate the relevance and viability of automatically incorporating argumentative information into FNC, setting the stage for a fully automated pipeline devoid of human annotation requirements.

5.5. Error Analysis

We first analyze the error distribution when relation classification is not used as an auxiliary task. Table 9 reports the F1-scores of the 6-way classification produced by the best MTL model (KGB+RC) and two baselines on the test set. It can be seen that without relation information, the improvement induced by KGB is mainly limited to the Pants-on-Fire and True labels, meaning that errors in intermediate labels persist despite the use of knowledge-enhanced LLMs. Indeed, 85% of the erroneous predictions for the intermediate labels remained the same in LG *vs.* KGB. These intermediate labels, or half-truths, are

Table 9

F1 scores for 6-way classification in two baselines *vs.* our best joint-learning model. Without relation information (KGB+RC), the improvement induced by KGB is mainly limited to the Pants-on-Fire and True labels. KGB has been compared with LG and KGB+RC with KGB.

Class	LG	KGB	KGB+RC
Pants-on-fire	0.35	**0.50**	**0.60**
False	0.29	0.32	**0.39**
Mostly-false	0.27	0.29	**0.40**
Half-true	0.26	0.25	**0.35**
Mostly-true	0.31	0.33	**0.33**
True	0.40	**0.38**	**0.45**
Avg	0.31	0.36	0.42

omnipresent and computationally more challenging to detect than other forms of disinformation [22], e.g., [42] even filter out half-truths before testing.

Concerning KGB+RC, we notice a strong correlation between the performance in relation classification and FNC. We observe that incorrect relation classification leads to a 45% error rate in FNC *vs.* 30% error rate when all the argument relations are correctly identified. For intermediate labels, the error rate rises from 40% to 60% when at least one error is made in relation classification. These results further confirm the importance of argument relations in FNC, particularly for intermediate labels. Example (4) shows a typical case of half-true news item where $Premise_1$ supports $\mathbf{Claim}_1$, while $Premise_2$ attacks $\mathbf{Claim}_1$. Most instances of this kind are correctly classified by KGB+RC only when all the relations are correctly classified.

(4) **[The economic turnaround started at the end of my term]**$_1$.
{During Crist's last year in office, Florida's economy experienced notable gains in personal income and industrial production, and more marginal improvements in the unemployment rate and in payroll employment}$_1$. *{But GDP didn't grow again until Scott took office}*$_2$.

It is important to highlight two common scenarios where the system tends to misclassify labels. The first occurs when argument components cannot be assessed in isolation. For instance, as demonstrated in Example (5) (True predicted as False), all the three premises must be considered together to accurately classify the claim. The second arises when resolving temporal relations is necessary. For instance, in Example (6) (False predicted as True), humans may clearly recognize that Reagan's presidency pertains to years before 2008 and 2009 based on the premise provided. However, this understanding may not have been adequately captured by the model.

(5) **[Donald Trump has changed his mind on abortion]**$_1$.
{As late as 2000, he wrote that he was pro-choice.}$_1$. *{By 2011, he said he was pro-life}*$_2$. *{Recently, he noted that he thinks exceptions for the life of the mother, incest and rape are appropriate}*$_3$.

(6) **[Ronald Reagan faced an even worse recession than the current one.]**$_1$.
{The misery index has been lower – 5.7 in 2008 and 11.8 so far in 2009}$_1$.

6. Concluding remarks

The main goal of this paper is to investigate whether the argumentative representation of evidence aids in the fake news classification task. To address this challenging issue, we present LIARArg, the first FNC dataset annotated with argument components and relations. Unlike LIAR-PLUS, in LIARArg we remove insufficient justifications, making it a solid benchmark for future research investigating how the internal structures of evidence can be better leveraged in the FNC task to improve the effectiveness of FNC models. Moreover, we propose a Multi-Task Learning framework to jointly learn FNC, CC and RC, as well as a COT-based framework to explicitly inject argumentative structures in a few-shot learning setting. Knowledge-enhanced embeddings are used to establish strong baselines for comparison.

The reader may argue that GPT-3 or Mistral 7B might have been trained on data comprising the datasets used in our experiments. As highlighted in [92], the potential risk of data leakage is indeed a growing concern in benchmark evaluation. Unfortunately, the training corpora for both LLMs are not publicly available, making it difficult to confirm or refute this hypothesis.

Our experiments show that argument relations, particularly fine-grained relations such as partial support and partial attack, significantly improve the performance of knowledge-enhanced FNC models. This enhancement is most notable in accurately determining intermediate truth labels, indicating a substantial advancement in the model's ability to discern complex, graded truth values. The best results are achieved using the Multi-Task Learning framework, outperforming both the SOTA approach on LIAR and most recent approaches in FNC based on knowledge-enhanced LLMs. Under the few-shot setting, COT-based methods yield results comparable to the best results with only 20 examples per label. To the best of our knowledge, this is the first approach showing that Argumentation Mining can be jointly trained with Fake News Classification to improve the latter's performance. Our work is also the first to exploit argument structures contained in evidence in a Chain-Of-Thought manner both in prompts and model outputs. Finally, we show that the integration of argumentation information into FNC is feasible without human annotations through a fully automated pipeline. This, along with prompting-based approaches, is a promising direction for future research to reduce the amount of annotated data to train fact-checking systems.

More specifically, future works focus on designing new annotation schemes for datasets so that models would be capable of considering multiple argument components simultaneously when classifying argument relations as in Example (5) where all the premises should be considered altogether to determine whether the news claim is attacked or supported, which is typical of temporal events. In terms of prompting, it would be particularly valuable to explore how the types of examples included in prompts interact with the performance of COT-based methods. Also, although weak attack and weak support have been added to the annotation scheme, it would be interesting to see if a more fine-grained typology of argument relations (e.g., classifying attack relations using Walton Schemes [79]) can be explored to further improve the performance of FNC models. Also, although COT-based models display the potential to produce argument-structure-like explanations (cf. Example (3) in Section 5.2), the extent to which these explanations are understandable to humans and how they can be used to enhance the transparency of automated FNC systems remains to be investigated through an extensive human evaluation. Finally, the combination of evidence retrieval, AM, and FNC presents a fascinating avenue for exploration. However, the corpora analyzed in this study are not particularly suited for such experiments because each news title in these datasets is typically linked to only one fact-checking article, rather than multiple articles. As a result, testing the relevance of retrieving multiple articles is not feasible. If we restrict retrieval to a single article, the correct one is usually retrieved, but this setup fails to reflect a realistic scenario. A

more appropriate approach, as in [83], would ideally retrieve multiple fact-checking articles related to the same news title to better simulate real-world conditions.

Acknowledgements

This work has been partially supported by the ANR project ATTENTION (ANR-21-CE23-0037) and the French government through the 3IA Côte d'Azur Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002.

References

[1] S. Abbas and H. Sawamura, A first step towards argument mining and its use in arguing agents and its, in: *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, Springer, 2008, pp. 149–157.
[2] P. Accuosto and H. Saggion, Mining arguments in scientific abstracts with discourse-level embeddings, *Data & Knowledge Engineering* **129** (2020), 101840.
[3] J. Ainslie, J. Lee-Thorp, M. de Jong, Y. Zemlyanskiy, F. Lebron and S. Sanghai, GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints, in: *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.
[4] T. Alhindi, S. Petridis and S. Muresan, Where Is Your Evidence: Improving Fact-checking by Justification Modeling, in: *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 85–90. doi:10.18653/v1/W18-5513.
[5] S.A. Aljawarneh and S.A. Swedat, Fake News Detection Using Enhanced BERT, *IEEE Transactions on Computational Social Systems* (2022).
[6] K.D. Ashley and V.R. Walker, Toward constructing evidence-based legal arguments using legal decision documents and machine learning, in: *Proceedings of the fourteenth international conference on artificial intelligence and law*, 2013, pp. 176–180.
[7] K. Atkinson, P. Baroni, M. Giacomin, A. Hunter, H. Prakken, C. Reed, G. Simari, M. Thimm and S. Villata, Towards artificial argumentation, *AI magazine* **38**(3) (2017), 25–36.
[8] C. Blackledge and A. Atapour-Abarghouei, Transforming Fake News: Robust Generalisable News Classification Using Transformers, in: *2021 IEEE International Conference on Big Data (Big Data)*, IEEE, 2021, pp. 3960–3968.
[9] F. Boltužić and J. Šnajder, Fill the Gap! Analyzing Implicit Premises between Claims from Online Debates, in: *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, 2016, pp. 124–133.
[10] S.R. Bowman, G. Angeli, C. Potts and C.D. Manning, A Large Annotated Corpus for Learning Natural Language Inference, in: *Conference on Empirical Methods in Natural Language Processing, EMNLP 2015*, Association for Computational Linguistics (ACL), 2015, pp. 632–642.
[11] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell et al., Language models are few-shot learners, *Advances in neural information processing systems* **33** (2020), 1877–1901.
[12] J. Burstein and D. Marcu, A machine learning approach for identification thesis and conclusion statements in student essays, *Computers and the Humanities* **37** (2003), 455–467.
[13] E. Cabrio and S. Villata, Five Years of Argument Mining: a Data-driven Analysis, in: *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, J. Lang, ed., ijcai.org, 2018, pp. 5427–5433. doi:10.24963/ijcai.2018/766.
[14] R.A. Caruana, Multitask Learning: A Knowledge-Based Source of Inductive Bias, in: *Machine Learning Proceedings 1993*, Elsevier, 1993, pp. 41–48. ISBN 978-1-55860-307-3. doi:10.1016/B978-1-55860-307-3.50012-5.
[15] W. Chen, H. Wang, J. Chen, Y. Zhang, H. Wang, S. Li, X. Zhou and W.Y. Wang, TabFact: A Large-scale Dataset for Table-based Fact Verification, in: *International Conference on Learning Representations*, 2019.
[16] R. Child, S. Gray, A. Radford and I. Sutskever, Generating long sequences with sparse transformers, *arXiv preprint arXiv:1904.10509* (2019).
[17] M. de Cock Buning, *A multi-dimensional approach to disinformation: Report of the independent high level group on fake news and online disinformation*, Publications Office of the European Union, 2018.
[18] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding, *arXiv preprint arXiv:1810.04805* (2018).

[19] V. Dua, A. Rajpal, S. Rajpal, M. Agarwal and N. Kumar, I-FLASH: Interpretable Fake News Detector Using LIME and SHAP, *Wireless Personal Communications* (2023), 1–34.
[20] L. Duong, T. Cohn, S. Bird and P. Cook, Low Resource Dependency Parsing: Cross-lingual Parameter Sharing in a Neural Network Parser, in: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, Association for Computational Linguistics, Beijing, China, 2015, pp. 845–850. doi:10.3115/v1/P15-2139.
[21] M. Elaraby and D. Litman, ArgLegalSumm: Improving abstractive summarization of legal documents with argument mining, *arXiv preprint arXiv:2209.01650* (2022).
[22] A. Estornell, S. Das and Y. Vorobeychik, Deception through Half-Truths, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, 2020, pp. 10110–10117.
[23] W. Ferreira and A. Vlachos, Emergent: A Novel Data-Set for Stance Classification, in: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, San Diego, California, 2016, pp. 1163–1168. doi:10.18653/v1/N16-1138.
[24] J.B. Freeman, *Argument Structure:: Representation and Theory*, Vol. 18, Springer Science & Business Media, 2011.
[25] D. Ghosh, A. Khanam, Y. Han and S. Muresan, Coarse-grained argumentation features for scoring persuasive essays, in: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2016, pp. 549–554.
[26] Z. Guo, M. Schlichtkrull and A. Vlachos, A Survey on Automated Fact-Checking, *Transactions of the Association for Computational Linguistics* **10** (2022), 178–206. doi:10.1162/tacl$_{a0}$0454.
[27] C. Hidey and M. Diab, Team SWEEPer: Joint Sentence Extraction and Fact Checking with Pointer Networks, in: *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, 2018, pp. 150–155.
[28] A.Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D.S. Chaplot, D.d.l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier et al., Mistral 7B, *arXiv preprint arXiv:2310.06825* (2023).
[29] R.K. Kaliyar, A. Goswami and P. Narang, FakeBERT: Fake News Detection in Social Media with a BERT-based Deep Learning Approach, *Multimedia tools and applications* **80**(8) (2021), 11765–11788.
[30] C. Kirschner, J. Eckle-Kohler and I. Gurevych, Linking the thoughts: Analysis of argumentation structures in scientific publications, in: *Proceedings of the 2nd Workshop on Argumentation Mining*, 2015, pp. 1–11.
[31] B. Koloski, T. Stepišnik Perdih, M. Robnik-Šikonja, S. Pollak and B. Škrlj, Knowledge Graph Informed Fake News Classification via Heterogeneous Representation Ensembles, *Neurocomputing* **496** (2022), 208–226. doi:10.1016/j.neucom.2022.01.096.
[32] J.R. Landis and G.G. Koch, The Measurement of Observer Agreement for Categorical Data, *Biometrics* **33**(1) (1977), 159–174. doi:10.2307/2529310.
[33] A. Lauscher, H. Wachsmuth, I. Gurevych and G. Glavas, Scientia Potentia Est - On the Role of Knowledge in Computational Argumentation, *Trans. Assoc. Comput. Linguistics* **10** (2022), 1392–1422. https://transacl.org/ojs/index.php/tacl/article/view/3967.
[34] J. Lawrence and C. Reed, Argument Mining: A Survey, *Computational Linguistics* **45**(4) (2020), 765–818. doi:10.1162/coli$_{a0}$0364.
[35] M. Liakata, S. Saha, S. Dobnik, C. Batchelor and D. Rebholz-Schuhmann, Automatic recognition of conceptualization zones in scientific articles and two life science applications, *Bioinformatics* **28**(7) (2012), 991–1000.
[36] C.-Y. Lin, C.-C. Chen, H.-H. Huang and H.-H. Chen, Argument-Based Sentiment Analysis on Forward-Looking Statements, in: *Findings of the Association for Computational Linguistics ACL 2024*, 2024, pp. 13804–13815.
[37] J. Ma, W. Gao, S. Joty and K.-F. Wong, Sentence-Level Evidence Embedding for Claim Verification with Hierarchical Attention Networks, in: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 2561–2571.
[38] J. Ma, C. Chen, C. Hou and X. Yuan, KAPALM: Knowledge grAPh enhAnced Language Models for Fake News Detection, in: *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023, pp. 3999–4009.
[39] T. Mayer, S. Marro, E. Cabrio and S. Villata, Enhancing evidence-based medicine with natural language argumentative analysis of clinical trials, *Artificial Intelligence in Medicine* **118** (2021), 102098.
[40] D. Mehta, A. Dwivedi, A. Patra and M. Anand Kumar, A Transformer-Based Architecture for Fake News Classification, *Social Network Analysis and Mining* **11**(1) (2021), 39. doi:10.1007/s13278-021-00738-y.
[41] K. Mishima and H. Yamana, A Survey on Explainable Fake News Detection, *IEICE Transactions on Information and Systems* **E105.D**(7) (2022), 1249–1257. doi:10.1587/transinf.2021EDR0003.
[42] R.A. Monteiro, R.L. Santos, T.A. Pardo, T.A. De Almeida, E.E. Ruiz and O.A. Vale, Contributions to the Study of Fake News in Portuguese: New Corpus and Automatic Detection Results, in: *Computational Processing of the Portuguese Language: 13th International Conference, PROPOR 2018, Canela, Brazil, September 24–26, 2018, Proceedings 13*, Springer, 2018, pp. 324–334.
[43] G. Morio, H. Ozaki, T. Morishita and K. Yanai, End-to-End Argument Mining with Cross-corpora Multi-task Learning, *Transactions of the Association for Computational Linguistics* **10** (2022), 639–658. doi:10.1162/tacl$_{a0}$0481.

[44] H. Nguyen and D. Litman, Argument mining for improving the automated scoring of persuasive essays, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32, 2018.
[45] Y. Nie, H. Chen and M. Bansal, Combining Fact Extraction and Verification with Neural Semantic Matching Networks, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, 2019, pp. 6859–6866.
[46] R.M. Palau and M.-F. Moens, Argumentation mining: the detection, classification and structure of arguments in text, in: *Proceedings of the 12th international conference on artificial intelligence and law*, 2009, pp. 98–107.
[47] A. Parikh, O. Täckström, D. Das and J. Uszkoreit, A Decomposable Attention Model for Natural Language Inference, in: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 2249–2255.
[48] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research* **12** (2011), 2825–2830.
[49] A. Peldszus and M. Stede, An annotated corpus of argumentative microtexts, in: *Argumentation and Reasoned Action: Proceedings of the 1st European Conference on Argumentation, Lisbon*, Vol. 2, 2015, pp. 801–815.
[50] A. Peldszus and M. Stede, An Annotated Corpus of Argumentative Microtexts, in: *Argumentation and Reasoned Action: Proceedings of the 1st European Conference on Argumentation, Lisbon*, Vol. 2, 2015, pp. 801–815.
[51] A. Peldszus and M. Stede, Joint prediction in MST-style discourse parsing for argumentation mining, in: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 938–948.
[52] K. Pelrine, J. Danovitch and R. Rabbany, The surprising performance of simple baselines for misinformation detection, in: *Proceedings of the Web Conference 2021*, 2021, pp. 3432–3441.
[53] M.E. Peters, M. Neumann, R. Logan, R. Schwartz, V. Joshi, S. Singh and N.A. Smith, Knowledge Enhanced Contextual Word Representations, in: *Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.
[54] D. Pomerleau and D. Rao, The Fake News Challenge: Exploring How Artificial Intelligence Technologies Could Be Leveraged to Combat Fake News., 2017.
[55] D. Pomerleau and D. Rao, The fake news challenge: Exploring how artificial intelligence technologies could be leveraged to combat fake news, *Fake news challenge* (2017).
[56] K. Popat, S. Mukherjee, A. Yates and G. Weikum, DeClarE: Debunking Fake News and False Claims using Evidence-Aware Deep Learning, in: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 2018.
[57] H. Rashkin, E. Choi, J.Y. Jang, S. Volkova and Y. Choi, 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*, 2017, pp. 2931–2937.
[58] S. Raza and C. Ding, Fake News Detection Based on News Content and Social Contexts: A Transformer-Based Approach, *International Journal of Data Science and Analytics* **13**(4) (2022), 335–362.
[59] C. Reed, S. Wells, J. Devereux and G. Rowe, AIF+: Dialogue in the Argument Interchange Format., *Frontiers in artificial intelligence and applications* **172** (2008), 311.
[60] F. Sadeghi, A.J. Bidgoly and H. Amirkhani, Fake news detection on social media using a natural language inference approach, *Multimedia Tools and Applications* **81**(23) (2022), 33801–33821.
[61] I. Salah, K. Jouini and O. Korbaa, On the use of text augmentation for stance and fake news detection, *Journal of Information and Telecommunication* **7**(3) (2023), 359–375.
[62] M.S. Schlichtkrull, V. Karpukhin, B. Oguz, M. Lewis, W.-t. Yih and S. Riedel, Joint Verification and Reranking for Open Fact Checking Over Tables, in: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 6787–6799.
[63] J. Schneider, Automated argumentation mining to the rescue? Envisioning argumentation and decision-making support for debates in open online collaboration communities, in: *Proceedings of the First Workshop on Argumentation Mining*, 2014.
[64] T. Schuster, A. Fisch and R. Barzilay, Get Your Vitamin C! Robust Fact Verification with Contrastive Evidence, in: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 624–643.
[65] G.K. Shahi and D. Nandini, *FakeCovid – A Multilingual Cross-domain Fact Check News Dataset for COVID-19*, 2020. doi:10.36190/2020.14.
[66] B. Shi and T. Weninger, Discriminative Predicate Path Mining for Fact Checking in Knowledge Graphs, *Knowledge-based systems* **104** (2016), 123–133.
[67] P. Sobhani, D. Inkpen and S. Matwin, From argumentation mining to stance classification, in: *Proceedings of the 2nd Workshop on Argumentation Mining*, 2015, pp. 67–77.
[68] C. Stab and I. Gurevych, Parsing Argumentation Structures in Persuasive Essays, *Computational Linguistics* **43**(3) (2017), 619–659.

[69] J. Thorne, A. Vlachos, C. Christodoulopoulos and A. Mittal, FEVER: A Large-scale Dataset for Fact Extraction and VERification, in: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 809–819. doi:10.18653/v1/N18-1074.
[70] J. Thorne, A. Vlachos, O. Cocarascu, C. Christodoulopoulos and A. Mittal, The FEVER 2.0 Shared Task, in: *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, 2019, pp. 1–6.
[71] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale et al., Llama 2: Open foundation and fine-tuned chat models, *arXiv preprint arXiv:2307.09288* (2023).
[72] J.E. Uscinski and R.W. Butler, The epistemology of fact checking, *Critical Review* **25**(2) (2013), 162–180.
[73] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò and Y. Bengio, Graph Attention Networks, in: *International Conference on Learning Representations*, 2018.
[74] O. Vinyals, M. Fortunato and N. Jaitly, Pointer Networks, *Advances in neural information processing systems* **28** (2015).
[75] J. Visser, J. Lawrence and C. Reed, Reason-Checking Fake News, *Communications of the ACM* **63**(11) (2020), 38–40. doi:10.1145/3397189.
[76] A. Vlachos and S. Riedel, Fact Checking: Task Definition and Dataset Construction, in: *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, 2014, pp. 18–22.
[77] D. Vrandečić and M. Krötzsch, Wikidata: a free collaborative knowledgebase, *Communications of the ACM* **57**(10) (2014), 78–85.
[78] H. Wachsmuth, K. Al Khatib and B. Stein, Using argument mining to assess the argumentation quality of essays, in: *Proceedings of COLING 2016, the 26th international conference on Computational Linguistics: Technical papers*, 2016, pp. 1680–1691.
[79] D. Walton, C. Reed and F. Macagno, *Argumentation Schemes*, Cambridge University Press, 2008.
[80] B. Wang, J. Ma, H. Lin, Z. Yang, R. Yang, Y. Tian and Y. Chang, Explainable Fake News Detection With Large Language Model via Defense Among Competing Wisdom, in: *Proceedings of the ACM on Web Conference 2024*, 2024, pp. 2452–2463.
[81] G. Wang, K. Harwood, L. Chillrud, A. Ananthram, M. Subbiah and K. McKeown, Check-COVID: Fact-Checking COVID-19 News Claims with Scientific Evidence, in: *Findings of the Association for Computational Linguistics: ACL 2023*, A. Rogers, J. Boyd-Graber and N. Okazaki, eds, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 14114–14127. doi:10.18653/v1/2023.findings-acl.888. https://aclanthology.org/2023.findings-acl.888.
[82] W.Y. Wang, "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)*, Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 422–426. doi:10.18653/v1/P17-2067.
[83] X. Wang, E. Cabrio and S. Villata, Argument-structured Justification Generation for Explainable Fact-checking, in: *2024 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, Bangkok, Thailand, 2024. https://inria.hal.science/hal-04862965.
[84] X. Wang, T. Gao, Z. Zhu, Z. Zhang, Z. Liu, J. Li and J. Tang, KEPLER: A unified model for knowledge embedding and pre-trained language representation, *Transactions of the Association for Computational Linguistics* **9** (2021), 176–194.
[85] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le and D. Zhou, Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, arXiv, 2023.
[86] C. Whitehouse, T. Weyde, P. Madhyastha and N. Komninos, Evaluation of fake news detection with knowledge-enhanced language models, in: *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16, 2022, pp. 1425–1429.
[87] H. Xu, J. Šavelka and K.D. Ashley, Using argument mining for legal text summarization, in: *Legal Knowledge and Information Systems*, IOS Press, 2020, pp. 184–193.
[88] Z. Yang, J. Ma, H. Chen, H. Lin, Z. Luo and Y. Chang, A coarse-to-fine cascaded evidence-distillation neural network for explainable fake news detection, *arXiv preprint arXiv:2209.14642* (2022).
[89] W. Yin and D. Roth, TwoWingOS: A Two-Wing Optimization Strategy for Evidential Claim Verification, in: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 105–114.
[90] Y. Zhang and Q. Yang, A Survey on Multi-Task Learning, *IEEE Transactions on Knowledge and Data Engineering* **34**(12) (2021), 5586–5609.
[91] Z. Zhang, W. Yu, M. Yu, Z. Guo and M. Jiang, A Survey of Multi-task Learning in Natural Language Processing: Regarding Task Relatedness and Training Methods, in: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2023, pp. 943–956.
[92] K. Zhou, Y. Zhu, Z. Chen, W. Chen, W.X. Zhao, X. Chen, Y. Lin, J.-R. Wen and J. Han, Don't Make Your LLM an Evaluation Benchmark Cheater, *arXiv preprint arXiv:2311.01964* (2023).