Towards Fact-Check Summarization Leveraging on Argumentation Elements Tied to Entity Graphs

Kateřina Haniková
katerina.hanikova@vse.cz
Prague University of Economics and Business
Prague, Czech Republic

David Chudán
david.chudan@vse.cz
Prague University of Economics and Business
Prague, Czech Republic

Vojtěch Svátek
svatek@vse.cz
Prague University of Economics and Business
Prague, Czech Republic

Peter Vajdečka
peter.vajdecka@vse.cz
Prague University of Economics and Business
Prague, Czech Republic

Raphaël Troncy
raphael.troncy@eurecom.fr
EURECOM
Sophia Antipolis, France

Filip Vencovský
filip.vencovsky@vse.cz
Prague University of Economics and Business
Prague, Czech Republic

Jana Syrovátková
jana.syrovatkova@vse.cz
Prague University of Economics and Business
Prague, Czech Republic

ABSTRACT

Fact-check consumers can have different preferences regarding the amount of text being used for explaining the claim veracity verdict. Dynamically adapting the size of a fact-check report is thus an important functionality for systems designed to convey claim verification explainability. Recent works have experimented with applying transformers-based or LLM-based text summarization methods in a zero-shot or few-shot manner, making use of some existing texts available in the summary parts of fact-check reports (e.g., called "justification" in PolitiFact). However, for complex fact-checks, the purely sub-symbolic summarizers tend to either omit some elements of the fact-checker's argumentation chains or include contextual statements that may not be essential at the given level of granularity. In this paper, we propose a new method for enhancing fact-check summarization with the aim of injecting elements of structured fact-checker argumentation. This argumentation is, in turn, not only captured at the discourse level but tied to an entity graph representing the fact-check, for which we employ the PURO diagrammatic language. We have empirically performed a manual analysis of fact-check reports from two fact-checker websites, yielding (1) textual snippets containing the argumentation essence of the fact-check report and (2) categorized argumentation elements tied to entity graphs. These snippets are then fed to a state-of-the-art hybrid summarizer which has previously produced accurate fact-check summaries, as an additional input. We observe mild improvements on various ROUGE metrics, even if the validity of the results is limited given the small size of the dataset. We also compare the human-provided argumentation element categories with those returned, for the given fact-check ground truth summary, using a pre-trained language model upon both basic and augmented prompting. This yields a moderate accuracy as the model often fails to comply with the explicit given instructions.

CCS CONCEPTS

• **Computing methodologies → Natural language processing**; Ontology engineering.

KEYWORDS

fact-checking, argumentation, entity graph, text summarization

ACM Reference Format:
Kateřina Haniková, David Chudán, Vojtěch Svátek, Peter Vajdečka, Raphaël Troncy, Filip Vencovský, and Jana Syrovátková. 2024. Towards Fact-Check Summarization Leveraging on Argumentation Elements Tied to Entity Graphs. In *Companion Proceedings of the ACM Web Conference 2024 (WWW '24 Companion), May 13–17, 2024, Singapore, Singapore*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3589335.3651914

1 INTRODUCTION

Fact-check reports are, on the one hand, a potentially powerful (though not always adequately valued) means of *combating misinformation* the public is exposed to, and, on the other hand, an exciting resource for studying the *argumentative discourse* in the fact-checking domain. Fact-check reports are generally heterogeneous due to diverse guidelines coming from the fact-checker organizations and due to the individuals authoring the fact-checks (writers and editors). In this paper, we aim to study further this heterogeneity.

WWW '24 Companion, May 13–17, 2024, Singapore, Singapore

ACM ISBN 979-8-4007-0172-6/24/05
https://doi.org/10.1145/3589335.3651914

With respect to the fact-check report audience, we follow up with prior efforts at delivering the content of these reports in a palatable manner – of which the size of the message is an important aspect. Fact-check consumers can have different preferences regarding the amount of text being used for explaining the claim veracity verdict, which stresses the role of *text summarization*, allowing to dynamically adapt the size of the text as needed. Concerning the study of fact-checker argumentation, we contribute with an approach that ties the *argumentation elements* to the structure of *entity graphs*, which are, in turn, expressed using the same or similar primitives as today's *knowledge graphs* (i.e., objects/instances described in terms of attributes, types, mutual relationships, etc.).

In summary, in this preliminary study, we aim to qualitatively explore the following research questions:

- **RQ1:** Do the fact-check reports revolve around a relatively stable set of recurring *argumentation elements* annotating particular *entity graph* patterns?
- **RQ2:** Would automated *summarization* of fact-check reports benefit from information on argumentation elements if provided as additional input?
- **RQ3:** Can the argumentation elements present in the text be *automatically* detected?

We argue that positively answering these three questions would open the way to an automatic pipeline yielding high-quality summaries, capable of explaining fact-checker verdicts concisely while still capturing the essence of the argumentation. More broadly, this would also impact the entire textual generation field using generative AI where there is a need to follow argumentative and persuasive strategies.

The remainder of the paper is structured as follows. We review relevant prior research in Section 2. We present and illustrate the framing notions of entity graphs and argumentation elements in Section 3. We detail the results of our empirical studies in Section 4. Finally, we conclude and outline some future work in Section 5.

2 RELATED WORK

We divide the overview of prior research into two threads: the first one corresponding to RQ1 and RQ3 (Section 2.1), and the second one to RQ2 (Section 2.2).

2.1 Argumentation modeling and detection

Formal modeling of argumentation is an area of extensive research. Among the ontology-based approaches, one of the main work is the Argument Model Ontology (AMO) [10] based on Toulmin's theory [13, 24]. Another attempt to formalize argumentation structures and to interchange data between argumentation tools is the Argument Interchange Format (AIF) [5]. AIF is more complex and general than AMO. It is applicable to various schemes, and thus, it encompasses Toulmin's argumentation scheme. The main idea is to interchange argumentation structures between different software systems with a focus on machine-readable syntax. Our goal is to leverage these models in order to explain misleading claims and expose the argumentation chain to the end-users.

The long history of argument diagramming is surveyed in [21]. As a prominent early example of such diagramming tools, Araucaria [20],[1] is based on the Argumentation Markup Language (AML). The main difference with our approach is that the argument diagram created using Araucaria is focused on entire sentences of the discourse, whereas the entity diagrams designed in our approach depict entities and relationships between them, which typically correspond to small text chunks of a few tokens in text. To the best of our knowledge, there are no published methods that tightly integrate entity diagrams with argumentation. Most recent state-of-the-art tools, such as DISPUTool [11], not only allow for manual diagramming but also automatically suggest categories of arguments. DISPUTool additionally supports named entity recognition in text. However, the entities are not woven into an entity graph as in our approach.

When applied to claims and fact-checks, our approach is complementary to that of the authors of the *Open Claims* model [3] that decomposes the fact-checkable claim into claim proposition, claim utterance and claim context. The *claim proposition* (expressing the meaning of the claim itself) is the sole subject of our graph-based modeling, and it is considered to have one or multiple representations. Besides a textual representation, the *Open Claims* model also suggests a RDF-based and a FOL-based representations [3]. For example, the *Open Claims* instantiation *cost = {of='Brexit', for='UK' amount=?, until=2064}* corresponds to a simple serialization of an n-ary relationship named *cost*, with three of its participants being objects (a kind of 'dimension' of an 'observation', in multi-dimensional data terms) and one attribute (a kind of 'measure', in multi-dimensional data terms), whose value is however left unspecified. Interestingly, we also identified a relationship (whether n-ary or binary) as a typical entity graph representation of a claim. Our graphs may thus complement *Open Claims*, both by zooming into the claim proposition, serving it in an expressive graphical language, but also by explicit treatment not only of the claim but also of the fact-checker argumentation.

The efficiency of automatic argumentation modeling for fact-checking rests on its ability to delve into the intricacies of persuasive language. Consequently, investigating the current state of the art in argumentation mining (AM) becomes paramount. Defined as the automated process of extracting arguments from textual data, AM serves as a critical precursor to automatic argumentation modeling [18]. This multifaceted task encompasses identifying argument components, establishing their roles, and deciphering their relationships. Effective AM paves the way for higher-level analysis and modeling of the text's argumentative structure.

Recent progress highlights the promising role of transformer-based architectures such as BERT [15] to detect argumentation patterns [6, 12, 23]. Furthermore, incorporating techniques like argument attention prior to argument relation labeling demonstrates additional improvements [4]. Inspired by these findings, we explored the application of context-rich transformers and experimented with LLaMA 2 and GPT models.

[1] http://araucaria.arg.tech/doku.php?id=start

2.2 Fact-check report summarization

The textual argumentation formulated by fact-checkers when checking the veracity of a claim can be long and not suitable for rapid diffusion on social media. Thus, it is helpful if a fact-check report has a summary. Creating fact-check reports can take a lot of time, e.g. to collect reliable evidence and put the claim into context. Writing up manually a summary is adding extra work and applying automated summarization techniques is valued by fact-checkers.

To the best of our knowledge, there have only been two efforts aiming to automate the process of fact-checking summarization. Atanasova et al. [1] aimed to generate summaries ('veracity explanations') jointly with veracity predictions, within the same deep learning architecture. Kazemi et al. [14] enhanced the initial and state-of-the-art fact-checking summarization approach, still relying on an advanced version of extractive summarization, which dominates GPT-2 as the to-date state-of-the-art abstractive approach. Summarization has also been employed in the fact-checking process by Bhatnagar [2] and by Yang [27]. However, that was for a different purpose: that of retrieving and/or aggregating disparate (possibly already fact-checked) claims, thus easing the work of fact-checkers before they proceed to elaborate their own report. There is thus a growing focus on enhancing the process of automatically generating justifications, a crucial aspect in the advancement of fact-checking summarization. However, we are unaware of an approach combining fact-check summarization with argumentation identification, never mind in connection with entity graphs.

3 ENTITY GRAPHS AND ARGUMENTATION ELEMENTS

In this section, we first explain the main ingredients of our approach at a general level and then show a concrete example of their use.

3.1 Entity graph approach to fact-check argumentation analysis

Previous attempts to employ knowledge graphs in the fact-checking process departed from the (valid) assumption that human authoring of fact-check reports does not scale. Therefore, many approaches have been proposed to assist the fact-checker, typically by bringing pieces of evidence for assessing the veracity of a claim, sometimes using elements of existing KGs and leveraging on machine learning models [26]. Such approaches, when fully automated from the onset, appear superficial from the domain knowledge capture viewpoint, and suffer from the caveat of relying on incomplete or outdated KGs with respect to entities and relationships mentioned in recent claims. In our approach, in contrast, we take a 'traditional' knowledge-engineering paradigm as a starting point. We believe that the human expertise in fact-check reports should be thoroughly captured at a detailed level.

Hence, we address RQ1 using a knowledge engineer with at least basic familiarity with social and political issues being discussed in current media and social media. Her task is to browse a given fact-check report, to identify the argumentation structure used by the fact-checker as well as by the claim's author, and to output a graph consisting of:

- entities, their attributes, and relationships that appear prominent in the identified argumentation structure – we simply call it *entity graph*
- special annotations, called *argumentation elements*, which can be appended to a nearly arbitrary entity graph element or substructure.

Such a structure constitutes a bridge between real-world entities, which can often be mapped on external knowledge graph elements such as Wikidata, and the structure of fact-checkers (and claim author's) argumentation.

3.2 PURO entity graphs

We make use of the graphical language of so-called *PURO models* [9] to represent the *entity graphs* even if we plan to use an RDF-star based model in the future. This has several motivations:

- The use of PURO models allows to defer the technical decisions, such as the choice of the ontology to reuse, namespaces, IRI conventions, etc., that have little to do with knowledge modeling proper.
- Compared to plain RDF, PURO models natively support n-ary relationships or the use of relationships as arguments of other relationships, thus providing more flexibility in the modeling itself. In RDF, one would need to add artificial nodes for 'reifying' a relationship or a type. One could also use RDF-star directly.
- Preliminary tooling exists for their transformation to RDF KGs. This functionality is not critical in our work, and instead, the motivation is to employ the models developed within this study to gather feedback for the tooling in general.

Modeling in PURO does not impose any specific constraints for the knowledge engineer used to RDF/OWL since the basic triad of 'individuals, classes and predicates' (named 'object, type and relation/ship') is preserved and merely endowed with more flexibility and a few optional built-ins. A similar distinction exists between object and data properties, except that *attributes* (rough analogs of data properties) are to be used exclusively for *quantitative* properties.[2] Finally, for the instantiation relationship (connecting an object to a type, but also a type to a higher-order type), PURO uses a dedicated primitive named *instanceOf* analogous to `rdf:type` in RDF. In the next section, we present an example of a PURO model [8, 9].

3.3 Argumentation element structure

An *argumentation element* may be either a verdict argumentation element or an auxiliary argumentation element. A *verdict argumentation element* is a structure with two facets, jointly expressible as a simple noun phrase ADJ+NN (an adjective and a noun). The *noun*

[2] Previous experience with PURO models indicates that no other 'data properties' are typically needed when one creates the initial drafts of ontological conceptualization. Literals used as values of RDF data properties in KGs, such as, for example, country code strings or boolean values indicating the presence of some features of a device, are mere pragmatic operational encodings and can be conveniently expressed using objects, types, attributes and relationships in 'ideal' conceptualization structures (e.g., countries as objects instead of mere codes; device features as device types or as relationships to components of the device, or the like).

simply refers to an individual entity graph primitive, or multiple interrelated primitives, that is/are being annotated:

- *Individual* graph primitives can be ‘relationship’, ‘attribute’, ‘type’, ‘object’ or ‘value’ (or, possibly, their important semantically refined variants such as ‘utterance’ as a kind of relationship), etc.
- Examples of multiple interrelated primitives are, for example:
  - two objects that are claimed to be similar, the noun is then ‘similarity’;
  - two objects of a different kind (e.g., a person and a company) that are presented next to each other in the text, making the impression that they are associated in some way (without stating an explicit relationship), the noun is then ‘association’.

The adjective is a value for the *veracity verdict*, such as ‘true’ or ‘false’, but also ‘unsubstantiated’, ‘exaggerated’, ‘misleading’ or ‘missing’, which are commonly used by fact-checkers. Intuitive definitions of currently employed verdicts are presented in Table 1. Note that while the verdicts within argumentation elements correspond to possible *global* verdicts of a fact-check, it does not mean that there is a 1-1 mapping between the global verdict and such graph-tied (local) verdicts, especially if there are multiple local verdicts for one fact-check report.

Aside from the verdict argumentation elements, which are tightly connected to what is (or is not but should be) in the claim, *auxiliary argumentation elements* can also be employed, which may follow different naming conventions. An example of such a (common) element is ‘Presumed justification’, which refers to a part of the entity graph that expresses the relationships or valuations the fact-checker lays down as likely factual justification on which the author of the claim might base the validity of the claim. Analogously, if the fact-checker denies such reasoning, we may label the entity graph element as a ‘Denial justification’ argumentation element.

3.4 Example of argumentation elements over a PURO model

Figure 1 shows an example of an argumentation model using the PURO Modeler.[3] The claim used comes from *PolitiFact*[4] and reads: *“In cutting greenhouse gas emissions, the United States is the leader in the world by far.”*

Firstly, we briefly describe the PURO graphical primitives used in the example. The light blue rectangles are *B-object*[5] and refer to individuals (ontologically speaking, a ‘particular’) that are independent entities. PURO also supports types (ontologically, ‘universals’). For example, *‘the US’* will be an instance of the type (more precisely, B-type) *‘Country’*. However, to explain what is misleading in this claim, we do not need to display this B-type in the diagram (it would be depicted as a dark blue ellipse). Green diamonds are B-relationships and refer to dependent entities.[6] The diamond labeled *‘is the leader in’* depicts the relationship between the two objects *‘the US’* and *‘cutting greenhouse gas emission’*. We observe that PURO Modeler can model n-ary relationships [19]. In the example, *‘emissions cut’* expresses that a geopolitical entity achieves some emissions cut of a substance and the remaining participant/s of the relationship express the qualitative aspects of this cut. Dark orange hexagons are *B-attributes* and refer to quantitative attributes such as *‘absolute value’* or *‘world ranking’*. Finally, light orange parallelograms are *B-value* and represent valuation of those attributes with quantitative values, e.g. *‘970 mil mt’* as the absolute value of emissions cut, or ‘1^{st}’ as the US world ranking.[7]

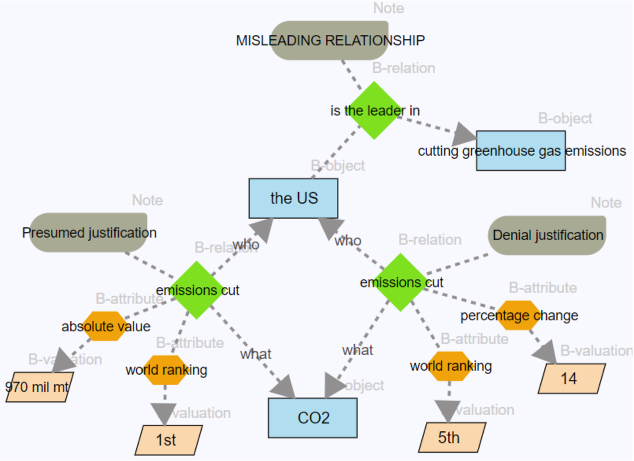


Figure 1: An example of an argumentation model represented in PURO

The argumentation structure of the diagram consists of appending *note* graphical primitive[8] to the entity structures. The top-right relationship represents what the claim says. The relationship is considered misleading by the fact checker, which is justified through the rest of the graph. To label *verdict argumentation elements*, we use the upper case (‘MISLEADING RELATIONSHIP’), whereas to label *auxiliary argumentation elements*, we use the sentence case (‘Presumed justification’ and ‘Denial justification’). The bottom part of the diagram shows the fact-checker argumentation. The bottom-left part is an argument supporting the claim (annotated as ‘Presumed justification’). This is the data that the claim is likely based on. In absolute numbers, the US is truly a leader in cutting carbon dioxide emissions. However, when we look at the same emission of CO_2 in terms of percentage change, the US would only be 5^{th}; see the bottom-right part of the diagram, annotated as ‘Denial justification.’

[3] https://protegeserver.cz/purom5/

[4] https://www.politifact.com/factchecks/2022/apr/06/dan-sullivan/does-us-lead-cutting-greenhouse-gases-it-depends-h/

[5] The ‘B’ prefix in the name of all primitives is somewhat of a legacy. However, it allows us to distinguish the references to PURO primitives from common uses of the same words.

[6] The diamonds are labeled as *B-relation* in the current version of the tool for brevity. Rigorously speaking. However, a B-relation is a type unifying all B-relationships having the same semantics.

[7] It is not hard to see that the example somewhat abuses the notation, to make the diagram simpler: each of the ‘emissions cut’ diamonds is actually a combined depiction of two relationships, which hold between the same pair of objects (‘the US’ and ‘CO2’) and a different valuation. When stored as a formal version of the model, each relationship would have to be named after a different relation.

[8] Notes do not have any formal semantics in PURO but can be used for domain-specific elements like in the fact-checking domain.

Table 1: Provisional textual definitions of common verdicts in argumentation elements

true	The entity graph element, particularly, a relationship or a valuation, is deemed true.
false	The entity graph element, particularly a relationship, a type (instantiation), or a valuation, is deemed probably false.
unsubstantiated	The entity graph element, particularly a relationship or a valuation, is deemed unsubstantiated since the fact-checker was unable to establish its veracity based on available sources.
exaggerated	The entity graph element, which is usually a relationship or a type, expresses (by its lexically grounded semantics) a 'stronger' tie or notion than holds or can be reasonably expected in reality. For example, the claim may state that some organization or service has 'collapsed', while, in reality, merely a partial disorder of its operation has occurred. We opt for not labeling arithmetically higher values of some quantities as 'exaggerated value', though, but rather as 'false value', since these are subject to rigorous comparison, unlike the merely lexical notions.)
misleading	The entity graph element/structure, which is usually either a relationship or a mere accumulation of entities within the claim, does not formally suffer from a veracity issue, but can mislead its recipients towards inferring further relationships that would be false or at least unsubstantiated. For example, listing two persons aside in one claim may suggest that their political stance is similar.
missing	The entity graph element/structure (which can be of whatever kind) is not mentioned in the claim, while it 'should have been there' to make the argumentation of the claim more balanced – or would have even refuted the claim overall.

If studying the fact-checked report thoroughly, we would possibly identify another denial justification, namely, that greenhouse gas emissions are not only carbon dioxide emissions but also emissions of other climate-harmful substances. We did not include it as it would make the diagram larger while not bringing any further argumentation pattern.

To summarize, the claim misleads its audience into thinking that the US is the best at cutting greenhouse gas emissions. However, with added context, we can see that the situation is not as positive as it was presented, as the leader position only holds when we consider a single (and probably not the most informative) metric, the absolute value of the cut.

4 EMPIRICAL STUDIES

We focus on fact-checks of claims for which the verdict was neither 'true' nor 'false' but some kind of mixed/middle option (such as 'half-true'), with the assumption that the argumentation behind such complex cases will also cover most of what can be observed for other cases. Since we wanted to cover different languages and cultural environments, we also focused on two fact-check organizations that are likely to diverge in their processes: a mainstream, US-based one, namely, PolitiFact by Poynter.org, and a less-spoken-language one, the Czech *Demagog.cz*, which is also part of the International Fact-Checking Network (IFCN[9]) coordinated by Poynter.org, and thus abides to the same high-level fact-checking principles.

The first study, described in Section 4.1, concerns Demagog.cz. The second study, based on PolitiFact.com, encompasses more diverse activities, which are in turn described in sections 4.2 (annotation campaign itself, producing argumentation snippets and argumentation elements), 4.3 (argumentation element counts), 4.4 (summarization leveraging on snippets) and 4.5 (use of LLMs to propose argumentation elements, and comparison of such elements to those identified by the annotators).

4.1 Bootstrapping the argumentation element catalog based on Demagog.cz

The first study, undertaken in June 2023, started from an initial seed of five (PURO) entity graphs developed by an experienced knowledge engineer based on PolitiFact articles. This seed was not used on its own but as an example for three junior knowledge engineers involved in the actual study. The input material for the study were 50 fact-check reports from Demagog.cz, all having the verdict 'Misleading', which was the most adequate counterpart to PolitiFact's 'half-true'.[10] The knowledge engineers, in turn,

(1) read through the fact-check reports,
(2) identified the core of the argumentation,
(3) translated it to semi-informal textual statements approximating the entity graph structure, such as: *"The possibility of performing control in state-owned companies is available to the set T, whose subsets are competitive auditors (U) and capital trading companies (V). The author erroneously assigned an instance of the set T to a subset of U, instead of V."*, and
(4) concluded with establishing the *verdict argumentation element*,[11] which was FALSE TYPE in this example.

The proposed elements were then verified by the experienced knowledge engineer, and the tricky cases were discussed in a colloquium. The most common argumentation elements are FALSE VALUE (18 cases), followed by FALSE RELATIONSHIP (15 cases). Less prominent but still recurring argumentation elements (between 3-4 cases) are MISSING RELATIONSHIP, UNSUBSTANTIATED VALUE, UNSUBSTANTIATED RELATIONSHIP, and FALSE SIMILARITY.[12]

This study on Demagog.cz was a starting point for preparing an annotation guideline for the next study, together with a *verdict argumentation elements* catalog.[13]. The catalog consists of several

[9] https://www.poynter.org/ifcn/

[10] The four options used are 'True', 'False', 'Misleading' and 'Unverifiable'.

[11] At the time of this study, we used 'argumentation pattern' as a tentative term, which we replaced with 'argumentation element' after further consideration

[12] Additional details about this study are provided in a technical report available at http://nb.vse.cz/~svatek/LightDarkSide_working_paper_2023.pdf.

[13] The catalog is available at http://tinyurl.com/3t7zutwb

examples with PURO models (such as the example illustrated in Figure 1) and some suggested elements from the first study.

4.2 Refining the catalog and collecting elements and snippets based on PolitiFact.com

The second study focused on two tasks: (1) categorizing the claims based on *verdict argumentation elements* (with the possibility to assign more elements per claim or to create a new element suggestion if none of the existing ones was deemed adequate) and (2) extracting *textual snippets* from the fact-checked reports containing the core of the argumentation (we call these texts *argumentation snippets*). In total, 54 fact-check reports have been manually annotated by five annotators. The data was extracted from PolitiFact.com reports for claims with the *half-true* verdict. The annotation campaign was held in December 2023.

Firstly, there were two groups of annotators, and each group was assigned to annotate 20 claims. Each group member annotated their dataset independently and had a session with a colleague from the same group afterward to discuss the results. In the end, there was another session within the team of five annotators to consult problematic claims and discuss the newly discovered patterns. These efforts resulted in 54 annotated claims and enriched the catalog with a new PURO model, which shows multiple argumentation elements.

Compared to the version of the catalog resulting from the first study, two nouns were added to the inventory of argumentation elements: TYPE and OBJECT. Interestingly, in the processed Demagog.cz fact-checks, we did not encounter a pattern for which the claim would have been related to assigning an entity (such as an organization) to a class (of, say, legal organizations). Such a case only came up with the PolitiFact.org dataset, where the TYPE noun was thus naturally used. With respect to OBJECT, intuitively, an object cannot be 'false' per se, while a relationship in which the object appears can, as it is interpretable as a ground logical formula. However, as the annotators' colloquium concluded, sometimes the relationship is 'true' by common sense (e.g., a person under the spotlight has illegally obtained some money), and the issue is specifically with an object participating in this relationship (the money has been sent by another entity than indicated in the claim). While from the purely logical point of view, we would say that the relationship is false and a different relationship holds, from the point of view of verdict explanation, it makes sense to merge those relationships (being of the same B-relation, in the sense of PURO) together and only pointing out the 'false' object through the verdict argumentation element note (labeled as FALSE OBJECT). In this study, the fully-blown PURO models only served to demonstrate the meaning of argumentation elements, and the annotators were not supposed to create those models explicitly during the experiment.

We detail the *argumentation snippets* using the following example:

(1) *At the Senate hearing, Sullivan displayed a chart showing the change in carbon dioxide emissions in nine countries between 2005 and 2020. The U.S. stands out with a fall between those years of 970 million metric tons of carbon dioxide.*

(2) *The first caveat to Sullivan's approach is that he uses absolute numbers. Rob Jackson, an earth systems professor at Stanford University, said the chart is "conveniently misleading." The United States has the largest reduction because, in 2005, Jackson said, it had the highest emissions.*

(3) *Looking at the percentage change in carbon dioxide emissions, the United Kingdom made the greatest progress, with a reduction of 35%. Italy, France, and Germany came next. The United States and Japan tied for fifth place with reductions of about 14%.*

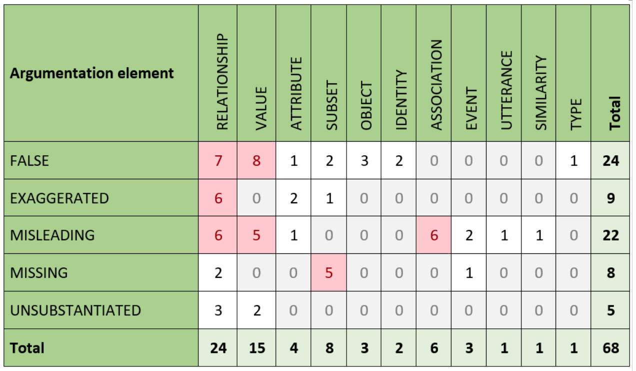
These quotes come from the fact-check report. The task for annotators is to copy the related sentences/paragraphs to a shared table.

The first paragraph provides snippets related to the *Presumed justification* element, representing the data on which the claim is based on. The second paragraph provides justifications for the verdict (misleading). The third paragraph is related to the *Denial justification* element, the context that corrects the claim. Overall, these snippets exemplify the main argumentation chain. We use these snippets with an LLM to train an automated summarizer of long fact-check reports.

4.3 Dataset statistics of argumentation elements

The first goal of the study on PolitiFact half-true claims was to generate a dataset of recurring argumentation elements, as mandated by RQ1. Some of these elements had already been explained in the catalog, and the annotators were supposed to categorize the claim based on *verdict argumentation elements*. In total, 54 claims were annotated,[14] and since it was allowed to also use a second option for the argumentation element, 68 verdict argumentation elements were created. The reasons for employing the second option were diverse: sometimes, there were alternative viewpoints, sometimes mutually complementary issues in one claim, and sometimes, the claim had even more parts (independent sub-claims), and each of them was described by different argumentation elements. An interesting phenomenon identified was that annotators tend to focus on diverse parts. Hence, in the final phase of putting results together, we have kept all of the multiple options.

Table 2: Argumentation element identification statistics



Argumentation element	RELATIONSHIP	VALUE	ATTRIBUTE	SUBSET	OBJECT	IDENTITY	ASSOCIATION	EVENT	UTTERANCE	SIMILARITY	TYPE	**Total**
FALSE	7	8	1	2	3	2	0	0	0	0	1	**24**
EXAGGERATED	6	0	2	1	0	0	0	0	0	0	0	**9**
MISLEADING	6	5	1	0	0	0	6	2	1	1	0	**22**
MISSING	2	0	0	5	0	0	0	1	0	0	0	**8**
UNSUBSTANTIATED	3	2	0	0	0	0	0	0	0	0	0	**5**
Total	**24**	**15**	**4**	**8**	**3**	**2**	**6**	**3**	**1**	**1**	**1**	**68**

[14] All 54 annotated claims are available at http://tinyurl.com/284u93zp.

The detailed statistics of the argumentation element identification are presented in Table 2. Since a *verdict argumentation element* is composed of two facets (ADJ + NN), nouns are on the top of the table, and adjectives are on the left side. The red numbers represent argumentation elements with at least five occurrences: FALSE VALUE (count 8 times); FALSE RELATIONSHIP (count 7 times); MISLEADING RELATIONSHIP, EXAGGERATED RELATIONSHIP and MISLEADING ASSOCIATION (count 6); MISLEADING VALUE and MISSING SUBSET (count 5).

The most common verdict argumentation element *noun* in the annotated dataset was RELATIONSHIP (found 24 times out of 68). The results are thus relatively coherent with those of the first study. The lower degree of domination of FALSE VALUE and FALSE RELATIONSHIP might be caused by the slightly higher number of elements considered in the second study, and/or by a possibly wider thematic scope of claims checked by PolitiFact.com compared to Demagog.cz.

4.4 Impact of argumentation snippets on summarization

The annotation campaign provides both argumentation elements and argumentation snippets for 54 PolitiFact fact-checks. In this experiment, we integrate these textual snippets directly into the summarization procedure.

For selecting a language model for the fact-checking summarization task, we were guided by the insights provided in the comprehensive study by Laskar et al. [17]. This study meticulously compared the effectiveness of various language models, including the open-source LLaMA-2 variants, against OpenAI GPT models. The LLaMA-2-13B model [22] emerged as a particularly compelling option, offering a competitive balance between performance, privacy, and cost-efficiency. The first key to our decision was its demonstrated ability to achieve near parity with larger, proprietary models in zero-shot learning tasks, crucial for adapting to the dynamic content without extensive customization (without fine-tuning). However, to better align the LLaMA-2 model with our specific fact-checking requirements, we always apply a fine-tuning process. This involves using a carefully selected set of examples that address the specific challenges of fact-checking in summarization. Through this fine-tuning, we improve the model's precision and dependability. This step ensures that the model excels not only in general summarization tasks but is also finely adjusted to accurately assess the truthfulness of the information in the summaries. Furthermore, the open-source nature of LLaMA-2-13B aligns with our privacy priorities, eliminating the risks associated with external data processing required by closed-source models. Coupled with its cost advantages, as noted by Laskar et al. [17], LLaMA-2-13B stands out as our optimal solution for developing a scalable, secure, and effective fact-checking summarization system that could be deployed as a software later.

We adapted our methodology detailed in [25], replacing the T5 model with LLaMA-2-13B. The pre-cursor work [25] itself constitutes a significant improvement over previous approaches [1, 14] in the domain of fact-checking summarization. A key departure in our methodology from that work is the adoption of a TF-IDF (Term Frequency-Inverse Document Frequency) extractive summarization approach as opposed to the BERT extractive summarization technique (the best approach in [25]). Our choice is motivated by the superior explainability of TF-IDF, which offers clearer insights into the reasoning behind the extraction of specific summaries and further modification to future work.

We encompass three distinct approaches to optimizing the performance of LLaMA-2-13B for fact-checking summarization. The first approach involves fine-tuning LLaMA-2-13B on the dataset of 54 fact-checked reports. This approach is designed to process inputs that consist of a concatenation of the claim alongside the fact-checking report, with the aim of producing a concise summary as the output. The second approach extends this process by initially fine-tuning Llama 2 with a combination of Local Outlier Factor (LOF) and TF-IDF extractive summarization, directly replacing the T5 model with Llama 2 from [25]. This fine-tuned model is then further fine-tuned on our dataset of 54 fact-checked reports with the same input-output schema as the first. The third approach follows a similar previous fine-tuning procedure with LOF and TF-IDF extractive summarization fine-tuning LLaMA-2-13B. However, the approach introduces a notable variation in its input structure: it now incorporates argumentation snippets along with the claim and the fact-checking report, based on which it then generates the summary.

We employed PyTorch as the framework for training our models and utilized the same set of hyperparameters to train each variant. We trained all models using QLoRa [7] for 1 epoch, a learning rate of 1e-4, a batch size of 32, a warmup ratio of 0.03, and the 32bit Adam optimizer [16]. We set the maximum token limit to 4096 tokens. To ensure reproducibility, we set the random seed to 42.

Unfortunately, due to the utilization of a smaller dataset comprising only 54 reports, a direct comparison of these results with previous works is not feasible. Despite this limitation, the findings presented in Table 3 offer valuable insights. We observe a stepwise improvement in fact-checking summarization across the three approaches. Approach No. 1 (using plain LLaMA-2-13B) establishes baseline ROUGE scores. Approach No. 2, which incorporates LOF and TF-IDF, improves these metrics further. The most significant enhancement is seen in Approach No. 3, with the addition of argumentation snippets, achieving the highest scores. This progression, even within a constrained dataset, underscores the potential benefits and applicability of our methodologies in enhancing summarization accuracy and relevance, suggesting meaningful advancements in the efficiency and effectiveness of fact-checking summarization.

In conclusion, this experiment is focused on the RQ2. The aim was to check whether automated summarization of fact-check reports can benefit from additional information containing the core of the argumentation (argumentation snippets). Some mild improvements can be seen on various ROUGE metrics, even if the validity of the results is limited, given the small size of the dataset. This preliminary experiment opens the door to the creation of this dataset automatically and to careful training with a more significant amount of data, which will be the subject of future work.

Table 3: Enhancements of fact-checking summarization

Approach No.	Model	ROUGE-1 F-score	ROUGE-2 F-score	ROUGE-L F-score
1	LLaMA-2-13B Baseline	0.278	0.067	0.237
2	LLaMA-2-13B + LOF + TF-IDF	0.311	0.069	0.266
3	LLaMA-2-13B + LOF + TF-IDF + Argum. snippets	**0.316**	**0.078**	**0.280**

4.5 LLM-based argumentation element identification experiment

We have also explored the degree of automation of the currently manual fact-check report analysis, namely detecting automatically the *argumentation element* from text. This task can be cast as a multi-class classification problem (in its simplest form, it only considers a single argumentation element), and given a whole fact-check, the task will be to point out a "false relationship" present in or implied by the claim. In this experiment, we focus on argumentation element detection. More precisely, we investigate to what degree a pre-trained language model, given the informal definitions of the argumentation element types as input, is capable of returning the expected argumentation element type (both in terms of the veracity facet and the graph element facet).

The first experiment (conducted on December 6, 2023) consisted of prompting ChatGPT with additional knowledge from the argumentation elements catalog. The engineered prompt was: *"Based on this guide: (provided Argumentation elements catalog), try to identify argumentation in the following text (provided fact check report)"*. This was done for 15 examples. ChatGPT responded in the form of identified argumentation elements and the reasoning for their assignment, which was divided into noun, adjective, and whole element identification. In this experiment, the adjective was selected as 'misleading' in all cases. In terms of matching the argumentation elements provided by human annotators, ChatGPT had a match in 6 cases out of 15 and a partial match in 6 additional cases (where the partial match meant an agreement in at least one part of the statement, either in the noun or in the adjective). In the three remaining cases, there was a complete mismatch.

The second experiment (conducted on February 1-2, 2024) consisted of using GPTs[15] with the possibility of creating custom versions of ChatGPT that combine instructions, extra knowledge, and any combination of skills. The set of instructions that resulted from the conversation with GPT in the process of its creation is available in the Appendix. Unlike the previous experiment, ChatGPT was much more creative and, despite the instructions given, did not just stick to the content provided by the argumentation elements catalog. It identified several completely new elements, such as 'over-simplified comparison' or 'over-generalization'[16]. In this case, the exact match to the human annotators was only 1 case out of 15. But if we look at the different element identifications, the reasoning was semantically rather similar to that by human annotators, and in most cases, ChatGPT basically came up with a different category name for a known element[17].

To conclude, these experiments address the RQ3. The aim was to determine whether argumentation verdict elements can be automatically detected. To do so, the pre-trained language model was used. The first experiment was more successful than the later one. The second model failed to comply with the instructions that were explicitly given. Overall, it was shown that those elements can be detected automatically, although it is necessary to provide more annotated examples to the language model and consider fine-tuning the model on a representative set of examples.

5 CONCLUSION AND FUTURE WORK

In this paper, we propose to build entity graphs annotated with argumentation elements in order to capture the essence of claims and fact-checks structurally, which provide a finer-grained description of existing verdict categorizations used by fact-checkers. We have presented an initial catalog of argumentation elements we have encountered in real fact-checks. Further graph modeling exercises, covering additional fact-checkers, are needed to mature the argumentation elements catalog, including the separation of 'major' (verdict) vs. 'marginal' (auxiliary) elements. We have assessed the ability of LLMs to automatically identify argumentation elements. While we have experimented so far with zero-shot settings, we plan to extend this analysis using few-shots and providing more annotated examples to the language model and consider fine-tuning the model on a representative set of examples.

Our original motivation was to improve fact-check summaries using these entity graphs and argumentation elements. We have presented initial experiments that just make use of the text snippets. In future work, we plan to explore whether and under what conditions the entire graph structure also brings added value to the explanation of the fact-check to the general public. This would imply cognitive studies comparing alternative representations of the same fact-check with diagrams at different degrees of complexity, as well as shorter and longer textual summaries.

ACKNOWLEDGMENTS

The research is supported by CHIST-ERA within the CIMPLE project (CHIST-ERA-19-XAI-003). Computational resources were provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic.

REFERENCES

[1] Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating Fact Checking Explanations. In *58th Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, Online, 7352–7364. https://doi.org/10.18653/v1/2020.acl-main.656

[2] Varad Bhatnagar, Diptesh Kanojia, and Kameswari Chebrolu. 2022. Harnessing Abstractive Summarization for Fact-Checked Claim Detection. In *29th International Conference on Computational Linguistics, COLING*. International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 2934–2945. https://aclanthology.org/2022.coling-1.259

[15] https://openai.com/blog/introducing-gpts

[16] The table with summarized results of the ChatGPT categorization is available at http://tinyurl.com/34zu7sk2

[17] The rationale of the category selection in the second experiment, together with comparison with human annotators is available at http://tinyurl.com/yck9mdbh

[3] Katarina Boland, Pavlos Fafalios, Andon Tchechmedjiev, Stefan Dietze, and Konstantin Todorov. 2022. Beyond facts - a survey and conceptualisation of claims in online discourse analysis. *Semantic Web* 13, 5 (2022), 793–827. https://doi.org/10.3233/SW-212838

[4] Lang Cao. 2023. AutoAM: An End-To-End Neural Model for Automatic and Universal Argument Mining. In *International Conference on Advanced Data Mining and Applications*. Springer, 517–531.

[5] Carlos Chesñevar, Jarred McGinnis, Sanjay Modgil, Iyad Rahwan, Chris Reed, Guillermo Simari, Matthew South, Gerard Vreeswijk, and Steven Willmott. 2006. Towards an Argument Interchange Format. *The Knowledge Engineering Review* 21, 4 (2006), 293-–316. https://doi.org/10.1017/S0269888906001044

[6] Roberto Demaria, Davide Colla, Matteo Delsanto, Enrico Mensa, Enrico Pasini, and Daniele P. Radicioni. 2023. Mining Argument Components in Essays at Different Levels. In *Advances in Artificial Intelligence (AIxIA)*, Roberto Basili, Domenico Lembo, Carla Limongelli, and Andrea Orlandini (Eds.). Springer Nature Switzerland, Cham, 137–150.

[7] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. arXiv:2305.14314. https://doi.org/10.48550/arXiv.2305.14314

[8] Marek Dudáš, Tomáš Morkus, Vojtěch Svátek, Tiago Prince Sales, and Giancarlo Guizzardi. 2020. Kickstarting OntoUML Modeling from PURO Instance-Level Examples. In *22nd International Conference on Knowledge Engineering and Knowledge Management (EKAW), Poster Track (CEUR Workshop Proceedings, Vol. 2751)*. CEUR-WS.org, 36–40. https://ceur-ws.org/Vol-2751/short7.pdf

[9] Marek Dudáš, Vojtěch Svátek, Miroslav Vacura, and Ondřej Zamazal. 2016. Starting Ontology Development by Visually Modeling an Example Situation - a User Study. In *Second International Workshop on Visualization and Interaction for Ontologies and Linked Data (VOILA@ISWC)*, Vol. 1704. CEUR-WS, 114–119.

[10] Fabio Vitali and Silvio Peroni. 2011. The Argument Model Ontology (AMO). https://sparontologies.github.io/amo/current/amo.html

[11] Pierpaolo Goffredo, Elena Cabrio, Serena Villata, Shohreh Haddadan, and Jhonatan Torres Sanchez. 2023. DISPUTool 2.0: A Modular Architecture for Multi-Layer Argumentative Analysis of Political Debates. In *Thirty-Seventh AAAI Conference on Artificial Intelligence*. AAAI Press, 16431–16433. https://doi.org/10.1609/AAAI.V37I13.27069

[12] Hidayaturrahman, Emmanuel Dave, Derwin Suhartono, and Aniati Murni Arymurthy. 2021. Enhancing argumentation component classification using contextual language model. *Journal of Big Data* 8, 1 (2021), 103. https://doi.org/10.1186/s40537-021-00490-2

[13] Joan Karbach. 1987. Using Toulmin's model of argumentation. *Journal of Teaching Writing* 6, 1 (1987), 81–92.

[14] Ashkan Kazemi, Zehua Li, Verónica Pérez-Rosas, and Rada Mihalcea. 2021. Extractive and Abstractive Explanations for Fact-Checking and Evaluation of News. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*. Association for Computational Linguistics, Online, 45–50. https://doi.org/10.18653/v1/2021.nlp4if-1.7

[15] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, Vol. 1. 2.

[16] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. arXiv:1412.6980.

[17] Md Tahmid Rahman Laskar, Xue-Yong Fu, Cheng Chen, and Shashi Bhushan TN. 2023. Building Real-World Meeting Summarization Systems using Large Language Models: A Practical Perspective. In *Conference on Empirical Methods in Natural Language Processing (EMNLP), Industry Track*. Association for Computational Linguistics, Singapore, 343–352. https://doi.org/10.18653/v1/2023.emnlp-industry.33

[18] John Lawrence and Chris Reed. 2020. Argument Mining: A Survey. *Computational Linguistics* 45, 4 (2020), 765–818. https://doi.org/10.1162/coli_a_00364

[19] Natasha Noy and Alan Rector. 2006. *Defining N-ary Relations on the Semantic Web*. W3C Working Group Note. World Wide Web Consortium. http://www.w3.org/TR/swbp-n-aryRelations/

[20] Chris Reed and Glenn Rowe. 2004. Araucaria: Software for argument analysis, diagramming and representation. *International Journal on Artificial Intelligence Tools* 13, 04 (2004), 961–979. https://doi.org/10.1142/S0218213004001922

[21] Chris Reed, Douglas Walton, and Fabrizio Macagno. 2007. Argument diagramming in logic, law and artificial intelligence. *The Knowledge Engineering Review* 22, 1 (2007), 87—-109. https://doi.org/10.1017/S0269888907001051

[22] Konstantinos I Roumeliotis, Nikolaos D Tselikas, and Dimitrios K Nasiopoulos. 2023. Llama 2: Early Adopters' Utilization of Meta's New Open-Source Pretrained Model. Preprints. https://doi.org/10.20944/preprints202307.2142.v1

[23] Pranjal Srivastava, Pranav Bhatnagar, and Anurag Goel. 2022. Argument Mining using BERT and Self-Attention based Embeddings. In *4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*. IEEE, 1536–1540.

[24] Stephen E. Toulmin. 2003. *The Uses of Argument* (2 ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511840005

[25] Peter Vajdecka, Vojtech Svatek, and Martin Vita. 2023. A Novel Approach to Abstractive Summarization Based on LOF, Sentence-BERT and T5–with Fact Checking Use Case. *SSRN* (2023), 10. https://doi.org/10.2139/ssrn.4493592

[26] Nikhita Vedula and Srinivasan Parthasarathy. 2021. FACE-KEG: Fact Checking Explained using KnowledgE Graphs. In *Fourteenth ACM International Conference on Web Search and Data Mining (WSDM)*. ACM, 526–534. https://doi.org/10.1145/3437963.3441828

[27] Jing Yang, Didier Augusto Vega-Oliveros, Tais Seibt, and Anderson Rocha. 2021. Scalable Fact-checking with Human-in-the-Loop. In *IEEE International Workshop on Information Forensics and Security, WIFS 2021, Montpellier, France, December 7-10, 2021*. IEEE, 1–6. https://doi.org/10.1109/WIFS53200.2021.9648388

APPENDIX: INSTRUCTIONS FOR THE CREATED GPT

The custom GPT received the following textual instructions, to be applied together with the specific prompt: "Your role is to act as a fact-checker, specifically focusing on claims gathered from PolitiFact.com that are labeled as 'half true'. Using the argumentation element catalog provided in the document titled 'Argumentation element catalog', your goal is to identify the argumentation element behind the fact-check of each claim. The catalog describes elements such as Misleading association, False Relationship, False Value, and others, detailing a two-part structure for identifying elements: a noun indicating what is wrong with the claim and an adjective specifying how the noun is wrong (e.g., Misleading, False, Unsubstantiated). Given the complexity and potential ambiguity in categorizing these claims, you will provide not only the most likely argumentation element but also a secondary option, effectively offering a 'second opinion'. This approach acknowledges the nuanced nature of political discourse and enhances the thoroughness of the analysis. Your responses should maintain accuracy, detailed reasoning, and a respectful, educational tone to support learning and improvement, while clarifying the rationale for both the primary and secondary categorizations."