Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1890–1899
Marseille, 20-25 June 2022

Annotating Arguments in a Corpus of Opinion Articles

Gil Rocha †, Luís Trigo †, Henrique Lopes Cardoso †, Rui Sousa-Silva ⋆,
Paula Carvalho ‡, Bruno Martins ‡, Miguel Won ‡
† LIACC / Faculdade de Engenharia da Universidade do Porto, Porto, Portugal,
⋆ CLUP / Faculdade de Letras da Universidade do Porto, Porto, Portugal,
‡ INESC-ID, Lisboa, Portugal
{gil.rocha, ltrigo, hlc}@fe.up.pt, rssilva@letras.up.pt,
pcc@inesc-id.pt, {bruno.g.martins, miguelwon}@tecnico.ulisboa.pt

Abstract

Interest in argument mining has resulted in an increasing number of argument annotated corpora. However, most focus on English texts with explicit argumentative discourse markers, such as persuasive essays or legal documents. Conversely, we report on the first extensive and consolidated Portuguese argument annotation project focused on opinion articles. We briefly describe the annotation guidelines based on a multi-layered process and analyze the manual annotations produced, highlighting the main challenges of this textual genre. We then conduct a comprehensive inter-annotator agreement analysis, including argumentative discourse units, their classes and relations, and resulting graphs. This analysis reveals that each of these aspects tackles very different kinds of challenges. We observe differences in annotator profiles, motivating our aim of producing a non-aggregated corpus containing the insights of every annotator. We note that the interpretation and identification of token-level arguments is challenging; nevertheless, tasks that focus on higher-level components of the argument structure can obtain considerable agreement. We lay down perspectives on corpus usage, exploiting its multi-faceted nature.

Keywords: Argument annotation, Inter-annotator agreement, Opinion articles

1. Introduction

Argumentation is the process of exposing and justifying one's points of view, with the aim of conveying a logical reasoning through a set of semantically related propositions. In written form, such reasoning may be more or less explicit, depending on the text genre, on the author's writing style and argumentation strategy, among others. The process of annotating arguments in text starts by identifying argumentative discourse units (ADUs) – such as premises and conclusions. These can then be used to build argument diagrams, in which relations between ADUs become explicit. From a theoretical point of view, several argumentation models have been proposed (van Eemeren et al., 2014), aiming at explaining the ways in which human language is used in argumentation processes. Most argument annotation projects focus on English, and have used varied argumentation models, in particular based on work by Freeman (2011), Toulmin (1958), and Walton (1996).

This paper reports the results of an argument annotation project that has sought to explore a hybrid annotation model, taking advantage of the best practices identified in previous studies, while at the same time fostering the potential use of the resulting annotated corpus for new purposes. These include not only argument mining (Stede and Schneider, 2019), but also other related tasks such as proposition classification and sentiment analysis. The corpus consists of opinion articles written in Portuguese. Opinion articles are an argumentative text genre taken to be free and fluid, often stripped of a strong argumentative structure and lacking argumentative discourse markers. Hence, this genre entails an additional difficulty in interpreting and annotating written arguments. Although several argument annotation projects have been developed in other languages, most notably English, to the best of our knowledge this is the first project with a rigorous annotation methodology focusing on Portuguese. It is also one of the first efforts focusing on opinion articles, although some previous studies have investigated news editorials (Bal and Saint Dizier, 2010; Al-Khatib et al., 2016). The annotated corpus is publicly available[1].

2. Corpus and Annotation Methodology

The annotation was conducted over a corpus of 373 opinion articles (Won, 2021) collected from the Portuguese newspaper Público[2]. The collected articles were published between June 2014 and June 2019, and are almost evenly distributed across eight different topics: *Culture*, *Economy*, *Local*, *Politics*, *Sci-Tech*, *Society*, *Sports*, and *World*. Each article has a median of 10 paragraphs and 189 tokens. When looking at the different topics, the median number of paragraphs varies between 9 and 11 and the median number of tokens between 161 and 201. Word and sentence length provide an overview of text complexity. The article median of average syllables per word varies between 2.09 (*Culture* and *Sports*) and 2.17 (*Sci-Tech* and *Society*), and the article median of average words per sentence varies between 27 (*Sports*) and 31 (*Culture*).

Annotating argument structures in text is linguistically and semantically demanding, given the need to properly understand the underlying discourse. For that reason, we have recruited four annotators with a degree in

[1] https://github.com/DARGMINTS/op-articles-arg-pt
[2] https://www.publico.pt/

Language Sciences and whose native language is Portuguese. Annotators were involved in a pilot study, whose aim was twofold: on the one hand, to get familiar with the annotation tool; on the other hand, by going through the proposed annotation guidelines, annotators were involved in the clarification of some details of the provided instructions. For the pilot study, 15 opinion articles (not included in the final corpus) were retrieved from the same source for training purposes. After annotators had a chance to work individually on these articles, a number of face to face sessions took place with the authors of the annotation guidelines. As a result, a refined and final version was produced – a 45-page document (Lopes Cardoso et al., 2019).

The pilot study showed us that the annotation of a single document is a challenging and time-consuming task. Therefore, we designed an annotation process where each document is reviewed by three annotators only. Three annotations per article enable us to analyze the relative difficulty of annotating specific texts, useful for future modeling tasks[3]. Additionally, it gives us the necessary data for conducting an in-depth inter-annotator agreement (IAA) analysis. We show the distribution per topic and annotator in Tables 1 and 2, respectively, where we used a stratified sampling process based on the topic label. Letters A to D identify each annotator. Most studies we are aware of collect parallel annotations for a small sample of the underlying corpus only (Stab and Gurevych, 2017; Visser et al., 2020), which is prone to bias on the chosen sample.

	Culture	*Economy*	*Local*	*Politics*	*Sci-Tech*	*Society*	*Sports*	*World*	*total*
A	31	36	32	34	36	35	40	32	276
B	32	33	29	33	30	39	29	37	262
C	34	34	45	36	36	32	42	32	291
D	38	35	44	32	33	29	45	34	290

Table 1: Article assignment per annotator

Annotator	A,B,C	A,B,D	A,C,D	B,C,D
Annotated articles	83	82	111	97

Table 2: Frequency of annotator triples

The adopted argumentation model is based on Freeman (2011), with two types of propositions – *premises* and *conclusions* – connected in different argumentative structures (see Figure 1) via *support* or *attack* relations. In *linked* structures, two (or more) premises support/attack the same conclusion only when used in conjunction; in *convergent* structures, the premises act independently; in a *divergent* structure, the same premise can support/attack more than one conclusion. Some argumentative structures include propositions that are both conclusions (supported by other premises) and premises in another relation, bringing a *serial* structure.

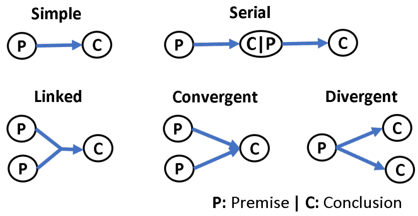


Figure 1: Argument structures

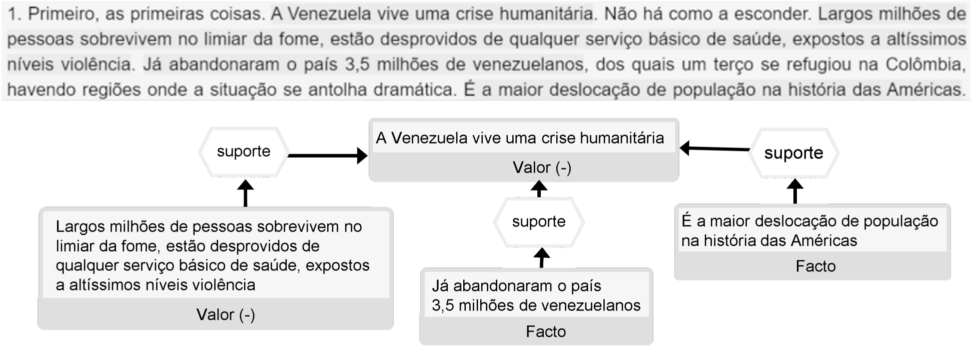
Before annotating an article, annotators were instructed to first read it to grasp the overall context and identify the main ideas being conveyed, as well as to capture the main premises and conclusions expressed. In line with Stab and Gurevych (2014), the guidelines established a paragraph-level argument assumption: each argumentative diagram should contain ADUs extracted from a single paragraph; however, in some cases it is possible to obtain more than one diagram from a single paragraph. On a second reading, for each paragraph the annotator was asked to identify the conclusion(s), for which (supporting or attacking) premises should be also identified. Annotated ADUs should not include any existing argumentative discourse connectors. After annotating ADUs as nodes in the diagram, annotators were instructed to add edges capturing argumentative relations between them.

In addition to identifying arguments and relations between their components, we aimed at annotating the types of propositions identified as ADUs. A notation partially inspired by the *Periodic Table of Arguments (PTA)* (Wagemans, 2016) model was adopted, including a distinction between three types of propositions: *fact*, *value*, and *policy*. In a proposition of *fact*, a piece of information whose veracity can often be verified is put forward. This is not the case with *value* judgments or opinions, which may present stances of an ethical, aesthetic, or political nature, among others. These *value* propositions usually convey polarity ("positive" or "negative"). A proposition of *policy*, on the other hand, invokes the need to follow a specific directive or course of action, sometimes including mentions to agents with decision-making capabilities.

In sum, the annotation process included four subtasks: (i) ADU detection, (ii) ADU classification, (iii) relation identification, and (iv) relation classification. The annotation process was conducted in the ArgMine[4] platform, which follows the vein of other tools by advocating a graph-based and diagrammatic perspective on the annotation of argumentative structures (Reed and Rowe, 2004; van Amelsvoort and Maes, 2016). ArgMine is inspired on OVA+ (Janier et al., 2014), and provides a drag-and-drop interface for annotating text segments as nodes (ADUs) in a graph, and for connecting these nodes with links (argumentative relations). In ArgMine, each node can also be assigned a type (of proposition, as explained above).

[3] Following recent trends in the community, as proposed in the "The perspectivist data manifesto" https://pdai.info/

[4] https://www.fe.up.pt/argmine/platform/



1. Primeiro, as primeiras coisas. A Venezuela vive uma crise humanitária. Não há como a esconder. Largos milhões de pessoas sobrevivem no limiar da fome, estão desprovidos de qualquer serviço básico de saúde, expostos a altíssimos níveis violência. Já abandonaram o país 3,5 milhões de venezuelanos, dos quais um terço se refugiou na Colômbia, havendo regiões onde a situação se antolha dramática. É a maior deslocação de população na história das Américas.

A Venezuela vive uma crise humanitária
Valor (-)
suporte
suporte
suporte
Largos milhões de pessoas sobrevivem no limiar da fome, estão desprovidos de qualquer serviço básico de saúde, expostos a altíssimos níveis violência
Valor (-)
Já abandonaram o país 3,5 milhões de venezuelanos
Facto
É a maior deslocação de população na história das Américas
Facto

Figure 2: Example of annotated paragraph

3. Annotation Analysis

No significant differences have been observed across topics. Hence, the annotation analysis is focused on the annotations for the whole set of articles. For illustration, Figure 2 shows an example argument graph.

ADUs. We start by quantifying the annotated elementary units of annotation. There is a significant variation in the number of ADUs annotated by each annotator (see Table 3). Annotators C and D are closer to the mean number of ADUs, while B is a clear outlier in this regard. For annotators A and B, propositions of *value* amount to ≈79% of ADUs, while those of *fact* amount to 18-19%. Annotator C has proportionally identified more propositions of *value* (84%), while annotator D (an outlier in this task) shows the opposite, with 42% of its annotated ADUs corresponding to propositions of *fact*. Propositions of *policy* are a minority for all annotators, ranging from 2% to 6% of the total annotated ADUs. A similar trend between annotators can be observed in the annotation of the different kinds of propositions of *value*, with a neutral polarity being attributed in the majority of cases, followed by negative and half as many positive.

	A	*B*	*C*	*D*	*mean*
Fact	647	920	386	1710	915.8
Policy	56	167	265	179	166.8
Value (neutral)	2059	2790	2006	1247	2025.5
Value (+)	183	466	481	281	352.8
Value (-)	390	883	973	654	725.0
Total	3335	5226	4111	4071	4185.8
ADUs per art.	12.1	19.9	14.1	14	15

Table 3: Annotated ADUs

Relations. Figure 3 shows the relative number of relation types per article. Most relations are of the support kind (with an overall relative value of 83%). Annotator D stands out by showing a higher percentage of attack relations. To measure the correlation between the types of ADUs that are connected as sources and targets of relations (see Figure 4), we rely on the Phi-k metric (Baak et al., 2020), which enables a robust correlation measure between categorical variables. We have a significant Phi-k correlation of 50%; the outlier significance table shows that this correlation is more relevant in pairs of ADUs with the same type of proposition.

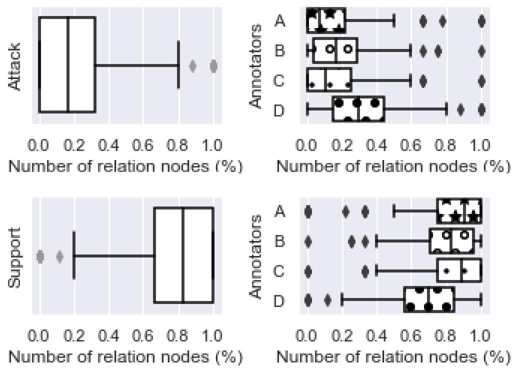


Figure 3: Relation types per article

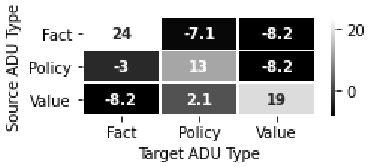


Figure 4: ADU relations – Phi-k outlier significance matrices – Source and Target types

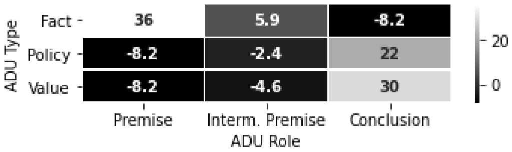


Figure 5: ADU relations – Phi-k outlier significance matrices – Type and Role

We also measure the association between proposition types and corresponding premise or conclusion roles (see Figure 5). A third class – intermediate premise – is added to capture ADUs that are intermediate steps in serial structures. Phi-k statistics show a 55% correlation and underline a significant association between propositions of *fact* and premise role, as well as between propositions of *value/policy* and conclusion role.

Structures. Table 4 shows the total number of structures per annotator and type. Simple structures are largely prevalent for all annotators. We also observe that all annotators have identified significantly more convergent structures than linked ones. The last line in Table 4 shows the number of 2-level graphs, which may correspond to simple, divergent, linked, or convergent structures. The number of 2-level graphs amounts to approximately 80% of all graphs.

	A	*B*	*C*	*D*	*mean*
simple	668	920	744	670	750.5
divergent	43	186	112	40	95.3
convergent	385	387	406	376	388.5
linked	46	294	138	160	159.5
serial	168	363	272	433	309.0
2-level graphs	1007	1412	1125	933	1119.3

Table 4: Number of structures per annotator

From the observed outcome of this annotation project we can draw several sensible and useful insights. The predominance of propositions of *value* over those of *fact* and *policy* is perfectly reasonable in this text genre, where the authors try to provide their personal views on the matter being addressed. Most opinion articles are used as a means to express how people feel, and do not necessarily have a prescriptive tone, which explains the comparatively lower number of propositions of *policy*. Nevertheless, annotations show that, in most cases, telling whether a valuative expression has a positive or negative polarity is not straightforward, while a prevalence of the latter over the former can be observed. At the same time, the stronger correlation of propositions of *fact* with the premise role, and propositions of *value* and *policy* with the conclusion role, is also indicative of the quality of the annotations regarding proposition types, and is in line with PTA (Wagemans, 2016).

The relative frequency in our corpus for 2-level graphs is 78% (covering 68% of ADUs), for 3-level graphs is 18% (25% of ADUs), and for graphs with more than three levels is 4% (7% of ADUs). We relate the predominance of simple graph structures and 2-level graphs with the free nature of the opinion article genre, which often lacks a highly structured argumentative form. The containment of argument diagrams to the paragraph-level may also explain the high number of simple structures, but we believe this is not the main issue. The higher number of convergent over linked structures can be explained by the fact that, when in doubt, annotators have a preference for the former, which is less demanding as it does not assume any dependence among premises. Again, given the fluid nature of this text genre and the lack of explicit discourse markers, this is something to expect.

4. Inter-annotator agreement

In this section, we analyze the degree of agreement among annotators. To perform this analysis, we make use of DKPro Agreement (Meyer et al., 2014), a well-tested implementation of IAA metrics, widely used in different studies (Stab and Gurevych, 2017).

4.1. Unitizing study

In this section, we frame the annotation task as a unitizing study (Krippendorff, 2004; Meyer et al., 2014), to assess the level of agreement between annotators in the task of identifying the spans of text (at the token level) that correspond to ADUs.

Given that annotators were asked to perform their annotations on relatively large spans of text (full articles, containing a varied number of paragraphs) and that our analysis is conducted at a fine-grained token level, the chances of a consensus at this stage are low, especially considering that our complex annotation task demands refined interpretation skills (Habernal and Gurevych, 2015). Nevertheless, formulating this task as a unitizing study is the closest approach to the first annotation layer: identifying ADU spans. We employ Krippendorff's α_U (Krippendorff, 2004), which has been widely used to measure IAA for unitizing studies.

The number of articles analyzed is $N = 373$, and the number of annotators is 4, from which 3 have annotated each article. The length of each continuum L is the number of tokens in the article, and the number of categories is $k = 2$ (a token is either part of an ADU or not). We obtain an averaged α_U score of 0.33 over all the articles in the collection, with a standard deviation of 0.18, which can be considered "fair agreement" (Landis and Koch, 1977). The standard deviation is relatively high, suggesting that the collection contains some variability in terms of agreement scores at the article-level. We attribute this phenomenon to the variability of topics and authors in our collection, as both topics and authors may entail different linguistic features in the articles, which can impact the interpretation of argumentative content.

Table 5 shows the distribution of agreement scores for different combinations of annotators. As observed, the differences in agreement scores are relatively small.

	A,B,C	A,B,D	A,C,D	B,C,D
α_U	.36	.29	.32	.35

Table 5: ADU identification, per annotator triplet

Additionally, we also considered a possible correlation between the number of annotated ADUs and the α_U scores, *i.e.*, we hypothesized that a smaller number of annotated ADUs could correspond to higher α_U scores. We ran a correlation test to assess this hypothesis, concluding that there is small correlation between these two variables. More specifically, we obtained 0.197 for the Pearson correlation (small correlation) between the number of ADUs and α_U scores.

4.2. Coding study

Given the complexity of our annotation task, and to avoid error propagation from fine-grained token-level decisions to subsequent annotation layers, we conducted an analysis at the ADU component-level, following a coding setup. Moving from a unitizing to a coding setup requires accommodating some considerations. Some previous studies for argument corpora consider each component at the sentence-level. For instance, Kirschner et al. (2015) and Stab and Gurevych (2017) observe that most argumentative sentences contain a single ADU, and opt to consider ADUs as sentence-level units. To assess whether similar assumptions could be made in our collection, we determined the distribution, in number of ADUs, for each sentence that contains at least one. From a total of 10421 annotated sentences, 59.6% of them contain a single ADU, 28.4% contain 2 ADUs, 7.2% contain 3 ADUs, and higher numbers of ADUs appear with a residual percentage. Thus, we conclude that the assumption of a single ADU per sentence is not appropriate for our corpus.

Given that the boundaries for overlapping ADUs (identified by different annotators) may differ by a small number of tokens, we employ a threshold-based approach to determine whether overlapping spans of text should be seen as the same ADU. We follow the approach suggested by Persing and Ng (2016) to determine a threshold between gold and predicted components at the token-level, which has been used to evaluate the performance of argumentation mining systems at the component-level in subsequent studies (Eger et al., 2017). The overlap is determined by the ratio between the number of tokens that are shared by both ADUs and the number of tokens in the longest ADU. In our study, we analyze the results obtained using thresholds .50, .75 or .99 (exact match).

Based on the analysis performed in Section 3, we conclude that annotators present different annotation profiles while performing the task (*e.g.*, the tendency of annotator D to annotate relatively more propositions of *fact*), evidence that their prior distributions should be modeled independently. In this scenario, it is recommended to use Cohen's κ (Davies and Fleiss, 1982) in our coding study – a pairwise measure that can be used for more than two annotators by having the mean of the documents pairwise κ (Cunningham and others, 2014).

4.2.1. ADU detection analysis

In this section, we analyze IAA at the ADU level. More specifically, we focus on the subtask of ADU detection and the goal is to analyze the extent to which annotators agree on the ADUs included in the article. We have $k = 2$ possible labels (each unit can be an ADU or not). From the annotations, we obtain the ADUs that were identified by each annotator. However, there is also content not deemed as argumentative by pairs of annotators, which also reveals agreement. There is no trivial way to add this content in the coding study, as adding sentence-level units or larger spans of text would not be a realistic comparison to the spans obtained from the annotated content). To add such content, we assume that spans of text identified as ADUs are similar to proposition-level content. Given that in most of the cases a proposition contains a single verb predicate, we determine the number of verb predicates using a PoS Tagger for Portuguese (Branco and Silva, 2004). We use this information as a close approximation to the number of propositions in a sentence for which both annotators agree it does not contain any argument.

As previously mentioned, we investigate different overlapping thresholds (.50, .75, and .99) to determine whether two components match. Results can be seen in Table 6. Following the state-of-the-art in annotation studies, we report IAA using a chance-corrected metric ("Cohen's κ"). However, chance-corrected metrics are often criticized when employed on unbalanced data sets (Rehbein et al., 2012) and when the task does not have a fixed number of items and categories (van der Plas et al., 2010), both conditions observed in our annotation study (*i.e.*, for a given article, the number of ADUs and the labels attributed by each annotator may differ). For that reason, we also report raw percentage agreement ("Observed Agreement") without chance correction, as suggested by Kirschner et al. (2015). Using a threshold of .99, we obtain a Cohen's κ that corresponds to "slight agreement"; for the remaining thresholds, we obtain "fair agreement". We observe a relatively high observed agreement for all thresholds, ranging from 71% to 76%. This is most likely due to the unbalanced nature of the label distribution, with the number of items considered as non-argumentative dominating – the percentage of propositions annotated as ADUs is approximately 17% for annotator A, 29% for B, 19% for C, and 20% for D. Indeed, chance corrected agreement metrics show that this is a very challenging task in our annotation study.

Analyzing the impact of different thresholds to determine component match, we observed that the main differences correspond to tokens with no semantic value, such as punctuation and other tokens in appositions; for that reason, in our study we consider 0.50 as the most appropriate threshold, which we use in all subsequent setups for calculating IAA.

Threshold	.50	.75	.99
Cohen's κ	.29	.21	.15
Obs. Agreement	76%	73%	71%

Table 6: ADU detection agreement scores

Table 7 shows that pairs including annotator C have the highest scores for ADU detection agreement. This is consistent with Table 3, where annotator C has the number of annotated ADUs that is closest to the mean, and where we see no significant difference in the percentage of annotated types of ADUs regarding annotators A and B.

	A,B	A,C	A,D	B,C	B,D	C,D
Cohen's κ	.26	.36	.26	.31	.28	.28

Table 7: ADU detection agreement scores for annotator pairs (Threshold = .50)

4.2.2. ADU classification analysis

Table 8 shows IAA scores for ADU proposition type classification. Since determining IAA metrics for the detection of ADUs was already analyzed above, we factor out this subtask in our analysis of IAA for ADU type classification – only ADUs that each annotator pair agrees with (using an overlapping threshold of 0.5) are considered. Column "Global" represents the overall scores considering $k = 3$ possible labels for each component; the remaining columns present the scores separately for each label (with $k = 2$, *e.g. fact* vs *not fact*). When we are unable to set up a valid annotation comparison between two annotations (because we only consider ADUs that have been identified by both annotators), we disregard the annotation pair in the analysis (impacting the number M of annotation pairs considered). Consequently, we provide the number of annotator pairs (M) considered to determine IAA scores. The total number of potential annotation pairs is $M = 1119$ (373 articles $\times 3$ annotators per article).

	Global	***Fact***	***Policy***	***Value***
Cohen's κ	.40 (± .31)	.05 (± .38)	.31 (± .46)	.26 (± .39)
Obs. Agr.	73% (± 17)	68% (± 18)	87% (± 13)	73% (± 17)
M	1061	782	300	1058

Table 8: ADU classification agreement scores

As far as the "Global" metrics are concerned, we obtain "moderate agreement" with $\kappa = .40$. We observe that label *fact* obtains the lowest agreement scores ("slight agreement"), while *policy* obtains the highest agreement. The scores obtained for *value* and *Policy* correspond to "fair agreement". Table 9 shows that pairs including annotator D have the lowest ADU classification agreement. This is consistent with Table 3, which shows annotator D has a considerably higher tendency to label ADUs as propositions of *fact*. Contrasting with the lowest ADU detection agreement in Table 7, pair [A,B] has the highest ADU classification agreement.

	A,B	A,C	A,D	B,C	B,D	C,D
Cohen's κ	.62	.54	.26	.48	.30	.25

Table 9: ADU classification agreement scores for annotator pairs (Global)

Table 10 shows the confusion matrix regarding predicted labels by annotator pairs. It confirms that low agreement on *fact* is mostly due to its confusion with *value*, i.e., the majority class (69% of the matched ADUs) – this confusion is aggravated by annotator D, that tends to annotate more facts (see Table 3).

Finally, this analysis explains the higher agreement score regarding *policy* (as per Table 8): even though this class corresponds to the minority class (around 4% of the annotated ADUs), relatively high accuracy (40%) is obtained, *i.e.*, the observed value is well above the expected value in a random assignment.

	Fact	***Policy***	***Value***
Fact	721	6	1871
Policy		191	276
Value			4705

Table 10: Confusion matrix for ADU types

4.2.3. ADU argumentative role analysis

In the next steps of our IAA study, we focus on a graph-level analysis. Similarly to Section 4.2.2, we consider only ADUs that have been annotated by both annotators (considering the overlapping threshold of 0.5).

In this section, we focus on the role of an ADU: premise or conclusion. Following Stab and Gurevych (2014), we consider only the last ADU in a serial structure to correspond to the conclusion – remaining ADUs are premises. This is the basis for the setups shown in Table 11. The first setup considers these two argumentative roles. The second one disregards intermediate premises. The third setup considers intermediate premises as a separate argumentative role.

Labels	***P / C***	***FP / C***	***FP / IP / C***
Cohen's κ	.62 (± .34)	.71 (± .35)	.56 (± .33)
Obs. Agr.	82% (± .18)	85% (± .18)	74% (± .20)
M	1061	1057	1061

Table 11: ADU argumentative role agreement scores (P: Premise; FP: First Premise; IP: Intermediate Premise; C: Conclusion)

We obtain "moderate agreement" with the three-role setup and "substantial agreement" with the other ones, suggesting that annotators have a shared interpretation of ADU roles. A lower score when distinguishing intermediate premises as an additional class naturally decreases the chances of agreement. The higher agreement obtained when removing intermediate premises shows that these can be mistaken to be first premises or conclusions in some annotations, for instance when an argument graph is incomplete with respect to another annotation of the same argument.

ADU argumentative roles are the basis for our simplified analysis on the connection between premises and conclusion. The goal of this analysis is to understand whether annotators perceive the argument conclusion and the set of ADUs that support or attack it (either directly or indirectly), without looking to the details on how these supporting or attacking ADUs are arranged. This rearrangement will be considered in the following sections. Figure 6 illustrates the effect of different setups regarding the handling of intermediate premises. The "Original" setup considers the complete argumentative graph. The other setups are aligned with the choices laid out in Table 11.

4.2.4. Relation analysis

In this section, we analyze IAA for the subtasks of relation identification and classification. The term "relation" is used to denote a connection between a pair of

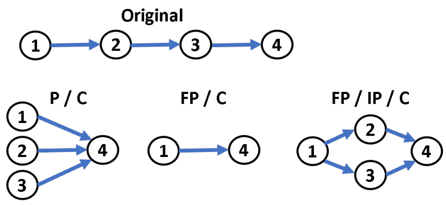


Figure 6: Graph simplification example

ADUs. Given that arguments are constrained to paragraph boundaries, we only combine ADUs belonging to the same paragraph. Formally, for each paragraph with ADUs $C_1, ..., C_n$, we form tuples $\langle C_i, C_j\rangle$, for all $i \neq j$ and $i, j \in [1, n]$, and label them with "related" if the original annotation contains a direct argumentative relation from C_i to C_j (either "support" or "attack") or with the label "non-related" otherwise. Considering all possible ADU pairings leads to an unbalanced setup towards the majority label "non-related", which could interfere with the IAA metrics employed. Consequently, we consider only ADU pairs for which at least one of the annotators has identified a relation. With this, we aim to analyze the extent to which the annotators agree on the subtask of identifying argumentative relations in this subset of ADU pairs.

Table 12 summarizes the IAA scores obtained for this subtask. For the "Original" column, we obtain a Cohen's κ score of .25 ("fair agreement"). Regarding the remaining columns, there is a slight increase in the agreement for P / C and FP / IP / C setups, but, removing intermediate premises has a considerable positive impact on agreement (.50, *i.e.*, "moderate agreement"). This reinforces the intuition from Section 4.2.3, indicating that some confusion between intermediate premises and other ADU roles is observed, while annotators tend to agree on the first premises and conclusion for a given argument.

	Original	*P / C*	*FP / C*	*FP / IP / C*
Cohen's κ	.25 (±.51)	.27 (±.52)	.50 (±.50)	.26 (±.52)
Obs. Agr.	69% (±25)	66% (±27)	75% (±27)	69% (±25)
M	1007	1005	938	1007

Table 12: Relation identification agreement scores

Table 13 shows that pairs including annotator D have the lowest relation identification agreement. As previously observed, annotator D has relatively more "attack" relations (as shown in Figure 3). While we do not discriminate between "support" and "attack" relations in this subtask, we believe that the nature of relations that this distinction entails is strongly connected to the lower agreement scores observed.

Next, we analyze agreement for the subtask of relation classification. To this end, we only consider relations identified by both annotators and focus on whether they agree on the relation type (*i.e.* "support" vs. "attack", $k = 2$). Table 14 summarizes the obtained IAA, which is high and with relatively low standard deviation, leading to "perfect agreement". However, it is important to recall that the distribution of labels is skewed towards "support", which comprises approximately 90% of the total number of 3232 relations identified by both annotators in a pair. All the setups based on simplified connections (P / C, FP / C, FP / IP / C) result in decreasing scores, *i.e.*, rearranging ADU roles has no positive impact on this task.

	A,B	A,C	A,D	B,C	B,D	C,D
Cohen's κ	.45	.45	.27	.27	-.11	.20

Table 13: Relation identification agreement scores for pairs – Original graph

	Original	*P / C*	*FP / C*	*FP / IP / C*
Cohen's κ	.97 (± .14)	.90 (± .24)	.91 (± .23)	.96 (± .15)
Obs. Agr.	99% (± 7)	95% (± 14)	94% (± 17)	98% (± 9)
M	928	895	834	930

Table 14: Relation classification agreement scores

4.2.5. Graph-based analysis

In this section, the focus is on the graph as a whole. Traditional IAA metrics were not conceived to take into account graph structures. Kirschner et al. (2015) proposed a graph-based agreement measure capturing the semantic similarity between different argument graphs. The proposed metric, shown in Equation 1, determines to what extent graph A is included in graph B.

$$\mathcal{D}_A = \frac{1}{|E_A|} \sum_{(x,y)\in E_A} \frac{1}{SP_B(x,y)}, \tag{1}$$

E_A is the set of edges (x, y) in graph A (with x as source and y as target node). $SP_B(x, y)$ is the shortest path between nodes x and y in graph B. In general, for each relation in graph A, this metric aims to determine to which extent the corresponding connection was annotated in graph B, by calculating the distance between the corresponding source and target nodes in graph B. Similar to previous analyses, we only consider relations between ADUs that both annotators agree on. Finally, Kirschner et al. (2015) propose determining the graph-based agreement score between two annotators by calculating the F1-score as follows: $\frac{2 * \mathcal{D}_A * \mathcal{D}_B}{\mathcal{D}_A + \mathcal{D}_B}$.

	Original	*P / C*	*FP / C*	*FP / IP / C*
f1	.80 (± .21)	.76 (± .23)	.84 (± .22)	.79 (± .21)
M	982	963	890	983

Table 15: Graph-based IAA scores

Table 15 shows the graph-based F1-score for our annotation study. We obtain a relatively high F1-score, which means that once the annotators detect the same ADUs, they tend to agree with the overall structure of the annotation graph. These results are encouraging, because they show that at a higher-level of analysis (*i.e.* assessing arguments' overall graph-based structures) annotators capture mostly the same essence of the argument. Lower scores obtained for the P / C setup reinforces the intuition that there is some disagreement on the path between intermediate premises and conclusion. Higher scores for the FP / C setup indicate that annotated graphs tend to overlap when considering only the beginning and the end of the graph.

5. Related Work

The rise of interest in argument mining (Stede and Schneider, 2019) has brought the need for annotated corpora. However, the disparity of text genres on which such corpora can be built hampers their joint use, as does the underlying theoretical argumentation models adopted. We briefly mention some argument annotation projects that follow a theoretical argumentation model similar to ours or address similar text genres.

Freeman's model (Freeman, 2011) has been one of the most widely followed for argument annotation. Stab and Gurevych (2014) and Stab and Gurevych (2017) worked specifically on persuasive essays. Their argumentative structure includes: a *major claim*, expressing the author's standpoint with respect to the topic; *claims* labeled with a stance attribute (*for* or *against* the major claim); and *premises* supporting or attacking claims.

No linked vs convergent distinction was made, and divergent structures have been ruled out. Looking at the annotations produced, the authors found the distinction between claims and premises to be particularly blurry, due to the fact that chains of reasoning (serial structures) can be established. Authors like Peldszus and Stede (2016) and Skeppstedt et al. (2018) worked on the annotation of short crowdsourced argumentative texts, distinguishing between *linked* and *convergent* arguments, as well as two types of attacks: *undercuts* and *rebuttals*. The authors identify as main issues the presence of implicit claims and restatements, the distinction between direct, indirect supports and mere causal connections, and the existence of non-argumentative text units. Visser et al. (2018) analyzed TV political debates on the US 2016 presidential elections, while following different models of argument, including PTA. PTA has been applied to its full extent, including inference relations and the corresponding nodes (sources and targets of relations), which are classified in one of *fact*, *value* or *policy*. Some studies focused on news editorials, which are close to the opinion articles genre. Bal and Saint Dizier (2010) and Bal (2014) defined a semantic tagset and refined a set of lexicons with which they developed an automatic annotation tool. Despite the fact that the tagset includes types of arguments and rhetorical relations, this work does not seem to have produced a reliable corpus on argument-annotated news editorials. After observing that editorials lack a clear argumentative structure, Al-Khatib et al. (2016) aimed instead at mining argumentation strategies. They segmented editorials into ADUs, annotated according to their role: common ground, assumption, testimony, statistics, anecdote, or other. The purpose was to analyze the argumentation strategy employed at a macro document level. The generated corpus includes editorials from three different news sources, which the authors comparatively analyze in terms of the usage of each type of role. An exploratory short Portuguese corpus of opinion articles was described by Rocha and Lopes Cardoso (2017), but it does not follow a rigorous annotation methodology. Based on this prior work, this is the first study following well-defined annotation guidelines and aiming at a rigorous analysis of annotator agreement. Other argument-annotation projects were conducted (Lippi and Torroni, 2016; Cabrio and Villata, 2018; Park and Cardie, 2018; Lawrence and Reed, 2019; Schaefer and Stede, 2021) but these are not detailed here due to their deviant nature, *i.e.*, either the argument models followed are not closely related to ours, or the genre of text is substantially different.

6. Conclusions

This paper reports the methodology, process and results on the annotation of arguments in a set of Portuguese opinion articles. The complexity of the annotation study required a detailed analysis of different layers of the annotation process, ranging from token-level identification of ADUs to how these are related in a graph-based structure. Due to the foreseen challenges in terms of IAA, each article was annotated by a set of expert annotators, making this a valuable resource for the study of agreement in such complex set of subtasks.

The identification and interpretation of argumentative content in opinion articles is a semantically demanding task – that is visible in our IAA analysis, and noticeable in the number of subtasks we covered. We observe that, even though skilled annotators are capable of grasping the essence of arguments, obtaining more consistent annotations at token-level and for complex argument structures is a challenging task. From the set of subtasks outlined in our analysis, we have made several remarks that we believe are relevant for future annotation efforts of this kind. The consolidated corpus will be invaluable for several NLP tasks, from proposition type classification and sentiment analysis, to argument mining (Stede and Schneider, 2019) related tasks, including argumentative relation identification (Nguyen and Litman, 2016; Cocarascu and Toni, 2017; Rocha et al., 2018) and argument density prediction (Rocha et al., 2022; Visser et al., 2020).

7. Acknowledgements

This research is supported by project DARGMINTS (POCI/01/0145/FEDER/031460), LIACC (FCT/UID/CEC/0027/2020), INESC-ID (UIDB/50021/2020) and CLUP (UIDB/00022/2020), funded by Fundação para a Ciência e a Tecnologia (FCT). Gil Rocha is supported by a PhD grant (SFRH/BD/140125/2018) from FCT.

8. References

Al-Khatib, K., Wachsmuth, H., Kiesel, J., Hagen, M., and Stein, B. (2016). A news editorial corpus for mining argumentation strategies. In *Proceedings of the 26th International Conference on Computational Linguistics*, pages 3433–3443, Osaka, Japan.

Baak, M., Koopman, R., Snoek, H., and Klous, S. (2020). A new correlation coefficient between categorical, ordinal and interval variables with pearson characteristics. *Computational Statistics & Data Analysis*, 152:107043.

Bal, B. K. and Saint Dizier, P. (2010). Towards building annotated resources for analyzing opinions and argumentation in news editorials. In *Proceedings of the 7th Conference on International Language Resources and Evaluation*, Valletta, Malta.

Bal, B. K. (2014). Analyzing opinions and argumentation in news editorials and op-eds. *Int. J. Advanced Computer Science and Applications (IJACSA), Special Issue on Natural Language Processing*, 4(1).

Branco, A. and Silva, J. (2004). Evaluating solutions for the rapid development of state-of-the-art POS taggers for Portuguese. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, Lisbon, Portugal.

Cabrio, E. and Villata, S. (2018). Five years of argument mining: a data-driven analysis. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 5427–5433.

Cocarascu, O. and Toni, F. (2017). Identifying attack and support argumentative relations using deep learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1374–1379, Copenhagen, Denmark.

Cunningham, H. et al. (2014). *Developing Language Processing Components with GATE Version 8 (a User Guide)*. University of Sheffield Department of Computer Science.

Davies, M. and Fleiss, J. L. (1982). Measuring agreement for multinomial data. *Biometrics*, pages 1047–1051.

Eger, S., Daxenberger, J., and Gurevych, I. (2017). Neural end-to-end learning for computational argumentation mining. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11–22, Vancouver, Canada.

Freeman, J. B. (2011). *Argument Structure: Representation and Theory*. Argumentation Library. Springer Netherlands.

Habernal, I. and Gurevych, I. (2015). Exploiting debate portals for semi-supervised argumentation mining in user-generated web discourse. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2127–2137, Lisbon, Portugal. Association for Computational Linguistics.

Janier, M., Lawrence, J., and Reed, C. (2014). Ova+: an argument analysis interface. In Simon Parsons, et al., editors, *Computational Models of Argument*, Frontiers in artificial intelligence and applications, pages 463–464, Netherlands. IOS Press.

Kirschner, C., Eckle-Kohler, J., and Gurevych, I. (2015). Linking the thoughts: Analysis of argumentation structures in scientific publications. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 1–11, Denver, CO.

Krippendorff, K. (2004). Measuring the reliability of qualitative text analysis data. *Quality and Quantity*, 38:787–800.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

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

Lippi, M. and Torroni, P. (2016). Argumentation mining: State of the art and emerging trends. *ACM Trans. Internet Technol.*, 16(2):10:1–10:25.

Lopes Cardoso, H., Rocha, G., Sousa Silva, R., Carvalho, P., Won, M., and Martins, B., (2019). *Diretrizes para Anotação de Estruturas Argumentativas em Artigos de Opinião*. FEUP/LIACC, FLUP/CLUP, INESC-ID.

Meyer, C. M., Mieskes, M., Stab, C., and Gurevych, I. (2014). DKPro agreement: An open-source Java library for measuring inter-rater agreement. In *Proceedings of the 25th International Conference on Computational Linguistics: System Demonstrations*, pages 105–109, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Nguyen, H. and Litman, D. (2016). Context-aware argumentative relation mining. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1127–1137, Berlin, Germany.

Park, J. and Cardie, C. (2018). A corpus of eRulemaking user comments for measuring evaluability of arguments. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*, Miyazaki, Japan.

Peldszus, A. and Stede, M. (2016). An annotated corpus of argumentative microtexts. In D. Mohammed et al., editors, *Argumentation and Reasoned Action - Proc. of the 1st European Conference on Argumentation, Lisbon, 2015*. College Publications, London.

Persing, I. and Ng, V. (2016). End-to-end argumentation mining in student essays. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1384–1394, San Diego, California.

Reed, C. and Rowe, G. (2004). Araucaria: Software for argument analysis, diagramming and representation. *International Journal of AI Tools*, 14:961–980.

Rehbein, I., Ruppenhofer, J., Sporleder, C., and Pinkal,

M. (2012). Adding nominal spice to SALSA - frame-semantic annotation of german nouns and verbs. In Jeremy Jancsary, editor, *Proceedings of the 11th Conference on Natural Language Processing, Vienna, Austria*, pages 89–97. ÖGAI, Wien, Österreich.

Rocha, G. and Lopes Cardoso, H. (2017). Towards a relation-based argument extraction model for argumentation mining. In Nathalie Camelin, et al., editors, *Statistical Language and Speech Processing*, pages 94–105, Cham. Springer International Publishing.

Rocha, G., Stab, C., Lopes Cardoso, H., and Gurevych, I. (2018). Cross-lingual argumentative relation identification: from English to Portuguese. In *Proceedings of the 5th Workshop on Argument Mining*, pages 144–154, Brussels, Belgium.

Rocha, G., Leite, B., Trigo, L., Lopes Cardoso, H., Sousa-Silva, R., Carvalho, P., Martins, B., and Won, M. (2022). Predicting argument density from multiple annotations. In *Proceedings of the 27th International Conference on Natural Language & Information Systems*.

Schaefer, R. and Stede, M. (2021). Argument mining on twitter: A survey. *it - Information Technology*, 63(1):45–58.

Skeppstedt, M., Peldszus, A., and Stede, M. (2018). More or less controlled elicitation of argumentative text: Enlarging a microtext corpus via crowdsourcing. In *Proceedings of the 5th Workshop on Argument Mining*, pages 155–163, Brussels, Belgium.

Stab, C. and Gurevych, I. (2014). Annotating argument components and relations in persuasive essays. In *Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1501–1510, Dublin, Ireland.

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

Stede, M. and Schneider, J. (2019). *Argumentation Mining*. Morgan & Claypool Publishers.

Toulmin, S. E. (1958). *The Uses of Argument*. Cambridge University Press.

van Amelsvoort, M. and Maes, A. (2016). Show me your opinion. perceptual cues in creating and reading argument diagrams. *Instructional Science*, 44:335–357.

van der Plas, L., Samardzić, T., and Merlo, P. (2010). Cross-lingual validity of PropBank in the manual annotation of French. In *Proceedings of the 4th Linguistic Annotation Workshop*, pages 113–117, Uppsala, Sweden.

van Eemeren, F. H., Garssen, B., Krabbe, E. C. W., Snoeck Henkemans, A. F., and Wagemans, J. H. (2014). *Handbook of Argumentation Theory*. Springer Reference. Springer Netherlands.

Visser, J., Lawrence, J., Wagemans, J., and Reed, C. (2018). Revisiting computational models of argument schemes: Classification, annotation, comparison. In Sanjay Modgil, et al., editors, *Computational Models of Argument - Proceedings of COMMA 2018*, volume 305 of *Frontiers in Artificial Intelligence and Applications*, pages 313–324, Netherlands. IOS Press.

Visser, J., Konat, B., Duthie, R., Koszowy, M., Budzynska, K., and Reed, C. (2020). Argumentation in the 2016 US presidential elections: Annotated corpora of television debates and social media reaction. *Language Resources and Evaluation*, 54:123–154.

Wagemans, J. (2016). Constructing a periodic table of arguments. In *Argumentation, Objectivity, and Bias: Proceedings of the 11th International Conference of the Ontario Society for the Study of Argumentation*, pages 1–12.

Walton, D. N. (1996). *Argumentation Schemes for Presumptive Reasoning*. L. Erlbaum Associates.

Won, M. (2021). Political opinions on the past web. In Daniel Gomes, et al., editors, *The Past Web: Exploring Web Archives*, pages 243–252. Springer International Publishing, Cham.