Which Argumentative Aspects of Hate Speech in Social Media can be reliably identified?

Damián Furman[1,2], **Pablo Torres**[3], **José A. Rodríguez**[3],
Diego Letzen[3], **Vanina Martínez**[4], **Laura Alonso Alemany**[5,6]
[1] Departamento de Computación, Universidad de Buenos Aires, Argentina
[2] Consejo Nacional de Investigaciones Científicas y Técnicas, Argentina
[3] Facultad de Filosofía, Universidad Nacional de Córdoba, Argentina
[4] Artificial Intelligence Research Institute (IIIA-CSIC), Barcelona, Spain
[5] Facultad de Matemática, Astronomía y Física, Universidad Nacional de Córdoba, Argentina
[6] Fundación Via Libre, Argentina

Abstract

With the increasing diversity of use cases of large language models, a more informative treatment of texts seems necessary. An argumentative analysis could foster a more reasoned usage of chatbots, text completion mechanisms or other applications. However, it is unclear which aspects of argumentation can be reliably identified and integrated in language models.

In this paper we present an empirical assessment of the reliability with which different argumentative aspects can be automatically identified in hate speech in social media. We have enriched the Hateval corpus (Basile et al., 2019) with a manual annotation of some argumentative components, adapted from Wagemans (2016)'s Periodic Table of Arguments. We show that some components can be identified with reasonable reliability. For those that present a high error ratio, we analyze the patterns of disagreement between expert annotators and errors in automatic procedures, and we propose adaptations of those categories that can be more reliably reproduced.

1 Introduction

With the impressive advances obtained in Large Language Models (LLMs), applications of automated language generation are quickly expanding to affect more and more areas of human activity, specially with the generalization of conversational chatbots. It is known that these models tend to amplify stereotypes, resulting in the naturalization of prejudices and finally the dehumanization of social groups in the form of hate speech.

Hate speech is a grave danger. The International Convention on the Elimination of all Forms of Racial Discrimination states that hate speech "*rejects basic human rights principles of human dignity and equality and seeks to degrade the position of individuals and groups in society's esteem*"[1].

Through the amplification provided by social media and LLMs, its effects are also amplified, as it can deepen prejudice and stereotypes (Citron and Norton, 2011). That is why great efforts have been made to detect and neutralize it. The most common form of neutralization to date has been banning hate speech from public forums. However, this strategy collisions with the right to freedom of expression. In addition, it is usually implemented by resorting to human moderators who are exposed to toxic content for long workdays.

Automatic argumentation analysis enables alternatives to censorship like argument retrieval and organization or automatic generation of counter-arguments. The recent developments of LLMs make these tasks more feasible. But, although they behave in a competent way from a purely conversational point of view, they do have not been designed to reason or argue. Moreover, they do not seem to be able to prevent harmful effects beyond very shallow guardrails, which is a critical concern when dealing with hate speech. That is why it seems necessary to enhance them beyond pure unannotated text, to obtain a more nuanced treatment of the argumentative dimension of texts.

The question remains, *how can we know which argumentative aspects will be useful for LLMs to improve their performance in nuanced, risky tasks like automatic generation of counter-arguments against hate speech in social media?*

In this work we present the Argumentation Structure Of Hate Messages Online (ASOHMO), a protocol to annotate argumentative information in hate tweets, and an annotated dataset of tweets to train automatic classifiers. These annotations are an adaptation of Wagemans (2016)'s proposal for hate speech in Twitter, where much of the argumentation refers to implicit elements, and one finds typos,

[1] United Nations Strategy and Plan of Action on Hate Speech: Detailed Guidance on Implementation for United Nations Field Presences, 2020.

Proceedings of the 4th International Workshop on Designing Meaning Representations, pages 136–153
June, 2023.

incomplete phrases and incoherent syntax. Despite this challenging context, by applying our protocol, we obtained substantial agreement between different human judges to identify the argumentative structure of tweets. We also found that LLMs can successfully detect some of these argumentative components, even when few annotated examples are provided, which seems to indicate that it is feasible to finetune them to address some specific argumentation tasks and domains.

The rest of the paper is organized as follows. In the next Section we discuss relevant work, including the foundational Wagemans (2016)'s proposal. Section 3 describes the categories that we distinguish in our annotation framework, and in Section 4 we present how they apply to hate speech in social media, more concretely, to the manual annotation of the Hateval corpus (Basile et al., 2019). Finally, in Section 5 we show how LLMs can identify some argumentative components, but not others, with varying degrees of success. We analyze the causes of low success and propose how to adapt the definition of the target argumentative aspects to improve their reliability of annotation, both manual and automatic.

2 Relevant Work

There are many different proposals on how to model the argumentative aspects of texts, even if we only consider those aimed or used for computational application. We are not providing an exhaustive overview of approaches here, but just some examples to motivate and frame the model of argument that we present in this work.

One of the main distinctions between proposals is whether they are general purpose or domain specific. Domain-specific approaches propose tailored categories, like Teufel et al. (1999)'s "*background*", "*aim*" or "*comparison*" for scientific papers, or Al-Khatib et al. (2016)'s "*anecdote*" or "*statistics*" for the argumentative analysis of editorials. They tend to achieve good inter-annotator agreement and good accuracy in automatic identification, but are not portable to different domains.

General-purpose argumentation models have very different approaches. Many computation-oriented proposals are based on Toulmin (2003)'s theory of practical argument. They distinguish between two main components of arguments, "*conclusion*" (also called "*claim*") and "*fact*" (also called "*justification*" or "*premise*"). They usually try to identify relations between components and between arguments, aiming to create a full argument tree that accounts for the argumentative structure of a text. This kind of model has been applied to essays (Stab and Gurevych, 2014) or user-generated discourse (Habernal and Gurevych, 2017). It is very general, thus easily portable to different domains. At the same time, it is not very stable, since inter-annotator agreement is not high, and the information it provides about the argument is not as rich as in the case of domain-specific approaches.

Another approach to modelling argument in texts are schemes. Argument schemes are "*patterns of informal reasoning*" (Walton et al., 2008) that "*represent forms of argument that are widely used in everyday conversational argumentation*" (Macagno et al., 2018). Argument Schemes specify a pattern of reasoning and a set of critical questions oriented to test the defeasibility conditions on the pattern. This pattern and critical questions provide very insightful, actionable information about the argument, which can be later used for applications like building a counter-argument.

Several authors have adapted Walton's schemes to specific purposes, even proposing alternatives to critical questions (Atkinson and Bench-Capon, 2018; Kökciyan et al., 2018). The main drawback of these proposals is that the inventory of scheme is very profligate, and it has become clear that, identifying a scheme within a given text becomes quite difficult, both manually and automatically.

2.1 The Periodic Table of Arguments

Trying to find a trade-off between the excessive detail of schemes and the scarce information provided by claim-premise approaches, Wagemans (2016) proposes an analytic approach to argument schemes, aimed to obtain the core schemes proposed by Walton et al. (2008), with fewer categories based on a limited set of general argument features.

This is a characteristic that we find particularly useful for building a simple system that is easy to annotate without an enormous effort and achieving a high level of agreement between human annotators, which leads us to expect higher reproducibility in inferred models. Moreover, an analytic approach allows determining which aspects of argumentation are more feasible to detect automatically, and identifying which particular aspects are more useful for a given application, such as components that could be used to elaborate a response.

All arguments under Wagemans's system have a premise and a conclusion labeled with one Type of statement each. But it goes beyond the mere premise-conclusion information. The PTA is a factorial typology of arguments that offers a comprehensive overview of the various types of arguments by describing them as a unique combination of three basic characteristics (Wagemans, 2019):

1. **first order or second order** argument. A common term between premise and conclusion transfers the acceptability from one to the other. If this common term is explicit, then it is a first order argument. If a reconstruction is needed, then it is a second order argument.

2. **predicate or subject** argument. If the common term is in the subject of the propositions making the premise and the conclusion, then it is a subject argument, otherwise, it is a predicate argument.

3. **policy, fact or value**. The conclusion and premise can be labeled each as a statement of policy (the speaker mandates or states that something should be done), a statement of value (the speaker issues an opinion about something), or a statement of fact (the speaker conveys a proposition as a true fact).

Visser et al. (2021) conducted an exhaustive research on annotating the US 2016 presidential debate corpus using both Walton's schemes and Wagemans's Periodic Table of Arguments. They reported a higher agreement for Wagemans's typology, specially without considering classification between first and second order arguments. Moreover, they sustain that for Wagemans's typology, "*the division into independent sub-tasks simplifies the annotation while maintaining reliability*".

We adapted Wagemans's proposal to hate speech on social media, with the goal of identifying elements that can be relevant to either a human or a machine in the task of analyzing or countering hate speech.

Focusing on hate speech on Twitter, we have to take into account that many argumentative hate tweets are based on assumptions justified by prejudice or context information that is difficult to recover. This means that in many cases, it is difficult to rebut them from the perspective of formal deductive logic. We believe that an approach based on informal logic, like the one proposed by Wagemans (2016), is more adequate to capture this kind of arguments that are organized with informal relations.

In the following Section we describe our approach. We provide an overview of other social media corpora annotated with argumentative information in Appendix H.

3 A Framework to Identify Argument Components on Twitter Hate Speech

The goal of our argumentation model is to provide an argumentative analysis that can help expose the core of the reasoning supporting a hate message. We believe that this can help both humans and automatic models to better address hate speech.

We are labeling two kinds of information: domain-specific and argumentative-general. Domain-specific information allows to exploit particular characteristics of hate messages on Twitter: they always mention a collective that is implicitly or explicitly associated with a negative property, action or consequence. Argumentative-general structure is based on a simplification of Wageman's proposal that is aimed to increase inter-annotator agreement. Reaching acceptable levels of inter-annotation agreement is very important to our purpose, as it indicates that the annotation process can be systematized and possibly automatized.

We created an annotation protocol[2] where both kinds of argumentative information are defined in a procedural manner. This protocol was applied by human analysts to annotate hate speech tweets, with five steps that are described as follows. The annotation team and environment are described in Appendix A.

3.1 Argumentative or Non-argumentative

Following (Wagemans, 2019), "*an argument (...) consist(s) of two statements, namely a conclusion – the statement that is doubted – and a premise*". We consider a tweet to be argumentative if it is possible to divide it in these two components. Examples of non-argumentative tweets can be found in Appendix E: exhortations to some action without justification, insults, name callings, support for a particular policy or description of facts without an explicit conclusion.

[2]Annotation guidelines can be found at `shorturl.at/cv458`.

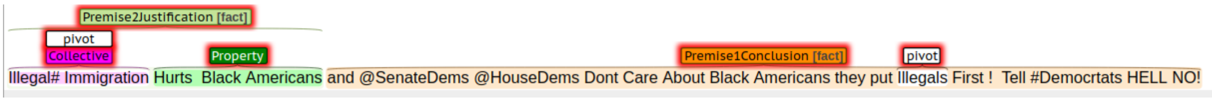

Figure 1: Example of labeled argumentative hate tweet.

3.2 Domain-specific components: Collective and Properties

All hate messages are directed towards a specific group by definition. Usually, the content of the message is to associate this group with a negative property or an undesirable action or consequence. If this property[3] is explicit, we label it.

3.3 General argumentative components: Justification and Conclusion

All argumentative tweets are labeled with one and only one Conclusion and Justification, though these can be separated in many non-contiguous parts inside the tweet. Annotators were instructed to choose the longest Conclusion and Justification that they could find, leaving out only hashtags indicating topics, links, user mentions or non-relevant words or information. Justifications may be arguments themselves, having their own inner structure involving different premises, but this is not annotated as we are only interested in capturing the main standpoint that the user wants to gain acceptability for.

When labeling these components, we are not considering the subject-predicate structure proposed by Wagemans. Visser et al. (2021) warned about how this model presupposes that premises and conclusions of arguments consist of complete categorical propositions comprising a clear subject-predicate structure, which is not always the case in user-generated, informal social media text.

3.4 Argumentative relation: Pivot

Following Wagemans, argumentative components transfer reason from one to the other. We assume that there can be textual cues of this transfer, in the form of an element that is common to both components. We call this element the pivot. We identify pivots as two sequences of words, one for each premise, that can be related to the element that those premises have in common.

This relation is generally not unique; the underlying common ground between the premises could be expressed in different forms or could present multiple aspects signaled by different words. Whenever this element is explicit in the text (it might be not), we annotate it.

The pivot holds a relation with Wagemans's categories of first and second order arguments. If an argument is considered first-order, it means that the common element between premises must be explicit (by definition, it must be either of the form A is X because A is Y or B is Z because C is Z). For a second-order argument, there might still be an explicit pivot or not.

3.5 Types of Proposition

Wagemans proposes "a characterization of the types of arguments based on the combination of the types of propositions they instantiate" (Wagemans, 2016). These types are taken from debate theory (Schut, 2014), where three distinctions on propositions are made: (1) policy (P), (2) value (V) and (3) fact (F).

We label our propositions using the same types and add to our annotation manual different guidelines on how to recognize each one: a policy proposition is a mandate often expressed as orders, imperatives, or actions that need to be accomplished in the public domain. Fact and value propositions were reported to be more difficult to differentiate. As a general criterion, a proposition to be labeled as value must have explicit markers of the speaker being involved in the assertion expressed (opinionated adjectives, verbs of thought, etc.). Otherwise, the premise is considered as fact. Examples can be seen in Appendix E.

4 The ASOHMO Corpus

We applied our argumentation model via the annotation protocol described in the previous Section to the HatEval 2019 corpus (Basile et al., 2019). Focusing on argumentative tweets, we did not annotate tweets labeled as "aggressive", consisting mostly of abusive language (name callings, insults, exhortations to action and other types of attacks), nor tweets targeted against specific individuals or women, as they were almost exclusively abusive and non-argumentative. After these filters, a corpus

[3] A property is anything that is associated with the targeted community, whether it is an adjective, a consequence, an action, etc.

	Argumentative	Domain-specific		Argument-general				
		Collective	Property	Pivot	Justif.	Concl.	Type of Conc.	Type of Just.
κ	.85	.64	.60	.52	.62	.64	.60	-.03

Table 1: Agreement scores between two annotators for 150 tweets.

of 970 tweets in English and 196 tweets in Spanish remained.

The dataset is released[4] for the free use of the scientific community, together with the scripts for reproducing experiments.

4.1 Inter-annotator Agreement

We calculated inter-annotator agreement to assess the reproducibility of the annotations and the feasibility of automatic identification. While the whole corpus was annotated by a single annotator, 150 tweets (15% of the corpus) were labeled by a second annotator[5]. Then, per-category agreement was calculated with Cohen's κ (Cohen, 1960). Agreement was calculated in a per-tweet basis for the Argumentative vs. Non-Argumentative category using a binary label, and for the Type of Conclusion and Justification categories, using one label with three possible values representing *fact*, *value* and *policy*. For all other categories, agreement was calculated in a per-word basis with a binary label assigned to each word, marking whether it belongs to the category or not.

In Table 1 we can see that annotators can reach a substantial level of reproducibility, around $\kappa = .6$ for Collective, Property, Justification, Conclusion and Type of Conclusion and .85 for the distinction between Argumentative or non-Argumentative tweets. In contrast, the Pivot presents a moderate level of inter-annotator agreement, and the Type of Justification presents no agreement at all.

To calculate agreement, we follow a criterion similar to that of Visser et al. (2021): while comparing two annotators, if at least 50% of the words in the smallest component marked by one of the annotators overlaps with words marked by the other one, then it is considered an agreement. For example, if one annotator marked "*the damage illegals do*" as a Property associated to a Collective and the other annotator marked just "*damage*" as a Property we consider that 100% of the words in the shortest "*damage*" in both examples and assume that all the other words are marked as not being part of the Property in both cases.

`@user @user` *<u>`sanctuary cities`</u>* *`are`* ***`against the law.`***`PLEASE SHUT` <u>`THEM`</u> `DOWN & ARREST/PROSECUTE ALL` **`CRIMINAL`** `GOVERNORS & MAYORS`

Figure 2: Disagreement concerning Pivot. One annotator is underlined, while the other is bolded. Justification is marked with italics and Conclusion with capitalization

When inspecting examples of disagreement between annotators for Pivot, as shown in Figure 2, we found that in many cases both annotations could be considered accurate, as there may be more than one possibility for annotators to tag. Furthermore, as the relation is very deep in the layers of meaning, annotators may interpret it as signalled by different surface features, and as a consequence they may tag different sequences of words while considering the same relation.

Finding patterns in the disagreements between annotators can be used to redefine categories (Teruel et al., 2018). In a second annotation phase, we will be redefining the Pivot category to obtain more agreement between annotators. We understand that this element is particularly challenging, because it signals a very deep relation and its correspondence with surface textual phenomena may not be direct, or multiple. That is why we plan to rethink it as a binary classification problem, where human judges are presented with one or more possibilities of Pivots for a given argument, and they have to say whether they consider any of them to be a valid Pivot for the example.

5 Automatic Identification of Arguments

We conducted several experiments to assess the feasibility that LLMs can automatically identify different argumentative aspects.

For each set of hyperparameters used, we fine-tuned the same language models using different random tweets for each partition, always respecting this proportion. We report the average of these three fine-tuned models' F1, Precision and Recall to detect or classify argument components. For

[4]https://github.com/ASOHMO/ASOHMO-Dataset

[5]The sample's size for the test is proportionally higher than many of the previous works: Bosc et al. (2016) used 100 tweets to calculate agreement over a dataset of 4000 whereas Dusmanu et al. (2017) used 100 tweets for its first dataset of 1887 tweets, 80 tweets for its second dataset of 1459 tweets and used the whole third dataset of 368 tweets.

	RoBERTa			BERTweet			XLM-RoBERTa-Mix			XLM-RoBERTa-XL		
	F1	Pr	Rec	F1	Pr	Rec	F1	Pr	Rec	F1	Pr	Rec
Arg./Non-Arg.	**.89**±.02	.84	.95	.88±.01	.84	.93	.87±.04	.84	.91	.84±.03	.84	.85
Justification	.73±.05	.69	.76	**.77**±.05	.75	.78	.76±.05	.71	.81	.75±.01	.71	.80
Conclusion	.55±.02	.60	.51	**.61**±.03	.64	.58	.60±.02	.59	.61	.54±.03	.59	.49
Type of Just.	.41±.09	.48	.39	**.42**±.09	.48	.41	.35±.05	.34	.37	.33±.03	.33	.35
Type of Conc.	.58±.05	.62	.57	**.65**±.11	.67	.65	.61±.06	.65	.62	.63±.02	.66	.62
Collective	**.59**±.03	.56	.64	.58±.05	.55	.62	**.59**±.06	.58	.60	.27±.07	.41	.21
Property	.46±.04	.52	.41	.47±.03	.50	.43	**.50**±.03	.57	.43	.42±.04	.42	.43
Pivot	**.45**±.04	.52	.41	.40±.08	.43	.39	.39±.08	.42	.38	.33±.08	.41	.27

Table 2: F1, precision and recall for the target class in the automatic detection of argument components in tweets. Each experiment was carried out with three randomized partitions, the mean and standard deviation of the F1 are presented. Best results for F1 for each category are highlighted in boldface.

multi-label classification, the macro average is calculated, otherwise, we report the score of the target class. We also report per-class F1 scores for the three possible Types of premises: Fact, Value and Policy.

Models. We fine-tuned the following LLMs:

RoBERTa (Liu et al., 2019): a BERT-like (Devlin et al., 2018) LLM, pre-trained with more data.

BERTweet (Nguyen et al., 2020): a RoBERTa-based LLM trained on data from Twitter.

XLM-Roberta (Conneau et al., 2019): a RoBERTa based multilingual LLM. We fine-tuned it with a Mixed Language (Mix) version using both English and Spanish for training and testing and with a Cross-Lingual version (XL) using English for training and Spanish for testing.

5.1 Predicting Individual Components

We trained different kinds of models to automatically recreate the annotation process one component at the time: one for sequence binary classification, to predict if a tweet is argumentative or not; five models for token classification, to predict for each word, if it is labeled as part of the collective, the property associated to that collective, the pivot, the justification or the conclusion, respectively; and two models for sequence classification, fed only with the correspondent text of the premise (Justification or Conclusion), to predict the Type associated with it (fact, value or policy). Results of this experiment are shown in table 2.

Distinguishing argumentative from non-argumentative tweets achieves a very satisfying .89 F1. In general, components with higher inter-annotator agreement perform better, with justifications identified with .77 F1. Components with low inter-annotator agreement are also identified with more errors: conclusions have an F1 of .61 ($\kappa = 64$), collective F1=59, ($\kappa = 64$), property F1=.50 ($\kappa = 60$) and finally pivots only reach an F1 of .45 ($\kappa = 52$).

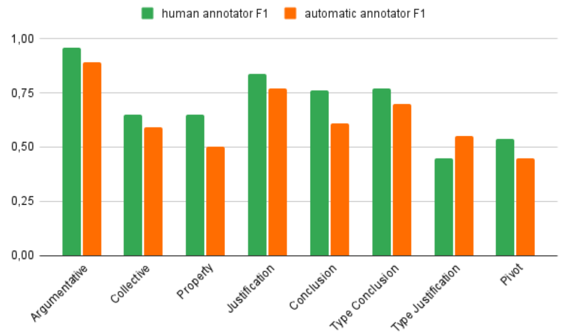

Figure 3: F1 score for "predictions" done by a human annotator and compared with predictions done by the best performing automatic classifiers (BERTweet for Justification, Conclusion and their Types - trained with all the premises -, Roberta for Argumentative and Pivot and XLM-Roberta for Collective and Property).

In Figure 3 we compared the F1 scores of the best performing models with inter-annotator agreement. We calculated the F1 score of the 150 tweets labeled by two judges using one as the ground truth and the other as the one being evaluated. We can see that both scores are highly correlated, although human annotators tend to agree slightly more than the automatic predictor with respect to the human ground truth, so there is still room for improvement for automatic predictors.

Analyzing predictions for the worst performant components (Property and Pivot), we can see that the models predicting Properties have a tendency to recognize any word with a negative charge, disregarding if it is referring to the Collective itself. Models recognizing Pivots sometimes find more than one possibility to label. Figure 4 shows how the model predicts the real pivot, but then also predicts another one not labeled on the original example that could be valid. This shows that, at least partially, some mistakes are made because of the subjective nature of the task and the multiple valid possibilities of labelling. To overcome this problem, annotators should consider the possibility of multiple pivots and try to label them all.

```
Salvini prosecuted for defending italian
sovereignity and finally preventing
hundreds of migrants to invade Italy
grande Salvini, help us preserve the
european culture against the invasion
#StopIslamization #ComplicediSalvini
#StopInvasion #RefugeesNotWelcome
```

Figure 4: Example of prediction of Pivot. Labeled justification is in blue, while conclusion is red. Real pivot is underlined, while predicted Pivot is bolded.

Regarding the different models, BERTweet achieves the best performance on most experiments involving Justifications, Conclusions or their types, and is close to the best results on other components. RoBERTa achieves higher results for Pivots.

Multilingual experiments achieve a performance similar to their monolingual counterparts for most components, specially Properties, indicating that training with mixed languages does not decrease, and can even improve, performance.

Results on cross-lingual experiments where models are trained with English and tested against Spanish, on the other hand, show different behavior depending on the component: for finding argumentative tweets, Justifications, Types of Justification and Types of Conclusion, results are similar to their counterparts on monolingual experiments. Collective, on the other hand, has a major drop in performance for all experiments compared to all other model settings. This is explained because of the very specific lexicon used for naming collectives, with lots of out of vocabulary and slang words. The pivot also suffers a drop in performance on both multilingual settings, but more so on cross-lingual.

5.2 Predicting Components Simultaneously

The goal in this case is to measure the performance of the models when simultaneously predicting components labeled on the same annotation step. We want to assess whether training with information about both components helps to improve the performance when predicting each of them individually or not. We ran an experiment to jointly predict Collective and Property and another for Justifications and Conclusions. Each word is assigned one of three labels, indicating if they belong to either of the two searched components or not.

Joint prediction of components labeled on the same annotation step produces almost the same results as predicting them individually. This has the advantage of consuming half of the resources and time; however, the definition of the problem changes, as each token can only be part of one or none component, but not both.

5.3 Predicting the Type of Premises

The Type of Conclusion or Justification (Fact, Policy or Value) should be independent of its premise (Justification or Conclusion), so in terms of semantic information, to predict this, it should not matter if models are trained with just one or both of them.

Moreover, using both kinds of premises increases the number of training examples and can help to overcome the unbalance between Fact and Policies (specially on Justifications, where facts are the vast majority).In Table 3 we can see that models trained to predict the Type of Premise with both Justifications and Conclusions perform much better than models trained with just one or the other. For Type of Justification, these models achieve F1 scores that are between 10 and 20 points higher. For Type of Conclusions, their F1 scores are around 5 points higher. When checking the per-class F1 scores, the improvement in performance is concentrated on the minority classes. For Type of Justification, both Value and Policy classes improve highly, and for Type of Conclusion the most difference is on the Value class.

5.4 Impact of training dataset size

We want to assess how much data is needed for the models to achieve an acceptable performance. For this purpose, we ran several experiments following the same settings as in 5.1 but using smaller portions of the original datasets. Our goal is to measure the impact of having smaller datasets for each component and the relative gain of adding new examples, considering that the task of labeling them is expensive. We used a random sample of 25%, 50% and 75% of the corpora used for training and compare the F1 scores with those obtained by the models trained with the whole corpus.

Figure 5 compares the F1 scores of the best performing models for each component in 5.1 with those obtained by the same models trained with smaller portions of the same datasets. On the left, we show the evolution of the F1 score when increasing the size of the training dataset. On the right, we show the percentage of improvement of the F1 score between each size of the dataset for each com-

	RoBERTa				BERTweet				XLM-RoBERTa-Mix				XLM-RoBERTa-XL			
	Macro	F	V	P	Macro	F	V	P	Macro	F	V	P	Macro	F	V	P
Models trained with both Justifications And Conclusions																
Type of Just	.49±.07	.92	.13	.41	.53±.08	.93	.19	.47	**.55**±.01	.94	.37	.34	.52±.17	.93	.41	.21
Type of Conc	.63±.14	.82	.22	.85	**.70**±.14	.85	.37	.87	.67±.16	.78	.37	.86	.57±.04	.78	.34	.60
Type of both	.66±.05	.90	.28	.79	**.69**±.12	.91	.34	.82	.67±.04	.88	.35	79	.60±.03	.89	.39	.53
Models trained with just one of them																
Type of Just	.41±.09	.95	.28	.00	**.42**±.09	.95	.13	.17	.35±.05	.97	.00	.08	.33±.03	.95	.0	.05
Type of Conc	.58±.05	.81	.07	.85	**.65**±.11	.84	.20	.89	.61±.06	.70	.28	.85	.63±.02	.78	.45	.65

Table 3: Results for identification of Type of premises tested against both Justification and Conclusion, only Justifications and only Conclusions. Results are compared against those achieved by the best performing model trained with only one of the two kinds of premises.

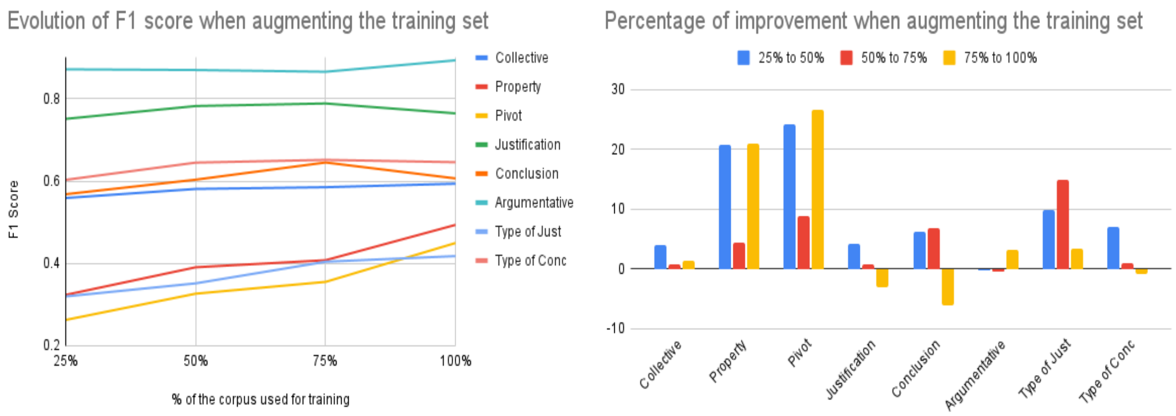

Figure 5: Evolution of F1 scores per argumentative component when increasing the size of the dataset used for training. The figure on the left shows F1 score absolute values, while the one in the right shows the percentage of the score wrt the final value obtained when reducing the dataset.

ponent. For example, the model predicting Pivot trained with 75% of the dataset achieved a score of 0.36 while the model trained with the whole dataset (100%) achieved a score of 0.45, which represents an improvement of 26.6% of this score.

When looking at performance of models trained with smaller fractions of the dataset (figure 5.1) we can see that those components with better scores can achieve similar results using fewer data, while components with worse performance (Property, Pivot and Type of Justification) are much more sensible to the amount of examples on the dataset. This could be considered as an indicator that the size of the dataset is enough for most components but for these last three, if more examples were added to the dataset performance could improve.

6 Discussion of results

We have seen that some argumentative aspects of hate speech in tweets can be successfully identified by Large Language Models (LLMs), namely, whether a tweet is argumentative or not and Justifications, Conclusions and the Type of Conclusions.

This kind of information may be useful to provide an argumentative analysis of tweets, possibly for argument retrieval. It is probably also useful to guide the (semi-)automatic generation of some counter-narratives, like those that are aimed to question the Justification or those aimed to some kinds of Conclusions, like Values or Policies.

Domain-specific argument information, like Collective and Property, are not very successfully identified. Different strategies, like Named Entity Recognition approaches, may yield better results.

Pivots, aiming to identify the relation between Justification and Conclusions, and a key component to reconstruct Wagemans's typology, cannot be successfully identified, either by humans or automatically. It seems that a different approach must be taken to identify them manually, possibly identifying all possible sequences of words that elicit a relation between Justification and Conclusion.

These results will be instrumental for the annotation of a bigger annotated corpus, specially for

Spanish, and to integrate these concepts into LLMs.

7 Summary and Future Work

We have presented an approach to determine which aspects of argumentative information from hate speech in social media is liable to be integrated into LLMs. We have adapted the analytic approach of an informal logic based on (Wagemans, 2016) and have developed annotation guidelines which have then used to enrich a reference dataset for hate speech with argumentative information.

We developed a robust annotation process and guidelines to obtain high agreement between annotators. Indeed, an initial assessment of inter-annotator agreement, shows agreement above $\kappa = .6$ for most categories, except the most interpretative ones. Considering we are dealing with user-generated text, we find this a very hopeful scenario. We are also working on adapting the categories with more disagreement, like Pivot, based on the patterns of the disagreemeent between annotators, so that in further annotation efforts they can be identified in a more reproducible ways, both by humans and automatic methods.

We show to which extent it is possible for Large Language Models to automatically identify the argumentative components, so that this kind of information can be integrated with purely data-driven approaches to enrich the analysis of text and produce more insightful, reasoned outputs.

Finally, the published dataset is also a contribution to the existing corpora of argument mining on social networks. It is publicly available at `https://github.com/ASOHMO/ASOHMO-Dataset`.

For future work, we plan to annotate bigger corpora, focusing on improving reliability on difficult, yet potentially useful, components, like Pivot. We also plan to add counter-narratives associated to each tweet and train models to automatically generate them. We want to assess to which extent the argumentative information helps in better generating automatic responses.

8 Limitations and Ethical Considerations

In the first place, we would like to make it clear that the human annotations presented here are the result of the subjectivity of the annotators. Although they have been instructed through a manual and training sessions, there are still significant variations between interpretations, and further researchers may assign different categories to examples.

Also, it is important to note that the automatic procedures obtained are prone to error, and should not be used blindly, but critically, with attention to possible mistakes and how they may affect users, groups and society.

Then, it is also important to note that the corpus used for this research is very small, specially in the Spanish part, so the results presented in this paper need to be considered indicative. A bigger sample should be obtained and annotated to obtain more statistically significant results.

The findings of this research can potentially inform the development and improvement of language models and chatbot systems. However, we emphasize the importance of responsible use and application of our findings. It is essential to ensure that the identified argumentative components are utilized in a manner that promotes reasoned usage and does not contribute to the spread of hate speech or harmful rhetoric. We encourage researchers and developers to consider the ethical implications and societal impact of incorporating argumentative analysis into their systems.

The data have been adequately anonymized by the original creators of the Hateval corpus.

Studying hate speech involves analyzing and processing content that may be offensive, harmful, or otherwise objectionable. We acknowledge the potential impact of working with such content and have taken steps to ensure the well-being of the research team involved. We have provided comprehensive guidelines and training to our annotators to mitigate any potential emotional distress or harm that may arise from exposure to hate speech. Additionally, we have implemented strict measures to prevent the dissemination or further propagation of hate speech during the research process.

Finally, we have not specifically conducted a study on biases within the corpus, the annotation or the automatic procedures inferred from it, nor on the LLMs that have been applied. We warn researchers using these tools and resources that they may find unchecked biases, and encourage further research in characterizing them.

Acknowledgments

Annotation was done using the brat annotation tool (Stenetorp et al., 2012). This work used computational resources from CCAD – Universidad Nacional de Córdoba (https://ccad.unc.edu.ar/), which are part of SNCAD – MinCyT, Argentina.

References

Khalid Al-Khatib, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein. 2016. A news editorial corpus for mining argumentation strategies. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3433–3443, Osaka, Japan. The COLING 2016 Organizing Committee.

Katie Atkinson and Trevor Bench-Capon. 2018. Taking account of the actions of others in value-based reasoning. *Artificial Intelligence*, 254:1–20.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. pages 54–63.

Muhammad Mahad Afzal Bhatti, Ahsan Suheer Ahmad, and Joonsuk Park. 2021. Argument mining on Twitter: A case study on the planned parenthood debate. In *Proceedings of the 8th Workshop on Argument Mining*, pages 1–11, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tom Bosc, Elena Cabrio, and Serena Villata. 2016. DART: a dataset of arguments and their relations on Twitter. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1258–1263, Portorož, Slovenia. European Language Resources Association (ELRA).

Danielle Keats Citron and Helen Norton. 2011. Intermediaries and hate speech: Fostering digital citizenship for our information age. *BUL Rev.*, 91:1435.

J. Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Phan Minh Dung. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77:321–357.

Mihai Dusmanu, Elena Cabrio, and Serena Villata. 2017. Argument mining on Twitter: Arguments, facts and sources. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2317–2322, Copenhagen, Denmark. Association for Computational Linguistics.

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

Nadin Kökciyan, Isabel Sassoon, Peter Young, Martin Chapman, Talya Porat, Mark Ashworth, Vasa Curcin, Sanjay Modgil, Simon Parsons, and Elizabeth Sklar. 2018. Towards an argumentation system for supporting patients in self-managing their chronic conditions.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Fabrizio Macagno, Douglas Walton, and Chris Reed. 2018. *Argumentation Schemes*, pages 517–574.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

Marco Polignano, Pierpaolo Basile, Marco De Gemmis, Giovanni Semeraro, and Valerio Basile. 2019. Alberto: Italian bert language understanding model for nlp challenging tasks based on tweets. In *6th Italian Conference on Computational Linguistics, CLiC-it 2019*, volume 2481, pages 1–6. CEUR.

Iyad Rahwan and Guillermo R Simari. 2009. *Argumentation in artificial intelligence*, volume 47. Springer.

Robin Schaefer and Manfred Stede. 2020. Annotation and detection of arguments in tweets. In *Proceedings of the 7th Workshop on Argument Mining*, pages 53–58, Online. Association for Computational Linguistics.

Robin Schaefer and Manfred Stede. 2022. German climate change tweet corpus (gercct).

& Wagemans Schut, D. 2014. Argumentatie en debat.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56, Doha, Qatar. Association for Computational Linguistics.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France. Association for Computational Linguistics.

Milagro Teruel, Cristian Cardellino, Fernando Cardellino, Laura Alonso Alemany, and Serena Villata. 2018. Increasing argument annotation reproducibility by using inter-annotator agreement to improve guidelines.

Simone Teufel, Jean Carletta, and Marc Moens. 1999. An annotation scheme for discourse-level argumentation in research articles. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 110–117, Bergen, Norway. Association for Computational Linguistics.

Stephen E. Toulmin. 2003. *The Uses of Argument*, 2 edition. Cambridge University Press.

Jacky Visser, John Lawrence, Chris Reed, Jean Wagemans, and Douglas Walton. 2021. Annotating argument schemes. *Argumentation*, 35.

Jean Wagemans. 2019. Four basic argument forms. *Research in Language*, 17:57–69.

Jean H. M. Wagemans. 2016. Constructing a periodic table of arguments.

Douglas Walton, Christopher Reed, and Fabrizio Macagno. 2008. *Argumentation Schemes*. Cambridge University Press.

APPENDIX

A Annotation team and environment

Two annotators (a philosopher and a computer scientist) have been trained with the guidelines described in section 3, with a three stage training process, where they labeled a first set of examples, discussed their difficulties, systematized further hints and criteria, updated the annotation manual and started again. We prioritized having the lesser amount of annotators doing the most possible amount of annotations. Our hypothesis is that the more annotators, the more difficult it is to reach a uniform criterion that can be understood in the same way by everyone. So fewer annotators doing more work should lead to more reliable annotations and to better inter-annotator agreement.

The average time for annotators to label a tweet is approximately 4 minutes per example. The annotation time changes depending on whether the tweet is argumentative or not. For argumentative tweets, the average time is around 5 minutes, while for non-argumentative tweets the average time is less than 1 minute.

The first annotator annotated 800 tweets in English and 196 in Spanish, while the second annotated 170 tweets in English.

B Corpus statistics

Table 4 show the percentage of tweets that are labeled as non-argumentative in English and in Spanish, and also the percentage of tweets in each language that have a pair of Collective and Property and a Pivot labeled. Considering only the non-targeted and non-aggressive hate tweets against immigrants from HatEval, the majority of tweets are labelled as Argumentative in both languages. Regarding the Collective-Property pair and the pivot, the table shows the percentage of the final dataset that have them labeled. Table 5 shows the percentage of Justifications and Conclusions that are labeled as **F**act, **P**olicy or **V**alue. Justifications have an ample majority of examples labeled as Fact, while the distribution between classes is more even when observing conclusions. In both cases, the "Value" class is the least frequent.

C Preprocessing

Preprocessing is very important when dealing with tweets, since they tend to have lots of non-alphanumeric characters, user handles (@user-

	Non-Arg	Collective	Pivot
English	25.3%	58.2%	45.1%
Spanish	26.5%	61.1%	37.5%

Table 4: Percentage of tweets labeled as Non-Argumentative and with Collective-Property and Pivot labeled

	Justification			Conclusion		
	F	P	V	F	P	V
English	93%	4%	3%	37%	57%	6%
Spanish	97%	2%	1%	56%	28%	16%

Table 5: Percentage of Justifications and Conclusions labeled as **F**act, **P**olicy or **V**alue

name), hashtags, emojis, misspellings, and other non-canonical text. Following (Nguyen et al., 2020) and (Polignano et al., 2019) we used a soft normalization strategy consisting of:

- Character repetitions are limited to a max of three
- User handles are converted to a special token `@usuario`
- Hashtags are replaced by a special token `hashtag` followed by the hashtag text and split into words if this is possible
- Emojis are replaced by their text representation using *emoji* library[6], surrounded by a special token `emoji`.

D Experiment settings

For all monolingual experiments we used 770 tweets of the English portion of the dataset as training (79%), 100 tweets as development (10.5%) and 100 tweets as test (10.5%). Multilingual experiments were twofold: using both English and Spanish for both training and testing, and using English for training and development and Spanish for test. In the first case, we used 770 English and 120 Spanish tweets as training (76.3% of the dataset), 100 English and 26 Spanish tweets as development (10.8%) and 100 English and 50 Spanish tweets as test (12.9%). In the second case, we used 850 English tweets as training (73% of the total), 120 English tweets as development (10%) and all the 196 Spanish tweets for testing (17%).

In all cases, we tried 5 different values for learning rate (1e-05, 2e-05, 5e-05, 5e-04 and 5e-06) and used the development dataset to implement early stopping with a maximum of 10 epochs. Table 6 shows the values for the hyperparameters used on all models trained with our examples.

Batch Size	16
Optimizer	AdamW
Dropout	0.1
Epochs	10
Weight Decay	0.01
Adam ϵ	1e-06
Adam $\beta 1$	0.9
Adam $\beta 2$	0.99

Table 6: Hyperparameters used for training all models used on our experiments

E Examples And Decisions From The Annotation Process

In the following section, we show examples of labeled tweets that illustrate particular decisions taken when defining the annotation protocol. Example 6 shows a frequent case of a non-aggressive, non-targeted and non-argumentative tweet, consisting on the expression of one or many stances or exhortation to one or many actions but without any explicit intention of connecting them.

Example 7 shows an example of a premise where a user states her opinion as a verifiable fact. Although it could be arguable that she is expressing a Value about a subject (immigration or assimilation), we consider all tweets that could be fact checked (specially if the user doesn't use explicit markers of her involvement in the statement) to be of type Fact. Example 8 shows an annotation of a Collective-Property pair. The Property is any negative concept, adjective, consequence or aspect of reality that is explicitly or implicitly associated with the target of the hate message. In this case, the tweet is stating that immigrants are not wanted by the people. The cases where there is no explicit association between Collective and Property are diverse, but we present three examples that we believe represent the majority of the cases. Example 9 shows a case where instead of defining a negative property associated with the targeted collective, the user defines a positive Property associated with the absence of that Collective. Example 10 shows a case where a negative Property is associated with the targeted Collective but in an indirect manner that must be reconstructed using contextual information not included in the hate message. In this case, the reference is made through the mention of "Operation SOAR", an operation made by the

[6] https://github.com/carpedm20/emoji/

ICE in the United States specifically targeting immigrants registered as sex offenders and by the hashtag "StopTheInvasion" referring to a narrative built against immigrants as if there were a coordinated plan to invade a country. Example 11 shows a case where the main standpoint of the tweet is an action that must be taken and there is no explicit mention of any Property. In these cases, the Property could potentially be reconstructed by appealing to find the motivation of these advertised actions, but it can not be labeled explicitly on the text of the tweet.

Example 12 shows a Justification labeled as Fact and a Conclusion labeled as Policy. The main standpoint of the message must be first identified as Conclusion, and then any part of the tweet that fulfills the role of providing reasons for that standpoint is identified as Justification. One typical pattern frequent in many tweets is to express a mandate or policy that must be followed, usually with the form of a phrase or hashtag using the imperative mode, and a Fact (or less frequently also another mandate) that supports and aims to explain why that mandate must be followed.

Example 13 show a tweet with a Pivot. In this case, the user binds the "money" as a cause of immigration to conclude that "money" is not needed. All tweets present some aspect that links Justification with Conclusions, but not always that relation is mentioned directly. Example 14 shows a tweet where no Pivot was labeled. The link between the premises relies on the implicit assumptions that the hate that they supposedly bring to the EU is against Christians and that because of that hate, Christians are not safe. But to recognize it, the relationship must be reconstructed using implicit information defined by the context, otherwise it is impossible to establish. In these cases, we do not label the Pivot for the sake of simplicity. We require that the relation between the two phrases constituting a Pivot is direct and easy to spot.

```
No to #EU migrant camps in Libya,
PM al-Serraj
```

Figure 6: Example of non-argumentative tweet

`@user Time to leave the uk commonwealth and Europe that would end immigration` **`people do not want more`** <u>`refugees`</u> `enough is enough`

Figure 8: Example of Collective and Property labeled. Collective is underlined while Property is bolded

```
Good this makes it a safe country
immigrants can now go home
```

Figure 9: Example of tweet without Collective and Property labeled. In this case, Property is associated with the absence of immigrants, therefore it is indirectly defined and not mentioned explicitly

```
Anyone who, ACTIVELY OR
PASSIVELY, subscribes to
immigration and especially
assimilation is joining the
battle to destroy White
```

Figure 7: Premise of a tweet labeled as “fact"

F Disagreement between annotators

In the following section, we analyze examples of disagreement between annotators to better understand the aspects that are most difficult to systematize about annotating argumentative components. Example 15 show a disagreement concerning Collective and Property. Here, one annotator didn't consider that there was a Collective and Property to label, while the other did. We found that most disagreements regarding these components are of this kind. If both annotators agree that the tweet has a Collective and Property to label, in most cases they agree also what parts of the text constitutes them. In the few cases where both annotators labeled a Collective and a Property, but they did not match exactly, they had a major overlap and only differed on adding a few more words at the beginning or at the end. Example 16 shows a disagreement of

```
ICE officers arrest 32 sex
offenders on Long Island as
part of 'Operation SOAR' :link:
#StopTheInvasion #SecureTheBorde
```

Figure 10: Example of tweet without Collective and Property labeled. In this case, the collective is not explicitly mentioned but referred through contextual information

```
Canada is an immigrant country
Don't change it to refugee
country please
```

Figure 11: Example of tweet without Collective and Property labeled. In this case, the focus of the message is put into an action that must be taken and not on associating the Collective with a Property

```
Victims of Illegal Alien Crime
describe heartbreak, frustration
#BuildTheWall #ProtectAmerica
#EndChainMigration
#EndIllegalBirthrightCitizenship
#NeverForget the American Victims
of Illegal Alien Migration
```

Figure 12: Example of labeling of Justification (in blue) and Conclusion (in red). Justification is labeled as Fact while Conclusion is labeled as Policy

Why do foreign individual dump
money (and refugees) into our
country? We don't need their
money and their programs.

Figure 13: Example of a tweet with a labeled Pivot. Justification is shown in blue while Conclusion is in red. Labeled Pivot is shown bolded

```
Nice tweet , Joyce, Truth is
they flee Iran etc but want to
bring their hate to the Eu even
in refugee camps Christians not
safe.
```

Figure 14: Example of a tweet without a Pivot labeled. The Justification is shown in Blue while the Conclusion is in Red. The link between the two premises relies on the relation between "hate" and "not safe"

@user @user The idea is to bring
in the "dreamers" **so that they
vote for Democrats** because
Dems know they have to import
their voters. That is literally
the only reason the Democrats
care about this issue. In the
meantime, YES THEIR PARENTS

Figure 15: Example of disagreement concerning the Collective and Property. One annotator did not label any of them. Collective labeled by the other annotator is underlined while Property is bolded

Mexico's not sending their
best. **They are dumping their
killers aka garbage on us.**
#StopTheInvasion #DeportThemAll
#NoAmnesty #BuildTheWall

Figure 16: Disagreement concerning the Property. Collective labeled by both annotators is shown in red. Property labeled by one annotator is bolded while the one labeled by the other annotator is underlined

such kind. Example 17 shows how both annotators agree on how to split the text but disagree on which part is the Justification and which is the Conclusion. To improve the annotation process, the guidelines should emphasize that the main standpoint of the tweet should be identified before labeling the Justification. Example 2 shows disagreement about labeling the pivot. In this case, each annotator found a different Pivot that could be considered correct. The annotation guidelines enforce each annotator to label only one Pivot but there are examples, like the one mentioned above, where multiple Pivots could be found. This indicates that there could be an opportunity of improving the system if we enforce annotators to label all possible Pivots. Example 18 shows a disagreement on the Type of a Justification. The premise has declarative sentences with informative content (like "It is the third anniversary of her death") mixed with mandates or actions that must be followed ("Remember Kate Steinle today" and "We must not forget"). Depending on the part of the sentence

G Analysis of differences between automatic classifications and ground truth

We analyze the errors made by automatic classifiers when recognizing argumentative components,

@user @user you come with the usual lies an insults. Fact is that mass immigration into Ireland has been going on for decades, most illegal and from other EU countries, still trans-formative. All the people seeking asylum

@user @user you come with the usual lies an insults. **Fact is that mass immigration into Ireland has been going on for decades, most illegal and from other EU countries, still trans-formative. All the people seeking asylum**

Figure 17: Disageement between annotators concerning Justification and Conclusion. Justification is bolded while Conclusion is underlined. While both annotators split the argument in the same fashion, they disagree on which part is the justification and which is the Conclusion

Remember Kate Steinle today. It is the third anniversary of her death We must not forget. #KateSteinle#IllegalAliens #OpenBorders#BuildThatWall #MondayMorning#ImmigrationReform #ImmigrationIsAWeapon

Figure 18: Example of disagreement concerning type of premise. Justification (bolded) was labeled as Fact by one annotator and as Policy by another

trying to determine possibilities of improvement either in the annotation process or in the settings of the task of automatic recognition.

Example 19 show an example of a non-argumentative tweet that was classified as argumentative by the automatic predictor trained as described in 5.1. The tweet has several hashtags calling for actions, but there is no explicit intention of using any of them as a justification of the others. The tweet refers to a mother who supposedly needs prayers, indicating that the author is aware of a context that is missing for us.

Example 20 shows a prediction done by a model trained following the settings described in 5.1. Here, the model correctly identifies a Collective mentioned in a xenophobe tweet, but there is no explicit Property assigned to them and because of this, it shouldn't have been labeled. Though this model was sometimes able to distinguish when the Collective should have been labeled or not, we found this error to be very frequent in experiments done with these settings. This led us to propose the experiment described in **??** separating the problem in two: first identifying if there is a pair of Collective and Property to label and then finding them on the tweet. When scoping the problem to find a Collective in a tweet that we know it is present, most errors produced by the automatic classifiers are discrepancies on the amount of words used to refer to the collective (like in example 22) or whenever the tweet mentions multiple collectives besides the target of the hate message (like example 21). We think that the first case reveals an opportunity for improvement on the annotation process, where sometimes a collective might have been labeled using one word and other times using many.

Example 23 show an incorrect prediction on the Property done by a model trained following the experiment described in 5.1. Although human trafficking could be considered as a negative consequence, the tweet does not explicitly associate it to a particular Collective. These models tend to identify phrases with negative connotations as Properties, disregarding if they are associated with the target group. This problem arises independently of the presence of a real Property and usually all words or phrases that could be considered as "negatives" are labeled by automatic predictors. Another error that automatic models are prone to are labeling bigger or smaller portions of text. Example 24 shows a prediction made by a model trained as described in section ??. The model correctly identified "illegally invade the U.S." as part of the Property, but missed the rest.

Regarding Pivots, we found that a common problem derivates from the incapacity of the models to jointly learn to find the pivot and the separation of the tweet into premises. Example 4 shows predictions made by a model trained following the settings described in 5.1 that found two words in different parts of the tweet that are directly related, but that are both within the justification, so they are not really a pivot between premises. A new setting for experimentation could provide the model with the information of where are the Justification and the Conclusion, and enforce to predict exactly one phrase within each of them. Another error found when predicting pivots comes from where multiple valid Pivots can be found within the premises. Example 27 shows prediction of a model also trained as described in 5.1 that found two valid Pivots: "Salvini-Salvini" and "invade-invasion". Each one of them could be considered a valid Pivot, though the only one that was labeled by the human annotator was "Salvini-Salvini". This phenomenon is related and could be considered as a consequence of the disagreement between annotators shown in the example 2. In order to avoid this kind of error, annotators should be instructed to label all the possible Pivots if there were more than one.

For Premises and Conclusions, we found also several cases where the model correctly divided the tweet in two premises but failed to assign the kind of the premise: if it was a Justification or a Conclusion. Example 34 shows a prediction done by a model trained to jointly predict both Justification and Conclusion at the same time, as explained in 5.2. Here, the model correctly identifies both parts of the argument but fails to correctly assign the Justification and Conclusion in itself. It is interesting to note that models predicting a single component as described in 5.1 do the same mistake when predicting Justification and Conclusion for this same example. This correlates with similar discrepancies between annotators shown in example 17.

For the Types of premises, models trained following the settings described in 5.1 usually fail to predict the minority classes ('Value' for Conclusions and 'Value' and 'Policy' on Justifications). On the contrary, performance on these classes improves when models are trained following the settings described in 5.3. We found that using both kind of premises for training instead of just one no only increases the amount of examples but also leverages the distribution among classes, which leds to a significant boost in performance, as shown in table 3. Example 35 shows a Justification predicted as Policy by a model trained using only justifications and then correctly predicted as Value by a model trained using both Justifications and Conclusions.

`Video: (part 1) London #BNP a frame trailer with patriotic sound system on the road in and around our capital city "say no to` **`immigration"`** `#Brexit #Immigration #ImmigrationBan #London #England #BrexitBorder #Brexiteer #Brexiteers #BrexitGoodNews #BrexitChaos`

Figure 20: Example of prediction of Collective from experiment described in 5.1. Though the model finds a mention of a Collective that seems to be accurate, there is no explicit Property associated so it shoudn't have been labeled

`At this time, w-organized crime/returning` **`jihadists`** `it's a matter of national security. #Italy #Salvini must ignore international social engineers/cultural` **`marxists`** `#V4 Itali Kurz others must challenge empty threats from un-eu` **`migration`** `pimps. What can they really do about it?`

Figure 21: Model predicting only on tweets that have a Collective, besides correctly finding 'immigration', also labeled 'jihadists' and 'marxists', which are being used as properties for either the target collective or other groups (like 'international social engineers')

`#Prayers for this mother #NoIllegals #SendThemAllBack w/ their families #NoDACA #BuildTheWallNow`

Figure 19: Example of Non-Argumentative tweet incorrectly labeled as argumentative by automatic model. The tweet refers to a context that is missing on the text

Chain migration imported 120K foreign nationals from terrorist-funding countries since 2005 - breitbart @user @user #EndChainMigration #EndDACA #NoAmesty #EndBirthrightCitizenshipForIllegalAliens #BuildTheWall #KeepAmericaSafe

Figure 22: Example of prediction of Collective from experiment described in 5.1. The prediction seems to be accurate, but it included the word "chain" associated with migration. Differences like this arise whenever there are frequent phrases like "Chain Migration" or "Illegal immigrants"

@user the disgrace is the illegal parent who brought their kids on their cirme spree to **illegally invade the U.S.** so taxpayers pay for their kids education wic and medicaid. We don't owe illegals our tax dollars #SendThemBack #WalkAway #Trump #MAGA #RedNationRising

Figure 24: Example of prediction of Property from experiment described in ??. Real Property is undelined while prediction is bolded. The model predicted just a portion of the real Property and left most of it unlabeled

Please dont call it ""rescue"" - it's **human trafficking** #PortsClosed #SendThemBack #BenefitSeekers

Figure 23: Example of prediction of Property. Predicted Property is bolded. There was no real property labeled in this example.

Americans agree with @user on immigration. We can not afford to give welfare to illegals while U.S. citizens are homeless **#VoteDemsOut #FamiliesBelongTogheterMarch**

Figure 25: Example of prediction of Conclusion. Real conclusion is underlined while predicted is bolded. Here, the two parts of the argument were correctly identified but predictor chose the conclusion incorrectly

Americans agree with @user on immigration. **We can not afford to give welfare to illegals while U.S. citizens are homeless** #VoteDemsOut #FamiliesBelongTogheterMarch

Figure 26: Example of prediction of Justification. Real justification is underlined, while predicted is bolded. Again, the two parts of the argument were correctly identified but predictor chose the incorrect half

Pressure on **Spain's** maritime border: Boatloads of #Illegal #Migrants Storm **Spanish** Tourist Beaches & Scatter #StopTheInvasion #Unregistered #UnVetted

Figure 27: Example of pivot predicted by model trained as described in section 5.1. Justification is in blue, while conclusion is in red. Although the words selected establish a relation between themselves, they are both part of the justification, so they are not really a pivot between both premises

Rich African Countries don't take in African Migrants. Rich muslim countries don't take in muslim migrants. Rich latin american countries don't take it latin migrants. **But white countries are supposed to acept them??**

Figure 28: The conclusion (bolded) was predicted as Fact though it is a Policy

Angry that UN @user does its job and checks Lebanon isn't coercing Syrian refugees into returning home, **Lebanon will stop giving residence permits to the agencys international staff**

Figure 29: This conclusion was predicted as Policy though it is a Fact

@user Amen: **See 'Canada in Decay' by Ricardo Duchesne for the similar reality of Canada.** We are not nations of immigrants.

Figure 30: The justification (bolded) was predicted as Fact though it is a Policy

`Good news. We are against illegal immigrants`

Figure 31: The justification (bolded) was predicted as Fact though it is a Value

`@user` <u>`Immigration in a picture`</u> `:link:` **`Some basic truths: Access to White people is not a human right.`**

Figure 32: Example of prediction of Justification and Conclusion. Predicted Conclusion is shown in blue while the real one is bolded. Predicted Justification is shown in red while the real one is underlined. Models were able to correctly divide the tweet in two premises but failed to correctly recognize Justification and Conclusion

`@user` <u>`Immigration in a picture`</u> `:link:` **`Some basic truths: Access to White people is not a human right.`**

Figure 33: Example of prediction of Conclusion. Predicted Conclusion is underlined while the real one is bolded.

`@user` **`Immigration in a picture`** `:link:` <u>`Some basic truths: Access to White people is not a human right.`</u>

Figure 34: Example of prediction of Justification. Predicted Conclusion is underlined while the real one is bolded.

`I do not want those vile thugs in our country`

Figure 35: Justification labeled as Value by human annotator. This premise was predicted as Policy by a model trained following the settings described in 5.1 and was correctly identified as Value by a model trained as described in 5.3

H Argument annotated social media corpora

There exist several datasets with argument annotations, but only a few of them annotate arguments on Twitter. DART relies on Argumentation Theory (Rahwan and Simari, 2009) finding relationships between tweets as a single unit, considered to be arguments within an Abstract Argumentation Framework (Dung, 1995). Tweets are considered as argumentative if they express opinion or claims showing stance about a particular topic, and then they are defined according to how they interact with other tweet-arguments. The work of Dusmanu et al. (2017) extends the *#Grexit* subset of DART (987 tweets) with another 900 labeled for argument detection and adds labels for factual arguments recognition and source identification. However, abstract frameworks do not consider the inner structure of arguments and are not useful in providing an argumentative analysis in the context of a single tweet.

Schaefer and Stede (2020) labeled 300 replies to context tweets about Climate Change in German language with claims and evidence to support the claims. This was later expanded to 1200 tweets and the annotation scheme was refined to focus on particular argument properties (Schaefer and Stede, 2022). This is the only work, to our knowledge, where spans are annotated within a tweet, but it is not a hate dataset and does not have domain specific information.

Finally, Bhatti et al. (2021) created a dataset of 24100 tweets searching two hashtags supporting and attacking Planned Parenthood. The whole tweet is assigned a single label (i.e., support or not the claim) and there is no argumentative structure segmentation within, so it is impossible to differentiate aspects of argumentative information.