Annotating Topics, Stance, Argumentativeness and Claims in Dutch Social Media Comments: A Pilot Study

Nina Bauwelinck and Els Lefever
LT3, Language and Translation Technology Team
Department of Translation, Interpreting and Communication – Ghent University
Groot-Brittanniëlaan 45, 9000 Ghent, Belgium
firstname.lastname@ugent.be

Abstract

One of the major challenges currently facing the field of argumentation mining is the lack of consensus on how to analyse argumentative user-generated texts such as online comments. The theoretical motivations underlying the annotation guidelines used to generate labelled corpora rarely include motivation for the use of a particular theoretical basis. This pilot study reports on the annotation of a corpus of 100 Dutch user comments made in response to politically-themed news articles on Facebook. The annotation covers topic and aspect labelling, stance labelling, argumentativeness detection and claim identification. Our IAA study reports substantial agreement scores for argumentativeness detection (0.76 Fleiss' kappa) and moderate agreement for claim labelling (0.45 Fleiss' kappa). We provide a clear justification of the theories and definitions underlying the design of our guidelines. Our analysis of the annotations signal the importance of adjusting our guidelines to include allowances for missing context information and defining the concept of argumentativeness in connection with stance. Our annotated corpus and associated guidelines are made publicly available.

1 Introduction

User-generated content (UGC) such as can be found in the comment sections of newspapers and social media sites is a valuable resource for the collection of argumentative texts written in natural language. According to Manosevitch and Walker (2009), user comments offer a "substantial amount of factual information, and [demonstrate] a public process of weighing alternatives via the expression of issue positions and supporting rationales". The field of argumentation mining, which forms a part of Natural Language Processing (NLP) research, uses this type of data as a resource to train and test automatic detection systems for the purpose of extracting the various components making up the argumentation expressed by the users (Park and Cardie, 2014; Villalba and Saint-Dizier, 2012). Training the systems requires annotating the data, for example labelling claims and reasons for those claims in the text. The various annotation tasks required for producing such data have proven to be very difficult for human annotators. Defining a good set of annotation guidelines is essential towards advancing the field of argumentation mining on UGC data such as social media comments. Currently, the myriad of theoretical perspectives on how to analyse argumentation as well as the unpredictable nature of UGC data have lead to a lack of current consensus on reliable guidelines for the various argumentation annotation tasks.

This paper presents a pilot annotation study for the identification of the topics, topic aspects and stance expressed by the comments, as well as the detection of argumentativeness and the main claim or conclusion presented. This study assesses the suitability of our current guidelines by measuring the Inter Annotator Agreement (IAA) for all tasks and analyses some specific cases which proved most challenging to our annotators. Our aim is to adjust the guidelines based on these results and analysis (Bauwelinck and Lefever, 2020), which will then serve as the basis for an extensive annotation study on a more substantial corpus and including more annotation tasks required for a full analysis of the

Proceedings of the 7th Workshop on Argument Mining, pages 8–18
Barcelona, Spain (Online), December 13, 2020.

argumentation presented in the comments (a.o., this will include premise and argumentative relation annotation).

In Section 2, we briefly discuss some of the relevant research. In Section 3, we give an overview of the theoretical frameworks which form the basis of our annotation scheme. In Section 4, we describe our pilot corpus and in Section 5 we give an overview of the annotation procedure, as well as more information on the rationale underlying our guidelines. In Section 6, we first present the results of the IAA study. In Section 7, we then present our analysis of the annotations. We end with Section 8 on concluding remarks as well as indications for future research.

2 Related Research

In the field of argumentation mining, many different problems related to the analysis of argumentation in texts are being treated as different subtasks for automatic detection. Many of these tasks relate more generally to the processing of various aspects of texts within the broader field of NLP. For example, as a preliminary step towards the more argumentation-specific detection tasks, the tasks of topic and stance detection are often performed. Users, especially when arguing on controversial topics, tend to emphasize specific aspects of the topic, a concept called "framing" (Entman, 1993). Therefore, both the more general topics and the more fine-grained topic aspects need to be identified. A major challenge still lies in determining how the different aspects relate to each other and to the major topic under discussion (Saint-Dizier, 2016). This challenge relates to the issue of determining how fine-grained the targets of the users' stance needs to be. There is still no consensus on this issue, but the findings have confirmed that the more fine-grained the stance target, the more difficult it is to automatically classify the stance, and stance targets that are defined too broadly may not be specific enough to become associated with each respective side (*pro/con*) of the debate target (Wojatzki and Zesch, 2016). Most authors therefore opt for a predetermined list of topics, in which topics either take the form of single words (more coarse-grained approach) or phrases (fine-grained). The latter approach was used by Saint-Dizier (2016), who defined controversial issues as evaluative statements (e.g., "Climate action is necessary") as this would aid in mining pro/con arguments for these specific controversies. As a preliminary step for the annotation of different argument components (Stab and Gurevych, 2014; Peldszus and Stede, 2015), the text needs to be split into smaller segments. This step is often skipped in argumentation mining research, in favour of departing from pre-segmented text (Ajjour et al., 2017). The segmentation of user-generated comments is not straightforward, as they contain many irregularities in the use of punctuation, capitalization and other orthographic markers of sentence boundaries. Determining the boundaries of argumentative segments is challenging but necessary for argumentation mining, as it forms the preliminary step both towards identifying the claim and premises as well as more fine-grained components of the argument. Textual indicators of argumentativeness such as discourse markers are often used as features for the automatic detection of argumentative segments (Eckle-Kohler et al., 2015). An important caveat of using argumentative words or phrases for the task of argumentativeness detection is that they may also occur in non-argumentative text (van Eemeren and Houtlosser, 2006). Argumentation mining research has produced some work on the linguistic cues which may be helpful for the identification of argumentativeness (Nguyen and Litman, 2015), especially the detection of discourse markers has been explored to serve as signallers of argumentativeness to aid the automatic detection of arguments (Somasundaran et al., 2008; Tseronis, 2011; Eckle-Kohler et al., 2015). The detection of segments representing the claim or conclusion of an argument in texts is an important prerequisite for applications involving fact checking (Vlachos and Riedel, 2014). It is a very challenging task, especially when applied to an extremely varied resource, like online discourse (Habernal and Gurevych, 2017). There is no current consensus on what exactly constitutes a claim, leading to many different annotation approaches (Daxenberger et al., 2017).

3 Theoretical Frameworks

For the annotation of topics in comments, it is difficult to predetermine a list of possible topics. The comments in our corpus were made in response to newspaper articles shared via Facebook, so we can

safely assume they will often refer to the newspaper content (Manosevitch and Walker, 2009). In order to capture all the available topic information, the distinction between structuring and interactional topics is useful. The *structuring topics* are those found in the surrounding context; the *interactional topics* are those which form the topics of discussion and are found in the immediate context of interaction (Stromer-Galley and Martinson, 2009). This distinction has been applied by Rowe (2015), who used the two categories for the topic annotation of user comments to newspaper articles shared via Facebook and comments made on the same articles, via the official newspaper website. In their study, the structuring topic is that which is reported on in the article and the interactional topics are present in the individual comments. This approach seems feasible for our corpus, since the data type is so similar.

For some interactional topics and aspects in the comment, the user expresses a stance reflecting his personal opinion towards the specific topic or aspect. Not all topics and aspects touched upon in the comment will be the target of the users' opinion, as some will only be mentioned in passing, or to set up the context for the argumentation. The typical stance annotation labels include *pro*, *contra* and *none*, where the last one represents a topic or aspect which is not used as a target for the opinion of the user (Küçük and Can, 2020). Parallel to the distinction between the broader (discussion-wide) structuring topic label and the narrower interactional topic label (based on the comment text itself), the distinction between a debate stance and explicit stance is made by Wojatzki and Zesch (2016). The debate stance is the stance towards the target of the whole debate. It is often implicit and inferrable from the explicit stance(s) which rely on textual evidence. This distinction has proven successful for stance annotation on noisy social media data and has even helped model implicit argumentation (Wojatzki and Zesch, 2016).

Given the difficulty of the automatic segmentation of argumentative texts (Al-Khatib et al., 2016) and our focus on the segment-level annotation of argumentativeness and claims, we also perform a preliminary manual segmentation step before handing the data to the annotators to avoid the error percolation which an added segmentation annotation task would inevitably introduce.

While discourse markers are considered useful indicators of argumentativeness (Eckle-Kohler et al., 2015), Tseronis (2011) proposed the more specific concept of the *argumentative marker* which is a lexical item signalling the presence of an argumentative move (e.g., marking a standpoint). The argumentative marker may consist of a single word, phrase or sentence. A distinction is made between those markers which are syntactically part of the main constituents of the phrase and those that are independent on a semantic and syntactic level (Tseronis, 2011). The concept of *shell language* developed by Madnani et al. (2012) is similar to the *argumentative markers*, in the sense that they may be used as signallers of argumentativeness and may consist of longer sequences of words than is the case for discourse markers. Shell language includes organizational phrases, such as ones marking the expression of an opinion (e.g., "I think that"), but also ones marking argumentative structures (such as "after all"). Du et al. (2014) have used shell language for the task of automatically separating topical contents from organizational language and have demonstrated the usefulness of this task, for instance to improve topic detection. They evaluated their fully unsupervised Shell Topic Model on argumentative UGC sourced from online forums. Ducrot (1982) contends that argumentativeness is present in an utterance even if it does not contain a linguistic expression it may explicitly be linked to. The first major challenge lies in defining what exactly constitutes argumentativeness. The distinction between the argumentative and informative components of utterances (Anscombre, 1995) provides one possible answer. The informative component corresponds to the propositional content of the sentence. The argumentative component signals the utterance's *argumentative orientation*: whether or not it has the potential of being used as a premise for a given conclusion. This perspective helps to circumvent the difficulty of reconstructing the intentions of the author of an argumentative text, as it emphasizes the text itself as the locus of the author's intention.

Biran and Rambow (2011)'s concept of the claim consists of any utterance conveying subjective information and anticipating the question "why are you telling me that?". Daxenberger et al. (2017) have criticized Biran and Rambow's (2011) annotated dataset of online comments from blog threads for containing noisy sentences annotated with claims. However, we hypothesize this is not necessarily caused by the definition of the claim concept. Instead, it is a characteristic of this type of data. Our annotation guidelines employ a very similar definition of the claim concept as the one used in Biran and

Rambow (2011).

The current pilot study represents the first phase of our aim to define the annotation guidelines on a corpus of 100 Dutch user-generated comments sourced from the social media platform Facebook. All comments were made in response to an online newspaper article being shared via the official Facebook page of a Flemish newspaper. In this first phase, we focus on measuring agreement and finding edge cases for the annotation tasks of topic and stance labelling, segmentation of the text into argumentative units and claim identification. The second phase will consist of an agreement study on the same corpus sample of 100 comments and will focus on the tasks of premise and argumentative relation identification.

4 Corpus

Facebook is by far the most popular platform for accessing news (Rowe, 2015). However, given the difficulty in collecting this data automatically and assumptions about the lack of argumentation in shorter texts (Habernal and Gurevych, 2017), this platform is rarely used to source data for argumentation mining purposes. Our corpus of 100 Dutch comments was sourced from the official Facebook page of Het Laatste Nieuws (HLN), a popular Flemish newspaper. As is the case for many news outlets (Rowe, 2015), the Facebook page is used to share news articles with a wider audience, often linking to the official website of the newspaper. The corpus was collected manually and contains comments made on posts published in the second and third weeks of June 2020. We did not predetermine a list of topics to filter the Facebook posts. Instead, we chose to collect the first ten comments of each most recent post on the page which was topically related to a political or controversial topic. This included news stories related to policy decisions, party politics, but also related to topics like health care, poverty and racism. We filtered out duplicate comments and comments in other languages than Dutch, but did not filter out multiple comments made by the same user. Aside from the 100 comments, we also collected the 30 associated article texts and gathered 30 screenshots of the Facebook posts commented upon.

5 Annotation

Six annotators participated in this study. All of them were Dutch native speakers and linguists who received the same annotation guidelines (Bauwelinck and Lefever, 2020) to follow[1]. They each annotated a total of 100 Dutch user comments, all 100 were the same set for each annotator and were used to measure inter-annotator agreement.

We performed two preliminary steps to prepare the corpus for the annotators. First, to prepare for the topic annotations we annotated the structuring topics and aspects contained in the Facebook posts and the news article texts. For the topic annotations of the Facebook posts, we labelled the topic information as present on screenshots of the post in question, thus including information like the title of the article, the accompanying image and description text in our decision making. For the topic annotations of the article texts, we limited ourselves to the title and the first three paragraphs of each article. The second preparatory step was to manually segment all the comments in order to prepare for the argumentativeness and claim annotation tasks, which were to be performed on a segment-level. We segmented the comments based on the shell language expressions we found (Madnani et al., 2012; Du et al., 2014), but only if they were also set apart from other segments in a syntactical way (Tseronis, 2011). If the expression occurred syntactically embedded in the phrase, we considered it a part of the larger segment. This approach allowed us to distinguish organizational segments from content segments. We did not consider markers of opinion (such as "I don't think that") as part of shell language, since we only wanted to focus on separating organizational phrases marking argumentative structures from content segments, leaving markers of subjectivity to remain a part of the content segments. Thus, our segmentation approach is more coarse-grained. Other researchers like Nguyen and Litman (2015) have used shell-like concepts for the task of detecting argument components in persuasive essays, also allowing shell language (in Nguyen and Litman's (2015) terms "argument words") to occur in the argument content (in Nguyen and Litman's (2015) terms "domain words").

[1]Both the corpus (consisting of comments, article texts and Facebook screenshots) and the associated annotation guidelines can be found at https://www.lt3.ugent.be/resources/platos_pilot-study/.

For the segmentation of comments containing rhetorical questions, we decided to follow the perspective of Speech Act Theory, which contends that this special type of questions only has the surface structure of a question, but realizes the speech act of making a statement (Walton, 2007). Thus, every rhetorical question which was followed by an explicit answer was considered as one segment together with that answer. Our approach resulted in a total number of 504 segments across the whole corpus. The highest number of segments for a single comment is 19; the lowest is 1. The comments in our corpus contained an average number of 5 segments.

We divided our annotators into two groups of three. Each group was assigned different annotation tasks:

1. The first group had to annotate the *topics* and *topic aspects* contained in the comment (*interactional topics*), as well as the *stances* expressed towards those topics and aspects. This task was performed on the level of the entire comment text. The annotators could select the topics and aspects from the list of structuring topics identified in the preliminary step for all the articles and Facebook posts. They were allowed to create additional topic and aspect labels if they could not find a suitable one in the list.

2. The second group was tasked with labelling the predefined segments of each comment as *argumentative unit (AU)* or *non-argumentative unit (NAU)*. Organizational elements which did not carry any argumentatively relevant information were to be marked as *NAU*. Then, the annotators had to indicate which segment best represented the *claim* of the argument. Only segments marked as *AU* were eligible to receive the claim label.

We provide an example of an annotated comment from our corpus to help clarify the annotation tasks. We refer the reader to the guidelines for a more detailed description of the labels. For the comment "And in the stores everyone touches all the fruit and vegetables with their hands better to monitor everything that is in the store, people even open packages and eat the product", the topic-aspect combinations *corona (measures, spread of disease)*, *enforcement* and *shops* were labelled by one of our annotators. Three stance labels were identified: *pro* towards the topic of *enforcement* and the aspect of *measures* and *contra* towards the aspect of *spread of disease*. The segmented comment was shown to the second group of annotators as follows: "1[And] 2[in the stores everyone touches all the fruit and vegetables with their hands] 3[better to monitor everything that is in the store,] 4[people even open packages and eat the product]". In a first round, the annotator determined for each segment whether it was argumentatively relevant or not, resulting in the following labels: argumentative (segments 2, 3 and 4); non-argumentative (segment 1). In a second round, the annotator was given the segments prelabelled for argumentativeness. The annotator then had to determine which of the segments labelled as argumentative best represented the central claim/conclusion of the user comment. In this example, segment 3 was labelled as the claim. All other segments marked argumentative in this comment are therefore seen to support or to be otherwise argumentatively relevant leading up to this claim.

6 Inter-annotator Agreement Results

For the first annotation tasks, viz. assigning topics, aspects and stance, the annotators could assign multiple labels to the same comment. As standard IAA scoring mechanisms (such as Cohen's pairwise kappa) assume the assignment of one category label per unit of annotation, these metrics are not suitable for measuring IAA for multiple labels per annotator. Therefore other metrics, such as Krippendorff's $Alpha$ (Krippendorff, 1970; Krippendorff, 2004), have to be used to calculate disagreement (or distance) between sets of assigned labels. Krippendorff's $Alpha$ considers difference in annotation on all possible annotation units, irrespective of the number of labels and the type of annotation (categorical, numeric, ordinal). To calculate the distance between two sets of annotation labels, we followed the implementation of Passonneau (2006), using MASI (Measuring Agreement on Set-valued Items). The resulting distance is 0 when sets are overlapping, and 1 when sets are disjoint.

The first annotation tasks appeared to be challenging, resulting in a total of 88 different topics, 245 aspects and 197 stance labels. As a lot of the disagreement for stance is caused by the choice of different

Alpha/MASI distance			#comments perfect agreement		
Ann1 - Ann2	Ann1 - Ann3	Ann2 - Ann3	Ann1 - Ann2	Ann1 - Ann3	Ann2 - Ann3
Topics					
0.36	0.64	0.66	55	20	19
Aspects					
0.61	0.76	0.81	29	14	10
Topic/Aspect_Stance					
0.66	0.84	0.82	30	12	11
Stance					
0.39	0.51	0.47	56	39	43

Table 1: Krippendorff's Alpha values using the MASI distance metric for each pair of annotators and the number of instances with total label agreement for topics, aspects, stance at the topic/aspect level and stance at the comment level.

aspects of the same topic, we also calculated the distance when only considering the sets of stance labels disregarding the specific topic or aspect they were attached to. Table 1 lists the distance between the sets of labels assigned per pair of annotators and the number of instances with total label agreement for topics, aspects, stance at the topic/aspect level and stance labels ignoring the aspect/topic.

In the following step, annotators were charged with labeling each segment as (1) either an argumentative unit (AU) or non-argumentative unit (NAU) and (2) a claim or no claim. As annotators could only assign one label per task, we could apply more traditional IAA metrics for these tasks such as Cohen's pairwise kappa (Cohen, 1960), which measures agreement between two raters, and Fleiss' kappa (Fleiss, 1971), that can be used for measuring agreement between multiple raters. Note that the Fleiss kappa is a multi-rater generalization of Scott's pi statistic, not Cohen's kappa. Table 2 lists the agreement scores for these two tasks. For the interpretation of the kappa scores, we refer to Landis and Koch (1977), who consider kappa scores ranging between 0.21 and 0.40 as *fair* agreement, between 0.41 and 0.60 as *moderate* agreement, between 0.61 and 0.80 as *substantial* agreement and between 0.81 and 0.99 as *almost perfect* agreement.

Cohen's kappa			Fleiss' kappa
Ann1 - Ann2	Ann1 - Ann3	Ann2 - Ann3	All annotators
Argumentativeness /vs/ non-argumentativeness			
0.74	0.82	0.71	0.76
Claim detection			
0.45	0.48	0.41	0.45

Table 2: Cohen's kappa agreement scores for pairs of annotators and Fleiss' kappa agreement for all three annotators for the tasks of argumentativeness and claim detection.

7 Analysis

7.1 Comment-level annotation tasks (topic, aspect, stance detection)

We reached moderate agreement for the topic and aspect labelling tasks, considering the Krippendorff's distance is rather high between all pairs of annotators. This is partly due to errors percolating from the topic labelling step to the aspect labels. When we consider only the stance labels per comment, without taking into account the topic and aspects they are linked to, we see that on average, almost half the comments show perfect agreement. In addition, annotators sometimes assign similar (sub)topics or additional (sub)topics, resulting in fairly large distance scores: e.g. [travel, Corona, aviation] vs [Corona, aviation] results in a distance of 0.55, and [police, politician] vs [politician] results in a distance of 0.665. Perfect agreement on the stance labelling task was reached for only 8 comments, all three annotators

agreeing on the "none" stance label for those comments. When considering partial agreement (defined here with the following condition satisfied: all three annotators share at least one stance label towards the same topic, e.g. "(politics)contra"), we noted partial agreement for 19 comments.

In the following, we briefly discuss some concrete annotation examples for the annotation tasks which were performed on the comment level. When considering the distance metric between the stance labels only, we noted some recurring labelling errors.

(1) Feeling ill = stay at home! Easy! There are people who can't stand wearing the mask due to breathing problems! Wearing a mask in this heat is asking for trouble for those people. Keep your distance, sanitize your hands and stay AT HOME when ill. EASY...[2]

In some cases, the annotator identified the stance towards a specific topic, while the others identified only the stance towards the broader related topic. In Example 1, Annotator 1 identified a contra stance towards the *mouth mask obligation* topic, while Annotator 2 identified only a pro stance towards the more general topic of *corona measures*. Annotator 3 did identify both stances. One possible explanation for this confusion is that the annotators may have made a distinction between main topics and more peripheral topics in the comment. In future guidelines, it will be necessary to emphasize the need for identifying both the stance towards the broader topics and more specific ones.

(2) Just have them all write a protest letter, you won't be receiving much mail.

Comments containing irony such as Example 2 seem to complicate the stance annotation in some cases (perfect agreement for stance labelling amounting only to 8% of the total of 25 ironic comments and partial agreement reaching 28% (7/25)). Two of the annotators indicated doubts in cases such as these and were subsequently instructed to use the *NONE* label in case of doubt. This instruction will also be added to the future guidelines. Additionally, we will ask the annotators to mark the instances of irony and context-related understandability issues, a strategy also applied by Wojatzki and Zesch (2016) in their stance annotation study.

(3) And the people who kept working from home every day and entertained the kids at the same time, also don't get anything from the government. The energy bill has nevertheless seriously increased in the past 3 months .. This decision was not well considered at all!

Some specific topic labels frequently caused disagreement. In Example 3, disagreement is caused by the fact that one annotator has interpreted the user's stance as being in favour of the government handing out a financial compensation to those in need, resulting in the stance label *(compensation)PRO*. Another annotator has interpreted the stance as being against the lack of compensation for specific groups of the population, resulting in the label *(compensation)CONTRA*. The third annotator has circumvented the issue by adding the extra labels *(government)CONTRA* and *(aid)PRO*. In any case, the future guidelines will still include the option of defining extra labels wherever called for.

(4) 1[Feeling ill = stay at home!] 2[Easy!] 3[There are people who can't stand wearing the mask due to breathing problems!] 4[Wearing a mask in this heat is asking for trouble for those people.] 5[Keep your distance, sanitize your hands and stay AT HOME when ill.] 6[EASY...]

7.2 Segment-level annotation tasks (argumentativeness, claim detection)

Considering the difficulty of the annotation task, we reached substantial agreement scores for the task of argumentativeness detection (0.76 Fleiss' kappa). For the argumentativeness labelling task, non-argumentative segments were defined as those having primarily an organizational function in the comment (for example, to introduce a particular component of the argumentation). Two annotators expressed their doubts as to how to annotate segments which in addition to the organizational function, also seemed to express the stance of the user, such as segments 2 and 6 in Example 4. To avoid mislabelling of such segments which are stance-bearing (and should thus be labelled as AU), we will add this specification to the guidelines.

[2] All examples from the corpus given here have been translated from the original Dutch.

(5) 1[She's right.] 2[But no matter what she says or does, it will never be enough.] [...]

Segments occurring at the very start of the comment caused disagreement for argumentativeness labelling, especially in cases like Example 5 where the segment appeared to interact with the surrounding context (for instance, the title or the content of the newspaper article). Since the information in these segments often consists of an evaluative comment on the article content, they are considered stance-bearing and should be labelled as AU. We found a total of 45 comments in our corpus in which the first segment interacted with the context. In 77.8% of these cases (35/45 comments), perfect agreement was reached across all three annotators for the argumentativeness labelling task. When we compare this to the 52% of comments reaching perfect agreement for this task on the whole corpus, we see that the influence of the lack of context information available to our annotators was moderate for this task on our current corpus. We will ensure such segments are flagged in future annotations. This will help us to determine which types of comments may require context features for the automatic detection system.

Additionally, it will be useful to ensure our future guidelines are more explicit on the possible guises of argumentativeness (defined in our guidelines as "all information relevant to the support or expression of the author's position"). Therefore, we will explicitly state that this then includes background information (e.g., "In 1400, a virus was considered an illness") and (parts of) personal narratives ("Just the other day I saw a whole family with no mouth masks on"). In our corpus, we found a total of 44 comments contained at least one of these less obviously argumentative segments (corresponding to 19% or 98/504 total segments in our corpus). Since the annotators themselves raised the question of whether or not such segments were to be considered argumentative fairly quickly, we were able to achieve perfect agreement across all three annotators for 80% (79/98) of segments. However, perfect agreement was only reached for 43% (19/44) of all comments containing this type of segment, meaning that there were few comments in which they were all captured. Therefore, we will include more examples of what sequences of less obviously argumentative segments may look like (such as the first segment of Example (6) in the guidelines). In Example (6), while both segments are argumentative (since together they form evidence for the user's claim that the experts are distributing confusing information about the usefulness of mouth masks), we found disagreement on the argumentative relevance of the first segment.

(6) 1[During the acute phase the experts said that wearing mouth masks wasn't useful .] 2[Now that the virus has been seriously reduced, we do have to wear them .] [...]

Maar (English: but) (29 instances) and *en (English: and)* (23 instances) were the most frequently occurring shell language expressions and proved indicative of argumentativeness (*But* preceded an AU segment for 23/29 instances and *And* preceded an AU segment for 15/23 instances for all annotators). This is not surprising, since *maar* (English: but) is a connective often used to signal a contrasting reason in argumentation and *en* (English: and) can be used to signal an additional reason. An important type of shell language expression we found in our corpus (*I shouldn't be saying this*; *There is not much more to say about this*) corresponds to the so-called "discourse markers of standpoint continuity", identified by Craig and Sanusi (2000) as a common characteristic of group discussions on controversial issues. They are commonly used to specify argumentative standpoints as well as to avoid disagreement while saving face (Craig and Sanusi, 2000). However, a clear distinction should be introduced in the guidelines between such markers and segments like *I do understand his point* or *I get that*, which are indicative of the stance of the author and should be annotated as AU (all three annotators currently annotated these segments as NAU). The presence of the verbs of understanding as well as the first person singular may help distinguish such segments.

We reached moderate agreement for the claim labelling task (0.45 Fleiss' kappa). For 33 comments full agreement was reached on the claim annotation task. Many of the claims identified are elliptical, e.g.: *Belgium doomed*, *Mayor not capable*. This is typical of the nature of our data. From Daxenberger et al.'s (2017) analysis of claim segments in various argumentation mining corpora, the presence of policy claims (Schiappa and Nordin, 2014) emerged as a common characteristic in multiple corpora (e.g., of the Wikipedia Talk Page Corpus (Biran and Rambow, 2011) and the Microtext corpus of Peldszus and Stede (2015)). In our corpus, such policy claims, which are often characterized by the presence of the

modal verb "should" (Daxenberger et al., 2017), were present in 32 comments, for example: *Everyone should just decide for themselves what they find important, taking care of yourself or blending in with the crowd*. Since the policy claim expresses a wish for things to be done differently, it may be expressed in the form of an advice (*We can stop driveling about the mouth masks now*), or as a strong imperative (*if your heart isn't in it, don't do it*), the latter of which seems to be particularly indicative of claim segments. Perfect agreement for the claim labeling task was reached on 44% of comments (14/32) (corresponding to 22 segments containing at least one segment with "should" or an imperative form.

Aside from the occurrence of policy claims, when comparing the claims annotated in our corpus to those found in the Web Discourse corpus (Habernal and Gurevych, 2017), our claims are more often anaphoric in the sense that they express the stance of the author, but without specific lexical reference to the given topic ("I personally find it really sad it's not obligatory,"). Many claims contain expressions signalling beliefs (such as "in my opinion", "personally", "I find"), which according to Daxenberger et al.'s (2017) analysis, is characteristic of the claims found in persuasive student essays (such as the corpus of Stab and Gurevych (2017)). In general, the claim of an argument is more likely to carry stance-taking words toward the topic. This aspect has been identifed as a useful feature for the automatic detection of claims (Ajjour et al., 2019).

The fact that we can find many correspondences with existing corpora of different genres and domains, strengthens our assumption that general markers of claim presence may still be found across genres. Adding these markers in the guidelines for the annotation can help create a more unified approach towards the annotation of claims, which appeared to be absent in the field (Daxenberger et al., 2017). We will investigate the presence and usefulness of such general markers of argumentativity in our more extended corpus, including more domains and platforms to source UGC data from.

8 Conclusion and Future Research

The insights we gained from this pilot annotation study will be used to improve our annotation guidelines. In this study, we pre-segmented all the comments in a preliminary step. This was done to avoid too much error percolation from the annotation results of the segmentation into the claim annotation task, which would make calculating the agreement on claim identification a difficult matter. However, we realize this pre-segmentation may introduce more bias when considering the annotation of other argumentative components like premises and relations between segments. Since the automatic segmentation of texts into argumentative and non-argumentative segments is still very challenging and advances in the field are still being developed, we will use Al-Khatib et al.'s (2016) rule-based algorithm for the automatic pre-segmentation of the corpus as a step prior to the manual annotation. In this way, the annotators will be asked to correct the automatic segmentation by merging incorrectly split segments.

We believe the low agreement results for the first set of annotation tasks (topic, aspect and stance detection) may be improved by reducing the number of possible structuring topic labels for the annotators to choose from. This will require pruning the list of structuring topics and aspects we identified to include only the most frequently occurring ones. Since we are aware of the risk of bias entering our annotation process in providing the structuring topic and aspect labels for the annotators to choose from when deciding on the interactional labels, we will include an evaluation step (removing duplicate labels and becoming aware of ambiguities in certain labels) in our revised annotation guidelines to be performed by a separate annotator. From our analysis of the results for the argumentativeness annotation task, we conclude that our guidelines need more incorporation of stance expressions as an important indicator of argumentativeness.

Since most of our annotators indicated that they had trouble annotating comments due to missing context, we want to explore the impact of context on the various annotation tasks we have performed in this study. First of all, our new guidelines will have to supply more examples of information that is considered argumentatively relevant, e.g., background or context setting information. In particular, we want to investigate whether there are "triggering devices" which are used to evoke context (Nyan, 2017) in the comment. Since the function of context is often defined as narrowing down the range of possible understandings of a text or utterance (Nyan, 2017), we are interested in studying how supplying

the annotators with various degrees of context information will for instance impact their understanding of the argumentativeness and the claim of a comment.

References

Y. Ajjour, W. Chen, J. Kiesel, H. Wachsmuth, and B. Stein. 2017. Unit segmentation of argumentative texts. In *Proceedings of the 4th Workshop on Argument Mining*, pages 118–128, Copenhagen, Denmark. Association for Computational Linguistics.

Y. Ajjour, M. Alshomary, H. Wachsmuth, and B. Stein. 2019. Modeling frames in argumentation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2922–2932, Hong Kong, China. Association for Computational Linguistics.

K. Al-Khatib, H. Wachsmuth, J. Kiesel, M. Hagen, and B. 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.

J.C. Anscombre. 1995. De l'argumentation dans la langue à la théorie des topoï. In *La théorie des topoï*, pages 11–47. Kimé, Paris.

N. Bauwelinck and E. Lefever. 2020. Annotation Guidelines for Labeling Topics, Aspects, Stance, Argumentativeness and Claims in Dutch social media comments, version 1.0. Technical report, Ghent University, LT3 15-01.

O. Biran and O. Rambow. 2011. Identifying justifications in written dialogs. In *Proceedings of the 2011 IEEE Fifth International Conference on Semantic Computing*, ICSC '11, pages 162–168, USA. IEEE Computer Society.

J. Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

R.T. Craig and A.L. Sanusi. 2000. 'i'm just saying...': Discourse markers of standpoint continuity. *Argumentation*, 14(4):425–445.

J. Daxenberger, S. Eger, I. Habernal, C. Stab, and I. Gurevych. 2017. What is the essence of a claim? cross-domain claim identification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2055–2066, Copenhagen, Denmark. Association for Computational Linguistics.

J. Du, J. Jiang, L. Yang, D. Song, and L. Liao. 2014. Shell miner: Mining organizational phrases in argumentative texts in social media. *2014 IEEE International Conference on Data Mining*, pages 797–802.

O. Ducrot. 1982. Note sur l'argumentation et l'acte d'argumenter in concession et consécution dans le discours. In *Cahiers de linguistique française, 4*, pages 143–163, Genève. Université de Genève.

J. Eckle-Kohler, R. Kluge, and I. Gurevych. 2015. On the role of discourse markers for discriminating claims and premises in argumentative discourse. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2236–2242, Lisbon, Portugal. Association for Computational Linguistics.

R. Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4):51–58.

J.L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382.

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

K. Krippendorff. 1970. Bivariate agreement coefficients for reliability of data. *Sociological methodology*, pages 139–150.

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

D. Küçük and F. Can. 2020. Stance Detection: A Survey. *ACM Computing Surveys*, 53(1).

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

N. Madnani, M. Heilman, J. Tetreault, and M. Chodorow. 2012. Identifying high-level organizational elements in argumentative discourse. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 20–28, Montréal, Canada. Association for Computational Linguistics.

E. Manosevitch and D. Walker. 2009. Reader comments to online opinion journalism: A space of public deliberation. In *10th International Symposium on Online Journalism*, Austin, Texas.

H. Nguyen and D. Litman. 2015. Extracting argument and domain words for identifying argument components in texts. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 22–28, Denver, Colorado, USA. Association for Computational Linguistics.

T. Nyan. 2017. Re-contextualising argumentative meanings: An adaptive perspective. *Argumentation*, 31:267–299.

J. Park and C. Cardie. 2014. Identifying appropriate support for propositions in online user comments. In *Proceedings of the First Workshop on Argumentation Mining*, pages 29–38, Baltimore, Maryland. Association for Computational Linguistics.

R. Passonneau. 2006. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

A. Peldszus and M. Stede. 2015. An annotated corpus of argumentative microtexts. In *Proceedings of the First Conference on Argumentation*, Lisbon, Portugal.

I. Rowe. 2015. Deliberation 2.0: Comparing the deliberative quality of online news user comments across platforms. *Journal of Broadcasting & Electronic Media*, 59(4):539–555.

P. Saint-Dizier. 2016. Challenges of argument mining: Generating an argument synthesis based on the qualia structure. In *Proceedings of the 9th International Natural Language Generation conference*, pages 79–83, Edinburgh, UK. Association for Computational Linguistics.

E. Schiappa and J.P. Nordin. 2014. *Argumentation: Keeping Faith with Reason*. Pearson.

S. Somasundaran, J. Wiebe, and J. Ruppenhofer. 2008. Discourse level opinion interpretation. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 801–808, Manchester, UK.

C. Stab and I. 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.

J. Stromer-Galley and A.M. Martinson. 2009. Coherence in political computer-mediated communication: analyzing topic relevance and drift in chat. *Discourse & Communication*, 3(2):195–216.

A. Tseronis. 2011. From Connectives to Argumentative Markers: A Quest for Markers of Argumentative Moves and of Related Aspects of Argumentative Discourse. *Argumentation*, 25(4):427–447.

F. van Eemeren and P. Houtlosser. 2006. Strategic maneuvering: A synthetic recapitulation. *Argumentation*, 20(4):381–392–802.

M.P.G. Villalba and P. Saint-Dizier. 2012. Some facets of argument mining for opinion analysis. *COMMA*, 245:23–34.

A. Vlachos and S. Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22, Baltimore, MD, USA. Association for Computational Linguistics.

D.N. Walton. 2007. *Dialog theory for critical argumentation*. John Benjamins, Amsterdam.

M. Wojatzki and T. Zesch. 2016. Stance-based Argument Mining – Modeling Implicit Argumentation Using Stance. In *Proceedings of the KONVENS*, pages 313–322.