Detecting Minority Arguments for Mutual Understanding: A Moderation Tool for the Online Climate Change Debate

Cedric Waterschoot and **Ernst van den Hemel**
KNAW Meertens Instituut
Oudezijds Achterburgwal 185
1012DK Amsterdam, The Netherlands
cedric.waterschoot@meertens.knaw.nl
ernst.van.den.hemel2@meertens.knaw.nl

Antal van den Bosch
Utrecht Institute of Linguistics OTS
Utrecht University
The Netherlands
a.p.j.vandenbosch@uu.nl

Abstract

Moderating user comments and promoting healthy understanding is a challenging task, especially in the context of polarized topics such as climate change. We propose a moderation tool to assist moderators in promoting mutual understanding in regard to this topic. The approach is twofold. First, we train classifiers to label incoming posts for the arguments they entail, with a specific focus on minority arguments. We apply active learning to further supplement the training data with rare arguments. Second, we dive deeper into singular arguments and extract the lexical patterns that distinguish each argument from the others. Our findings indicate that climate change arguments form clearly separable clusters in the embedding space. These classes are characterized by their own unique lexical patterns that provide a quick insight in an argument's key concepts. Additionally, supplementing our training data was necessary for our classifiers to be able to adequately recognize rare arguments. We argue that this detailed rundown of each argument provides insight into where others are coming from. These computational approaches can be part of the toolkit for content moderators and researchers struggling with other polarized debates.

1 Introduction

Even though a consensus has existed within the scientific community on the topic of human-caused climate change for some time, the online debate remains very polarized. Online comment spaces are typically overwhelmed with a large quantity of contributions. This information flood hinders the promotion of mutual understanding and inclusivity in debate spaces. Additionally, climate change presents a splintered debate with niche opinions and many viewpoints. The recognition of these niche arguments are vital to support the moderator in adhering to the heterogeneous discussion that climate change presents. This setting presents opportunities for mutual understanding by improving issue awareness and the quality of deliberation.

In this paper, we construct a twofold approach to support mutual understanding in the online climate change discussion. First, we aim to classify posts for the argument they present. Second, we dive deeper into singular arguments to create an overview of the lexical patterns in each argument-specific sub-corpus. We conclude the paper by discussing the limitations of modeling nuanced argumentation by a computational method and link our approach to fields struggling with content moderation and polarized debates.

2 Background

2.1 Argument Mining & Stance Detection

Our application falls under the umbrella of 'argument mining' and 'stance detection'. Within Natural Language Processing, argument mining is defined as the automated identification and extraction of argumentation found in natural language (Lawrence and Reed, 2019). Following the stark increase in the availability of textual data found on online fora and social media platforms, argument detection tasks have been receiving a lot of attention. The related task of stance detection is aimed at classifying the stance of the producer of a

Proceedings of the 29th International Conference on Computational Linguistics, pages 6715–6725
October 12–17, 2022.

piece of text towards the target topic (Küçük and Can, 2020). This result is often performed over three classes: *in favour ('Pro'), against ('Con') or neutral*.

To define an argument, researchers often look towards the Toulmin model of argument (Toulmin, 2003). Toulmin defined a formal argumentative model comprising of the following five elements: *claim, data, warrant, qualifier, and rebuttal* (Toulmin, 2003). However, textual data from social media or comment platforms tend to fall short of fulfilling these formal requirements due to their briefness and elliptic nature. Researchers have therefore labeled tweets as argumentative when a portion of the formal argumentative structure was present (Bosc et al., 2016). These portions can be a premise, a, conclusion, or the connecting relationship between these two argumentative parts. In this paper, we follow the same operationalization of the definition of arguments.

One factor further complicating these tasks is the influence of context. Context may affect whether an utterance is interpreted as argumentative or not (Carstens and Toni, 2015). Typically, the classification tasks are restricted to features intrinsic to the sentence, post, or utterance, and are blind to context; therefore, resulting models may not be robust across different contexts (Lawrence and Reed, 2019). What makes the contextual factor challenging is the fact that not all content and context is expressed explicitly (Moens, 2018). A lot of this knowledge and expression remains "in the mind of communicator and audience" (Moens, 2018). Some have even argued that, in particular cases, content can be less important than the context it resides in (Opitz and Frank, 2019).

Related to the contextual factor is the importance of previous knowledge in stance detection and annotation. The complexity of stance-taking includes cultural and social aspects (ALDayel and Magdy, 2021). Personal opinions and the aforementioned non-personal aspects make stance detection a non-trivial task (Du Bois, 2007).

In recent years, a range of work has focused on argument detection in online content. The first step of these approaches often relates to making the distinction between argumentative and non-argumentative samples. Addawood and Bashir (2016) perform such a classification while subsequently classifying the evidence type presented within argumentative tweets with Support Vector Machines (SVM) and Decision trees. Naderi and Hirst (2016) created a corpus of parliamentary discourse labelled as 'Pro' and 'Con' on the subject of gay marriage, alongside pre-defined argumentation specific to the topic. Cross-topic experiments pose even greater challenges than single topic argument classification. Stab et al. (2018) annotated and classified web texts across eight different topics based on the three stance classes (pro, con and neutral). 'Pro/Con' classification on unseen topics has also been done using BERT models, which improved F_1-scores compared to attention-based neural networks (Reimers et al., 2020). In this paper, we follow the methodology set out in the existing literature by creating a single-topic corpus (Naderi and Hirst, 2016; Bosc et al., 2016). The annotation scheme is based on pre-defined arguments in the discussion that are already explored in the wider literature on the selected topic of climate change.

2.2 Climate change argumentation

In the upcoming paragraphs, we outline the specific arguments that have been defined in the literature. Argumentation is divided between 'Pro', i.e. those acknowledging climate change is a human-caused threat, and 'Con', arguments that deny climate change as a problem caused by human action.

The latter seems to be the most diverse cluster. Rahmstorf (2003) proposes a three-way distinction in climate change denial arguments: *(1) Impact scepticism, (2) Trend scepticism and (3) Attribution scepticism*. Trend scepticism rejects the warming trend all together, while attribution sceptics question whether human activity is the cause (Rahmstorf, 2003). The former seems to be an idea that is disappearing (Rahmstorf, 2003; Dunlap and McCright, 2012). On the other hand, impact scepticism states that the consequences from climate change might not be that bad (Rahmstorf, 2003). Examples of this argument are statements detailing that a warmer climate is desirable or that we can simply mitigate the effects. Dunlap and McCright (2012) detail the same three movements against human-caused climate change: (1) no warming, (2) not caused by human activity and, (3) the 'non-problemacity' of climate change. The latter focus of 'non-problemacity' seems to be based on a dominant social paradigm that our species is able to exert control over nature (McCright and Dunlap, 2003). This control directly leads to the conclusion that climate change cannot pose a threat (Bord

Stance	Argument (labels)	Explanation
Con	Impact scepticism	Denial of consequences
	Attribution scepticism	Denial of human influence
	Trend scepticism	Denial of warming trend
	No consensus	Denial of consensus among scientists
	Bad science	Accusation of bad models/forecasts used in science
	Conspiracy theories	Umbrella category for all conspiracy-related content
Pro	Anthropogenic climate change (ACC)	Climate change is caused by human activity
None	No argument	No relevant argument is present / post is off-topic

Table 1: Climate change argumentation & annotation scheme

et al., 2000; Poortinga et al., 2011).

Aside from these three forms of scepticism, climate change denial also focuses on the scientific community. More specifically, the existence of a scientific consensus is often questioned (Leiserowitz et al., 2010). We label this argument *No consensus*. Interestingly, a consensus among scientists has long existed (Doran and Zimmerman, 2009; Oreskes, 2005). While it is uncertain as to why this consensus is questioned, a potential explanation lies in the fact that the scientists have long shied away from making dramatic warnings or conclusions in publications (Brysse et al., 2013). A second science-focused argument against climate change takes aim at the science and models themselves, which we label as *'Bad science'*. The claim posits that the complexity and uncertainty surrounding the climate system is a hurdle for scientists to make rigid forecasts (Poortinga et al., 2011). Pinpointing the exact cause for every reasoning disputing human-caused climate change is difficult if at all possible. However, a number of sources can be found, including organized anti-environmental movements like those found in the U.S. in the 1990s (McCright and Dunlap, 2003), unreliable or incomplete interpretation of scientific evidence (Whitmarsh, 2011) or online content like videos found on Youtube (Allgaier, 2019). These sources are often presented as 'manufacturers of doubt' (Van Linden et al., 2015). A final category arguing against climate change is the *conspiracy-related* class. Content related to conspiracy theories often emerge in polarized debate in the online sphere, even in good-faith discussions (Samory and Mitra, 2018). Similar to the definition of argument, we define 'conspiracy' loosely by not requiring all elements of a conspiracy theory, *agent, action and target*, to be explicitly present (Samory and Mitra, 2018). References to conspiracies in user comments tend to be compact and make use of the most common denominator words for a conspiracy, and further rely on context to complete the conspiratory content.

Those arguing that the current climate crisis is caused by human activity find themselves in a more unified environment, which we label under the term anthropogenic climate change (ACC). By the late 1980s, and after the vast accumulation of evidence, the majority of academics had concluded that anthropogenic climate change was occurring (Leiserowitz, 2007). The argument is in practice quite straight-forward and is reflected in the literature in the form of surveys of experts (Doran and Zimmerman, 2009) or literature reviews of the field (Oreskes, 2005). Additionally, references are often made to the reports from the Intergovernmental Panel for Climate Change (IPCC) (Masson-Delmotte et al., 2021).

2.3 Deliberation on online platforms

This paper focuses on mutual understanding in the climate change debate in the setting of online comment platforms. In the previous paragraphs, we outlined the polarized argumentation that occurs in the discussion. Briefly, mutual understanding is established through comprehension of what others are trying to do or say as well as why (Margaret, 1994). Exposure to other opinions can improve out-group tolerance, which in turn can facilitate mutual understanding (Mutz and Mondak, 2006; Andersen and Hansen, 2007). Evidence indeed shows that these heterogeneous environments are important for facilitating deliberative qualities (Suiter et al., 2016). A vital part of this process is the exposure to conflicting views, which promotes debate participation (Suiter et al., 2016). Online platforms can develop this deliberative atmosphere further. Hearing out marginalized argumentative camps through active facilitation may fundamentally improve the deliberative properties of a discussion (Strandberg et al., 2017). Experimental evidence indeed suggests that opinion polarization can be deconstructed through

the implementation and facilitation of deliberative norms, as is the goal in moderated comment spaces (Grönlund et al., 2015). Thus, designing online fora with deliberative norms in mind, such as inclusion, justification, and equality of discussion, can result in a suitable comment space for mutual understanding in the climate change discussion (Wright and Street, 2007).

3 Methodology

3.1 Data collection & annotation

We accessed a large dataset of comments from the platform NUjij, the discussion platform of online Dutch newspaper NU.nl. All contributions were posted in 2020, are in Dutch and include comments that were removed by moderators. The presence of these comments can be vital for our focus on minority classes, as we need training data for rare or unwanted arguments as well. First, we filtered out all comments that were not placed under articles with the tag *climate*. These tags originate from the journalists and editors themselves. This initial filtering step resulted in a comment pool of $43,106$ posts. From this climate dataset, we randomly sampled $3,000$ posts for our initial annotation. Furthermore, we sampled 500 extra comments to create a separate validation dataset that will be used to validate each model in upcoming sections [1]. Annotation was done following the scheme presented in Table 1. To derive inter-annotator agreement, subsets of the original data were labelled by two additional annotators. A subset of the original dataset (n=250) was given to two independent annotators. To inform their choices, we created a document with the argumentation scheme. This sheet included clear explanations for each argument that we derived from the climate change literature, alongside examples of comments that contained the argument. These examples were not part of the annotated data. Following this procedure, we achieved a Krippendorf's alpha of 0.73.

3.2 Argument classification

Our particular task consists of a multiclass classification with eight different labels (see Table 1). We split the original dataset containing $3,000$ posts into a training (80%) and test set (20%). This test set remained constant over all versions in this paper, similar to the validation data. As a classifier, we used a pre-trained Dutch transformer-based language model, RobBERT, and finetuned it on the training data (Delobelle et al., 2020). More specifically, we employed the version aimed at sequence classification, which adds a linear classification head on top of the pooled output (Wolf et al., 2020; Delobelle et al., 2020). The final models had a batch size of 32, a learning rate of $5e^{-5}$, optimized with AdamW (Loshchilov and Hutter, 2019) and were trained for ten epochs. The best performing classifiers were achieved after two epochs.

3.3 Minority argument supplementation

During the annotation process, it became clear that certain argumentation classes were extremely rare in 'natural' discussion (Table 2). The bulk of comments were either 'no argument/ off-topic' or 'anthropogenic climate change'. The scarcity made classification of these nuanced cases difficult. With the specific goal of finding minority arguments to boost heterogeneous debate, it was vital to obtain and annotate more of these scarce comments. We opted for an active learning approach to get a better grip on minority classes and to counter possible frequency-related bias in our classification results.

In order to obtain more minority class comments for our training data, we employed a 'query-by-committee' active learning strategy (Zhao et al., 2006). The goal is to filter out more minority arguments that will subsequently be added to the training data to finetune RobBERT further (Figure 1). First, we extract the BERT embeddings from our primary RobBert model (finetuned on only the original data) as input for the first active learning committee. The committee is a collection of five classifiers implemented with Scikit-learn (Pedregosa et al., 2011; Danka and Horvath): *(1) Random Forest, (2) Support Vector Machine (SVM) (radial), (3) SVM (polynomial), (4) SVM (linear) and, (5) gradient boosting classifier*. Each learner within the committee starts with 10 labelled posts as initial training data. With every iteration, a new post from the original data is queried based on the disagreement within the committee, calculated with Kullback-Leibler divergence (Zhao et al., 2006). This sample is subsequently added to the training data. This process is repeated for 250 iterations.

Such a trained committee can be used for prediction, but more importantly for our application, we extracted the uncertainty measure for unseen posts. In this case, the uncertainty is computed as

[1] Supplementary materials found at: github.com/Cwaterschoot/Minority_Argumentation

$1 - class\_probability$. This process is visualized in Figure 2. Each learner in the committee assigns probabilities to every comment for each of the eight classes. We obtain the $class\_probability$ by averaging these probabilities per class across the five learners, resulting in eight probability scores. We take the argument with the highest average class probability for the uncertainty calculation. For example, a comment that is difficult to classify may only have a class probability score of 0.3, which equals a high uncertainty score equal to 0.7 (Figure 2).

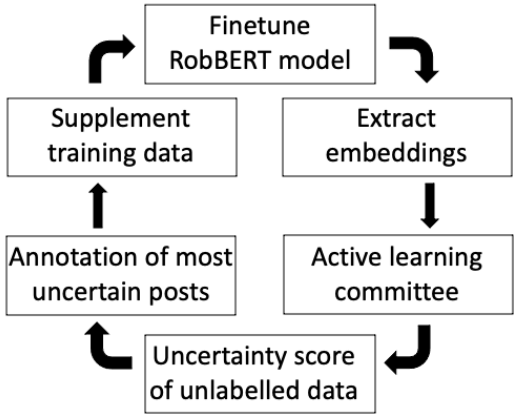

Figure 1: Active learning approach and supplementation of training data

We randomly sampled $10,000$ unseen posts from the climate tagged dataset (containing in total $43,106$ comments) and extracted the uncertainty for each comment. Subsequently, we annotated the $1,000$ most uncertain posts from this collection. Table 2 indicates that this task achieved our goal, namely to relatively increase the number of arguments from minority classes compared to the non-argumentative/off-topic category. Uncertain posts were annotated by a human annotator. Predictions from the committee were disregarded. We repeated this circular procedure a second time *(wave 2)* to add more argumentative posts to the training data (Table 2).

After each wave of newly annotated data, we continued finetuning RobBERT using the previous version as the starting point (see Figure 1). Following this looping procedure, we obtained three versions:(1) *v1* based on the original data, (2) *v2* consisting of v1 supplemented with the first wave and, (3) a fine-tuned version of v2 using both waves of uncertain posts (*v3*). As stated in the previous section, these versions have a linear classification head. Additionally, we extracted the embeddings from all three RobBERT models as input for an active learning committee. Naturally, both the v1 and v2 embeddings are paired with the committees we had used to obtain the uncertain posts. To classify comments based on the v3 embeddings, we trained a third committee following the exact same procedure.

Argument	Original	Wave 1	Wave 2
Impact scepticism	0.02	0.05	0.04
Attribution scepticism	0.03	0.09	0.11
Trend scepticism	0.01	0.01	0.015
No consensus	0.01	0.01	0.004
Bad science	0.01	0.04	0.057
Conspiracy theories	0.01	0.04	0.042
ACC	0.19	0.40	0.30
No argument/off-topic	0.72	0.36	0.42

Table 2: Original data versus uncertain posts. Numbers are fractions of 1 (e.g. $0.72 = 72\%$)

3.4 Patterns in argumentation

Previous sections outlined the automatic annotation of incoming comments for the argument it presents in order to aid moderators in balancing the discussion. Additionally, we aim to boost mutual understanding by diving deeper into what each argument brings to the table. It is important to comprehend the different viewpoints and arguments.

Unique patterns for each argument, i.e. those that have significant presence in one argument compared to all others, were analysed. First, we lowercased the entire corpus and removed stopwords. Subsequently, the corpus was split based on the eight argumentative classes. We used Colibri Core to collect recurring patterns in each subcorpus (van Gompel and van den Bosch, 2016). Following the procedure outlined in van Gompel and van den Bosch (2016), the first step was to class encode the corpus. Subsequently, we created an unindexed pattern model entailing the word n-grams occurring at least twice and with a maximum length of eight tokens. We compared the collection of patterns belonging to a single argument with the seven other argumentative subcorpora taken together. To make this comparison, we utilized the log-likelihood (L) function outlined by Rayson and Garside (2000).

4 Results

4.1 Argument classification

The automatic labelling of posts for the argument it presents may assist moderators in maintaining the

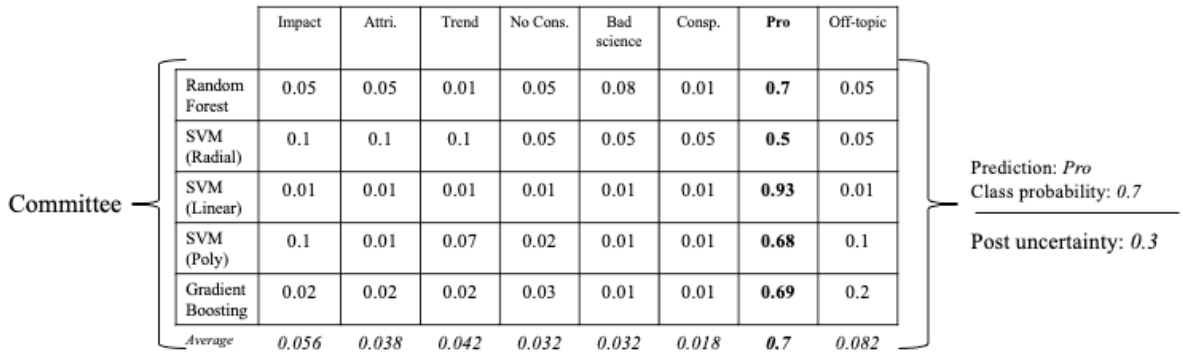

Figure 2: Calculating uncertainty using the active learning committee (fictional post)

desired form of discussion. As outlined earlier, we finetuned a total of three RobBERT models alongside active learning committees that have been used to tag unseen posts for classification uncertainty. Additionally, these committees are used as a classifier on top of the embeddings from each RobBERT model. Each committee consists of five learners and predict arguments by averaging class probabilities within the committee.

Version	Precision	Recall	F1
RobBERT v1 (original data)	0.65	0.51	0.55
RobBERT v1 + committee	0.75	0.50	0.58
RobBERT v2 (original + wave 1)	0.65	0.62	0.62
RobBERT v2 + committee	0.81	0.60	0.64
RobBERT v3 (original + wave 1&2)	0.88	0.68	0.75
RobBERT v3 + committee	0.94	0.67	**0.78**
Random forest (Baseline) [2]			0.25

Table 3: Classification scores on validation data (macro scores)

Table 3 displays the classification metrics on the validation data. Classifying comments using the linear head on top of RobBERT underperforms the committee with each version. The latter improves the macro F1-score score by two to three percentage points by boosting the macro precision score slightly at the expense of the macro recall. RobBERT v3 paired with the committee of classifiers, which is trained on the original training data supplemented with two waves of uncertain posts, outperformed all other versions and achieves a macro F1-score of 0.78.

We constructed the active learning approach to improve the recognition of minority arguments. Table 4 shows that certain arguments like *'Consensus denial', 'Bad science' and 'Conspiracy theories'* posed severe problems for earlier versions. The third iteration of models, which included two waves of uncertain posts in the training data, produced improved F1-scores on the validation set (Table 4). The precision scores for each argument reaches very high levels. This is due to the fact that certain classes have a small number of comments in the data. Impact scepticism is found in 9 comments in the validation data, which is still more than trend scepticism ($n = 2$) and no consensus ($n = 3$). These minority arguments can lead to precision scores that are misleadingly high. For example, one post labelled trend scepticism is the only comment that gets labelled as such by the classifier, leading to a perfect precision score, while recall (0.5 for each version) lacks due to the fact that the other comment belonging to the trend scepticism class is never correctly detected.

Figure 3 shows a two-dimensional representation of the embeddings extracted from RobBERT v3. The arguments, including the relatively rare ones, form noticeable clusters in the embedding space. In the next section, we look at the language and patterns within each argument. Patterns that are distinctively found in a single argument make these arguments distinguishable.

4.2 Argument vocabulary

We previously focused on the computational recognition of climate change argumentation presented in online comments. Additionally, Figure 3 showed that the arguments form visible clusters in the embedding space, hinting at unique vocabulary and patterns within the arguments. We particularly aimed to recognize minority standpoints in order to present the whole range of online opinions. Subsequently, it is important that users and moderators understand what is being said. The discussion of polarized discussion may be boosted by not only

[2] Three classes were not predicted

Version	Impact	Attribution	Trend	Consensus	Bad science	Conspiracy	Pro
RobBERT v1	0.47	0.68	0.67	0.2	0.42	0.4	0.63
RobBERT v1 + committee	0.62	0.70	0.67	0.4	0.33	0.43	0.62
RobBERT v2	0.63	0.69	0.67	0.4	0.45	0.5	0.71
RobBERT v2 + committee	0.75	0.67	0.67	0.67	0.2	0.59	0.72
RobBERT v3	**0.8**	0.72	0.67	**0.8**	0.67	0.71	0.74
RobBERT v3 + committee	**0.8**	**0.79**	0.67	**0.8**	**0.71**	**0.75**	**0.77**

Table 4: F1-score per minority argument on validation data

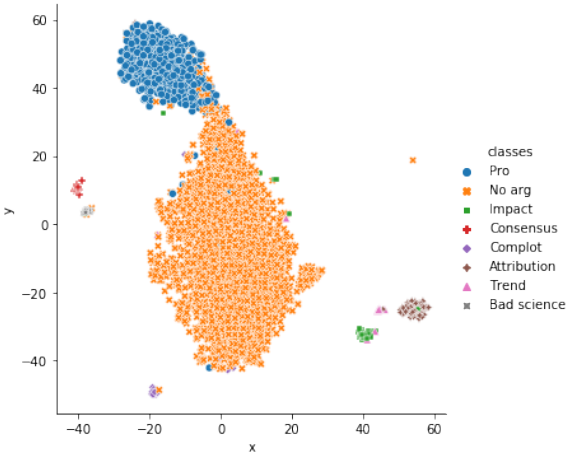

Figure 3: TSNE visualisation of RobBERT v3 embeddings

trying to comprehend others, but to invite them into an heterogeneous debate environment. To achieve this understanding of varying argumentative positions, we have derived the vocabulary and patterns within each argument to showcase what sets each of them apart. Table 5 presents the collected patterns with the highest log-likelihood per argument.

The first class, *no argument/off-topic*, contains a wide collection of patterns. Posts within this category talk about a variety of topics, including the energy transition in The Netherlands and the potential effect of the growing population and consumption (Table 5). These posts are often adjacent to the discussion at hand, but do not present actual argumentation aimed at the cause of climate change.

The argument in favour of human-caused climate change (*ACC*) has a different focus. The political aspect comes to the forefront, expressed in terms like 'voting', 'political' and 'importance'. These comments attempt to rally readers to take action. Another distinctive pattern in this argument details the global component of climate change. Commenters write about the need of unison action.

In our annotation scheme, climate change scepticism is broken down into three subcategories: *Impact, Attribution and Trend*. The latter has very clear patterns that sets this argument apart from the others. Trend sceptics on the comment platform often point towards cold winters, volcanic activity and the existence of ice sheets to reject the warming trend. Furthermore, these sceptics call human-caused climate change a religion and garbage. Generally, it seems that this scepticism is the most straight-forward rejection of human-caused climate change. On the other hand, attribution sceptics seem to be focused on the historical aspect of climate change. We recognize patterns like '[from] all times', 'billion years' and 'million years ago' (Table 5). Alongside this focal point, some attribution sceptics seem to concede that human influence might speed up natural processes ('human influence', 'speed up', 'partly'). These natural processes include the position of the planet relative to the sun and ice age cycles. These topics are not found in other arguments. The third and last argument within the scepticism umbrella, impact, mainly revolves around language claiming it is not necessary to worry about climate change ('worry', 'whine', 'be okay', 'measures' and 'say with certainty').

The two arguments rejecting climate science also rely on specific patterns. On the one hand, we see the dismissal of *scientific consensus* based on very distinctive patterns, e.g. 'prove hypothesis', 'expert' and 'consensus' (Table 5). The accusation of *bad science* revolves around the overarching notion of taking it with a 'grain of salt'. We detect among the patterns 'prediction', 'assumption', 'theories' and 'fearmongering'. These posts urge readers not to take these scientific models too seriously, as they are based on theories and assumptions which do not correspond to real-life circumstances.

The final argumentative class to break down into patterns are the *conspiracy-related content*. We detected conspiracy terms like 'propaganda', 'hoax', 'money' and 'manipulated'. Other unique content references to the 'paris accord' of 2015 and 'acid rain', an environmental issue that received a lot of

Argument	Terms
No argument / off-topic	trash, solution, electricity, somewhere, powerplant, overpopulation, advantage, most people
Impact	worry, previous years, whine, be okay, good economy, say with certainty, measures, stop
Attribution	all times, billion years, speed up, earth sun, human influence, ice age, partly, million years ago
Trend	cold winter, volcanic, tree rings, religion, garbage, ice sheet, every year again
No consensus	consensus, prove, 40 years, 0 co2, assumption, prove hypothesis, phenomenon, expert
Bad science	prediction, assumption, study, grain of salt, case scenario, fearmongering, theories
Conspiracy	paris accord, pro, farmer, propaganda, acid rain, hoax, money, independent, manipulated
ACC (Pro)	use, less people, whole world, houses, voting, political, importance, inhabitants, 3 degrees

Table 5: Argument vocabulary: patterns with highest log-likelihood per argument

attention over the past decades.

5 Discussion

We presented an approach to automatically label online comments for the argument it entails, combined with a deeper dive into each argument in the discussion. In the upcoming paragraphs, we go through some methodological considerations and discuss our approach through the lens of content moderation. Furthermore, we reflect on the usefulness of our approach for other fields that struggle with mutual understanding and opinion polarization.

Translating detailed and nuanced concepts of argumentation into a computational labelling task requires generalization. Poortinga et al. (2011) make the useful distinction between scepticism, uncertainty, and ambivalence. In our annotation scheme, we did not make this specific contrast. Whereas clear-cut scepticism can be rare, as is shown in our data, uncertainty about the anthropogenic causes of climate change might be much more widespread (Whitmarsh, 2011). The dichotomy between uncertainty and scepticism may be an important aspect for mutual understanding and working towards the comprehension and acceptance of human-caused climate change. Unstable or uncertain beliefs can change through contact with scientific cues and information (Jenkins-Smith et al., 2020). In this paper, uncertainty is included within the argument classes, even though the label refers to scepticism. Future research could include this distinction in the methodology to encompass the nuance of polarized debates into the computational approach.

Researchers in the field of content moderation and digital journalism struggle with the concept of mutual understanding, as well as with the implementation of computational technologies (Binns et al., 2017; Ruckenstein and Turunen, 2020). The growing quantity of contributions threatens real-time curation efforts by human moderators. Automatic applications like the one we have presented in this paper are an avenue for assisting human moderators in curating the online comment space (Ruckenstein and Turunen, 2020). While the moderators manage ongoing, interactive processes that are highly dependent on context, computational systems can assist this operation, for example in the form of argument classification and summaries.

Furthermore, research fields that specifically deal with polarized topics struggle with safeguarding civil discussion and mutual understanding. The climate change debate certainly falls within this category. Additionally, online debates on the topic of vaccination lack mutual understanding as well (Jiang et al., 2021). This discussion often lacks heterogeneous discussion due to so-called 'echo chamber' effects (Schmidt et al., 2018). Computational moderation tools, like the one presented in this paper, are an asset for those invested in promoting mutual understanding in these polarized discussions. This approach can be expanded beyond the topic of climate change into other polarized topics. Clearly defined arguments are needed. An example of such a discussion is vaccination, in which clear pro and con sides can be detected (Jiang et al., 2021). Domain-specific research is a requirement to create annotation schemes that adequately entail all minority arguments.

6 Conclusion

In this paper, we created a twofold approach to develop a moderation tool aimed at the climate change debate on online platforms. First, we trained classifiers that label comments for the argument they present. Certain minority arguments, like trend scepticism and accusations of bad science, were very rare. An active learning approach was constructed with the goal of collecting more minority arguments to supplement into our train-

ing dataset. Our best model, after two waves of uncertain posts, achieved a macro F1-score of 0.78. Second, we dove deeper into singular arguments by extracting the lexical patterns that characterize each class. The arguments formed clusters in the embedding space, indicating that each reasoning may be characterized by specific vocabularies. These patterns serve as a swift and understandable view into each argument. Additionally, we formulated methodological considerations regarding the nuance in the annotation scheme and linked our approach to research fields that struggle with moderating online content while safeguarding understanding among participants. The computational approach presented in this paper serves an assisting role to the human moderator, who in turn can deal with the contextual factors.

Acknowledgements

This publication is part of the project Better-MODS with project number 410.19.006 of the research programme 'Digital Society - The Informed Citizen' which is financed by the Dutch Research Council (NWO).

References

Aseel Addawood and Masooda Bashir. 2016. "What Is Your Evidence?" A Study of Controversial Topics on Social Media. *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 1–11.

Abeer ALDayel and Walid Magdy. 2021. Stance detection on social media: State of the art and trends. *Information Processing and Management*, 58(4):102597.

Joachim Allgaier. 2019. Science and Environmental Communication on YouTube: Strategically Distorted Communications in Online Videos on Climate Change and Climate Engineering. *Frontiers in Communication*, 4(July):1–15.

Vibeke Normann Andersen and Kasper M Hansen. 2007. How deliberation makes better citizens: The Danish Deliberative pool. *European Journal of Political Research*, 2007:531–556.

Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. 2017. Like trainer, like bot? Inheritance of bias in algorithmic content moderation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10540 LNCS:405–415.

Richard J. Bord, Robert E. O'Connor, and Ann Fisher. 2000. In what sense does the public need to understand global climate change? *Public Understanding of Science*, 9(3):205–218.

Tom Bosc, Elena Cabrio, and Serena Villata. 2016. Tweeties Squabbling : Positive and Negative Results in Applying Argument Mining on Social Media. *Computational Models of Argument*, 0:21–32.

Keynyn Brysse, Naomi Oreskes, Jessica O'Reilly, and Michael Oppenheimer. 2013. Climate change prediction: Erring on the side of least drama? *Global Environmental Change*, 23(1):327–337.

Lucas Carstens and Francesca Toni. 2015. Towards relation based argumentation mining. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 29–34, Denver, CO. Association for Computational Linguistics.

Tivadar Danka and Peter Horvath. modAL: A modular active learning framework for Python. Available on arXiv at `https://arxiv.org/abs/1805.00979`.

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. RobBERT: A Dutch RoBERTa-based language model. *Findings of the Association for Computational Linguistics Findings of ACL: EMNLP 2020*, pages 3255–3265.

Peter T. Doran and Maggie Kendall Zimmerman. 2009. Examining the scientific consensus on climate change. *Eos*, 90(3):22–23.

John W. Du Bois. 2007. The stance triangle. *Stancetaking in Discourse*, pages 139–182.

Riley E. Dunlap and Aaron M. McCright. 2012. Organized Climate Change Denial. *The Oxford Handbook of Climate Change and Society*, (January 2012).

Kimmo Grönlund, Kaisa Herne, and Maija Setälä. 2015. Does Enclave Deliberation Polarize Opinions? *Political Behavior*, 37(4):995–1020.

Hank C. Jenkins-Smith, Joseph T. Ripberger, Carol L. Silva, Deven E. Carlson, Kuhika Gupta, Nina Carlson, Ani Ter-Mkrtchyan, and Riley E. Dunlap. 2020. Partisan asymmetry in temporal stability of climate change beliefs. *Nature Climate Change*, 10(4):322–328.

Xiaoya Jiang, Min Hsin Su, Juwon Hwang, Ruixue Lian, Markus Brauer, Sunghak Kim, and Dhavan Shah. 2021. Polarization Over Vaccination: Ideological Differences in Twitter Expression About COVID-19 Vaccine Favorability and Specific Hesitancy Concerns. *Social Media and Society*, 7(3).

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Comput. Surv.*, 53(1).

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

Anthony Leiserowitz. 2007. Fighting climate change: Human solidarity in a divided world International Public Opinion, Perception, and Understanding of Global Climate Change. page 40.

Anthony Leiserowitz, Edward W. Maibach, Connie Roser-Renouf, Geoff Feinberg, and Peter Howe. 2010. Climate Change in the American Mind: Americans' Global Warming Beliefs and Attitudes in June 2010. *Yale Project on Climate Change Communication*, (June):1–9.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. *ICLR 2019*.

Tan Margaret. 1994. Establishing Mutual Understanding in Systems Design: An Empirical Study. *Journal of Management Information Systems*, 10(4):159–182.

V. Masson-Delmotte, P. Zhai, A. Pirani, S.L. Connors, C. Pean, Berger S., N. Caud, Y. Chen, L. Goldfarb, M.i. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J.B.R. Matthews, T.K. Maycock, T. Waterfield, O. Yelekci, R. Yu, and B. Zhou. 2021. Ipcc, 2021: Climate change 2021: The physical science basis. contribution of working group i to the sixth assessment report of the intergovernmental panel on climate change. Technical report.

Aaron M. McCright and Riley E. Dunlap. 2003. Defeating Kyoto: The Conservative Movement's Impact on U.S. Climate Change Policy. *Social Problems*, 50(3):348–373.

Marie Francine Moens. 2018. Argumentation mining: How can a machine acquire common sense and world knowledge? *Argument and Computation*, 9(1):1–4.

Diana Mutz and Jeffrey Mondak. 2006. The workplace as a context for cross-cutting political discourse. *The Handbook of Discourse Analysis*, 68(1):398–415.

Nona Naderi and Graeme Hirst. 2016. Argumentation mining in parliamentary discourse. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9935 LNAI:16–25.

Juri Opitz and Anette Frank. 2019. An argument-marker model for syntax-agnostic proto-role labeling. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 224–234, Minneapolis, Minnesota. Association for Computational Linguistics.

Naomi Oreskes. 2005. Scientific consensus on Climate Change. *Science*, 306(January):2004–2005.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Wouter Poortinga, Alexa Spence, Lorraine Whitmarsh, Stuart Capstick, and Nick F. Pidgeon. 2011. Uncertain climate: An investigation into public scepticism about anthropogenic climate change. *Global Environmental Change*, 21(3):1015–1024.

Stephan Rahmstorf. 2003. The climate sceptics. *The state of science*.

Paul Rayson and Roger Garside. 2000. Comparing Corpora using Frequency Profiling. *WCC '00: Proceedings of the workshop on Comparing corpora*, pages 1–6.

Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2020. Classification and clustering of arguments with contextualized word embeddings. *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pages 567–578.

Minna Ruckenstein and Linda Lisa Maria Turunen. 2020. Re-humanizing the platform: Content moderators and the logic of care. *New Media and Society*, 22(6):1026–1042.

Mattia Samory and Tanushree Mitra. 2018. 'The Government Spies Using Our Webcams'. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–24.

Ana Lucía Schmidt, Fabiana Zollo, Antonio Scala, Cornelia Betsch, and Walter Quattrociocchi. 2018. Polarization of the vaccination debate on Facebook. *Vaccine*, 36(25):3606–3612.

Christian Stab, Tristan Miller, and Iryna Gurevych. 2018. Cross-topic Argument Mining from Heterogeneous Sources Using Attention-based Neural Networks. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Kim Strandberg, Staffan Himmelroos, and Kimmo Grönlund. 2017. Do discussions in like-minded groups necessarily lead to more extreme opinions? Deliberative democracy and group polarization. *International Political Science Review*, 40(1):41–57.

Jane Suiter, David M. Farrell, and Eoin O'Malley. 2016. When do deliberative citizens change their opinions? Evidence from the Irish Citizens' Assembly. *International Political Science Review*, 37(2):198–212.

Stephen Toulmin. 2003. *The Uses of Argument*. Cambridge University Press.

Maarten van Gompel and Antal van den Bosch. 2016. Efficient n-gram, Skipgram and Flexgram Modelling with Colibri Core. *Journal of Open Research Software*, 4.

Sander L.Der Van Linden, Anthony A. Leiserowitz, Geoffrey D. Feinberg, and Edward W. Maibach. 2015. The scientific consensus on climate change as a gateway belief: Experimental evidence. *PLoS ONE*, 10(2):2–9.

Lorraine Whitmarsh. 2011. Scepticism and uncertainty about climate change: Dimensions, determinants and change over time. *Global Environmental Change*, 21(2):690–700.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Scott Wright and John Street. 2007. Democracy, Deliberation and Design: the case of online discussion forums. *New Media & Society*, 9(5).

Yue Zhao, Ciwen Xu, and Yongcun Cao. 2006. Research on Query-by-Committee Method of Active Learning and Application. *International Conference on Advanced Data Mining and Applications*, pages 985–991.