Topic-specific social science theory in stance detection: a proposal and interdisciplinary pilot study on sustainability initiatives

Myrthe Reuver[1], Alessandra Polimeno[3*], Antske Fokkens[1], Ana Isabel Lopes[2]
[1] Computational Linguistics & Text Mining Lab, Vrije Universiteit Amsterdam
[2] Communication Science Department, Vrije Universiteit Amsterdam
[3] Utrecht Data School, Utrecht University
[1] firstname.lastname@vu.nl, [2] a.i.loureiro.lopes@vu.nl, [3] aapolimeno@gmail.com
[*] work completed while employed by the Vrije Universiteit Amsterdam

Abstract

Topic-specificity is often seen as a limitation of stance detection models and datasets, especially for analyzing political and societal debates. However, stances contain topic-specific aspects that are crucial for an in-depth understanding of these debates. Our interdisciplinary approach identifies social science theories on specific debate topics as an opportunity for further defining stance detection research and analyzing online debate. This paper explores *sustainability* as debate topic, and connects stance to the sustainability-related Value-Belief-Norm (VBN) theory. VBN theory states that arguments in favor or against sustainability initiatives contain the dimensions of feeling *power* to change the issue with the initiative, and thinking whether or not the initiative tackles an urgent *threat* to the environment. In a pilot study with our *Reddit European Sustainability Initiatives corpus*, we develop an annotation procedure for these complex concepts. We then compare crowd-workers with Natural Language Processing experts' annotation proficiency. Both crowd-workers and NLP experts find the tasks difficult, but experts reach more agreement on some difficult examples. This pilot study shows that complex theories about debate topics are feasible and worthwhile as annotation tasks for stance detection.

1 Introduction

Online platforms see people discussing politicians (i.e., Emmanuel Macron), political issues (i.e., immigration), and cultural debates (i.e., feminist messages in the movie *Barbie*). Stance models usually classify written arguments in such debates into whether they are in favor or against the topic under discussion (Küçük and Can, 2020). The task of stance detection is often conceptualized as topic-*independent*: in datasets and papers, a stance in favour of feminism is seen as conceptually similar as one in favour of immigration.

However, it has been shown that stance models are in fact topic-*dependent*: Transformer models trained on detecting different stances in one topic do not necessarily work on unseen topics (Reuver et al., 2021b; Thorn Jakobsen et al., 2021). Recent work (Ajjour et al., 2023) attempts to tackle this limitation of topic-independent stance modelling by diversifying the number of topics in stance detection datasets, while Beck et al. (2023) update Transformer models' access to knowledge of topic context to improve cross-topic stance detection.

Instead of seeing specialization into one topic as a limitation, we argue that this topic-specificity of debates can also be an asset for stance detection research. Social science theories can play a crucial role in this challenge. Such theories can be used to develop topic-specific stance data and models, which increases the impact of stance detection on socially relevant research questions. This approach also tackles limitations of work assuming stance is topic-independent, such as models not fully capturing the underlying socio-cultural dimensions of specific topics (Reuver et al., 2021b). Social science theory can lead computational argumentation researchers to dimensions of stance that are unique for specific debate topics. These dimensions can then be annotated, and this knowledge of theories help models (and humans) navigate the unique dimensions of the debate.

We argue that defining topic-specific aspects of the debate helps analyzing, modelling, and interpreting the stances in such debates. As a case study, we apply Value-Belief-Norm (VBN) theory (Stern et al., 1999) of environmental debates to stances in environmental debates. We develop an annotation framework and test-drive this by annotating a dataset of 91 Reddit comments reacting to sustainability initiatives with stance, threat, and power. We then analyze the advantages and disadvantages of this approach for stance research on sustainability, and also on other debate topics.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 101–111
July 5–10, 2020.

This paper has the following contributions:
(1) we identify **topic-specific stance detection with social science theory** as an avenue for research in computational argumentation;
(2) we present an **annotation pipeline** for theory-driven stance detection for sustainability debates, and our findings from pilot annotations;
(3) we release **a dataset of Europe-centered debates on sustainability on Reddit**, with a small subset annotated with this annotation pipeline.

2 Topic-dependence and Theory in Stance

Stance detection (Küçük and Can, 2020) is a task in computational argumentation or argument mining (Lawrence and Reed, 2020) consisting of classifying arguments into pro, con, or neutral towards an idea or discussion topic. Stance detection has been used to measure support on social media for topics (Grčar et al., 2017; Scott et al., 2021). These topics are for instance vaccination, but also debate statements such as "we should abolish free speech."

Recent work has indicated that stance models are topic-*dependent* despite being designed as topic-*independent*. Reuver et al. (2021b) found that cross-topic capabilities of Transformer stance models are dependent on topical cues, and that model errors are related to a lack of understanding of socio-cultural dimensions in debates such as gun control and abortion. Thorn Jakobsen et al. (2021) found that these models learn topic-dependent signals, and use mostly topic-dependent words not related to stance as a topic-independent concept (e.g. word 'gun' rather than argumentation-related words).

Earlier work has claimed high cross-topic stance performance, but these performances have still been highly topic-dependent. Some research uses topic dependence in stance for these results, by measuring similarity between two discussion topics, and using the most related topics for cross-topic stance detection. This obtains F1 scores between .67 and .80 on stance detection in unseen topics (Xu et al., 2018; Wei and Mao, 2019; Liang et al., 2021). However, Allaway et al. (2021) do not consider topic-relatedness when modelling and obtain much more modest scores of F1 = .49 and .54 on unseen topics. A similar result can be seen in Reuver et al. (2024), where strategies for few-shot cross-topic stance detection with Transformer models lead to inconsistent performance (between F1 = .344 and F1 = .766) and are largely dependent on dataset choice rather than choices made in model design.

Approaches to improve these non-robust cross-topic capabilities of stance detection models go into two related, but distinct, directions. One is a data-centric approach that can be summarized as **improving the debate topic diversity in datasets**. Earlier work already mentioned how claims and arguments as defined in datasets are topic and context-dependent (Levy et al., 2014). Recently, Ajjour et al. (2023) have developed an ontology for defining diversity of debate topics in computational argumentation datasets.

Another direction is a more model-centric approach that can be summarized as **improving the models' use of topic knowledge**. Earlier work has also looked into improving world knowledge use in stance models (Zhang et al., 2020; Clark et al., 2021). Beck et al. (2023) recently designed a Transformer model architecture that uses real-world knowledge for classification decisions, in the form of a context encoder that "injects" domain-relevant world knowledge into stance models. See Lauscher et al. (2022) for an overview of using knowledge in computational argumentation.

While both directions have promising results, we argue there is another option for overcoming the weaknesses of topic-independent stance detection: designing datasets as well as models with relevant social science theory on the specific debates. This work argues a *debate topic* is broader than specific individual texts or statements (such as "climate change is bad"), but more narrow than what other works call *domain* (which often resorts to categories such as "legal", "social media", etc.). We define a topic as a specific area of socio-cultural discussion, with its own dimensions and aspects of debate such as "climate change", or "immigration". Our definition of *topic* most closely responds to the Level 1 and 2 topics in Ajjour et al. (2023)'s argument topic ontology. Stances in such topics have unique, topic-specific aspects, that can be captured by social science theory on the debate topic in question.

Recently, stance detection work has attempted to include dimensions of opinion beyond simply support or reject, such as argument type (Draws et al., 2022) and underlying values (Kobbe et al., 2020). These variables add underlying reasons *why* an idea is supported or rejected, often a neglected aspect of stance (Joseph et al., 2021; Scott et al., 2021). However, theories on the individual debate topics are often neglected in this exploration of aspects related to stance and arguments.

2.1 Social Science Theory in NLP and Stance

Previous work has outlined how a connection with social science literature and specifically theory can improve NLP tasks, analyses based on them, as well as the theory itself. Radford and Joseph (2020) describe how the traditional Machine Learning pipeline of prediction-based modelling can be enhanced by using theories that are based on the social data or social phenomenon being modelled. These theories can influence relevant sample selection, but theory can also influence the selection of research problems, design of task instructions, as well as how a successful outcome is measured. McCarthy and Dore (2023) argue that theories from the social sciences can help in connecting NLP to relevant research problems. Their work covers an extensive analysis of trends in NLP publishing, and concludes that NLP work in *ACL venues is not *grounded* in the theory about the social phenomena in text it models.

Other work has specifically connected different tasks in computational argumentation to social science theories. Lauscher et al. (2020) research theory in argument quality assessment by an extensive annotation study using theories of argument properties. Vecchi et al. (2021) find that the social science theory of deliberative quality helps solve a definition problem when trying to define and then detect argument quality. Additionally, Reuver et al. (2021a) use the theory of deliberative democracy to identify argument-related NLP tasks relevant to solving a societal problem (non-diverse news recommenders threatening democracy).

However, to our knowledge no work has yet connected social science theory on specific debate topics to the gaps in topic-independent stance detection, and the benefits of topic-dependent stance detection. We will illustrate this connection with a case study on sustainability initiatives.

3 Case Study: Sustainability Initiatives

A stance on sustainability initiatives can be defined as an argument in favor or against initiatives such as renewable energy in local communities (Hewitt et al., 2019) or sustainable behavior at music festivals (Bär et al., 2022). Other work within computational argumentation has looked into sustainability, for instance by annotating evidence that supports sustainability claims in scientific papers (Fergadis et al., 2021), or by detecting sustainable diet patterns in tweets (Hansen and Hershcovich, 2022).

However, the tasks and annotation variables (such as stance, claim-evidence pairs, and argument units) in these earlier papers are very similar to other computational argumentation literature. We would like the annotation procedure to be influenced directly by the sustainability literature in social science. What can a social science theory about sustainability debates tell us about stances in this debate, and how to analyze debates on sustainability?

3.1 Theory and Stances on Sustainability

One theory connected to sustainability and stance is the Value-Belief-Norm theory (VBN) (Stern et al., 1999) of environmentalism. We select this theory for its connection to both stance (support/rejection of initiatives) and the debate topic (sustainability).[1] This theory claims individuals who *support* a sustainability initiative have three things in common: one, they **value** the object under discussion. Two, they believe this object (in this case, the environment or society) is under **threat**. Three, they believe their actions can help restore the desired object (they feel **power to restore**). With these three conditions met, individuals will support a climate initiative. For instance, an initiative to incentivize the consumption of locally produced food might attract arguments that express a negative stance towards it. According to this theory, this negative stance does not mean that people do not support the environment (not valuing the object). A negative stance could mean consumers do not think non-local food production affects the environment (no threat to the desired object) or because they do not believe individuals changing food habits has a collective effect (no power to restore the object). This makes a stance more complex: a negative stance on a climate-related issue does not imply a negative stance on the climate or sustainability.

3.2 VBN aspects in Sustainability Stances

Consider some example arguments.[2] One specific initiative that is debated is *"Spanish should eat less meat to limit climate crisis, says minister"*. One commenter says: *"He's right. High levels of meat consumption and bio industry is a threat to all of humanity."*. This specific comment not only

[1] We realise this is not the only theory related to stance or sustainability: Future work could implement this approach with other theories.

[2] Examples come from our corpus on sustainability initiatives, and are also in our annotation guidelines, see below.

supports the initiative (has a positive **stance**), but also directly mentions **threat** (this issue directly threatens the environment, a valued object).

Another discussion topic has comments more clearly mentioning the **power** dimension of the stance of the commenter. On the topic *Recycling rate of plastic packaging waste*, one commenter mentions "Recycling plastic is mostly pointless. Far better to reduce the use of plastics in packaging as much as possible.", mentioning how individual action after the production process is pointless (lacks **power**). The commenter mentions a negative stance towards recycling, but clearly does support the goal of reducing plastic waste. The next section outlines our annotation pipeline and dataset for these concepts.

4 Data

Our dataset on European sustainability discussions, mostly in English, is obtained from the Reddit.com sub-communities (Proferes et al., 2021) called *europe*, *europeanunion*, and *europes*. We identify any sustainability discussion posted from 2017 to 2022 to contain five years of comments[3]. Our dataset consists of 2.073 discussions with 46.285 comments. Nearly half (922) of these have one or more comments. We release the entire corpus, without annotations, as the *Reddit European Sustainability Initiatives corpus*[4], for non-commercial research-use only under CC-BY-NC licence.[5].

4.1 Annotation

We test both crowd and expert annotation of comments on a small subset of our data, and make our annotation guidelines and task design public - see Appendix B and also our GitHub repository.[6] We also release the annotated dataset, for non-commercial research-use only under CC-BY-NC licence.[3]

Crowd Task A non-expert crowd of 5 annotators hired through annotation platform Prolific annotated 91 random comment-topic text pairs on whether it contained a sustainbility initiative, whether the comment expressed a stance towards it, and power/threat towards the environment. More details on task design are in Appendix B.

Crowd characteristics Our 5 annotators from recruitment platform Prolific were self-reported fluent speakers of English from the US, UK, Canada and Ireland. Pay was US $16 an hour, above minimum wage in highest-paying area Canada. We selected annotators with > 95% approval rating for > 100 previous tasks (Douglas et al., 2023).

Inter-Annotator Agreement We report moderate agreement (Landis and Koch, 1977) for annotation whether thread titles contain **sustainability initiatives** to discuss (Fleiss κ = .47). A similar pilot annotation study on annotating debatable claim vs no claim on 100 social media comments (Bauwelinck and Lefever, 2020) reports a comparable Fleiss κ of .45. Despite its imperfections, percentage agreement is a commonly used agreement measure for stance detection datasets (Ng and Carley, 2022). On average, 89% annotators agree per item (range: 60% to 100%) for annotating the presence of a sustainability initiative.

For **stance**, we initially see a Fleiss κ of .31, which is considerably lower. However, one annotator shows a pattern of unreliability and consistently chooses the positive stance class in the last third of annotation decisions. Removing this annotator increases the Fleiss κ to .39, close to moderate agreement. Stance has an average of 68% annotators agreeing per item (40% to 100%).

Agreement for **threat** is only moderate: Fleiss κ = .33. However, there is a strong difference per item: on average, 60% of annotators agree per item for threat, but some items nearly have complete disagreement, with on 3 items even only 33% agreement. The **power** agreement is also only moderate: Fleiss κ of .29. However, we again see a large difference per item. On average, 60% of annotators agree per item, but for 4 items the majority agrees only with 33%.

Expert Annotation NLP experts from the author's university attempted to improve the threat and power annotation. Four annotators annotated all 91 examples for power. This led to a Fleiss κ of .26: very similar to the crowd annotators. On average, there was 66% agreement over items - slightly higher agreement than the crowdworkers. However, again it shows 4 items with agreement of 33%. For threat (3 annotators), this led to a Fleiss κ of .18,

[3]We scraped with a manual keyword list expanded with pre-trained word embeddings, see Appendix A.

[4]A basic topic model analysis as well as qualitative analysis of this corpus is in Appendix C.

[5]https://huggingface.co/datasets/Myrthe/RedditEuropeanSustainabilityInitiatives

[6]https://github.com/myrthereuver/TopicSpecific_Stance_SocialScience

which is considerably lower than the crowd workers - but could be attributed to fatigue, as annotators annotated this variable after power, and the session was long. On average, there was 59% agreement over items, which is similar to the crowdworkers.

4.2 Per-item annotation differences

Annotating power and threat is more difficult in some comments than others. A deeper look into these comments shows why. One item that had low agreement (33%) from both experts and crowdworkers is one where on the topic *"Climate change: The rich are to blame, international study finds"*, a commenter appears to respond sarcastically: *"Incredible, truly incredible ..did they hire Sherlock for this one ?"*. This added sarcasm makes it hard to differentiate whether this commenter thinks climate change is a serious threat, for both expert and non-expert annotators. The annotation instructions do explicitly ask annotators to attempt to consider sarcasm and commenters' intent when annotating, but disagreement about intent is still possible.

A comment only crowdworkers struggled to get agreement on, is a complex comment on the initiative to use leaf plates. The commenter makes a multi-sentence argument: *'This makes no sense. A ceramic plates using hot water from a zero carbon source would last millions of cycles where as these leaf plates require some kind of glue from an outside sources. I doubt these lasted long and how do they preserve the leafs autumn when all the leafs on the trees have disintegrated away.'*. Crowd annotators struggle to obtain agreement, but experts are correctly able to parse that this does mean the commenter expresses that the environment is threatened (the need to save trees and reduce carbon).

5 Discussion

Our pilot study gave several insights. Firstly, we note that irony is a specific issue in argument annotation. This has been noted by earlier work integrating social science theory in computational argumentation studies, e.g. in a tutorial on the topic by Lapesa et al. (2024). Lauscher et al. (2020) also find in annotation experiments for argument quality that even experts struggle with annotating and interpreting irony in arguments when annotating with complex theories.

Secondly, using theories in natural language processing can also help connecting a theory to a phenomenon, and finding gaps between these (Radford and Joseph, 2020). Responsibility is a dimension which is not part the VBN theory or of our annotation pilot, but in annotation we found it was a clear dimension in the debate: in multiple discussions, commenters mentioned that while they supported the initiative (e.g. nudging people to produce less waste), they felt others (either the rich elite, people in China or America, or companies) were mainly responsible for climate-related problems. These comments are in line with a different, but not mutually exclusive, theory about climate debate: that of *social identity theory*, were people feel pushed to blame outside groups (Post et al., 2019). This connection may be interesting for future work on sustainability and stance.

Another question is whether the VBN theory applies to other debate topics. We note that 'power' and 'threat' may relate to stances in especially other policy-related debates. However, the two dimensions in this theory are also different, and it seems the threat dimension is more applicable to debates on debates that feel existential (e.g. *is immigration a threat?*). The power dimension (*Do we have power to restore the desired state?*), is more related to feeling whether people have influence on the outcome with their own actions, which is more applicable to debates with a central role for individual action, i.e. donating money, or voting.

Topic-specific aspects also exist beyond sustainability. Another debate topic is COVID-19 policies, popular in stance detection research (Hossain et al., 2020; Glandt et al., 2021). Topic-independent pro/con stances ignores the COVID19-specific issue of whether people disagree because the measure is too strict, or not strict enough. Without this topic-specific aspect, there are limitations to interpreting stances in this debate (Scott et al., 2021).

6 Conclusion

We propose to integrate topic-specific social science theories in stance detection, improving some weaknesses of topic-independent conceptualizations of stance detection. As a case study, we use Value-Belief-Norm theory (Stern et al., 1999) for stances on sustainability, and apply this theory to a pilot annotation task on 91 comments in our *Reddit European Sustainability Initiatives corpus*. The aspects are difficult to annotate, but experts annotate some difficult examples better than crowdworkers. Topic-specific theories improve stance understanding - for both models and humans.

Limitations

We identified several limitations of our study that may lead to our results not being representative beyond this study. We invite future work to improve on these limitations.

Small Sample Due to time as well as funding constraints, our annotated sample is somewhat small, with 91 comments on 86 unique sustainability initiatives. Future work may address this concern by increasing the size of the data, both in size (a larger dataset) and in scope (more topics, language, and contexts, see below).

Only One Debate Topic This work is limited by only analyzing our proposed approach to one overarching discussion topic: that of sustainability initiatives. Our findings may not generalize well to other debate topics.

Only One Language and Debate Context Additionally, this topic and our dataset is limited to not only one language (English) (Bender, 2019) but also one socio-cultural context (Europe-focused online debates). This may mean our findings do not generalize well to user-generated textual debate in other contexts. Similarly, we analyze debates on Reddit.com, which is a very specific debate context: its norms, nuances, and specifics (Proferes et al., 2021) may make results on this data not applicable to other platforms.

Online Stance not Representative of Offline Opinions The detection of online stances is often used to predict stances of people in offline settings. However, research has shown that this has limited validity: Joseph et al. (2021) find a limited connection between people's survey responses and the same individuals' online stance-taking on social media. This may also mean that theories on offline stance-taking may not connect well to stance-taking behaviour on online platforms, as these debate contexts (online debate measurement vs offline questioning) lead to different outcomes of opinion measurement even for the same participants, which may lead to different conclusions about the debate from researchers in the social sciences than from computational researchers. We therefore also caution against any research using stance models as the sole measurement of public or individual opinion.

Ethics Statement

The data used in this project was scraped from Reddit in December 2022 with the PushShift API, before Reddit's PushShift API restrictions from April 2023 onwards. We ensure the data is released for non-commercial use only. This is also in-line with Reddit users' concern of their data being used for training commercial LLMs or other technology.

This paper concerns debate on sensitive, political topics. We completed an ERB check from the Social Science department at the Vrije Universiteit Amsterdam, which indicated we could proceed with our scraping and analyses without harm. We encourage other authors to also seek approval and a check on ethical and legal concerns before proceeding with scraping or analyzing data. We do not process identifying information on users such as usernames or post history, and neither do we release such data.

Additionally, we employ human annotators during our study. We are aware of the power dynamics and precarity involved in annotation platform work, but found it necessary for our study. We paid our annotators a fair wage, used fair attention and test tasks, and paid all annotators completing the task.

Acknowledgements

Alessandra Polimeno's contributions as well as crowd annotator costs were funded by a research voucher grant for the project *Reasons for online (dis)trust in sustainable initiatives* awarded to Myrthe Reuver and Ana Isabel Lopes by the Network Institute at the Vrije Universiteit Amsterdam. Myrthe Reuver and Antske Fokkens were also funded by the Netherlands Organisation for Scientific Research (NWO) through the the *Rethinking News Algorithms* project via the Open Competition Digitalization Humanities & Social Science grant (406.D1.19.073).

We thank lab members at CLTL for participating in the expert annotator sessions, and feedback sessions on earlier versions of this paper.

A preliminary version of this paper was presented at the non-archival International Communication Association (ICA) 2023 conference. We would like to thank all reviewers, whose comments improved this version and earlier versions of this paper. All remaining errors or unclarities are ours.

References

Yamen Ajjour, Johannes Kiesel, Benno Stein, and Martin Potthast. 2023. Topic ontologies for arguments. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1411–1427, Dubrovnik, Croatia. Association for Computational Linguistics.

Emily Allaway, Malavika Srikanth, and Kathleen Mckeown. 2021. Adversarial learning for zero-shot stance detection on social media. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4756–4767.

Sören Bär, Laura Korrmann, and Markus Kurscheidt. 2022. How nudging inspires sustainable behavior among event attendees: A qualitative analysis of selected music festivals. *Sustainability*, 14(10):6321.

Nina Bauwelinck and Els Lefever. 2020. Annotating topics, stance, argumentativeness and claims in dutch social media comments: a pilot study. In *Proceedings of the 7th Workshop on Argument Mining*, pages 8–18.

Tilman Beck, Andreas Waldis, and Iryna Gurevych. 2023. Robust integration of contextual information for cross-target stance detection. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 494–511, Toronto, Canada. Association for Computational Linguistics.

Anya Belz, Craig Thomson, Ehud Reiter, Gavin Abercrombie, Jose M Alonso-Moral, Mohammad Arvan, Jackie Cheung, Mark Cieliebak, Elizabeth Clark, Kees van Deemter, et al. 2023. Missing information, unresponsive authors, experimental flaws: The impossibility of assessing the reproducibility of previous human evaluations in nlp. *arXiv preprint arXiv:2305.01633*.

Emily Bender. 2019. The# benderrule: On naming the languages we study and why it matters. *The Gradient*, 14.

Thomas Clark, Costanza Conforti, Fangyu Liu, Zaiqiao Meng, Ehsan Shareghi, and Nigel Collier. 2021. Integrating transformers and knowledge graphs for Twitter stance detection. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 304–312, Online. Association for Computational Linguistics.

Johannes Daxenberger, Steffen Eger, Ivan Habernal, Christian Stab, and Iryna 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.

Benjamin D Douglas, Patrick J Ewell, and Markus Brauer. 2023. Data quality in online human-subjects research: Comparisons between mturk, prolific, cloudresearch, qualtrics, and sona. *Plos one*, 18(3):e0279720.

Tim Draws, Oana Inel, Nava Tintarev, Christian Baden, and Benjamin Timmermans. 2022. Comprehensive viewpoint representations for a deeper understanding of user interactions with debated topics. In *ACM SIGIR Conference on Human Information Interaction and Retrieval*, pages 135–145.

Aris Fergadis, Dimitris Pappas, Antonia Karamolegkou, and Harris Papageorgiou. 2021. Argumentation mining in scientific literature for sustainable development. In *Proceedings of the 8th Workshop on Argument Mining*, pages 100–111.

Kyle Glandt, Sarthak Khanal, Yingjie Li, Doina Caragea, and Cornelia Caragea. 2021. Stance detection in covid-19 tweets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Long Papers)*, volume 1.

Miha Grčar, Darko Cherepnalkoski, Igor Mozetič, and Petra Kralj Novak. 2017. Stance and influence of twitter users regarding the brexit referendum. *Computational social networks*, 4:1–25.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Marcus Hansen and Daniel Hershcovich. 2022. A dataset of sustainable diet arguments on Twitter. In *Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*, pages 40–58, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Richard J Hewitt, Nicholas Bradley, Andrea Baggio Compagnucci, Carla Barlagne, Andrzej Ceglarz, Roger Cremades, Margaret McKeen, Ilona M Otto, and Bill Slee. 2019. Social innovation in community energy in europe: A review of the evidence. *Frontiers in Energy Research*, 7:31.

Tamanna Hossain, Robert L. Logan IV, Arjuna Ugarte, Yoshitomo Matsubara, Sean Young, and Sameer Singh. 2020. COVIDLies: Detecting COVID-19 misinformation on social media. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

Kenneth Joseph, Sarah Shugars, Ryan Gallagher, Jon Green, Alexi Quintana Mathé, Zijian An, and David Lazer. 2021. (mis)alignment between stance expressed in social media data and public opinion surveys. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 312–324, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jonathan Kobbe, Ines Rehbein, Ioana Hulpuș, and Heiner Stuckenschmidt. 2020. Exploring morality in

argumentation. In *Proceedings of the 7th Workshop on Argument Mining*, pages 30–40, Online. Association for Computational Linguistics.

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

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Gabriella Lapesa, Eva Maria Vecchi, Serena Villata, and Henning Wachsmuth. 2024. Mining, assessing, and improving arguments in NLP and the social sciences. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024): Tutorial Summaries*, pages 26–32, Torino, Italia. ELRA and ICCL.

Anne Lauscher, Lily Ng, Courtney Napoles, and Joel Tetreault. 2020. Rhetoric, logic, and dialectic: Advancing theory-based argument quality assessment in natural language processing. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4563–4574.

Anne Lauscher, Henning Wachsmuth, Iryna Gurevych, and Goran Glavaš. 2022. Scientia potentia est—on the role of knowledge in computational argumentation. *Transactions of the Association for Computational Linguistics*, 10:1392–1422.

John Lawrence and Chris Reed. 2020. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context dependent claim detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1489–1500, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Bin Liang, Yonghao Fu, Lin Gui, Min Yang, Jiachen Du, Yulan He, and Ruifeng Xu. 2021. Target-adaptive graph for cross-target stance detection. In *Proceedings of the Web Conference 2021*, pages 3453–3464.

Sara Marjanovic, Karolina Stańczak, and Isabelle Augenstein. 2022. Quantifying Gender Biases Towards Politicians on Reddit. *PLoS ONE*. To appear.

Arya D McCarthy and Giovanna Maria Dora Dore. 2023. Theory-grounded computational text analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1586–1594.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Lynnette Hui Xian Ng and Kathleen M Carley. 2022. Is my stance the same as your stance? a cross validation study of stance detection datasets. *Information Processing & Management*, 59(6):103070.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, Doha, Qatar. ACL.

Senja Post, Katharina Kleinen-von Königslöw, and Mike S Schäfer. 2019. Between guilt and obligation: Debating the responsibility for climate change and climate politics in the media. *Environmental Communication*, 13(6):723–739.

Nicholas Proferes, Naiyan Jones, Sarah Gilbert, Casey Fiesler, and Michael Zimmer. 2021. Studying reddit: A systematic overview of disciplines, approaches, methods, and ethics. *Social Media+ Society*, 7(2):20563051211019004.

Jason Radford and Kenneth Joseph. 2020. Theory in, theory out: The uses of social theory in machine learning for social science. *Frontiers in Big Data*, 3.

Myrthe Reuver, Antske Fokkens, and Suzan Verberne. 2021a. No NLP task should be an island: Multidisciplinarity for diversity in news recommender systems. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, pages 45–55, Online. Association for Computational Linguistics.

Myrthe Reuver, Suzan Verberne, and Antske Fokkens. 2024. Investigating the robustness of modelling decisions for few-shot cross-topic stance detection: A preregistered study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9245–9260, Torino, Italia. ELRA and ICCL.

Myrthe Reuver, Suzan Verberne, Roser Morante, and Antske Fokkens. 2021b. Is stance detection topic-independent and cross-topic generalizable?-a reproduction study. In *Proceedings of the 8th Workshop on Argument Mining*, pages 46–56.

Kristen Scott, Pieter Delobelle, and Bettina Berendt. 2021. Measuring shifts in attitudes towards covid-19 measures in belgium. *Computational Linguistics in the Netherlands Journal*, 11:161–171.

Paul C Stern, Thomas Dietz, Troy Abel, Gregory A Guagnano, and Linda Kalof. 1999. A value-belief-norm theory of support for social movements: The case of environmentalism. *Human ecology review*, pages 81–97.

Terne Sasha Thorn Jakobsen, Maria Barrett, and Anders Søgaard. 2021. Spurious correlations in cross-topic argument mining. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 263–277, Online. Association for Computational Linguistics.

Eva Maria Vecchi, Neele Falk, Iman Jundi, and Gabriella Lapesa. 2021. Towards argument mining for social good: A survey. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1338–1352.

Penghui Wei and Wenji Mao. 2019. Modeling transferable topics for cross-target stance detection. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1173–1176.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Chang Xu, Cecile Paris, Surya Nepal, and Ross Sparks. 2018. Cross-target stance classification with self-attention networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 778–783.

Bowen Zhang, Min Yang, Xutao Li, Yunming Ye, Xiaofei Xu, and Kuai Dai. 2020. Enhancing cross-target stance detection with transferable semantic-emotion knowledge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3188–3197, Online. Association for Computational Linguistics.

Appendix

A Data Scraping

We identified relevant discussions in the reddit boards (sub-communities) europe, europeanunion, and europes and define a list of 10 keywords, then extend it with with word2vec embeddings (Mikolov et al., 2013) of the Google News corpus and the Glove embeddings (Pennington et al., 2014) on the GigaWord corpus. This process led to a keyword list of 38 words: ["climate change", "climate goals "," climate activists", "climate top", "climate target", "climate crisis", "climate crises ", "climate protesters", "sustainable", "sustainability", "carbon emissions", "co2 emissions", "green energy", "green shift ", "green energy", "global warming", "global temperature", "circular economy", "recycling", "recycle", "recyclables", "recyclable", "e-waste", "waste disposal", "landfills", "landfilling", "landfill", "carbon neutrality", "carbon neutral", "biodiversity", "biodiversity conservation", "biodiversity loss", "deforestation", "desertification", "renewable energy", "ecology threats", "ecology protection", "ecology-friendly"]

We scraped discussions from 2017 to 2022 with these keywords using the Pushshift Reddit API. We filter comments of bots (common on Reddit for automatic moderation) by means of a regular expression and rule-based method (Marjanovic et al., 2022), and remove empty or deleted discussions.

B Annotation Details

Crowd Annotation set-up We annotate stance of the comment towards the Reddit topic text in [comment - topic text] pairs. Stance can be *SUPPORT*, *REJECT*, or *NEUTRAL* towards the initiative in the topic text.

When the comment expresses a stance, we add two dimensions: *threat* and *power*. These aspects also have three classes: absence (no mention of this aspect in the stance), positive presence, and negative presence. Positive for threat means explicit recognition of the initiative reacting to a threat. Negative presence of threat means that the comment explicitly mentions the initiative does **not** react to a threat. Positive for power means that the comment mentions feeling power to alleviate this threat. Negative presence of power means explicitly expressing a lack of power on the issue.

We use a simple task design. First, annotators decide whether the topic text contains sustainability action, initiative or statement one can agree or disagree with.[7] Then, they annotate the stance of comments towards these initiatives. Lastly, the 68[8] comment-topic pairs determined to have a sustainability by the majority were annotated for the threat and power dimension.

The authors of this paper annotated 13 examples, with 7 used as training material for annotators and 6 used as quality check items during the task. To assure data quality, the task contained 2 attention checks per batch of around 20 items.

Task Design and Format Our task design used a Qualtrics survey adapted to ask the same questions over different texts with a Loop & Merge Field, in a random loop for each participant. Two attention checks early in the task removed participants not reading the task items, which removed one participant in the threat & power task.

Increasing data quality was achieved with 5 random expert-annotated items interspersed through the annotation task, with reminders of reasoning behind annotation decisions provided.

The task flow was as follows: 3 instruction slides, then 5 annotation blocks with 8 to 25 items, each followed by an attention item. The Qualtrics template is released in our GitHub repository[9], both as word file and as .qsf file ready to import into Qualtrics. We release these files inspired by research on the (non) reproducibility of human evaluation & annotation tasks noted by Belz et al. (2023).

C Analysis of Corpus

C.1 Methods

SentenceBERT Clustering Our initial exploratory analysis consisted of exploring clusters of arguments in order to identify the main topics being discussed in the Reddit Communities. We use the SentenceBERT architecture (embedding texts in a shared dimensional space) with MiniLM version 2 as pre-trained embeddings, with batch size 64.

Our initial clustering algorithm was the basic Community Detection embedded into SentenceBERT. We set this to a minimum community size

[7] A narrow definition of policy claim / debate topic such as "X should Y" (Daxenberger et al., 2017) does not capture the real-world stance-taking reactions people show online to utterances such as questions, announcements of protests, and quotes on sustainability.

[8] There were 71 items in total, but 3 items had an annotation error in the threat/power task.

[9] https://github.com/myrthereuver/TopicSpecific_Stance_SocialScience

of 50, and indicated that communities should have a cosine similarity threshold of at least .60. Comments not within this boundary are discarded. This divides up the large embedding space with 46.285 arguments into 25 clusters.

BERTopic Our second, more extensive exploratory analysis consists of BERTopic (Grootendorst, 2022), a BERT-based topic model technique based on Huggingface Transformers (Wolf et al., 2020). This out-of-the-box approach uses the SentenceBERT bi-encoding approach outlined above to embed sentences, and adds HBDSCAN as clustering algorithm and UMAP as dimensionality reduction to create and unsupervised clustering approach. The clusters receive "labels" that function as topic names with TF-IDF weighting of most prominent words per cluster. BERTopic is slightly non-deterministic due to the UMAP dimensionality reduction algorithm having a stochastic aspect: however, we found our results to be relatively stable across 3 runs due to the more deterministic results of both SentenceBERT text representation as well as HBDSCAN clustering.

C.2 Results

SBERT + miniLM The input for our clustering analysis were the 46.285 comments found after our preprocessing procedure, and the goal was to find whether there were broad trends and themes in comments. Cluster size varies between 1.549 texts (Cluster 1) and 50 texts (cluster 25). Note that these are only groups that have large enough clusters to all fall within a cosine similarity boundary of .60. A manual inspection of clusters shows that many of these clusters are specific topics and argument types. The largest cluster (1.549 comments) identifies a group of similar comments on renewable energy and specifically nuclear energy as a solution. The second-largest cluster (526 elements) instead focusses on discussions and comments on China versus the west when it comes to CO2 emissions per capita. Another cluster finds all comments related to recycling and waste use, and interestingly does so from various different discussions, also discussions nuclear energy where commenters mention nuclear waste. A smaller cluster (76 texts) focusses on the difference between weather and climate. More detailed results can be found in our GitHub repository. [10]

BERTopic Our second preliminary analysis consisted of a BERTopic model. This model allows us to see broad trends and themes across the discussions. The input for our BERTopic model were the 2.073 individual discussions found in our preprocessing step, to see whether the discussions could be grouped into broader themes. The BERTopic model identified 19 topics. The outlier group (573 discussions) consisted mostly of general discussions on climate change and co2 emissions, and because of its lower semantic coherence should not be considered in further analysis (Grootendorst, 2022). The largest cluster (127 discussions) was one on recycling, waste, and landfills, and another large group (127 discussions) discussed student protests and activists. Most topics consisted of broader themes such as heatwaves and increased hot weather in summer (35 discussions), or a broad initiative like the circular economy (31 discussions), but smaller clusters sometimes discussed very specific incidents in the news, such as a Norwegian ban on palm oil (27 discussions) and a courtcase against Shell in the Netherlands (26 discussions). These two incidents seemed to attract attention in the discussion boards.

Brief Qualitative Analysis Our annotation process as well as clustering experiments found a variety of reasons why people agreed or disagreed with sustainable initiatives, indicated by the different topics brought up in the discussion. Clustering results indicate that a basic pro/con stance analysis of arguments in sustainable discussions does not do justice to the actual discussion - commenters mention many different aspects of arguments, even the same argument aspects (waste, activism) across different topics and stances in these discussions.

BERTopic models allowed us to find prominent sustainability discussions. One finding is that discussions on activism and activists as well as protests are relatively common. We also found this during our annotation process, so much so that we added "activists" as an actor of sustainability initiatives. Additionally, we found that some specific initiatives in the news (a ban on palm oil and a court case against shell) attracted more comments than others.

[10] https://github.com/myrthereuver/TopicSpecific_Stance_SocialScience