Comprehensive viewpoint representations for a deeper understanding of user interactions with debated topics

Draws, Tim; Inel, Oana; Tintarev, Nava; Baden, Christian; Timmermans, Benjamin

DOI
10.1145/3498366.3505812

Publication date
2022

Document Version
Final published version

Published in
CHIIR 2022 - Proceedings of the 2022 Conference on Human Information Interaction and Retrieval

Citation (APA)
Draws, T., Inel, O., Tintarev, N., Baden, C., & Timmermans, B. (2022). Comprehensive viewpoint representations for a deeper understanding of user interactions with debated topics. In *CHIIR 2022 - Proceedings of the 2022 Conference on Human Information Interaction and Retrieval* (pp. 135-145). (CHIIR 2022 - Proceedings of the 2022 Conference on Human Information Interaction and Retrieval). Association for Computing Machinery (ACM). https://doi.org/10.1145/3498366.3505812

Comprehensive Viewpoint Representations for a Deeper Understanding of User Interactions With Debated Topics

Tim Draws
Delft University of Technology
Delft, The Netherlands
t.a.draws@tudelft.nl

Oana Inel
Delft University of Technology
Delft, The Netherlands
o.inel@tudelft.nl

Nava Tintarev
Maastricht University
Maastricht, The Netherlands
n.tintarev@maastrichtuniversity.nl

Christian Baden
The Hebrew University of Jerusalem
Jerusalem, Israel
c.baden@mail.huji.ac.il

Benjamin Timmermans
IBM
Amsterdam, The Netherlands
b.timmermans@nl.ibm.com

ABSTRACT

Research in the area of human information interaction (HII) typically represents viewpoints on debated topics in a binary fashion, as either *against* or *in favor of* a given topic (e.g., the feminist movement). This simple taxonomy, however, greatly reduces the latent richness of viewpoints and thereby limits the potential of research and practical applications in this field. Work in the communication sciences has already demonstrated that viewpoints can be represented in much more comprehensive ways, which could enable a deeper understanding of users' interactions with debated topics online. For instance, a viewpoint's *stance* usually has a degree of strength (e.g., mild or strong), and, even if two viewpoints support or oppose something to the same degree, they may use different *logics of evaluation* (i.e., underlying reasons). In this paper, we draw from communication science practice to propose a novel, two-dimensional way of representing viewpoints that incorporates a viewpoint's stance degree as well as its logic of evaluation. We show in a case study of tweets on debated topics how our proposed viewpoint label can be obtained via crowdsourcing with acceptable reliability. By analyzing the resulting data set and conducting a user study, we further show that the two-dimensional viewpoint representation we propose allows for more meaningful analyses and diversification interventions compared to current approaches. Finally, we discuss what this novel viewpoint label implies for HII research and how obtaining it may be made cheaper in the future.

CCS CONCEPTS

• **Information systems** → **Document representation**.

KEYWORDS

viewpoint, stance, label, debated topic, crowdsourcing

ACM Reference Format:
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 *Proceedings of the 2022 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '22), March 14–18, 2022, Regensburg, Germany*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3498366.3505812

CHIIR '22, March 14–18, 2022, Regensburg, Germany

ACM ISBN 978-1-4503-9186-3/22/03.
https://doi.org/10.1145/3498366.3505812

1 INTRODUCTION

Increasing amounts of human information interaction (HII) research now concern viewpoints on debated topics. For example, attitude change in web search [2, 17, 21, 42, 46, 59], fake news [9, 45, 55], and interactions with debated topics on social media or the web in general [36, 37] have been recent subjects of study. This type of research has followed calls for combating bias on the web [7, 43] and makes important contributions towards a diverse, enriching, and safe web experience for users.

A yet unresolved but essential question is how to *represent* viewpoints on debated topics. For instance, when studying attitude change in web search, each user and search result needs to receive some label that reflects the viewpoint they hold or express [17]. Earlier work in HII has predominantly done this by assigning binary or ternary viewpoint labels (e.g., *against*/*in favor*, *democrat*/*centrist*/*republican*) [24, 46, 61]. Such labels can broadly categorize viewpoints while allowing for cheap computation of metrics and algorithms (e.g., related to ranking fairness [16, 60, 62]). Moreover, they are relatively easy to obtain via crowdsourcing [38, 39] or automatic *stance detection* methods [1, 49, 58].

Despite their advantages, binary and ternary viewpoint representations reduce viewpoints to members of extremely broad categories. Recent research in the communication sciences has argued that viewpoints are complex constructs with multiple dimensions and can vary in a plurality of ways [4–6, 8]. For example, consider a set of tweets related to the feminist movement. Whereas one tweet may strongly favor feminism, another tweet may only slightly support it. Merely classifying such documents as either *against* or *in favor* removes any notion of a *degree* to which a viewpoint may oppose or support a topic. Moreover, two tweets may support feminism for different reasons, such as the morality of treating women and men equally or the economic benefits of empowering women. This latent richness of viewpoints is almost entirely lost when classifying subjects using a binary or ternary scheme.

Obtaining a deep understanding of user interactions with debated topics may require information such as whether a user moved

from "strongly opposing" to "somewhat opposing" feminism after a web search session. Similarly, interventions for diverse news reading could more effectively expose users to alternative perspectives when knowing which reasons for opposing or supporting a topic the user has already considered. Enabling such advanced analyses could unlock greater potential for research and practical applications in HII. To this end, we argue that more comprehensive viewpoint representations are needed.

Recent research in HII has already begun to represent viewpoints in alternative formats. For instance, viewpoints have been represented on ordinal scales [16, 17], continuous scales [30], or as topic-specific perspectives [12, 14]. Mulder et al. [40] drew from the communication sciences to operationalize *framing*, a concept that represents viewpoints in four different dimensions (i.e., *problem definition*, *causal attribution*, *moral evaluation*, and *treatment recommendation*). They used different automatic methods to compute a distance function that considers these four dimensions to gauge the viewpoint similarity between news articles. This earlier work shows – albeit operationalized by a distance function instead of a label – that the richer notions of viewpoints handled in the communication sciences can be practically applied in HII. However, to the best of our knowledge, no currently existing method translates comprehensive viewpoint representations into practical viewpoint labels applicable to user interactions with debated topics. HII lacks a standard, go-to framework that is easy to use but significantly more comprehensive than currently used methods. This paper aims to fill this research gap. Four research questions guide our work:

- **RQ1.** What label represents viewpoints in a comprehensive yet relatively simple and topic-independent fashion?
- **RQ2.** Can crowd workers reliably assign our proposed viewpoint label to textual documents?
- **RQ3.** Do cognitive biases affect crowd workers when assigning our proposed viewpoint label?
- **RQ4.** Is our proposed viewpoint representation more meaningful compared to binary viewpoint labels?

To address **RQ1**, we drew from work in the communication sciences and developed a topic-independent, two-dimensional viewpoint representation that incorporates a viewpoints' *stance* (i.e., the degree to which it supports or opposes a claim) and *logic of evaluation* (i.e., its perspective or underlying reason; see Section 3). We then tasked crowd workers to assign our novel viewpoint label to tweets on several debated topics. Analyses of this crowdsourcing task suggest that crowd workers can perform this task reliably (**RQ2**), and there was no evidence that cognitive biases would have affected the results (**RQ3**; see Section 4). We further demonstrate in a qualitative viewpoint diversity analysis of the tweets that our proposed viewpoint label leads to more meaningful insights compared to a standard approach such as a binary *against/in favor* viewpoint label (**RQ4**; see Section 5). Finally, we report on a user study where participants saw sets of tweets, diversified either based on our proposed viewpoint representation or a binary viewpoint label. Exploratory results of this user study suggest that users judge sets of tweets as more viewpoint-diverse when the sets are diversified based on our proposed label compared to the baseline – as long as these sets do not contain too many extreme viewpoints (**RQ4**; see Section 6).

Supplementary material related to this research (e.g., annotated data sets, task screenshots, user study material, and analysis code) is openly available at https://osf.io/pjws9/.

2 RELATED WORK

Recent years have seen a stark increase of research concerning viewpoints and debated topics in HII. Inspired by calls to combat bias on the web [7, 43], such research has explored user interactions with debated topics in web search [2, 17, 21, 25, 42, 46, 59], social media [37], and the web in general [36]. These efforts are supported by other lines of research that aim to automatically classify documents into different viewpoint categories [1, 49, 58], detect fake news [9, 45, 55], measure viewpoint-related ranking bias [16, 30], or re-rank recommended items based on viewpoint diversity [40, 54]. An essential part of research concerning debated topics in HII is how to *represent* viewpoints.

2.1 Viewpoint Representation in HII

HII research typically represents viewpoints in binary (e.g., *against/in favor*) or ternary (e.g., *democrat/centrist/republican*) fashions [24, 46, 61]. For instance, Gezici et al. [24] used *against/in support* as well as *liberal/conservative* viewpoint categories, and Yom-Tov et al. [61] classified users and documents into the political leanings *democrat*, *centrist*, or *republican*. These simple taxonomies allow for extensive research concerning user interactions with debated topics as they enable cheap computation of metrics and algorithms. Moreover, they are comparatively easy to obtain using crowdsourcing, which also forms the basis for automatic *stance detection* methods [1, 32, 39, 49, 52, 58] (see Section 2.3).

Recent work in HII has explored alternatives to binary viewpoint representations; e.g., by representing stances on ordinal [16, 17] or continuous scales [30]. However, despite adding more nuance to the *against/in favor* dichotomy, such labels are still lacking crucial information about the underlying reasons of viewpoints (e.g., a moral perspective of gender equality). This notion of *perspective* as a dimension next to a viewpoint's stance has already been explored [12, 14] but often faced the limitation of these perspectives being highly topic-dependent (e.g., the debated topics *atheism* and *feminist movement* have vastly different perspective spaces).

2.2 Viewpoint Representation in the Communication Sciences

Viewpoint diversity in public discourse is a long-standing subject of study in the communication sciences [4–6, 33–35, 47, 48, 56] that has already been applied to information access systems [26, 27, 41, 57]. Compared to HII, the communication sciences have also brought forward more advanced viewpoint representations. There, for instance, a common way to explore viewpoints is *framing*, whereby a viewpoint is usually analyzed on four different dimensions: *problem definition* (i.e., what is happening), *causal attribution* (i.e., who is responsible for the problem), *moral evaluation* (i.e., whether the problem is good or bad), and *treatment recommendation* (i.e., suggestions in response to the problem) [20].

More recent work has combined framing with the notion of *interpretative repertoires* to propose a topic-independent way of representing viewpoints [5, 6]. In this method, each *frame* (i.e., a

Table 1: The seven logics of evaluation we consider for our proposed viewpoint label, adapted from Baden and Springer [5]. Each logic represents a particular orientation of what is desired and can be used to either support or oppose a given claim.

Logic of evaluation	Good is…	Examples
Inspired	…what is true, divine, and amazing	Righteous, pre-ordained, beautiful; false, uncreative, dull
Popular	…what is popular or what the people want	Preferred, popular, favourite; resented, feared, isolated
Moral	…what is social, fair, and moral	Solidary, responsible, just; inhumane, asocial, egoistic
Civic	…what is legal, accepted, and conventional	Legal, agreed, common; scandalous, deviant, inappropriate
Economic	…what is profitable and creates value	Beneficial, economic, affordable; wasted, costly, unproductive
Functional	…what works	Effective, necessary, quick; dysfunctional, inefficient, useless
Ecological	…what is sustainable and natural	Sustainable, organic; unnatural, irreversible

viewpoint based on the four dimensions mentioned above) is seen as an instance of a more general way of interpreting the world (i.e., the interpretative repertoire). Building on the idea of "common worlds" proposed by Boltanski and Thévenot [8], Baden and Springer [5, 6] view frames as commensurable if they refer to the same repertoire commonly used in argumentation (e.g., referring to belief systems, morality, or economic factors). For example, consider the phrases "*feminism is on the rise because women should be treated equally*" and "*stop attacking feminists, they are the ones who fight for fair treatment*". These two phrases express different frames but have the same *logic of evaluation* (i.e., good is what is social, fair, and moral). This logic of evaluation is a key aspect of interpretative repertoires and offers a topic-independent way to represent perspectives behind the stances of viewpoints (see Table 1).

A drawback of analyzing viewpoints using framing or interpretative repertoires is that it usually requires a trained expert who performs manual annotation. This is impractical for HII and related fields that need to obtain viewpoint representations at scale to enable cheap computation of metrics and algorithms. Although first attempts have been made to analyze the viewpoint diversity of content in hybrid [4] or automatic ways [40], to the best of our knowledge, no currently existing method can reliably and cheaply obtain viewpoint labels that at least approximate the comprehensiveness of those typically handled in the communication sciences.

2.3 Annotation of Viewpoint Representation

Extensive experiments have been performed to collect binary viewpoint annotations on news articles and tweets, by means of expert annotators or through crowdsourcing [13, 23, 32, 39, 52]. In addition, labels such as *neutral*, *neither in favor nor against*, or *I don't know* are used to identify texts that do not take a stance, are unclear, unrelated, or ambiguous. In general, the agreement percentages and the inter-rater reliability (IRR) values are substantial. For instance, Mohammad et al. [39] report an agreement percentage of 73% regarding the stance of the tweets, while Li et al. [32] report Krippendorff's α values of 0.60 and 0.81, when considering ternary and respectively, binary representations. Burscher et al. [10] used two trained annotators to identify pre-determined *frame types* (i.e., *conflict*, *morality*, *economic consequences*, and *human-interest*) in 156 political news articles. IRR c.f. Krippendorff's α ranged from 0.21 (*morality*) to 0.58 (*economic consequence*). Thus, human annotation of viewpoint labels is feasible but its difficulty increases with label complexity. Furthermore, to the best of our knowledge, annotations for *logics of evaluation* have so far only been performed by experts in communication science and not yet by crowd annotators.

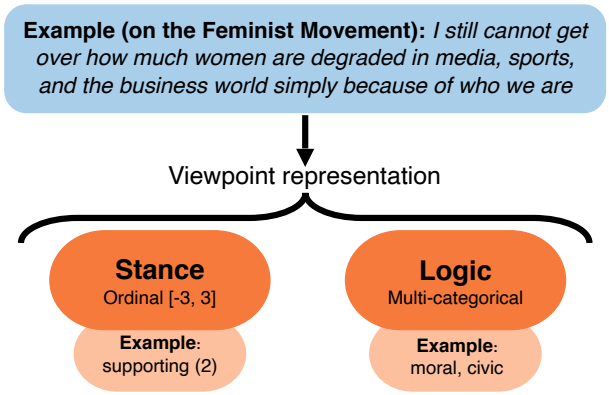

Figure 1: Proposed viewpoint representation at the example of a tweet from the *SemEval 2016 Stance Detection* data set [38]. A viewpoint is evaluated on two dimensions: *stance* (i.e., on a seven-point ordinal scale ranging from "strongly opposing" to "strongly supporting" a topic) and *logic of evaluation* (i.e., in a multi-categorical format to include all logics present; see Table 1).

3 NOVEL VIEWPOINT REPRESENTATION

We propose a novel viewpoint label for HII that improves upon binary viewpoint labels by reflecting a viewpoint's *stance* on a more nuanced level and a viewpoint's *logic of evaluation* as a second dimension (see Figure 1). Thereby, our proposed viewpoint representation is more comprehensive compared to existing methods. We detail the two dimensions of our proposed representation below.

Stance. The first dimension in our proposed viewpoint representation is a viewpoint's *stance*; i.e., its moral evaluation of the topic at hand. For example, consider the tweet displayed in Figure 1. This tweet is clearly *in favor* of the feminist movement and was therefore classified accordingly in the *SemEval 2016 Stance Detection* data set [38] (see Section 4.1). In our proposed framework, however, stances are represented on a seven-point ordinal scale ranging from "strongly opposing" (-3) to "strongly supporting" (3; adopted from Draws et al. [16, 17]). This representation reflects a viewpoint's general orientation similar to the standard binary approach but

also the *degree* to which a viewpoint opposes or supports a topic. For instance, we may label the tweet in Figure 1 as "supporting" (2), meaning that it takes a clear stance in favor of feminism but does not do so to an extreme extent.

Logic of evaluation. The second dimension of our proposed viewpoint representation is a viewpoint's *logic of evaluation* (or simply *logic*), a construct that we borrow from the communication sciences [5, 6, 8]. A viewpoint's logic of evaluation reflects the general perspective behind the stance: it describes *why* a stance is taken. For example, the statements "women should be treated fairly" and "empowering women would benefit the economy" both arguably support feminism but do so for different reasons. Whereas the first one refers to fairness (i.e., using a *moral* logic), the second one refers to value creation (i.e., using an *economic* logic). Baden and Springer [5] mention seven different logics that a viewpoint can include: *inspired*, *popular*, *moral*, *civic*, *economic*, *functional*, and *ecological* (see Table 1). Each of these seven logics represents a particular maxim according to which a problem may be evaluated. For instance, an *ecological* logic is employed when the viewpoint refers to something that is supposedly (not) sustainable or natural; e.g., opposing feminism by expressing that "equal treatment of men and women is unnatural". Classifying viewpoints into logics of evaluation thus allows for entirely topic-independent descriptions of the latent perspectives that viewpoints embody. Note that any document may refer to one or several of the seven logics. For example, the example tweet in Figure 1 refers to a *moral* logic (i.e., arguing that women are treated unfairly) and a *civic* logic (i.e., suggesting that this is not acceptable). This type of information is lacking when using a standard binary viewpoint label.

4 OBTAINING VIEWPOINT LABELS

In this section, we report on a crowdsourcing study in which we collected viewpoint labels according to our proposed framework (see Section 3) for 169 tweets from the *SemEval 2016 Stance Detection* data set [39, 53]. We describe the data, task setup, and process of collecting the annotations. Furthermore, we analyze whether workers were able to assign viewpoint labels reliably and whether they were influenced by cognitive biases when annotating.

4.1 Data

One of the most utilized data sets for stance classification is the *SemEval 2016 Stance Detection* data set, which consists of 4,870 tweets on six different debated topics: *atheism*, *climate change*, *Donald Trump*, *feminist movement*, *Hillary Clinton*, and *legalization of abortion* [39, 53]. It was originally created for the *SemEval 2016 Stance Detection Challenge* [38], which invited contributors to create automatic methods for classifying tweets into four viewpoint categories: *in favor*, *neutral*, *against*, and *none* (i.e., no viewpoint). All tweets in the *SemEval 2016 Stance Detection* data set are annotated for their viewpoint (i.e., using the same four categories) and relevance concerning the target topic.

We aimed to collect annotations according to the viewpoint representation we propose in Section 3 for a subset of the tweets contained in the *SemEval 2016 Stance Detection* data set. Specifically, we selected all 169 tweets that at least 90% of the original annotators judged as relevant to the topics *atheism* (16), *Donald Trump* (54), or *feminist movement* (99). We chose these three topics to limit expenses (i.e., allowing for more annotations per tweet) while maintaining topical diversity (i.e., they cover diverse topics such as religion, politics, and social and political movements) and relevance in online discussions and information sharing platforms.

4.2 Prior Considerations

Aside from collecting our proposed viewpoint label for the 169 tweets in our final data set, we also aimed to investigate whether cognitive biases can affect crowd workers when assigning these labels. Draws et al. [15] propose a 12-item checklist to document, assess, and mitigate cognitive biases in crowdsourcing. We applied this checklist and concluded that two different cognitive biases might affect crowd workers in our task. First, we were concerned about a *halo effect*, in which irrelevant pieces of information affect crowd workers' annotations. We were particularly concerned that crowd workers with pre-existing solid knowledge on the topic at hand might rate viewpoints as more extreme (i.e., more readily placing tweets into the "opposing camp" or "supporting camp"). Second, we suspected that the *confirmation bias* could affect crowd workers if they had a tendency to label tweets in line with their personal stance (i.e., looking for attitude-confirming evidence). We thus decided to incorporate measurements of personal knowledge and stance concerning the given topic in our task design.

4.3 Task Setup

We designed a human intelligence task (HIT) to obtain viewpoint annotations in our proposed format. A research ethics committee at our institution had approved the task before data collection and all crowd workers agreed to an informed consent. First, crowd workers were presented with one of the three topics (i.e., *atheism*, *Donald Trump*, or the *feminist movement*) and asked for their personal knowledge and stance on it (see Section 4.2). We measured these constructs on seven-point Likert scales ranging from "non-existent" to "expert" (knowledge) and from "strongly opposing" to "strongly supporting" (stance). Crowd workers then saw one of the 169 tweets in our data set, relevant to the same topic.

The main task for crowd workers was to evaluate the viewpoint expressed in the tweet in the three subsequent steps: they (1) described the expressed viewpoint in their own words, (2) judged its stance regarding the topic on a seven-point Likert scale ranging from "strongly opposing" to "strongly supporting", and (3) selected which logic(s) applied (see Section 3). In step (3), the seven logics were displayed as completions of a sentence; e.g., *"Fundamentally, the viewpoint contained in the tweet is that Feminist Movement is (not) in line with… what is social, fair, or moral."* (i.e., indicating a moral logic, c.f., Table 1). Crowd workers could obtain more information (including examples) about any given logic by hovering over the respective option. In this last step, participants first selected the viewpoint's main logic by choosing one of the seven categories and then had the option to select any other logic that may also apply. We also added a mandatory attention check (i.e., an item where we explicitly told crowd workers which option to select) and an option to give feedback in open text form. We published the task on *Amazon Mechanical Turk* (MTurk).[1]

[1] https://www.mturk.com

4.4 Human Annotators

Crowd annotators. A total of 66 crowd annotators annotated our HITs (i.e., consisting of one tweet). They had a *Master* status on MTurk, a HIT approval rate of at least 95%, and at least 500 accepted HITs. Furthermore, we only allowed crowd workers from a selection of 30 countries that either has English as its main language (e.g., The United States) or that has high English proficiency according to the EF English Proficiency Index[2] (e.g., The Netherlands and Denmark). These constraints ensured a high-quality pool of annotators with good English understanding (i.e., our tweets are in English). Furthermore, we excluded nine annotations for which the crowd worker failed the mandatory attention check. The final sample consisted of 1197 annotations from 66 different crowd annotators. Crowd workers were allowed to submit as many HITs as they wished and were rewarded with $0.50 for each completed HIT. Each tweet received between six and eight annotations (mean = 7.08, sd = 0.30). On average, crowd workers reported a good knowledge across the three topics *atheism* (mean = 1.70, sd = 1.24), *Donald Trump* (mean = 1.91, sd = 1.05), and the *feminist movement* (mean = 1.54, sd = 1.19).[3] Regarding personal stance, they slightly supported *atheism* (mean = 0.62, sd = 1.98), opposed *Donald Trump* (mean = −1.49, sd = 1.96), and were approximately neutral towards *feminism* (mean = −0.14, sd = 2.07).

Expert annotators. To evaluate the quality of the crowd annotations, we created a ground truth data set consisting of 34 tweets (i.e., 20% of the tweets used in our study). We aimed to avoid bias by randomly selecting the tweets for each of the three topics of interest and proportional to the total number of tweets for a given topic (i.e., 4 on *atheism*, 11 on *Donald Trump*, and 19 on the *feminist movement*). First, two expert annotators, authors of the paper with a background in computer science and familiar with the logics depicted in Table 1, independently annotated the 34 tweets. The two experts annotated the 34 tweets using the same task that was provided to the crowd annotators, i.e., on MTurk. We computed the inter-rater reliability of the two experts concerning tweets' stances and logics using Krippendorff's α [29]. The reasons for choosing this metric were three-fold: it is (1) applicable on both ordinal and nominal values (i.e., our data is ordinal - stance and nominal - logics), (2) deals with missing data (not all annotators annotate all examples), and (3) generalizes to any number of annotators. Regarding stance, the two experts had a high IRR score of 0.84, while in terms of logics their agreement varied from almost no agreement (e.g., *popular*, *functional*, *inspired*, *civic*, and *financial* logics have α values below 0.07) to high agreement (e.g., 0.58 for *moral*) and perfect agreement (e.g., 1.0 for *ecological*). The two experts then discussed the annotations with a third expert who has a background in communication science (also co-author of this paper). In the discussion session, all 34 tweets in the ground truth were individually discussed until an agreement was reached regarding the applicable stances and logics.

4.5 Crowd Annotation Aggregation and Quality

As described in Section 4.3, we asked crowd annotators to judge the *stance* and the *logic(s)* of each tweet. In this section, we report on the aggregation of the crowd annotations to identify the collective stance and logics for each tweet, as well as on the quality of the annotations gathered in our crowdsourcing study.

4.5.1 Tweet Viewpoint Stance. To aggregate stance annotations, we represented the seven options from the Likert scale as integers [−3, 3] and assigned each tweet the median annotation value (i.e., rounded to integer).

Crowd annotators largely agreed on the extent to which a tweet opposes or supports a particular stance. Their IRR score on the tweet stance (c.f. Krippendorff's α) is 0.69 on the entire data set and 0.72 on the expert-annotated data set of 34 tweets. We also compared the aggregated crowd and expert stance on the tweets. In this case, the IRR score c.f. Krippendorff's α is 0.84, further emphasizing the crowd's reliability in annotating stances of tweets using our more complex representation, i.e., on an ordinal scale ranging from −3 to 3. The crowd's micro F1-score in terms of stance was 0.53 when using the ordinal scale ranging from −3 to 3 and 0.97 when using a ternary scale (against, neutral, in favor). The aggregated stance labels from crowd workers matched the stance indication contained in the original *SemEval 2016 Stance Detection* data set in 97% of cases. Five tweets that had all originally been classified as *in favor* of feminism or atheism were annotated as *neutral* (0) in our data (e.g., "*Just been putting the finishing touches to a feminist-themed cryptic crossword... Standard. #crosswords*").

4.5.2 Tweet Viewpoint Logic. Annotating logics to each tweet was the more difficult task for crowd workers, as the interpretation of logics could be somewhat subjective and ambiguous. Moreover, a given tweet may contain multiple different logics with different degrees of relevance or intensity, so attaching a single logic to each tweet is not optimal. These observations led us to analyze the crowd annotations regarding the logic(s) of the tweets with the disagreement-aware metrics called *CrowdTruth* [18, 19], which compute quality scores for input units (i.e., tweets), crowd annotators, and target annotations (i.e., the seven logics).[4] When applying the metrics, we considered the main logic as well as all additional logics that a crowd annotator selected.

The CrowdTruth metrics assume that the three main components of the crowdsourcing task (i.e., tweet, crowd annotators, and logics) are mutually dependent. For instance, a difficult tweet can make crowd annotators disagree, but this does not necessarily mean that their answers' quality is poor (i.e., annotators can fill in each others' gaps by adding logics that others have missed). Thus, c.f. the CrowdTruth metrics, the quality of a tweet is weighted by the quality of the crowd annotators that annotated the tweet and of the target annotations, i.e., the logics, and vice versa. The answers of a crowd annotator who constantly disagrees with the other crowd annotators will have a lower weight in the final aggregation of answers. These quality scores are computed in a loop, using a dynamic programming approach, until convergence. Each quality score ranges from 0 and 1, where higher values indicate higher quality or clarity.

Upon applying the CrowdTruth metrics, we thus had (1) crowd annotators quality scores, (2) tweet quality scores, and (3) tweet-logic scores. A tweet-logic score is computed for each tweet and

[2] https://www.ef.com/wwen/epi/

[3] We here represent the seven Likert points as integers on an ordinal scale [−3, 3].

[4] https://github.com/CrowdTruth/CrowdTruth-core

each logic, expressing the likelihood of the logic to be expressed by the tweet. We evaluated the crowd's performance in terms of the micro-F1 score [44], using the 34 tweets for which we collected ground truth data from expert annotators (see Section 4.4).[5] For this, we use the tweet-logic score as a threshold to differentiate between positive and negative samples (i.e., logics expressed and not expressed in a tweet). We experimented with threshold values between 0 and 1, in increments of 0.01, and computed the crowd's micro-F1 score for each such threshold. We generally observe that, the lower the threshold (i.e., considering more logics to be expressed in a tweet), the higher the crowd micro-F1 score. For example, the micro-F1 score is equal to 0.67 at a threshold of 0.01, and equal to 0.02 at a threshold of 1. Based on this analysis, we considered a threshold of 0.25 as optimal (micro-F1 = 0.61), to have a more balanced performance concerning recall and precision, and still eliminate logics that are considered applicable by only a few crowd annotators or crowd annotators with low-quality scores. The final viewpoint label per tweet thus comprised of two dimensions: the median stance annotation and a vector of all logics that passed the aforementioned threshold (see Figure 1 for an example).

Compared to stance, logic annotations generated substantially more disagreement, resulting in much lower Krippendorff's α values (0.23 or lower on both the main and the expert-annotated data set). The crowd agreed most on the *moral* and *functional* logics. When compared to the expert logics on the 34 tweets in our ground truth, we observe similar agreements as for the experts. Specifically, we found perfect agreement for the *ecological* logic, moderate to high agreement for the *moral* (Krippenforff's $\alpha = 0.58$) and *popular* ($\alpha = 0.36$) logics, and low agreement for the other logics.

4.6 Gauging the Annotation Difficulty

To better understand the difficulty of our task, we had eight different crowd workers annotate between one and 103 tweets *twice*. We ensured here that there was always a considerable amount of time and other HITs between the first and second annotation of a tweet. We found that workers were largely consistent in their two annotations of the same tweet. Overall, annotators did not diverge more than one point on the stance scale in 89% of cases and assigned precisely the same set of logics in 38% of cases. The average Jaccard distance of logics annotation pairs was 0.44, indicating that workers may often have missed or added a logic in their second annotation compared to their first, but usually annotated with some degree of overlap.[6] For example, if a worker first only assigned [*inspired*] to a tweet but annotated [*inspired, popular*] at the second time, the Jaccard distance between the two annotations was 0.5. This also shows that the low inter-rater reliability scores reported in the previous paragraph may give a somewhat misleading image regarding the task difficulty. In sum, workers were fairly consistent when annotating a tweet for the second time but may have missed certain logics that other crowd workers detected.

Checking for Cognitive Biases. As explained in Section 4.2, we tested whether specific cognitive biases (i.e., the *halo effect* and the *confirmation bias*) influence crowd workers when assigning our proposed viewpoint label. The halo effect we were concerned about would have taken place if workers' knowledge of the topic at hand had influenced the variance of their stance annotations (i.e., placing tweets in either *extremely opposing* or *extremely supporting* "camps"). However, we found no evidence of this effect from a Spearman correlation analysis between workers' self-reported knowledge on their assigned topic and the standard deviation of their stance annotations ($\rho = 0.14$, $p = 0.5$).[7] A confirmation bias in our task could have meant that crowd workers look for information that confirms their pre-existing beliefs and thus annotate stances in line with their personal stance. However, we also found no evidence for a confirmation bias from a Spearman correlation analysis between workers' self-reported stance on their assigned topic and their mean stance annotation ($\rho = 0.04$, $p = 0.8$).

5 ANALYZING VIEWPOINT DIVERSITY

This section presents a viewpoint diversity analysis of the data described in Section 4.1 using the two-dimensional viewpoint labels we collected (see Section 4). The aim of this analysis is to obtain insights into the discussions surrounding the three debated topics (i.e., *atheism*, *Donald Trump*, *feminist movement*) and to showcase the depth of understanding that our proposed viewpoint representation provides. For each topic, we analyze (1) the stance distribution, (2) the logics distribution, and (3) how the different logics relate to each other within the online discussions.

5.1 Method

We analyze the viewpoint diversity of the tweets in our data set in a qualitative fashion. Aside from their raw format, we examine the data using two different visualizations. Figure 2 shows per topic the relative frequency of the seven different logics across all tweets. We compute the relative frequency by dividing the number of tweets in which a logic appears by the total number of tweets within that topic. This provides a visual overview of the relative importance of the different logics.

We also investigate structural similarities between the logics. Figure 3 shows per topic a network plot of how similar the tweets are in terms of the logics they use. We first computed Jaccard similarity matrices of all tweets for each of the three topics based on the logics they refer to. We then created the networks using the similarity matrices as weight matrices. Each node in a network represents a single tweet and is colored according to its stance. Stronger edges indicate stronger similarities between tweets. However, to maintain a good overview, we omitted all edges with Jaccard similarities of 0.4 or lower. The networks visualize how people on different sides of the debated topics argue by showing how tweets of different stances cluster together in terms of the logics they use.

5.2 Results

Atheism. Of the 16 relevant tweets in our data set, only two were labeled as either "somewhat opposing" or "strongly opposing" atheism. The remaining 14 tweets received "neutral" (1), "supporting" (9), or "strongly supporting" (4) labels. The left-hand panel of Figure 2 shows the relative importance of the different logics that

[5] We compute micro-F1 scores because we deal with a multi-label classification problem, where logics are not equally represented across the data set. We also consider all logics equally important, and we are interested to see how the crowd performs across logics.
[6] There was no overlap concerning logics annotations in 24% of cases.
[7] To ensure independence of observations for this analysis, we here only considered one stance (on one topic) per crowd worker.

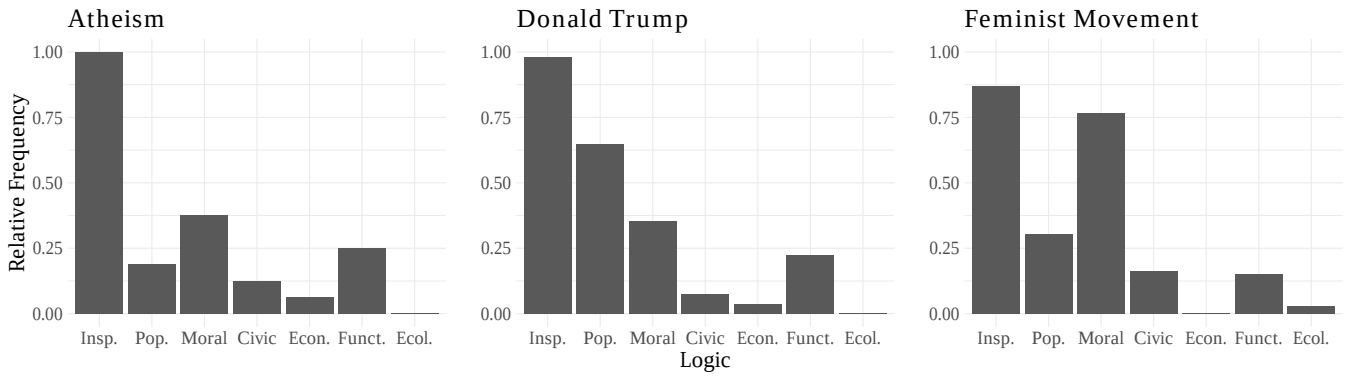

Figure 2: Relative frequency of the seven different logics across the topics *atheism*, *Donald Trump*, and *feminist movement*.

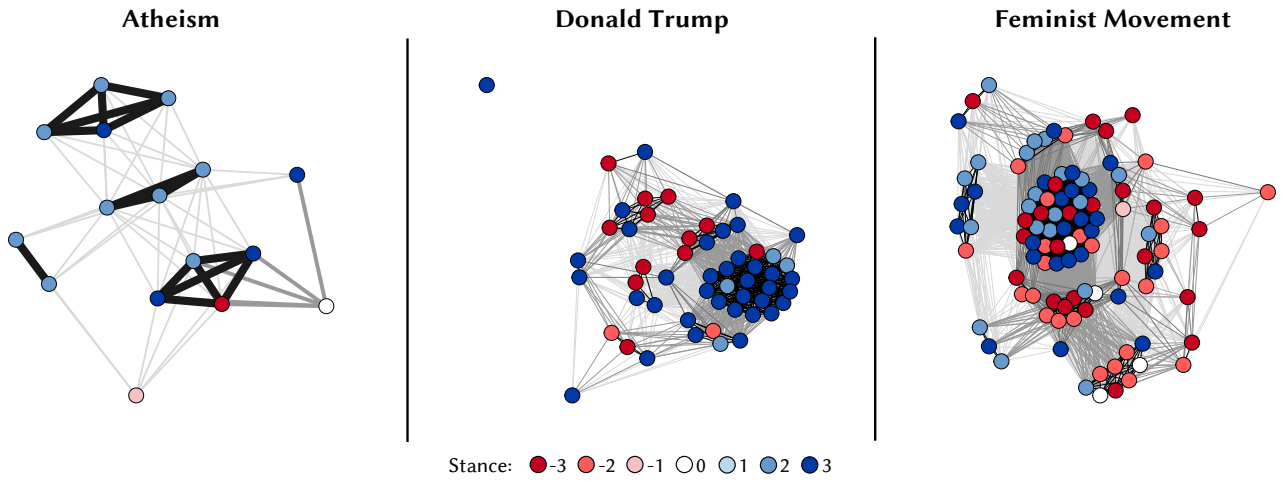

Figure 3: Network plots of all tweets divided into the three topics *atheism* (left-hand panel), *Donald Trump* (central panel), and the *feminist movement* (right-hand panel). Each node is a single tweet, whereby its color indicates the stance. Edges indicate the Jaccard similarity between tweets based on the assigned logics (i.e., the stronger the edge, the greater the similarity).

were used when discussing *atheism*. Whereas the *inspired* logic was found in every tweet, all other logics appeared in 0% to 31% of tweets. What types of viewpoints users expressed in their tweets becomes more clear when looking at the network plot in the left-hand panel of Figure 3. The network plot shows tree main clusters of at least three tweets that evaluate atheism by referring to an *inspired* logic (e.g., "*[...] which god? Yours? not mine. oh wait i don't have one. #LoveWins*"), or by combining an *inspired* logic with either a *functional* logic (e.g., "*If God = Miraculous And Miracles = Impossible Then God = Impossible #logic #reason #science #RT*") or a *moral* logic (e.g., "*Serious question for my atheist libertarians: How can rights exist without God? #ChristianLibertarian*").

Donald Trump. Our data set contains 54 tweets from 2016 that evaluated *Donald Trump*. Compared to the other topics, the discussion around Donald Trump is much more polarized, as 89% of tweets are either strongly supporting or strongly opposing Donald Trump. The barplot in the central panel of Figure 2 shows that the *inspired* and *popular* logics were used most often. Conversely, only a few tweets in our data set express viewpoints that refer to an *economic* or *ecological* logic. The network plot in the central panel of Figure 3 shows that most tweets are highly similar to each other. The largest cluster consists almost entirely of tweets that strongly support Donald Trump and represents a combination of the *inspired* and *popular* logics (e.g., "*[...] We have got to take our country back. It's time! Win it Mr. Trump*"). Tweets in a similar cluster that is almost entirely in favor of Donald Trump combine the *inspired* and *popular* logics with a *functional* logic (e.g., "*[...] Hell I'm from the UK and I believe realDonaldTrump would make an amazing WORLD Leader*"). Arguments on the opposing side, in contrast, were usually made by taking a *moral* aspect into account (e.g., "*Donald Trump needs to stop embarrassing himself. Racist assholes...*").

Feminist Movement. The majority of tweets in our data set (99) evaluate the *feminist movement*. Here, the stance distribution is comparatively balanced with 46% supporting, 4% neutral, and 50%

opposing tweets, only half of which are at the extreme ends of the stance spectrum. The barplot in the right-hand panel of Figure 2 shows that feminism was discussed using similar logics compared to the other topics, but that the *moral* logic is noticeably more important here. This is also reflected in the network plot displayed in the right-hand panel of Figure 3. The largest cluster contains tweets that combine *inspired* and *moral* logics to argue on both sides of the spectrum (e.g., "*I think it's okay for a woman to take a mans name if she wants to. #genderequality*"). Many tweets that support feminism argue exclusively using a moral logic ("*I shouldn't have to be holding a man's hand to be left alone on the street. #catcalling #streetharassment #equality*"). On the other hand, tweets opposing *feminism* tend to use the *inspired* logic more often (e.g., "*All the feminist block me because I speak true.*") and sometimes combine that with other logics such as the *popular* one (e.g., "*[...] Most feminists don't know what they are fighting for?! Most ego maniac's who want they're 15 minutes of fame. #c4news*").

6 USER EVALUATION OF VIEWPOINT LABEL

We have shown that our proposed viewpoint label is obtainable via crowdsourcing with acceptable reliability (see Section 4) and that it enables in-depth viewpoint analyses (see Section 5). It is yet unclear whether this approach can also help to create noticeably superior outcomes from the user's perspective. To test whether using our proposed viewpoint representation can more meaningfully organize online discussions (i.e., addressing **RQ4**), we conducted a user study. We presented users with sets of tweets that were diversified based on either our proposed viewpoint label or a binary viewpoint label and asked them which set was more viewpoint-diverse. The user study had been preregistered before any data collection.[8]

6.1 Method

6.1.1 Data. For this user study, we considered tweets that were part of the data set described in Section 4.1 and that related to the topic *feminist movement*. We only focused on the *feminist movement* here because the other two topics had comparably few relevant tweets and skewed stance distributions, which hindered diversification efforts. We further excluded five feminism-related tweets that had received a neutral stance label (4) or that were the only ones in their stance category (i.e., one *somewhat opposing* tweet).

6.1.2 Sets of Tweets. We assembled a total of 10 different sets of tweets from the data set described above. Each set contained six tweets on the *feminist movement* and was created using one of two different sampling algorithms. The first algorithm diversified tweets using our proposed viewpoint label: after sampling one random tweet as the first element in the set, this algorithm added the five remaining tweets by always picking the tweet with the maximum average Jaccard distance to the tweets that were already in the set. It did this in such a way that the stance distribution was as balanced as possible, i.e., including at least one of the available stance categories. The second algorithm diversified tweets based on the original binary label contained in the *SemEval 2016 Stance Detection* data set and therefore randomly sampled three *against* and three *in favor* tweets to create a set. We created five such sets per algorithm.

6.1.3 Procedure. The user study consisted of two steps. First, participants read an informed consent and stated their gender and age group (i.e., both from multiple choices). Second, we presented participants with a scenario: they were co-organizing a debating event aiming to bring people of diverse viewpoints together. It was explained that two methods are being tested to diversify the table seat allocations based on attendees' recent tweets on the feminist movement. Participants then saw two random sets of tweets (i.e., one per sampling algorithm; in random order) graphically arranged in a circle to imitate a table seat allocation (see our repository for screenshots). A border surrounding each tweet was colored red (*against*) or blue (*in favor*) depending on the tweet's stance label in the original *SemEval 2016 Stance Detection* data set. We asked participants to judge which table had a greater viewpoint diversity and shortly explain their choice in an open text field.

6.1.4 Analysis. Our hypothesis for this study was that users would judge tweet sets created with the sampling algorithm based on our proposed viewpoint label as more diverse. To test this hypothesis, we conducted a binomial test with a test value of 0.5 (i.e., testing the null hypothesis that users choose tables at random).

6.1.5 Participants. We conducted a power analysis before data collection to gauge the required number of participants for this user study. Using the software *G*Power* [22], we specified that we expect a medium effect size (i.e., Cohen's $g = 0.15$), handle a significance threshold of $\alpha = 0.05$, and aim for a statistical power of $\beta = 0.8$ in a two-tailed binomial test. This resulted in a required sample size of 90 participants, which we thus recruited from *Prolific*.[9] All participants were native English speakers above 18 years of age. We paid $0.70 per participation (an average of $10.33 per hour), while allowing each participant to only judge one pair of tweet sets.

6.2 Results

Among the 90 participants we had recruited, 58 (64%) were female, 31 (34%) were male, and one (1%) was non-binary. Participants' age distribution was somewhat skewed towards younger ages, with only 7 participants being older than 44 years of age. Most participants (56%) judged the sets of tweets that had been diversified based on our proposed viewpoint label as more diverse than the sets sampled based on a binary label. However, the binomial test was not significant ($p = 0.34$). We thus did not find any evidence for a difference between the two types of tweet sets.

6.2.1 Exploratory analysis. To help explain why we did not find a significant difference between the two types of tweet sets, we collected additional data and conducted a second, exploratory analysis. One potential reason we suspected could have led to the insignificant results was an overestimation of the effect size in our initial required sample size computation (see Section 6.1.5). To address this potential issue of insufficient power, we adjusted the sample size calculation to detect a smaller effect (i.e., Cohen's $g = 0.1$) rather than a medium effect. We thus recruited an additional 110

[8] Preregistering our user study meant publicly declaring the research question, hypothesis, procedure, and analysis plan prior to any data collection. The preregistration is available at: https://osf.io/cn8qa/.

[9] https://prolific.co

participants (i.e., raising the sample size to 200), who went through the same procedure as the first 90.

Another suspected reason for the insignificant result concerned spurious variation in the tweets. Upon closer examination of the results, we noticed that most participants judged four out of the five tweet sets diversified based on our proposed viewpoint representation as more diverse. However, for one particular tweet set pair, our diversification was judged as more diverse only five out of eighteen times. Participants stated that this set contained many extreme opinions and that therefore it did not seem like a good discussion would result from this set. Indeed, our method had assembled a set containing four extreme viewpoints (i.e., *strongly opposing* and *strongly supporting*) and only two mild viewpoints. This was different in all other sets, which had no more than 50% extreme viewpoints. We therefore excluded data from participants who had annotated this set from this exploratory analysis.

Ninety-seven (61%) out of the remaining 159 participants judged the sets diversified using our proposed viewpoint label as more diverse, a proportion significantly higher than random ($p = 0.007$).[10] Note that these analyses are exploratory as we conducted them outwith the preregistration and after examining the main results.

7 DISCUSSION

We have proposed a novel viewpoint representation for HII that overcomes the limitations of currently used binary viewpoint labels in two crucial ways. First, instead of classifying viewpoints into broad stance categories, it represents a viewpoint's stance on a more nuanced, seven-point ordinal scale ranging from "strongly opposing" to "strongly supporting". Second, it includes a viewpoint's logic(s) of evaluation (i.e., a notion that we borrow from the communication sciences), representing underlying reasons or perspectives using seven general categories. Our proposed viewpoint representation thus incorporates important aspects of viewpoints identified by the communication sciences in two dimensions while remaining topic-independent (**RQ1**; see Section 3). We have shown that workers can assign this novel viewpoint label with satisfactory reliability (**RQ2**) and found no evidence for an influence of cognitive biases (i.e., the *halo effect* and the *confirmation bias*) in this context (**RQ3**; see Section 4). Furthermore, in a viewpoint diversity analysis of tweets and a user study, we have demonstrated that our proposed viewpoint representation, while subtle, is more comprehensive and meaningful compared to binary viewpoint labels (**RQ4**; see Sections 5 and 6). Our exploratory analyses further suggest that the diversification algorithm must be tuned correctly concerning stance; i.e., including too many extreme opinions from either side of the spectrum may lead users to find the diversification less meaningful.

7.1 Guidelines for Obtaining Viewpoint Labels

Our crowdsourcing study has shown that workers are sufficiently reliable when annotating our proposed viewpoint label. However, especially with respect to assigning logics of evaluation or when dealing with ambiguous tweets, this task can be difficult. Worker feedback on our task included comments such as "*The tweet doesn't really mention the logic behind the support.*"; "*This one doesn't seem to make any sort of argument.*" and "*Really have to read between the lines with this one honestly.*" Based on our experience, we therefore propose a set of guidelines that requesters should follow when aiming to obtain annotations of our proposed viewpoint label:

(1) Given the difficulty of the task and in line with earlier work on this topic [15], we recommend setting the worker requirements rather high; e.g., *Master* workers from MTurk.
(2) While crowd workers seem to have no trouble annotating stance even on a seven-point ordinal scale, the logic(s) of evaluation can be hard to interpret. Requesters should ensure that all logics are well-explained and include several examples as well as relevant words to look for (see Table 1).
(3) We recommend collecting at least six annotations per document. Disagreement might still be high in this case, but we found that crowd workers fill in each other's gaps by identifying logics that others may have missed. When aggregating six or more annotations in a weighted fashion, the final labels are comparable with expert evaluations (see Section 4).
(4) When collecting difficult viewpoint representations such as logics of evaluation, requesters should consider training campaigns for crowd workers to build a pool of knowledgeable and reliable annotators over time.
(5) Asking the crowd workers to justify their answer or describe the viewpoint in their own words has been shown to increase the quality of their annotations [31]. In a workflow setting [11], a crowd worker could use such rationales to approve or reject a certain logic of evaluation provided by a different crowd worker.

7.2 Implications

The two-dimensional viewpoint representation we propose has implications for HII research and practical applications that concern user interactions with debated topics. For instance, it may lead to a better understanding of attitude change in web search by providing insight into nuanced shifts of stance. The dynamics of discussions on social media may similarly be studied in more depth when considering which logics of evaluation drive conversations (e.g., to automatically determine where exactly people of different stances disagree). From a practical point of view, ranking bias metrics and re-ranking algorithms may take both dimensions of our proposed viewpoint representation into account, e.g., for a richer notion of viewpoint diversity in a list of recommended news items. In the same way, user interface interventions that aim to mitigate user biases in content consumption could benefit from comprehensive viewpoint representations by taking nuanced stances and logics into account when highlighting, hiding, or explaining documents.

7.3 Limitations and Future Work

Crowd annotators were reliable in annotating tweets' stances. While the logics of evaluation generated more disagreement, workers were still able to perform as well as expert annotators, whose annotations were also often in disagreement. The discussion session conducted by the experts, however, proved beneficial to reach consensus, and we consider the lack of discussion among crowd annotators as a limitation. As future work, we plan to incorporate collaborative workflows [11, 28] in our crowdsourcing tasks. One approach could be to ask crowd workers to choose all logics that apply and provide

[10] Without removing the problematic set of tweets, the binomial test was not significant even in the larger sample of 200 participants ($p = 0.1$).

a rationale for each, either in a free-text fashion or by highlighting the words in the text that support their decision. Then, a second annotator could approve or reject these. In related studies, asking crowd annotators to provide rationales for their annotations proved useful for increasing quality [31].

Another limitation of our approach is that crowdsourcing studies can become expensive when large amounts of data need to be annotated and even more so when the task is difficult. To lower the cost, automatic methods such as sentiment analysis, topic modeling, or stance detection could be used as preprocessing methods (i.e., the crowd then only validates the output of the automated methods). Researchers could also use such automatic methods to generate two-dimensional representations of viewpoints. Studying their suitability, however, is part of future work. Similarly, future work could build machine learning models for our novel viewpoint representation. We hypothesize that the aforementioned crowd annotators' rationales could be helpful for learning the logic-specific language and improving the performance of such tools [3].

Finally, although we did evaluate a diversification algorithm based on our two-dimensional viewpoint representation against one using a binary label, future studies could also investigate the use of our method in mitigating specific user biases (e.g., the confirmation bias) in online information seeking [50, 51].

8 CONCLUSION

In this paper, we proposed a novel, two-dimensional viewpoint representation for HII, inspired by research from the communication sciences. The proposed two-dimensional viewpoint representation consists of a viewpoint's *stance* on a nuanced level which reflects the degree to which a viewpoint opposes or supports a topic, and a viewpoint's *logic of evaluation*, which reflects the perspective behind the stance. We efficiently collected such viewpoint's stances and logics in a crowdsourcing study with acceptable reliability. In a viewpoint diversity analysis and user study, we further showed that our proposed viewpoint representation can be more meaningful in representing diverse opinions on a topic compared to binary viewpoint labels (i.e., *against/in favor*). We hope that our work enables researchers and practitioners to represent viewpoints in a more detailed fashion, eventually leading to a better understanding and more effective interventions related to user interactions with debated topics on the web.

ACKNOWLEDGMENTS

This activity is financed by IBM and the Allowance for Top Consortia for Knowledge and Innovation (TKI's) of the Dutch ministry of economic affairs. We also wish to thank Francesco Barile, Amir Fard, Rishav Hada, Shabnam Najafian, Alisa Rieger, and Tjitze Rienstra for their feedback on an earlier draft of this paper.

REFERENCES

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

[2] Ahmed Allam, Peter Johannes Schulz, and Kent Nakamoto. 2014. The Impact of Search Engine Selection and Sorting Criteria on Vaccination Beliefs and Attitudes: Two Experiments Manipulating Google Output. *Journal of Medical Internet Research* 16, 4 (April 2014), e100. https://doi.org/10.2196/jmir.2642

[3] Ines Arous, Ljiljana Dolamic, Jie Yang, Akansha Bhardwaj, Giuseppe Cuccu, and Philippe Cudré-Mauroux. 2021. MARTA: Leveraging Human Rationales for Explainable Text Classification. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 7 (May 2021), 5868–5876.

[4] Christian Baden, Neta Kligler-Vilenchik, and Moran Yarchi. 2020. Hybrid Content Analysis: Toward a Strategy for the Theory-driven, Computer-assisted Classification of Large Text Corpora. *Communication Methods and Measures* 14, 3 (July 2020), 165–183. https://doi.org/10.1080/19312458.2020.1803247

[5] Christian Baden and Nina Springer. 2014. Com(ple)menting the news on the financial crisis: The contribution of news users' commentary to the diversity of viewpoints in the public debate. *European Journal of Communication* 29, 5 (Oct. 2014), 529–548. https://doi.org/10.1177/0267323114538724

[6] Christian Baden and Nina Springer. 2017. Conceptualizing viewpoint diversity in news discourse. *Journalism* 18, 2 (Feb. 2017), 176–194. https://doi.org/10.1177/1464884915605028

[7] Ricardo Baeza-Yates. 2018. Bias on the web. *Commun. ACM* 61, 6 (May 2018), 54–61. https://doi.org/10.1145/3209581

[8] Luc Boltanski and Laurent Thévenot. 2006. *On Justification: Economies of Worth*. Vol. 27. Princeton University Press.

[9] Bjarte Botnevik, Eirik Sakariassen, and Vinay Setty. 2020. *BRENDA: Browser Extension for Fake News Detection*. Association for Computing Machinery, New York, NY, USA, 2117–2120. https://doi.org/10.1145/3397271.3401396

[10] Björn Burscher, Daan Odijk, Rens Vliegenthart, Maarten De Rijke, and Claes H De Vreese. 2014. Teaching the computer to code frames in news: Comparing two supervised machine learning approaches to frame analysis. *Communication Methods and Measures* 8, 3 (2014), 190–206.

[11] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. 2017. Revolt: Collaborative Crowdsourcing for Labeling Machine Learning Datasets. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 2334–2346. https://doi.org/10.1145/3025453.3026044

[12] Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. 2019. Seeing Things from a Different Angle:Discovering Diverse Perspectives about Claims. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 542–557. https://doi.org/10.18653/v1/N19-1053

[13] Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. 2020. Will-They-Won't-They: A Very Large Dataset for Stance Detection on Twitter. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 1715–1724. https://doi.org/10.18653/v1/2020.acl-main.157

[14] Tim Draws, Jody Liu, and Nava Tintarev. 2020. Helping users discover perspectives: Enhancing opinion mining with joint topic models. In *2020 International Conference on Data Mining Workshops (ICDMW)*. IEEE, Sorrento, Italy, 23–30. https://doi.org/10.1109/ICDMW51313.2020.00013

[15] Tim Draws, Alisa Rieger, Oana Inel, Ujwal Gadiraju, and Nava Tintarev. 2021. A Checklist to Combat Cognitive Biases in Crowdsourcing. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 9, 1 (Oct. 2021), 48–59. https://ojs.aaai.org/index.php/HCOMP/article/view/18939

[16] Tim Draws, Nava Tintarev, Ujwal Gadiraju, Alessandro Bozzon, and Benjamin Timmermans. 2021. Assessing Viewpoint Diversity in Search Results Using Ranking Fairness Metrics. *ACM SIGKDD Explorations Newsletter* 23, 1 (May 2021), 50–58. https://doi.org/10.1145/3468507.3468515

[17] Tim Draws, Nava Tintarev, Ujwal Gadiraju, Alessandro Bozzon, and Benjamin Timmermans. 2021. This Is Not What We Ordered: Exploring Why Biased Search Result Rankings Affect User Attitudes on Debated Topics. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, Virtual Event Canada, 295–305. https://doi.org/10.1145/3404835.3462851

[18] Anca Dumitrache, Lora Aroyo, and Chris Welty. 2018. Capturing ambiguity in crowdsourcing frame disambiguation. In *Sixth AAAI Conference on Human Computation and Crowdsourcing*.

[19] Anca Dumitrache, Oana Inel, Lora Aroyo, Benjamin Timmermans, and Chris Welty. 2018. CrowdTruth 2.0: Quality Metrics for Crowdsourcing with Disagreement. (2018). https://arxiv.org/abs/1808.06080

[20] Robert M Entman. 2003. Cascading activation: Contesting the White House's frame after 9/11. *Political Communication*, 20, 4 (2003), 415–432.

[21] Robert Epstein and Ronald E. Robertson. 2015. The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences* 112, 33 (Aug. 2015), E4512–E4521. https://doi.org/10.1073/pnas.1419828112

[22] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods* 39, 2 (2007), 175–191.

[23] William Ferreira and Andreas Vlachos. 2016. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, 1163–1168.

https://doi.org/10.18653/v1/N16-1138

[24] Gizem Gezici, Aldo Lipani, Yucel Saygin, and Emine Yilmaz. 2021. Evaluation metrics for measuring bias in search engine results. *Information Retrieval Journal* 24, 2 (April 2021), 85–113. https://doi.org/10.1007/s10791-020-09386-w

[25] Amira Ghenai, Mark D. Smucker, and Charles L.A. Clarke. 2020. A Think-Aloud Study to Understand Factors Affecting Online Health Search. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*. ACM, Vancouver BC Canada, 273–282. https://doi.org/10.1145/3343413.3377961

[26] Natali Helberger. 2019. On the Democratic Role of News Recommenders. *Digital Journalism* 7, 8 (Sept. 2019), 993–1012. https://doi.org/10.1080/21670811.2019.1623700

[27] Natali Helberger, Kari Karppinen, and Lucia D'Acunto. 2018. Exposure diversity as a design principle for recommender systems. *Information, Communication & Society* 21, 2 (Feb. 2018), 191–207. https://doi.org/10.1080/1369118X.2016.1271900

[28] Joy Kim, Sarah Sterman, Allegra Argent Beal Cohen, and Michael S. Bernstein. 2017. Mechanical Novel: Crowdsourcing Complex Work through Reflection and Revision. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 233–245. https://doi.org/10.1145/2998181.2998196

[29] Klaus Krippendorff. 2004. Reliability in Content Analysis.: Some Common Misconceptions and Recommendations. *Human Communication Research* 30, 3 (July 2004), 411–433. https://doi.org/10.1111/j.1468-2958.2004.tb00738.x

[30] Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar, Saptarshi Ghosh, Krishna P. Gummadi, and Karrie Karahalios. 2019. Search bias quantification: investigating political bias in social media and web search. *Information Retrieval Journal* 22, 1-2 (April 2019), 188–227. https://doi.org/10.1007/s10791-018-9341-2

[31] Mucahid Kutlu, Tyler McDonnell, Tamer Elsayed, and Matthew Lease. 2020. Annotator rationales for labeling tasks in crowdsourcing. *Journal of Artificial Intelligence Research* 69 (2020), 143–189.

[32] Yingjie Li, Tiberiu Sosea, Aditya Sawant, Ajith Jayaraman Nair, Diana Inkpen, and Cornelia Caragea. 2021. P-Stance: A Large Dataset for Stance Detection in Political Domain. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics, Online, 2355–2365. https://doi.org/10.18653/v1/2021.findings-acl.208

[33] Felicia Loecherbach, Judith Moeller, Damian Trilling, and Wouter van Atteveldt. 2020. The Unified Framework of Media Diversity: A Systematic Literature Review. *Digital Journalism* 8, 5 (May 2020), 605–642. https://doi.org/10.1080/21670811.2020.1764374

[34] Andrea Masini and Peter Van Aelst. 2017. Actor diversity and viewpoint diversity: Two of a kind? *Communications* 42, 2 (Jan. 2017). https://doi.org/10.1515/commun-2017-0017

[35] Andrea Masini, Peter Van Aelst, Thomas Zerback, Carsten Reinemann, Paolo Mancini, Marco Mazzoni, Marco Damiani, and Sharon Coen. 2018. Measuring and Explaining the Diversity of Voices and Viewpoints in the News: A comparative study on the determinants of content diversity of immigration news. *Journalism Studies* 19, 15 (Nov. 2018), 2324–2343. https://doi.org/10.1080/1461670X.2017.1343650

[36] Dana McKay, Stephann Makri, Marisela Gutierrez-Lopez, Andrew MacFarlane, Sondess Missaoui, Colin Porlezza, and Glenda Cooper. 2020. We are the change that we seek: information interactions during a change of viewpoint. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*. 173–182.

[37] Florian Meier and David Elsweiler. 2019. Studying Politicians' Information Sharing on Social Media. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval* (Glasgow, Scotland UK) *(CHIIR '19)*. Association for Computing Machinery, New York, NY, USA, 237–241. https://doi.org/10.1145/3295750.3298944

[38] Saif M. Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 Task 6: Detecting Stance in Tweets. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval '16)*. San Diego, California.

[39] Saif M. Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2017. Stance and Sentiment in Tweets. *Special Section of the ACM Transactions on Internet Technology on Argumentation in Social Media* 17, 3 (2017).

[40] Mats Mulder, Oana Inel, Jasper Oosterman, and Nava Tintarev. 2021. Operationalizing Framing to Support Multiperspective Recommendations of Opinion Pieces. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Virtual Event Canada, 478–488. https://doi.org/10.1145/3442188.3445911

[41] Judith Möller, Damian Trilling, Natali Helberger, and Bram van Es. 2018. Do not blame it on the algorithm: an empirical assessment of multiple recommender systems and their impact on content diversity. *Information, Communication & Society* 21, 7 (July 2018), 959–977. https://doi.org/10.1080/1369118X.2018.1444076

[42] Alamir Novin and Eric Meyers. 2017. Making Sense of Conflicting Science Information: Exploring Bias in the Search Engine Result Page. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*. ACM, Oslo Norway, 175–184. https://doi.org/10.1145/3020165.3020185

[43] Eli Pariser. 2011. *The filter bubble: What the Internet is hiding from you*. Penguin UK.

[44] Rafael B Pereira, Alexandre Plastino, Bianca Zadrozny, and Luiz HC Merschmann. 2018. Correlation analysis of performance measures for multi-label classification. *Information Processing & Management* 54, 3 (2018), 359–369.

[45] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic Detection of Fake News. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, 3391–3401. https://aclanthology.org/C18-1287

[46] Frances A. Pogacar, Amira Ghenai, Mark D. Smucker, and Charles L.A. Clarke. 2017. The Positive and Negative Influence of Search Results on People's Decisions about the Efficacy of Medical Treatments. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*. ACM, Amsterdam The Netherlands, 209–216. https://doi.org/10.1145/3121050.3121074

[47] Mauro P. Porto. 2007. Frame Diversity and Citizen Competence: Towards a Critical Approach to News Quality. *Critical Studies in Media Communication* 24, 4 (Oct. 2007), 303–321. https://doi.org/10.1080/07393180701560864

[48] Cornelius Puschmann. 2019. Beyond the Bubble: Assessing the Diversity of Political Search Results. *Digital Journalism* 7, 6 (July 2019), 824–843. https://doi.org/10.1080/21670811.2018.1539626

[49] Rezvaneh Rezapour, Ly Dinh, and Jana Diesner. 2021. Incorporating the Measurement of Moral Foundations Theory into Analyzing Stances on Controversial Topics. In *Proceedings of the 32st ACM Conference on Hypertext and Social Media*. ACM, Virtual Event USA, 177–188. https://doi.org/10.1145/3465336.3475112

[50] Alisa Rieger, Tim Draws, Mariët Theune, and Nava Tintarev. 2021. This Item Might Reinforce Your Opinion: Obfuscation and Labeling of Search Results to Mitigate Confirmation Bias. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media* (Virtual Event, USA) *(HT '21)*. Association for Computing Machinery, New York, NY, USA, 189–199. https://doi.org/10.1145/3465336.3475101

[51] Alisa Rieger, Mariët Theune, and Nava Tintarev. 2020. Toward natural language mitigation strategies for cognitive biases in recommender systems. In *2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence*. 50–54.

[52] Parinaz Sobhani, Diana Inkpen, and Xiaodan Zhu. 2017. A Dataset for Multi-Target Stance Detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Association for Computational Linguistics, Valencia, Spain, 551–557. https://aclanthology.org/E17-2088

[53] Parinaz Sobhani, Saif M. Mohammad, and Svetlana Kiritchenko. 2016. Detecting Stance in Tweets And Analyzing its Interaction with Sentiment. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics (*Sem)*. Berlin, Germany.

[54] Nava Tintarev, Emily Sullivan, Dror Guldin, Sihang Qiu, and Daan Odjik. 2018. Same, Same, but Different: Algorithmic Diversification of Viewpoints in News. In *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization*. ACM, Singapore Singapore, 7–13. https://doi.org/10.1145/3213586.3226203

[55] Nguyen Vo and Kyumin Lee. 2018. The Rise of Guardians: Fact-Checking URL Recommendation to Combat Fake News. In *The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval* (Ann Arbor, MI, USA) *(SIGIR '18)*. Association for Computing Machinery, New York, NY, USA, 275–284. https://doi.org/10.1145/3209978.3210037

[56] Paul S Voakes, Jack Kapfer, David Kurpius, and David Shano-yeon Chern. 1996. Diversity in the news: A conceptual and methodological framework. *Journalism & Mass Communication Quarterly* 73, 3 (1996), 582–593.

[57] Sanne Vrijenhoek, Mesut Kaya, Nadia Metoui, Judith Möller, Daan Odijk, and Natali Helberger. 2021. *Recommenders with a Mission: Assessing Diversity in News Recommendations*. Association for Computing Machinery, New York, NY, USA, 173–183. https://doi.org/10.1145/3406522.3446019

[58] Rui Wang, Deyu Zhou, Mingmin Jiang, Jiasheng Si, and Yang Yang. 2019. A Survey on Opinion Mining: From Stance to Product Aspect. *IEEE Access* 7 (2019), 41101–41124. https://doi.org/10.1109/ACCESS.2019.2906754

[59] Ryen White. 2013. Beliefs and biases in web search. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. ACM, Dublin Ireland, 3–12. https://doi.org/10.1145/2484028.2484053

[60] Ke Yang and Julia Stoyanovich. 2017. Measuring Fairness in Ranked Outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*. ACM, Chicago IL USA, 1–6. https://doi.org/10.1145/3085504.3085526

[61] Elad Yom-Tov, Susan Dumais, and Qi Guo. 2014. Promoting Civil Discourse Through Search Engine Diversity. *Social Science Computer Review* 32, 2 (April 2014), 145–154. https://doi.org/10.1177/0894439313506838

[62] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. FA*IR: A Fair Top-k Ranking Algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, Singapore Singapore, 1569–1578. https://doi.org/10.1145/3132847.3132938