Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 4126–4134
Marseille, 20-25 June 2022

Towards Modelling Self-imposed Filter Bubbles in Argumentative Dialogue Systems

Annalena Aicher[1], Wolfgang Minker[1], Stefan Ultes[2]
[1] Institute of Communications Engineering, Ulm University, Albert-Einstein-Allee 43, 89075 Ulm
[2]Mercedes-Benz Research and Development, Sindelfingen, Germany
annalena.aicher@uni-ulm.de

Abstract

To build a well-founded opinion it is natural for humans to gather and exchange new arguments. Especially when being confronted with an overwhelming amount of information, people tend to focus on only the part of the available information that fits into their current beliefs or convenient opinions. To overcome this "self-imposed filter bubble" (SFB) in the information seeking process, it is crucial to identify influential indicators for the former. Within this paper we propose and investigate indicators for the the user's SFB, mainly their *Reflective User Engagement* (RUE), their *Personal Relevance* (PR) ranking of content-related subtopics as well as their *False* (FK) and *True Knowledge* (TK) on the topic. Therefore, we analysed the answers of 202 participants of an online conducted user study, who interacted with our argumentative dialogue system BEA ("Building Engaging Argumentation"). Moreover, also the influence of different input/output modalities (speech/speech and drop-down menu/text) on the interaction with regard to the suggested indicators was investigated.

Keywords: Confirmation Bias, Spoken Dialogue Systems (SDS), Computational Argumentation, Human-Computer Interaction (HCI), User Study, User Modelling

1. Introduction

Conversations display a natural way for humans to resolve different points of view and build an opinion. Most popular virtual agents are trained to handle simple conversations, e.g. travel inquiries, still they are inept to demanding conversations (Saha et al., 2020). Especially, dialogue systems that exchange arguments and can converse with humans via natural language display a big challenge in artificial intelligence. Another challenge displays the tendency of people which tend to focus on a biased subset of sources that repeat or strengthen an already established or convenient opinion (Pariser, 2011). In order to avoid this process of intellectual isolation, we investigate and analyze possible indicators for a self-imposed filter bubble (SFB) (Ekström, 2021). Our user study gives an insight into which indicators could be suitable to model an SFB and thus, provides a first step towards our aim to break the former in an engaging argumentative dialogue. Hence, the user shall be able to scrutinize arguments on both sides of a controversial topic in a natural and intuitive way. To this end our system engages in a cooperative dialogue with a user in order to support an unbiased and critical reflected opinion building process. In particular, we focus on four main indicators: *Reflective User Engagement (RUE), Personal Relevance (PR), True Knowledge (TK)* and *False Knowledge (FK)*. The *RUE* describes the critical-thinking and open-mindedness demonstrated by the user, following our definition in previous work (Aicher et al., 2021a). The *PR* refers to the user individual assessment of the relevance of subtopics with regard to the topic of the discussion. *True Knowledge* is defined as the information the user already has on a topic, which is consistent and also present in the system's database. *False Knowledge* on the other hand is described as the user's information on a topic which contradicts the information in the system's database.

The remainder of the paper is as follows: an overview over existing literature is given in Section 2. After Section 3 introduces the framework and architecture of our argumentative dialogue system (ADS), we propose four indicators to model the SFB of a user in the context of ADS in Section 4. Section 5 describes the experimental setting of the user study we conducted to investigate the previously defined indicators. Subsequently, the according results are discussed in Section 6. We close with a conclusion and a brief discussion of future work in Section 7.

2. Related Work

As we pursue a cooperative exchange of arguments our system does not try to persuade or win a debate against the user unlike most approaches to human-machine argumentation. Those approaches utilize different models to structure the interaction and are embedded in a competitive scenario. Slonim et al. (2021) use a classical debating setting. Their IBM Debater is an autonomous debating system that can engage in a competitive debate with humans via natural language. Another speech-based approach was introduced by Rosenfeld and Kraus (2016) presenting a system based on weighted Bipolar Argumentation Frameworks (wBAG). Arguing chatbots such as Debbie (Rakshit et al., 2017) and Dave (Le et al., 2018) interact via text with the user. Another menu-based framework that incorporates the beliefs and concerns of the opponent was presented by Hadoux and Hunter (2021). In the same line, (Chalaguine and Hunter, 2020) used a previously

crowd-sourced argument graph and considered the concerns of the user to persuade them. A persuasive prototype chatbot is introduced by (Chalaguine and Hunter, 2021) to convince users to vaccinate against COVID-19 using computational models of argument. Furthermore, (Fazzinga et al., 2021) illustrate an approach towards a dialogue system architecture that uses argumentative concepts to perform reasoning and provide answers consistent with the user input, which is illustrated by the example of a user requiring information about COVID-19 vaccines.

In contrast to our system, none of the aforementioned ADS tries to cooperatively engage the users to explore arguments and stating their preferences in natural language. We modified and extended our previously introduced menu-based argumentative dialogue system BEA (Aicher et al., 2021b) ('Building Engaging Argumentation'[1]) such that it is able to interact via speech. Therefore, also the set of possible user actions to fit the new flexibility and graphical user interface were completely revised respectively.

In the context of information seeking and opinion building, especially regarding sources such as search engines or social media platforms, two important phenomena have to be distinguished, which both might lead to a bias. On the one hand, due to filter algorithms information content is selected based on previous online behavior. Thus, the users are separated from information disagreeing with their viewpoints and isolated in cultural/ideological bubbles, so-called "Filter Bubbles" (Pariser, 2011). On the other hand, "confirmation bias, a term typically used in the psychological literature, connotes the seeking or interpreting of evidence in ways that are partial to existing beliefs, expectations, or a hypothesis in hand" (Nickerson, 1998, p. 175). Allahverdyan and Galstyan (2014) describe confirmation bias as the tendency to acquire or evaluate new information in a way that is consistent with one's preexisting beliefs. Additionally, Jones and Sugden (2001) show that a positive confirmation bias, in both information acquisition and information use, is present in an experiment in which individuals choose what information to buy, prior to making a decision.

To resolve the confirmation bias of a user in the context of decision making processes Huang et al. (2012) propose the usage of computer-mediated counter-argument. Schwind and Buder (2012) regard preference-inconsistent recommendations as a promising approach to trigger critical thinking. Still, if too many counter-arguments are introduced this could lead to unwanted effects negative emotional consequences (annoyance, confusion) (Huang et al., 2012). Consequently, Huang et al. (2012) stress the need for an intelligent system which is able to adapt the frequency, timing and choice of the counter-arguments. To provide such a system, it is crucial to develop and find a model, which can be adapted to the user.

Approaches like the one introduced by Del Vicario et al. (2017), study online social debates and try to model and describe the related polarization dynamics based on confirmation bias. In contrast, we aim to model the cause of this confirmation bias, the so-called "self-imposed filter bubble" (SFB) (Ekström, 2021). To the best of our knowledge, we are the first to investigate potential indicators to describe and model this phenomenon in context of an argumentative cooperative dialogue. This cooperative setting is motivated by the findings of Villarroel et al. (2016) stating that a consensual dialogue is much more likely to resolve diverging perspectives on evidence and repair incorrect, partial and subjective readings of evidence than a persuasive one.

3. System Framework and Architecture

In the following, the architecture of BEA is outlined. After describing the dialogue framework and model, the interface and NLG/NLU architecture are introduced. An overview over the whole architecture of BEA is given in Figure 3

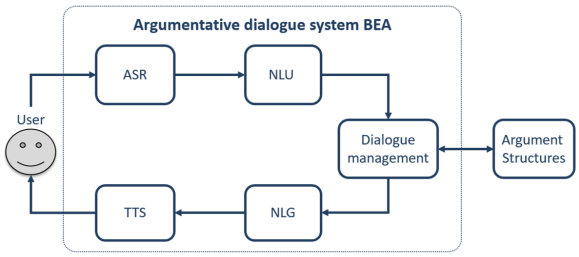

Figure 1: Architecture of BEA. After the user's spoken input is processed by the automated speech recognition module (ASR), it is passed to the Natural Language Understanding unit, which extracts the respective information. This abstractly represented information can be processed by the dialogue management, which decides a suitable corresponding system response by interacting with an argument structure. Once an appropriate response is selected it is processed by a Natural Language Generation (NLG) module which formulates its textual representation and finally presented to the user in natural language by Text-to-Speech (TTS) module. In case of the baseline system the ASR and TTS modules were omitted.

3.1. Dialogue Framework and Model

In order to be able to combine the presented system with existing argument mining approaches to ensure its topic flexibility, we follow the argument annotation scheme introduced by Stab and Gurevych (2014). It distinguishes three different types of components (Major Claim, claim, premise), which are structured in the form of bipolar argumentation trees depicted in Figure 2. The overall topic of the debate is formulated

[1]BEA engages in a deliberative dialogue with a human user in order to support their opinion building process by incrementally presenting automatically extracted arguments.

as the *Major Claim* φ_0 representing the root node in the graph. *Claims* (C1 and C2) on the other hand are assertions which formulate a certain opinion targeting the *Major Claim* but still need to be justified by further arguments, *premises* (P1 and P2) respectively. We consider two relations between these argument components (nodes), *support* (green arrows) or *attack* (red arrows). Each component apart from the Major Claim φ_0 (which has no relation) has exactly one unique relation to another component. This leads to a non-cyclic tree structure, where each node or "parent" (C1 and C2) is either supported or attacked by its "children". If no children exist, the node is a leaf (e.g. P1, P2 and P3) and marks the end of a branch.

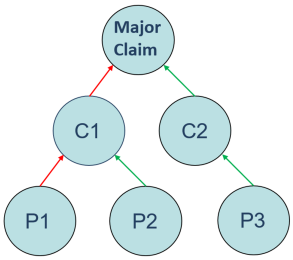

Figure 2: Visualization of argument tree structure. The major claim is the root node, which is supported by the claim C2 (denoted by a green arrow) and attacked by claim C1 (denoted by a red arrow). The respective leaf nodes are the premises P1, P2 and P3.

The interaction between the system and the user is separated in turns, consisting of a user action and corresponding system answer. Table 1 shows the possible actions (moves) the user is able to choose from. In general, we distinguish three main types of moves, apart from the *exit* move (termination of the conversation): navigation moves, feedback moves and status quo moves. The determiners display the moves' availability depending on the position of the current argument φ_i (root/parent/leaf node).

In analogy to our previous approach (Aicher et al., 2021b), explicit user feedback ($prefer(\varphi_i)$,$reject(\varphi_i)$) is used to estimate the (overall) preference considering wBAGs (Amgoud and Ben-Naim, 2016; Amgoud and Ben-Naim, 2018). Thus, to each φ_i a *weight* ω_i is assigned. The *strength* of an argument component φ_i is determined by its weight ω_i and the strength of its attackers and supporters. If the user performs a feedback move on φ_i, its weights are updated such that:

$$\begin{aligned} \textit{prefer}: &\quad \omega_i = \omega_{i,\max} + \frac{n_v}{n_a}\left(1 - \omega_{i,\max}\right) \\ \text{reject}: &\quad \omega_i = 0 \\ \text{indifferent}: &\quad \omega_i = 0.5, \end{aligned}$$

where $\omega_{i,max}$ denotes the maximum strength of all siblings of argument i. Here, n_v describes the number of sibling arguments of argument i which have already been presented to the user and n_a denotes the total number of all sibling arguments. The preference update takes into account how many siblings have already been heard in relation to the ones available. The nodes are updated recursively until the Major Claim φ_0 is reached. The thereby calculated user stance can be accessed via the actions $stance(\varphi_i, i = 0, ..., n$ on every argument component of the tree.

In the herein presented study, a sample debate on the topic *Marriage is an outdated institution* is chosen (Rach et al., 2018), which suits the argument scheme described above. It serves as knowledge base for the arguments and is taken from the *Debatabase* of the idebate.org[2] website. It consists of a total of 72 argument components (1 Major Claim, 10 Claims and 61 Premises) and their corresponding relations are encoded in an OWL ontology (Bechhofer, 2009) for further use. Due to the generality of the annotation scheme, the system is not restricted to the herein considered data. In general, every argument structure that can be mapped into the applied scheme can be processed by the system.

3.2. Interface and NLU Framework

The graphical user interface (GUI) of BEA is illustrated in Figure 3. The interface can either provide a drop-down menu or speech input as needed. To detect possible differences between both modalities, we conducted our user study with two groups for each modality (see Section 5). In the drop-down system users can choose their action by clicking, whereas in the speech system a NLU framework introduced by Abro et al. (2021) processes the spoken user utterance. This input is captured with a browser-based audio recording that is further processed by the `Python` library `SpeechRecognition` using Google Speech Recognition. Its intent classifier uses the BERT Transformer Encoder presented by Devlin et al. (2018) and a bidirectional LSTM classifier. The system-specific intents are trained with a set of pre-defined sample utterances. To increase the robustness of the NLU these utterances were extended by expressions of participants of a previous user study (Aicher et al., 2022). After a user intent is recognized, the spoken system response is presented using the Speech Synthesis of Web Speech API.

In the speech-based system, instead of the drop-down menu displayed in Figure 3, a button with "Start Talking" is shown. The button is pressed to start and stop the speech recording. Except for this difference both systems share a similar architecture. The dialogue history shows the system's responses left-aligned and corresponding user answers right-aligned. A progress bar

[2] `https://idebate.org/debatabase` (last accessed 23th June 2021).

Move	Description	Determiners
$why_{pro}(\varphi_i)$	Ask for a pro argument on current φ_i	If supporting child exists
$why_{con}(\varphi_i)$	Ask for a con argument on current φ_i	If attacking child node exists
$level_{up}$	Returns to previous φ	If $\varphi_i \neq \varphi_0$
$jump\_to(\varphi_i)$	Jump to φ_i	
$prefer(\varphi_i)$	Feedback to prefer φ_i	If $\varphi_i \neq \varphi_0$
$prefer(\varphi_i > \varphi_j)$	Feedback to prefer φ_i over φ_j	If siblings of φ_i are preferred
$reject(\varphi_i)$	Feedback to reject φ_i	If $\varphi_i \neq \varphi_0$
$indifferent(\varphi_i)$	Feedback to be indifferent about φ_i	If $\varphi_i \neq \varphi_0$
$stance(\varphi_i)$	Ask for own stance on current φ_i	
$stance(\varphi_0)$	Ask for own stance on current φ_i	
$number_{visited}$	Ask for number of heard arguments	
$moves_{available}(\varphi_i)$	Ask for available moves depending on φ_i	Speech I/O setting
$exit$	End conversation	$number_{visited} \geq = 10$

Table 1: Description of the thirteen moves with corresponding determiners.

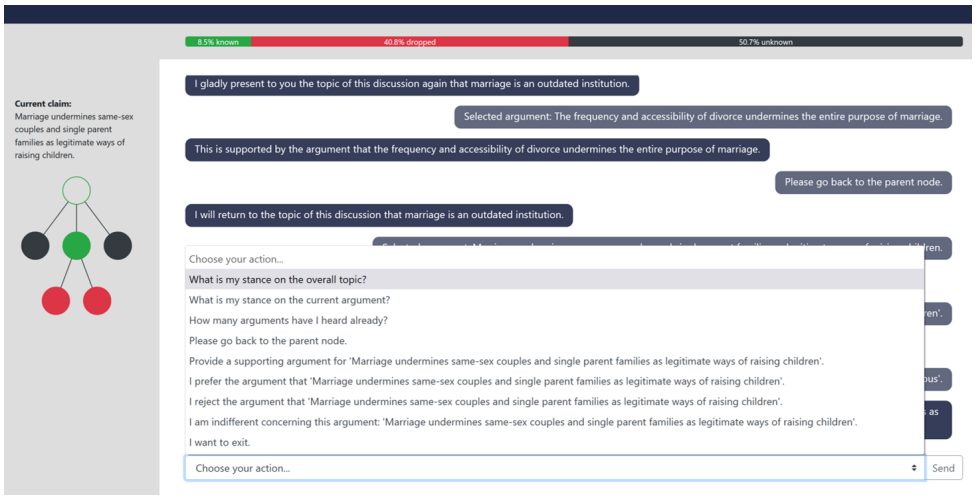

Figure 3: GUI of the baseline system with unfolded drop-down menu. Above the drop-down menu the dialogue history is shown. On the left side the sub-graph of the current branch is visible.

above the dialogue history shows the number of arguments that were already discussed and how many are still unknown for the user at each stage of the interaction. This provides a visual cue of the length of the ongoing conversation to the user. Furthermore, on the left side the sub-graph of the bipolar argumentation tree structure (with the displayed claim as root) is shown. The current position (i.e. argument) is displayed with a white node outlined with green line. Already heard arguments are shown in green and skipped arguments in red. Nodes shown in grey are still unheard.

The natural language generation is based on the original textual representation of the argument components. As described in (Aicher et al., 2021c) the annotated sentences were slightly modified to form a stand-alone utterance serving as a template for the respective system response. Additionally, a list of natural language representations for each type of system move was defined. During the generation of the utterances, the explicit formulation and introductory phrase is chosen from this list randomly.

4. Indicators to Model Self-imposed Filter Bubble Model

In the following, we motivate the choice of our SFB indicators Note that we do not claim the indicators or our model to be complete but a first approach to model SFBs. As previously mentioned we focus mainly on four indicators: *Reflective User Engagement (RUE), Personal Relevance (PR), True Knowledge (TK)* and *False Knowledge (FK)*. We motivate this choice building upon findings in well-established state-of-the-art literature.

The reflective user engagement *RUE* describes the critical-thinking and open-mindedness demonstrated by the user when exploring a controversial topic (Aicher et al., 2021a). Both, critical-thinking and open-mindedness appear as frequently suggested

starting points to counteract various types of biases (Alsharif and Symons, 2021; Schwind and Buder, 2012; Macpherson and Stanovich, 2007). In (Aicher et al., 2021a) we presented an approach to determine the RUE by taking into account the polarity and number of arguments a user has heard on a topic. Thus, the *RUE* is very likely to have a big influence on the user's SFB and henceforth, is explored with regard to its suitability to be incorporated into a SFB model. To get a more detailed information, we distinguish between a RUE for pro and for con arguments.

The elaboration likelihood model (ELM) (Petty et al., 2009), a well-established framework in persuasion research, suggests that attitude change occurs as a result of two different information processing modes – central vs. peripheral. Westerwick et al. (2017) state that if users process information via the central route, they engage carefully and thoroughly with the information, reflect on it, connect it with pre-existing cognitions and integrate it into their overall cognitive network. But when lacking the motivation and ability for such effortful consideration, recipients may engage in peripheral processing, not scrutinize the message content much. Furthermore, they claim that the respective route depends on the individuals' motivation (e.g. personal relevance) and ability (e.g. preexisting knowledge) (Westerwick et al., 2017). Building upon this argumentation, we consider *Personal Relevance (PR), True Knowledge (TK), False Knowledge (FK)* as possible additional indicators of the SFB we want to explore in the following study. The *PR* refers to the user individual assessment of how relevant a subtopic (cluster) is with regard to the topic of the discussion. The herein presented topic consists of ten main subtopics: *Law, Alternative Relationships and Parenthoods, Children, Divorce, Remarriage, Harmful Relationships, Relationship Stability, Religion, Expectations and Commitment* and *Social Acceptance*. *(F)K* is defined as the user's pre-existing correct and respectively, incorrect information. Without loss of generality, we assume the system's database to contain only correct information and consequently, information factually contradicting the former as incorrect.

5. User Study Setting

The user study was divided into two parts, one was focused on the change of the herein explained SFB indicators during the interaction, and the other focused on a detailed system evaluation with respect to the users' perception of BEA. Since the latter exceeds the scope of this paper, only the aspects relevant to the herein presented results are discussed.

The study was conducted online via the crowdsourcing platform "Crowdee" (`https://www.crowdee.com/`, 12-29th November 2021) with participants from the UK, US and Australia. All 292 participants were non-experts without a topic-specific background. After an introduction to the system (short text and demo video), the users' task was defined as listening to enough arguments to build a well-founded opinion on the topic. The first 139 participants interacted with BEA via drop-down menu input, the other 153 via speech. We consider the interaction length (time and number of heard arguments) as an additional indicator for user interest (Yi et al., 2014). Taking into account that uninterested users might want to quit the interaction, the participants were allowed to end the dialogue whenever they felt like having heard enough arguments (minimum: ten arguments) to build a well-founded opinion. Before and after the interaction with BEA the participants had to answer a questionnaire concerning the SFB indicators. In the follow-up of the interaction the questionnaire containing control questions and questions on the general perception of the system and its quality was posed.

Analyzing the questionnaire answers and feedback, 90 participants seemed to have issues or their data showed anomalies. Their data was excluded according to previously defined exclusion criteria: Contradictory answers in control question in the questionnaire, taking less than $30\,\mathrm{sec}$ to read through introduction and watch the introduction videos, taking less than $120\,\mathrm{sec}$ to answer 40+15 questions in the final questionnaire and feedback indicating that problems occurred during the interaction or participants reported that they did not knew what to do. This leads to a total number of data records of 202 participants (menu: 104, speech: 98) which were used in the following evaluation.

6. Results and Discussion

In average the participants interacted with BEA for 31:45 minutes (menu: $27.57\,\mathrm{min}$; speech: $35.34\,\mathrm{min}$). In the speech(menu) system 17(10) of 98(104) participants and thus, 17.3% (9.6%) heard only the minimum number of arguments. In total 27 participants (13.4%) quit the system after a minimum of ten presented arguments.

To determine whether the difference between before and after the interaction with the system is significant, we used the nonparametric Wilcoxon signed rank test (Woolson, 2007) for paired samples for all indicators.

In the following, the results for both and the separate systems (menu, speech) are discussed, as also the influence of different modalities on the SFB indicators in question is analyzed. The participants' ratings regarding both systems are shown in Table 2 and the corresponding separated results in Table 3. For a better readability the results for the *Personal Relevance* are displayed separately in Table 4.

6.1. Reflective User Engagement

Regarding the indicator Reflective User Engagement (RUE) participants had to rate two statements on a five-point Likert scale, before and after the interaction (5 = Extremely interested, 4 = Very interested, 3 = Moderately interested, 2 = Slightly interested, 1 = Not at all

Category	Question	M_{pre}	M_{post}
RUE pro	How much are you interested in hearing arguments which support the topic's claim?	**3.5** (0.974)	**3.35*** (1.027)
RUE con	How much are you interested in hearing arguments which attack the topic's claim?	3.44 (0.956)	3.29 (1.064)
True Knowledge	How big do you consider your current knowledge on this topic after the interaction?	**3.32 (0.913)**	**3.46**** (0.847)
False Knowledge	How big do you estimate the percentage of arguments BEA provided that contradict the arguments you have known so far?	2.85 (0.754)	2.83 (0.917)
Opinion	What is your current opinion about this claim?	**2.65** (1.168)	**2.94**** (1.232))
Interest	How interesting is the topic for you?	3.4 (0.963)	3.32 (1.115)

Table 2: Participants' ratings regarding both systems. M_{pre} denotes the mean before and M_{post} after the interaction. The differences that are statistically (highly) significant with ($\alpha < 0.01$) $\alpha < 0.05$ are marked with (**) *.

Category	Menu		Speech	
	M_{pre}	M_{post}	M_{pre}	M_{post}
RUE pro	3.41 (1.011)	3.36 (1.042)	3.58 (0.930)	3.35 (1.016)
RUE con	**3.42** (0.992)	**3.18*** (1.077)	3.46 (0.921)	3.40 (1.043)
True Knowledge	**3.21** (0.889)	**3.41*** (0.796)	3.43 (0.931)	**3.51** (0.900)
False Knowledge	2.80 (0.805)	2.80 (0.885)	2.90 (0.696)	2.86 (0.952)
Opinion	**2.62** (1.160)	**2.89**** (1.264)	**2.68** (1.181)	**3.00**** (1.201)
Interest	3.31 (1.034)	3.48 (1.005)	**3.49** (0.876)	**3.14**** (1.201)

Table 3: Participants' ratings on 5-point Likert scale separated in speech and menu system. M_{pre} denotes the mean of both system before and M_{post} after the interaction. The differences that are (highly) statistically significant with ($\alpha < 0.01$) $\alpha < 0.05$ are marked with (**) *.

interested) as shown in Table 2. As the polarity of the arguments the participants are interested in is of importance for the RUE, we distinguish between the interest for pro/con arguments with regard to the Major Claim. Interestingly, for both systems the results show a decrease of *RUE pro* and *RUE con* before and after the interaction. In particular this decrease is significant for *RUE pro* ($p < 0.01$) for both systems and with regard to *RUE con* ($p < 0.01$) for the menu system. This might be explained by the fact that the interest in new pro/con arguments tends to slightly saturate during the conversation. This meets our expectation that the more information is presented, it gets likelier users lose interest, especially if the ADS makes no efforts to keep up the motivation to engage in the discussion. Interestingly, the RUE ratings match the information-seeking behavior within the conversation, as for the menu/speech system 14,58%/22,67% of all performed moves requested contradicting and 28,47%/30,93% supporting arguments. Between $RUEpro$ and $RUEcon$ no significant difference is perceivable which indicates that the user themselves rated their interest rather equal, even tough the number between heard pro and con arguments is significantly different. This underpins our expectation that self-assessment should not mainly be considered for the RUE, but rather the actual user behaviour. To incorporate the latter, the implicit RUE calculation suggested in (Aicher et al., 2021a) shall be extended and investigated.

Regarding the users' interest in the topic itself, it is noticeable that even though the speech system users show a significant ($p < 0.01$) decrease in interest, the menu users reported an increase in interest. Consistent to the interest in pro/con arguments, the interest in the topic itself decreases in both systems and even significantly with regard to the speech system. Concerning the users' opinion, which significantly changes in both systems from a slight rejection of the major claim to a neutral position towards it (5 = Strongly agree, 4 = Agree, 3 = Neutral, 2 = Disagree, 1 = Strongly disagree). As significantly more pro arguments have been heard it is plausible that this might have led to an opinion change. This furthermore implies that a balanced exploring of arguments and thus the consideration of the RUE is important when aiming to break the users' SFBs and helping to form a well-founded opinion.

6.2. True and False Knowledge

With regard to the *True Knowledge* and *False Knowledge* participants had to rate the statements in Table 2 on a 5-point Likert scale ($5 \leq 100\%, 4 \leq 75\%, 3 \leq 50\%, 2 \leq 25\%, 1 = 0\%$). A significant ($p < 0.05$) change is perceivable in the *TK* for both systems and in particular, in the menu system. This meets our ex-

Category	both		Menu		Speech	
	$M_{\textbf{pre}}$	$M_{\textbf{post}}$	$M_{\textbf{pre}}$	$M_{\textbf{post}}$	$M_{\textbf{pre}}$	$M_{\textbf{post}}$
Alternative Relationships & Parenthoods	3.85 (0.973)	3.93 (0.995)	3.94 (0.933)	3.98 (0.975)	3.74 (1.008)	3.88 (1.018)
Children	4.00 (0.946)	4.03 (0.930)	4.03 (0.990)	4.10 (0.865)	3.97 (0.902)	3.96 (0.994)
Divorce	3.9 (0.995)	3.9 (1.007)	3.92 (1.031)	3.87 (1.080)	3.87 (0.959)	3.94 (0.929)
Remarriage	3.61 (1.007)	3.74 (1.000)	3.63 (1.089)	3.77 (0.997)	3.59 (0.918)	3.70 (1.007)
Harmful Relationships	3.62 (1.217)	3.63 (1.195)	3.51 (1.269)	3.52 (1.300)	3.73 (1.154)	3.76 (1.065)
Law	3.62 (1.217)	**3.77*** (0.988)	3.54 (1.088)	**3.74*** (1.005)	3.70 (0.911)	3.80 (0.973)
Relationship Stability	3.89 (0.913)	3.99 (0.898)	3.87 (0.996)	4.00 (0.914)	3.92 (0.821)	3.98 (0.885)
Religion	3.42 (1.284)	3.52 (1.227)	3.45 (1.336)	**3.66** (1.179)	3.39 (1.232)	3.38 (1.264)
Expectations & Commitment	3.97 (0.837)	3.98 (0.869)	4.03 (0.897)	4.03 (0.886)	3.90 (0.766)	3.93 (0.853)
Social Acceptance	3.44 (1.078)	3.52 (1.003)	3.45 (1.096)	3.56 (1.041)	3.43 (1.065)	3.49 (0.966)

Table 4: Participants' ratings of the indicator *PR* on 5-point Likert scale. M_{pre} denotes the mean of both system before and M_{post} after the interaction. The differences that are (highly) statistically significant with ($\alpha < 0.01$) $\alpha < 0.05$ are marked with (**) *.

pectation, as the system provides at least ten arguments to the user and thus, it is very likely that new information is provided, which is captured with this indicator. Considering the *FK* no change with regard to the menu and a slight decrease in the speech system was observable. As contradiction might be interpreted differently than false, it will need further exploration, if the user with regard to single arguments question their veracity. Still, we conclude that also the *FK* captures a result consistent with our expectation.

6.3. Personal Relevance:

Considering the indicator *Personal Relevance*, the users had to rate the subtopics shown in 4 regarding the statement "I think this aspect is personally relevant in the discussion if 'Marriage is an outdated institution."' on a 5-point Likert scale (5 = Strongly agree , 4 = Agree, 3 = Neutral, 2 = Disagree, 1 = Strongly disagree) before and after the interaction. Interestingly, except for the subtopics *Children, Religion* in the speech group and *Divorce* in the menu group show slightly decreasing ratings, whereas all other subtopics are rated higher after the interaction. A significant ($p < 0.05$) increase is perceivable for the subtopic *Law*. When comparing the increased ratings with the subtopic-relation of the heard arguments, one can observe a correlation. Thus, we perceive a tendency that the more arguments are heard on a specific subtopic, the more relevant this subtopic gets for the user. This fits our expectation that when consuming new information on new subtopics, the perception of relevance, especially if the former is convincing, will increase.

In conclusion, all discussed indicators which we extracted from literature research, showed a consistent behavior in the self-rating setting with real users and can be mapped onto interconnected aspects of argument exploration coherent with the SFB.

7. Conclusion and Future Work

In this work, we introduced potential indicators to model users' self-imposed filter bubbles and analysed self-ratings of human users with regard to these indicators. A significant change was perceivable with regard to the interest in pro arguments (*RUE pro*) and the *TK*. The *PR* increased in all subtopics, which was even significant with regard to the subtopic *Law*. Regarding the differences between speech and menu input/output we could see that all indicators may vary in strength, but go in the same direction. This is crucial as we aim for a SFB model which is invariant to different input/output modalities. All indicators showed consistent behaviour and already recognizable significant differences before and after the interaction consistent to our expectations. This indicates that These indicators enable us to detect changes occurring in the interaction. Thus we propose *Reflective User Engagement, Personal Relevance, Knowledge* and *False Knowledge* as suitable (but not limited thereto) dimensions to model a user's SFB.

These findings will serve as a starting point for further exploration in a user study, where the change of each indicator shall be tracked in detail during the dialogue. Moreover, in future work we want to explore the SFB indicators in more detail and merge them to model SFBs. As the herein presented results are based on self-ratings at distinct times, in a next step we will investigate methods to determine this indicators implicitly and continuously during the interaction, e.g. by incorporating our RUE calculation (Aicher et al., 2021a).

Especially, we aim to look into the change and behaviour of these indicators during the interaction and identify which further factors might be of importance, such as user trust (e.g. in argument sources), communication styles (the way arguments are presented) and a virtual agent interface. With the help of the SFB model, the system shall be trained via Reinforcement Learning approaches to be able to adapt to the user and engage the user to recognize and overcome their SFB. Therefore, the herein presented findings take us a step closer towards our aim to provide an ADS that helps users to build an opinion and fosters critical and reflective thinking and open-mindedness.

8. Acknowledgements

This work has been funded by the DFG within the project "BEA - Building Engaging Argumentation", Grant no. 313723125, as part of the Priority Program "Robust Argumentation Machines (RATIO)" (SPP-1999).

9. Bibliographical References

Abro, W. A., Aicher, A., Rach, N., Ultes, S., Minker, W., and Qi, G. (2021). Natural language understanding for argumentative dialogue systems in the opinion building domain. *arXiv*, arXiv:2103.02691.

Aicher, A., Minker, W., and Ultes, S. (2021a). Determination of reflective user engagement in argumentative dialogue systems. page 8.

Aicher, A., Rach, N., Minker, W., and Ultes, S. (2021b). Opinion building based on the argumentative dialogue system bea. In *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction: 10th International Workshop on Spoken Dialogue Systems*, pages 307–318. Springer Singapore.

Aicher, A., Rach, N., Minker, W., and Ultes, S. (2021c). Opinion building based on the argumentative dialogue system bea. *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction: 10th IWSDS*, pages 307–318.

Aicher, A., Gerstenlauer, N., Feustel, I., Minker, W., and Ultes, S. (2022). Towards building a spoken dialogue system for argument exploration. In *Proceedings of the 13th International Conference on Language Resources and Evaluation (LREC 2020)*. European Language Resources Association (ELRA).

Allahverdyan, A. E. and Galstyan, A. (2014). Opinion dynamics with confirmation bias. *PloS one*, 9(7):e99557.

Alsharif, H. and Symons, J. (2021). Open-mindedness as a corrective virtue. *Philosophy*, 96(1):73–97.

Amgoud, L. and Ben-Naim, J. (2016). Evaluation of arguments from support relations: Axioms and semantics. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence, IJCAI-16*, pages 900—-906.

Amgoud, L. and Ben-Naim, J. (2018). Weighted bipolar argumentation graphs: Axioms and semantics. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 5194–5198.

Bechhofer, S. (2009). Owl: Web ontology language. In *Encyclopedia of Database Systems*, pages 2008–2009. Springer.

Chalaguine, L. A. and Hunter, A. (2020). A persuasive chatbot using a crowd-sourced argument graph and concerns. In *COMMA*.

Chalaguine, L. and Hunter, A. (2021). Addressing popular concerns regarding covid-19 vaccination with natural language argumentation dialogues. In Jiřina Vejnarová et al., editors, *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, pages 59–73, Cham.

Del Vicario, M., Scala, A., Caldarelli, G., Stanley, H. E., and Quattrociocchi, W. (2017). Modeling confirmation bias and polarization. *Scientific reports*, 7(1):1–9.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ekström, A. (2021). The self-imposed filter bubble hypothesis. Student Paper.

Fazzinga, B., Galassi, A., and Torroni, P. (2021). An argumentative dialogue system for covid-19 vaccine information. In *Logic and Argumentation*, pages 477–485, Cham.

Hadoux, E. and Hunter, Anthony, e. a. (2021). Strategic argumentation dialogues for persuasion: Framework and experiments based on modelling the beliefs and concerns of the persuadee. In *arXiv*, volume 2101.11870.

Huang, H.-H., Hsu, J. S.-C., and Ku, C.-Y. (2012). Understanding the role of computer-mediated counter-argument in countering confirmation bias. *Decision Support Systems*, 53(3):438–447.

Jones, M. and Sugden, R. (2001). Positive confirmation bias in the acquisition of information. *Theory and Decision*, 50(1):59–99.

Le, D. T., Nguyen, C.-T., and Nguyen, K. A. (2018). Dave the debater: a retrieval-based and generative argumentative dialogue agent. *Proceedings of the 5th Workshop on Argument Mining*, pages 121–130.

Macpherson, R. and Stanovich, K. E. (2007). Cognitive ability, thinking dispositions, and instructional set as predictors of critical thinking. *Learning and Individual Differences*, 17(2):115–127.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2):175–220.

Pariser, E. (2011). *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin.

Petty, R. E., Briñol, P., and Priester, J. R. (2009). Mass media attitude change: Implications of the elabora-

tion likelihood model of persuasion. In *Media effects*, pages 141–180. Routledge.

Rach, N., Langhammer, S., Minker, W., and Ultes, S. (2018). Utilizing argument mining techniques for argumentative dialogue systems. In *Proceedings of the 9th International Workshop On Spoken Dialogue Systems (IWSDS)*, May.

Rakshit, G., Bowden, K. K., Reed, L., Misra, A., and Walker, M. A. (2017). Debbie, the debate bot of the future. In *Advanced Social Interaction with Agents - 8th International Workshop on Spoken Dialog Systems*, pages 45–52.

Rosenfeld, A. and Kraus, S. (2016). Strategical argumentative agent for human persuasion. In *ECAI'16*, pages 320–328.

Saha, T., Saha, S., and Bhattacharyya, P. (2020). Towards sentiment-aware multi-modal dialogue policy learning. *Cognitive Computation*, pages 1–15, 11.

Schwind, C. and Buder, J. (2012). Reducing confirmation bias and evaluation bias: When are preference-inconsistent recommendations effective–and when not? *Computers in Human Behavior*, 28(6):2280–2290.

Slonim, N., Bilu, Y., Alzate, C., Bar-Haim, R., Bogin, B., Bonin, F., Choshen, L., Cohen-Karlik, E., Dankin, L., Edelstein, L., et al. (2021). An autonomous debating system. *Nature*, 591(7850):379–384.

Stab, C. and Gurevych, I. (2014). Annotating argument components and relations in persuasive essays. In *COLING*, pages 1501–1510.

Villarroel, C., Felton, M., and Garcia-Mila, M. (2016). Arguing against confirmation bias: The effect of argumentative discourse goals on the use of disconfirming evidence in written argument. *International Journal of Educational Research*, 79:167–179.

Westerwick, A., Johnson, B. K., and Knobloch-Westerwick, S. (2017). Confirmation biases in selective exposure to political online information: Source bias vs. content bias. *Communication Monographs*, 84(3):343–364.

Woolson, R. (2007). Wilcoxon signed-rank test. *Wiley encyclopedia of clinical trials*, pages 1–3.

Yi, X., Hong, L., Zhong, E., Liu, N. N., and Rajan, S. (2014). Beyond clicks: Dwell time for personalization. In *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys '14, page 113–120, New York, NY, USA. Association for Computing Machinery.