How Will I Argue? A Dataset for Evaluating Recommender Systems for Argumentations

Markus Brenneis
Heinrich-Heine-Universität
Markus.Brenneis@hhu.de

Maike Behrendt
Heinrich-Heine-Universität
Maike.Behrendt@hhu.de

Stefan Harmeling
Heinrich-Heine-Universität
Stefan.Harmeling@hhu.de

Abstract

Exchanging arguments is an important part in communication, but we are often flooded with lots of arguments for different positions or are captured in filter bubbles. Tools which can present strong arguments relevant to oneself could help to reduce those problems. To be able to evaluate algorithms which can predict how convincing an argument is, we have collected a dataset with more than 900 arguments and personal attitudes of 600 individuals, which we present in this paper. Based on this data, we suggest three recommender tasks, for which we provide two baseline results from a simple majority classifier and a more complex nearest-neighbor algorithm. Our results suggest that better algorithms can still be developed, and we invite the community to improve on our results.

1 Introduction

Argumentation is an important tool of human communication and interaction. Arguments allow us to justify our views and opinions and persuade others. They also play an important role when it comes to decision-making. Not only in terms of law and justice (Collenette et al., 2020; Bench-Capon and Modgil, 2009), but also for each and every personal decision we make on a daily basis.

Taking a position on a controversial issue can be difficult, especially when there are many pro and contra arguments to consider. Finding the arguments that are most important and convincing for oneself is an important aspect in the process of decision-making. For a wide range of fields, recommender systems already facilitate our decisions, using collaborative and content-based filtering algorithms (Schafer et al., 2007), filtering the great load of information that can be found online (Bobadilla et al., 2013). A recommender system for argumentations could help users to make decisions more confidently and also gain a better understanding of the whole issue discussed. First applications like the *Predictive and Relevance based Heuristic agent* (Rosenfeld and Kraus, 2016) and our platform *deliberate* (Brenneis and Mauve, 2020) were presented to address this task. They try to present arguments to users which are most relevant for them.

But large-scale datasets to systematically test and evaluate such recommender systems for argumentations outside a laboratory setting are missing. In this work, we provide a dataset including more than 900 arguments and 600 user profiles, obtained as part of a larger study on political opinion-forming. In this study, we let participants interact with our platform *deliberate*, exposing them to arguments we gathered beforehand, concerning two different controversial questions on nutrition policy. The participants could rate the overall strength of the displayed arguments, indicate whether they find them convincing, and add own arguments. They were exposed to the topics at different points of time, such that the user profiles grow over time and the dataset can be used to test predicting future user behavior.

The dataset we provide here should serve to test and evaluate metrics and algorithms for argument recommender systems. As a baseline, we provide our results from two different algorithms on three different tasks which are predicting the user conviction towards an argument, the assigned strength of an argument, and the top-3 convincing arguments. The baseline results are obtained using a plain majority classifier and the existing recommender algorithm of *deliberate* to test its performance. To our knowledge, we provide the first large-scale dataset on the task of argument recommendation which contains user attitudes at different points of time.

The paper is structured as follows. In Section 2, the theoretical basics on argumentation and the

Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 360–367

terms used in this paper are defined. The data we collected is described in detail in Section 3. Section 4 introduces the three challenges and subtasks for argument recommendation we propose in this work, for which we provide two baseline results which are subsequently discussed. Section 5 gives an overview of related research, and finally, we summarize our work and look at future work.

2 Definitions

In this paper, we use terms based on the IBIS model (Kunz and Rittel, 1970), but our dataset can also be interpreted in bipolar Dung-style (Dung, 1995) argumentation frameworks. The atomic building blocks of argumentations are textual *statements*. Two statements, called *premise* and *conclusion*, form an *argument*. The premise can either support or attack the conclusion. A controversial statement which is argued about is called *position*, e.g., "plastic packaging for fresh food should be prohibited," and is typically an action which can be performed. Positions do not have a conclusion, but they can be used as conclusions when arguing why the position is sensible or not.

All statements define an argumentation graph where statements are nodes and the edges are arguments, i.e., they represent the argumentative relation between statements. For simplicity, user-interfaces like *deliberate* often call the premises themselves *arguments* to hide the technical definition of *argument* from the user. When the conclusion talked about is fixed, an argument can be uniquely identified by its premise.

Individual persons can have different *opinions* on the statements in an argumentation, e.g., agree or disagree with them with different *strengths* (i.e., the person can be (un)sure about their opinion). In real-world applications, a person's opinion on a statement can be unknown, leading to sparse data.

Furthermore, a person can consider an argument more or less convincing than another argument with the same conclusion; we call this *weight*, and we use a value from the interval $[0, 6]$ to represent it, where higher values correspond to stronger weights; this interval directly corresponds to the Likert scale we used during data collection.

We call the collection of weights and opinions of a person in an argumentation *attitude*. A person's attitude and their user name form a *user profile*.

S will refer to a set of statements. For a statement $s \in S$ which is an argument's premise, $c(s) \in \{0, 1\}$ indicates whether the argument is considered convincing (1) or not (0) by a user, and $w(s) \in [0, 6]$ is the associated weight. Predicted values for conviction and weight produced by a prediction algorithm are referred to as $\hat{c}(s)$ and $\hat{w}(s)$, respectively. The set of all user profiles is called U and can be represented as big sparse matrix with user profiles in the rows (i.e., in our case, with columns for the user name, position agreement strength, and, for each argument, columns for premise conviction and argument weight). Table 1 summarizes our notation.

3 Description of the Dataset

We present our new argumentation dataset with arguments on two different positions on nutrition policies in Germany (see Table 2): The prohibition of plastic packaging and the prohibition of genetic engineering. In contrast to other argumentation corpora, we also include the opinions and argument weights of different persons gathered at different points of time as part of an empirical study on political opinion-forming using our argumentation tool *deliberate* (Brenneis and Mauve, 2020).

The two discussed issues have been identified as the most topical and polarizing ones from a preselected set of controversial questions through a pre-test survey before our main study. In the original main study, we examined whether the use of artificial intelligence methods to pre-select arguments participants can see has an impact on the political opinion forming of individuals in the field of nutrition policies.

Now, we first explain the general data collection and the demographics of the participants. Afterwards, we expound on the pieces of information collected for our data set. Finally, we explain how the dataset looks like and where to obtain it.

3.1 Data collection & Participants

The main study was carried out over a period of four months, including three waves of data collection in August 2020 (T_1), October 2020 (T_2) and December 2020 (T_3). A pretest was conducted in April 2020 (T_0). The study participants were selected from the German online population, representative regarding age, gender, and education, and have agreed to the data publication. For the recruiting process and conducting our online study, we commissioned a German market-research company.

Table 1: Notation used throughout the paper.

S	set of statements
$c(s)$	individual's conviction in argument given by premise s (0 or 1)
$w(s)$	individual's integer conviction weight for corresponding argument (0–6)
$\hat{c}(s)$	algorithm's prediction for $c(s)$
$\hat{w}(s)$	algorithm's prediction for $w(s)$
U	set of user profiles
S_u	subset of statements for which the ratings of user u are known
$T_1 \rightarrow T_2$	predicting data from T_2 using data known at time point T_1
$T_2 \rightarrow T_3$	predicting data from T_3 using data known at time point T_2

In total, we had 674 participants whose data is included in our dataset: 264 in the pre-test T_0 and 410 in T_1, from which 121 dropped out in T_2 and 60 in T_3. The age span reaches from 18 to 74 with an average age of 46.5, which is slightly above the average age (44.5 (Statistisches Bundesamt (Destatis))) of the German population. 52.23% of the study participants were male (in comparison to 49.35% in the German population (Statistisches Bundesamt (Destatis))), 47.48% female (50.65% in the population). 42.14% had at least a high school degree, which exceeds the average for the population as a whole where only 33.5% have at least a high school degree (Statistisches Bundesamt (Destatis)).

Besides working with the argumentation tool, participants were presented a questionnaire which embedded the discussion software and collected, i.a., demographic information.

3.2 Data Collected by Us

Throughout each wave, the participants were exposed to arguments concerning the two different issues on nutrition policies. For each position discussed, a set of at least 18 supporting and 18 attacking arguments has been provided by us beforehand. We chose the arguments from a pre-selection of arguments on both topics that were clearly identifiable as pro or con in a pre-test. Other arguments could be added by the participants and the participants provided their attitudes on these positions and arguments.

For example, one statement arguing in favor of genetic engineering which was provided by us is "Genetic engineering is used to improve plants just like classical breeding, which is not prohibited." Participants who were presented that statement as supporting argument had to indicate *whether* they consider this statement to be a convincing argument for genetic engineering (binary decision) and *how much* they are convinced (Likert scale from *not convincing at all* (0) to *very convincing* (6)).

Overall, the following pieces of information were collected:

- T_0: Pre-test data with 264 participants; opinions and opinion strengths on positions about *plastic packaging* and *genetic engineering*; attitudes on at least 7 randomly selected arguments per topic.

- T_1: first main experiment with 410 participants; attitudes (opinions and opinion strengths) on *plastic packaging* and *genetic engineering* (no arguments involved).

- T_2: second main experiment with 289 participants (subset of users from T_1); attitudes (i.e. opinions and weights) on *plastic packaging* and on 3 randomly selected supporting, and 3 randomly[1] selected attacking arguments; users were able to contribute own arguments for/against the issue or other arguments (which were not included in the randomly selected arguments); attitude on *genetic engineering* (possibly changed since T_1).

- T_3: third main experiment with 229 participants (subset of users from T_2); attitudes on *genetic engineering* and 3 randomly selected supporting and 3 randomly selected attacking arguments; users were again able to contribute own arguments; attitude on *plastic packaging*.

To clarify, the settings in T_2 and T_3 only differ in the position being argued about. The opinions on all positions (whether a participant is for allowance or prohibition and how strong their opinion is) have

[1] Due to a technical problem, 8 of 36 arguments were not included in the random selection.

Table 2: Positions and number of records in the dataset; the number of arguments is split in the number of arguments provided by us beforehand and the number of new arguments entered by users (each counted as the number of unique premise statements).

Position	Number of Arguments	No. of User Profiles T_0	T_1	T_2	T_3
Should plastic packaging for fresh food such as fruit and vegetables be allowed or prohibited in Germany?	36+521	264	410	289	
Should the growing of genetically modified plants for food production be allowed or prohibited in Germany?	38+351	264	410		229

been collected at every time point, i.e. it was possible for participants to change their minds between each poll.

Arguments added by the users could be directly for/against the position discussed, or for/against other arguments.

Having collected the data at different points of time has several practical advantages: First, the data from T_0 and T_1 can be used to tackle the cold-start problem (Schafer et al., 2007) when predicting attitudes from T_2 and T_3, since the users' opinions on the positions is known from T_1. What is more, we can realistically check the performance of a real-world recommender system over time: The dataset considers that we might have incomplete information about persons (e.g., no argument attitude information for the new users in T_1), and we take into account that people might change some of their attitudes over time.

3.3 Content of the Dataset

Our complete dataset is freely available online[2] as CSV files, and the argumentation data is also provided in AIF (Chesnevar et al., 2006) for easy use in standard applications for argumentation frameworks. The dataset published in this work is part of a larger dataset with more experimental groups; we only publish the data of the group that was exposed to randomized arguments to ensure the data is not biased. The original statements are in German, but an English translation is supplied for better understanding of the dataset.

To get a feeling of how the data looks like, we describe the T_0 data (which is not part of any test set): There are 264 user profiles. In the context of the positions, 81% of the users support the prohibition of plastic packaging, 74% are in favor of the prohibition of genetic engineering. For the plastic topic, all pro-prohibition arguments are considered convincing by 81%; for genetic engineering, the number is 67%. The arguments against prohibition are convincing for 36%, or 41%, respectively.

The average length of the arguments in the initial argumentation pool compiled by us is 15.7 words (standard deviation 4.7). The mean length of the users' arguments is 10.4 words (standard deviation 7.3).

In the dataset provided, the user profiles are stored as a sparse matrix. The matrix for T_0 has 264 rows and 151 columns, of which at least 31 have a value (user name, opinion and strength on 2 positions, and at least 7 arguments per position with conviction and weight). The matrix for T_1 comprises all the user profiles from T_0 and, in addition, the profiles of new users from T_1, resulting in a matrix with 674 rows ($264 + 410$ users), and 151 columns. For T_2, the matrix contains a subset of updated rows of T_1; the users at T_2 are a subset of the T_1 users, i.e., users who left the empirical study between T_1 and T_2 are removed, leaving 553 rows; as new arguments were added, the matrix has 407 columns. Analogously, the matrix for T_3 is an update of the T_2 matrix and comprises 493 rows and 495 columns (note that there are not opinions for all statements, but only for a total of 247, as statements added by users from other experimental groups are also included).

4 Challenges and Baseline Results for Recommender Systems

Based on our dataset, we introduce three different classification and recommendation tasks where the opinions on statements and weights of arguments have to be predicted. We provide baseline results from a majority classifier and a neighbor-based recommendation algorithm to get a first feeling for the hardness of the tasks.

[2] https://github.com/hhucn/argumentation-attitude-dataset

4.1 Challenges

We propose the following three tasks on our dataset to show its applicability for further research on argument recommender systems:

1. Predicting a user's conviction
2. Predicting the argument weights
3. Predicting the most convincing arguments

For each task, it is possible to predict data from T_2 (for the *plastic packaging* topic) based on the data known at T_1 (i.e., including the data from T_0, which solves the cold-start problem), as well as the data from T_3 (*genetic engineering*) based on T_2. We will refer to those variants as $T_1 \rightarrow T_2$, or $T_2 \rightarrow T_3$, respectively. For dealing with sparse data, we follow an approach mentioned by Herlocker et al. (2004) for all tasks: We "ignore recommendations for items for which there are no ratings." The set of statements we evaluate a user $u \in U$ on with this approach is denoted as S_u. All prediction tasks are described in detail in the following.

4.1.1 Prediction of Conviction (PoC)

Based on the given data at time point T_i, predict whether the user considers an argument convincing (1) or not (0) for each user and each premise statement which was provided by us and for which the user opinion is known at time point T_{i+1}. The evaluation measure for this task is the mean accuracy: The accuracy for each user is calculated and then averaged over all users.

$$acc = \frac{\sum_{u \in U} \frac{\sum_{s \in S_u} [c(s) = \hat{c}(s)]}{|S_u|}}{|U|} \quad (1)$$

This tasks tests how good an algorithm can predict whether a user considers an argument the user has not seen before convincing.

4.1.2 Prediction of Weight (PoW)

Based on the given data at time point T_i, predict the weight for an argument (value in the interval $[0, 6]$) for each user and each argument which was provided by us and the user's weight is known for at time point T_{i+1}. We use the averaged root mean squared error as evaluation measure. This way, algorithms which produce some very bad predictions are punished.

$$rmse = \frac{\sum_{u \in U} \sqrt{\frac{\sum_{s \in S_u} \left(w(s) - \hat{w}(s)\right)^2}{|S_u|}}}{|U|} \quad (2)$$

Algorithms which perform well on this task are able to select arguments which are better suited to convince users.

4.1.3 Prediction of Statements (PoS)

Based on the given data at time point T_i, predict up to three statements the user considers convincing for each user and each premise statement which was provided by us and the user opinion is known for at time point T_{i+1}. We evaluate the macro precision on the created set of recommendations S_{u3} (which is commonly referred to as precision@3 (Silveira et al., 2019)).

$$p@3 = \frac{\sum_{u \in U} \frac{\sum_{s \in S_{u3}} [c(s) = \hat{c}(s)]}{|S_{u3}|}}{|U|} \quad (3)$$

In case S_{u3} is empty, that user is skipped in the evaluation. The goal of this task is measuring the quality of an algorithm's top recommendations, i.e., cases in which the algorithm is very sure that the user is convinced of a statement.

Many other tasks, e.g., predicting the opinion on positions, could also be looked at, but we limit ourselves to those three tasks in this paper. We think that the proposed tasks are important for applications which want to suggest interesting or persuasive arguments to a user.

Our dataset contains appropriate training data for the tasks we propose above, as well as a validate–test split (50%/50%): For each of the variants $T_1 \rightarrow T_2$, and $T_2 \rightarrow T_3$, the training data comprises the user profiles known at the points of time T_1, or T_2, respectively. The validation and test data contain the data of participants at T_2, or T_3, respectively, randomly assigned to either the validation or test dataset.

4.2 Baseline Results

We provide baseline results from a simple majority classifier and a more sophisticated nearest-neighbor (NN) classifier. The majority classifier always predicts the most common opinion of all users for which the opinion to be predicted is known (PoC) or considers the averaged weight (PoW and PoS).

The NN classifier was also used in our original research study to predict arguments that the users would most likely find convincing. We used it in some experimental groups, whereas other groups were confronted with randomly chosen arguments. We originally chose that algorithm on a best-guess basis because of a lack of suitable evaluation data

Table 3: Searched hyperparameter space.

n:	$5, 10, 20, 30, 40, 50, 100, 500$
α:	$0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$
depth:	$1, 2$

for comparing different algorithms before carrying out our study. Using our dataset, we can now quantify how good that algorithm actually is. By publishing our results we want to motivate other researchers to outperform our baseline results, and we provide an evaluation data set for future experiments that are similar to our own experiment.

The NN classifier uses the collaborative-filtering based recommendation algorithm from our argumentation tool *deliberate* (Brenneis and Mauve, 2020). To predict a value v, it first determines the n nearest users for whom the value to predict is known, using our pseudometric for weighted argumentation graphs (Brenneis et al., 2020). The pseudometric considers the attitudes of users and gives a higher weight to attitudes closer to the root of an argumentation (depending on a parameter α, where a lower α emphasizes positions over deeper statements in the argumentation tree, similar to the PageRank algorithm (Page et al., 1999)). Then, the value v of those nearest users is averaged, weighted by the calculated distance to each user.

The values for the hyperparameters have been chosen based on the results on the validation set. The search space is depicted in Table 3; all possible combinations were evaluated. The parametrizations used for each task are presented in Table 4.

Table 5 depicts the results on the test sets for both algorithms. From the results we can see that the NN algorithm performs better for all tasks and dataset combinations. The difference for the $T_2 \rightarrow T_3$ variant is always bigger than the difference for $T_1 \rightarrow T_2$. In the following section the results are discussed and analyzed in further detail.

The code to reproduce our results is provided together with our dataset.

4.3 Discussion of Baseline Results & Evaluation

From the increasingly greater difference of the NN algorithm, compared to the majority algorithm from $T_2 \rightarrow T_3$ to $T_1 \rightarrow T_2$, we can anticipate an NN algorithm to perform better on all tasks, if more thorough user profiles are available (remember that only two data points are known for participants in T_1). On the other hand, the description of our T_0 data has also shown that the arguments related to *genetic engineering* are considered less convincing on average than those for/against *plastic packaging*; this might be a disadvantage for the majority classifier when predicting the *genetic engineering* data for $T_2 \rightarrow T_3$. This could also explain why both algorithms perform worse when evaluated on data from T_3.

Although the NN approach outperforms the majority classifier, the difference is still quite small. It is certainly possible to build better predictors, maybe incorporating linguistic information of the arguments, e.g., the appearance of certain keywords, for instance "nature." Another approach would be using different metrics for the NN classifier or applying a completely different machine learning method, e.g., decision trees or neural networks.

We chose evaluation measures which seemed sensible for us in our applications contexts, i.e., within the use case of the software *deliberate*. But depending on the application, other evaluation measures might be more sensible, like utility and novelty (Silveira et al., 2019), which might need more data on how a user consumed an argument (comparable to the click-through rate for search engine results).

The way we handled sparse data for the evaluation can also be discussed. Herlocker et al. (2004), who suggested "to ignore recommendations for items for which there are no ratings" for sparse data, also point out a disadvantage of this method, namely "that the quality of the items that the user would actually see may never be measured." We do not think that this is a big issue in our evaluation context, since we basically evaluate the system on six randomly selected items per user for which the ratings are known.

5 Related Work

Similar datasets have been published before, and similar recommender tasks have been considered.

Habernal and Gurevych (2016) suggested the task of predicting convincingness of web argument pairs. They annotated and published a large-scale dataset of 16k argument pairs on 32 topics for the task of convincingness prediction and argument ranking. Different from our work, the task was not predicting the attitudes for each user for a given argument, but compare arguments in pairs and de-

Table 4: Hyperparameters for the nearest-neighbor classifier for each task, determined with the validation sets.

Task	n	α	depth of statements considered
PoC	20	0.5	2
PoW	100	0.5	1
PoS	10	0.5	2

Table 5: Results of our baseline methods on the test sets for the three different tasks for each dataset combination. NN always outperforms Majority.

Task	PoC (acc)		PoW ($rmse$)		PoS (p@3)	
Algorithm	$T_1 \rightarrow T_2$	$T_2 \rightarrow T_3$	$T_1 \rightarrow T_2$	$T_2 \rightarrow T_3$	$T_1 \rightarrow T_2$	$T_2 \rightarrow T_3$
Majority	.793	.639	1.80	1.95	.846	.627
NN	**.804**	**.675**	**1.74**	**1.82**	**.856**	**.677**

termine their objective convincingness.

Rahman et al. (2019) presented a dataset with 16 positions on 4 issues, for which 309 students gave their attitudes by adding arguments and indicating their level of agreement with that argument on a scale from -1 (total disagreement) to 1 (total agreement). Using the information about argument agreement, the agreement with the position was calculated. In our work, however, we explicitly ask for the agreement with a position, which allows a user to have an opinion which is inconsistent with their arguments. The authors also compared different algorithms for predicting user opinions on positions, where the best algorithm was a kind of soft cosine measure, which exploited feature similarity using position correlation.

Rosenfeld and Kraus (2016) tested different recommender agents in laboratory argumentation settings where arguments probably used next in a discussion were suggested. Different features were considered, i.a., the distance of arguments in the argumentation graph, a calculated argument strength, and the current context in the discussion. Several machine learning algorithms like SVMs and neural networks were evaluated. This is different from our work because we only recommend statements which are a premise for a given statement, although considering a broader suggestion strategy, which suggests statements from a different context, might be more appropriate for specific applications.

Chalaguine and Hunter (2020) presented a chat bot which should select appropriate counter-arguments, using cosine and concern similarity, with the goal of persuading a human to change their opinion. They compared their algorithms with a random baseline and got significantly better-than-random results for selecting relevant arguments. A crowd-sourced dataset with arguments about UK university fees was used (Chalaguine and Hunter, 2019). In contrast to our work, this dataset only contains arguments, but no user profiles with the attitudes of different persons on the arguments. The same applies to other corpora, like the Internet Argument Corpus (Walker et al., 2012).

6 Conclusion and Future Work

In our work, we introduce an extensive dataset which contains more than 900 arguments for two political positions and the user attitude data from more than 600 individuals, collected at different points of time. This dataset can be used for evaluating argument recommender systems, which can, e.g., be used to help people finding personally relevant arguments in discussions with many arguments. We suggest three different recommender tasks and provide baseline results from a simple majority predictor and a more sophisticated nearest-neighbor algorithm, which yields better results.

Our baseline results can still be improved on, and we invite everyone to develop better algorithms. Possible first improvements are considering linguistic information, and using different metrics for the nearest-neighbor classifier. What is more, other tasks could be defined on our dataset, e.g., predicting T_3 data from T_1 or non-convincing arguments. Furthermore, we want to research the effects of different recommendation strategies for argumentation on the formation of opinion when they are used to pre-filter content a user can see. Other evaluations in terms of novelty and utility should also be considered in the future.

Acknowledgments

We thank Marc Feger for translating the dataset. This publication has been created in the context of the Manchot research group *Decision-making with the help of Artificial Intelligence*, use case politics.

References

Trevor Bench-Capon and Sanjay Modgil. 2009. Case law in extended argumentation frameworks. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, pages 118–127.

Jesús Bobadilla, Fernando Ortega, Antonio Hernando, and Abraham Gutiérrez. 2013. Recommender systems survey. *Knowledge-based systems*, 46:109–132.

Markus Brenneis, Maike Behrendt, Stefan Harmeling, and Martin Mauve. 2020. How Much Do I Argue Like You? Towards a Metric on Weighted Argumentation Graphs. In *Proceedings of the Third International Workshop on Systems and Algorithms for Formal Argumentation (SAFA 2020)*, number 2672 in CEUR Workshop Proceedings, pages 2–13, Aachen.

Markus Brenneis and Martin Mauve. 2020. deliberate – Online Argumentation with Collaborative Filtering. In *Computational Models of Argument*, volume 326, page 453–454. IOS Press.

Lisa Andreevna Chalaguine and Anthony Hunter. 2019. Knowledge acquisition and corpus for argumentation-based chatbots. In *CEUR Workshop Proceedings*, volume 2528, pages 1–14. CEUR Workshop Proceedings.

Lisa Andreevna Chalaguine and Anthony Hunter. 2020. A persuasive chatbot using a crowd-sourced argument graph and concerns. *Frontiers in Artificial Intelligence and Applications*, 326(Computational Models of Argument):9–20.

Carlos Chesnevar, Sanjay Modgil, Iyad Rahwan, Chris Reed, Guillermo Simari, Matthew South, Gerard Vreeswijk, Steven Willmott, et al. 2006. Towards an argument interchange format. *The knowledge engineering review*, 21(4):293–316.

Joe Collenette, Katie Atkinson, and Trevor Bench-Capon. 2020. An explainable approach to deducing outcomes in european court of human rights cases using adfs. *Frontiers in Artificial Intelligence and Applications*, 326:21–32.

Phan Minh Dung. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321–357.

Ivan Habernal and Iryna Gurevych. 2016. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599, Berlin, Germany. Association for Computational Linguistics.

Jonathan L Herlocker, Joseph A Konstan, Loren G Terveen, and John T Riedl. 2004. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, 22(1):5–53.

Werner Kunz and Horst W. J. Rittel. 1970. *Issues as elements of information systems*, volume 131. Citeseer.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab. Previous number = SIDL-WP-1999-0120.

Md Mahfuzer Rahman, Joseph Sirrianni, Xiaoqing (Frank) Liu, and Douglas Adams. 2019. Predicting opinions across multiple issues in large scale cyber argumentation using collaborative filtering and viewpoint correlation. *The Ninth International Conference on Social Media Technologies, Communication, and Informatics*, pages 45–51.

Ariel Rosenfeld and Sarit Kraus. 2016. Providing arguments in discussions on the basis of the prediction of human argumentative behavior. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 6(4):1–33.

J. Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen. 2007. Collaborative filtering recommender systems. In *The adaptive web*, pages 291–324. Springer.

Thiago Silveira, Min Zhang, Xiao Lin, Yiqun Liu, and Shaoping Ma. 2019. How good your recommender system is? a survey on evaluations in recommendation. *International Journal of Machine Learning and Cybernetics*, 10(5):813–831.

Statistisches Bundesamt (Destatis). Gesellschaft und Umwelt.

Marilyn A. Walker, Jean E. Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A corpus for research on deliberation and debate. In *LREC*, volume 12, pages 812–817. Istanbul.