Empirical Methods for Modelling Persuadees in Dialogical Argumentation

Anthony Hunter*, Sylwia Polberg*
*Department of Computer Science, University College London, London, United Kingdom

***Abstract*—For a participant to play persuasive arguments in a dialogue, s/he may create a model of the other participants. This may include an estimation of what arguments the other participants find believable, convincing, or appealing. The participant can then choose to put forward those arguments that have high scores in the desired criteria. In this paper, we consider how we can crowd-source opinions on the believability, convincingness, and appeal of arguments, and how we can use this information to predict opinions for specific participants on the believability, convincingness, and appeal of specific arguments. We evaluate our approach by crowd-sourcing opinions from 50 participants about 30 arguments. We also discuss how this form of user modelling can be used in a decision-theoretic approach to choosing moves in dialogical argumentation.**

I. Introduction

Persuasion is an activity that involves one party trying to get another party to believe (or not believe) something or do something (or not do something). Consider, for example, a doctor convincing a patient to drink less, a road safety expert persuading drivers to not text while driving, or an online safety expert getting users of social media sites not to reveal too much personal information. Hence, it is an important and multifaceted human facility.

As computing becomes involved in every sphere of life, so too is persuasion a target for applying computer-based solutions. Persuasion technologies have come out of developments in human-computer interaction research (see for example the influential work by Fogg [1]) with a particular emphasis on addressing the need for systems to help people make positive changes to their behaviour, particularly in healthcare and healthy life-styles. Interestingly, argumentation is not central to the current manifestations of persuasion technologies [2]. Rather, there is an emphasis on either helping users to explore their issues (e.g. game playing) or helping users once they are persuaded to do something (e.g. diaries for recording calorie intake for weight management).

To address the lack of explicit argumentation in persuasion technologies, we have been developing a framework for persuasion in argumentation dialogues with an emphasis on behaviour change applications [3]. A **system** (the *persuader* running for example as an app) enters into a dialogue with a **user** (the *persuadee* using the app) to persuade them to accept a specific argument called the **persuasion goal** (e.g. eat more fruit because it is healthy for you). A **dialogue** is a sequence of moves where the possible moves at each step of the dialogue depend on a **protocol** and may include positing an argument, asking a query, answering a query, and more. We assume that a dialogue concerns an argument graph, i.e. a directed graph where each node denotes an argument and each arc denotes an attack as proposed by Dung [4]. In Figure 1, we give some examples of arguments in the form of such a graph.

By choosing appropriate arguments to present to the user, the system may raise the user's belief in the persuasion goal. However, for the system, there is a problem of how to communicate with the user and get his/her arguments, which is necessary in order to support a fair and frank persuasion dialogue. We assume the system cannot understand arguments presented in natural language, given the complexity of processing arguments in free text. Hence, the interface with the user is restricted. One solution is for the system to give a menu of arguments that the user might believe, and the user presents agreement/disagreement in each argument by giving it a score (as in a Likert scale [5]). This score is in the unit interval and denotes the belief the user has in the argument.

Example 1. *Suppose the system gives argument* A *in Figure 1 as its persuasion goal. It is aware of two potential counterarguments* B *and* C*. So it presents these in a menu, and asks the user for his/her degree of belief in them. If the user declares belief greater than 0.5 in* B *(resp.* C*), then the system presents* D *and/or* E *(resp.* F*) with the aim of lowering the user's belief in* B *(resp.* C*) and increasing the user's belief in* A*.*

Asking the user about which arguments s/he believes allows for a dialogue to be tailored to the user. However, a user might be asked too many questions, which might cause the user to disengage. To address this, we can construct a **user model** that contains information such as the belief that the user has in some of the arguments. The system can then harness the user model to choose its moves, as in the next example.

Example 2. *Suppose the system gives argument* A *in Figure 1 as its persuasion goal, and presents* B *and* C *in a menu. Also suppose the user expresses belief of 0.9 in* B *and belief of 0.1 in* C*. Suppose the system does not want to present both* D *and* E *as a counterargument to* B*, and it wants to choose the argument that is most likely to be believed. Also, suppose it does not want to ask another question. If it has a user model, that says for example that* D *has belief of 0.8, and* E *has belief of 0.1, then the system will present* D*.*

So far we have focused on belief in arguments as beliefs are a primary feature of theoretical models of behaviour change.

This research is funded by EPSRC Project EP/N008294/1 "Framework for Computational Persuasion".

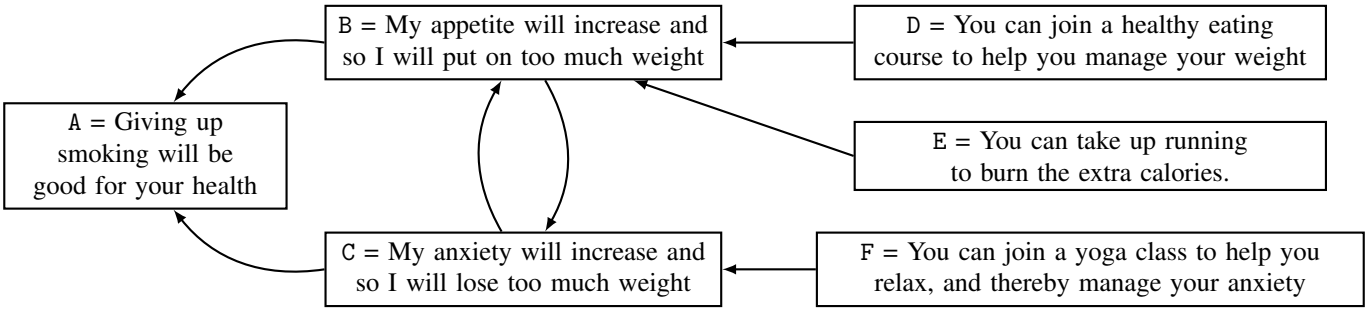

Fig. 1: Example of argument graph for persuasion. Each node denotes an argument, and each arc denotes one argument attacking another argument. The argument graph contains the arguments known (but not necessarily believed) by the system. Argument A could be a persuasion goal and so B and C are potential counterarguments for the user.

There is substantial evidence in the behaviour change literature that shows the importance of the beliefs of a persuadee in affecting the likelihood that the persuadee will be persuaded by a specific behaviour change intervention (see for example the review by Ogden [6]). In behaviour change, beliefs concern a variety of issues including: causes of the healthcare problem; risks to the agent from the healthcare problem; benefits to the agent from resolving the healthcare problem; opportunities for the agent for resolving the healthcare problem; capacity of an agent for resolving the healthcare problem; and the views of the agent's family of the healthcare problem.

To represent and reason with belief in arguments, we can use the epistemic approach to probabilistic argumentation [8]–[11]. Applying this approach to modelling persuadee's beliefs in arguments has produced methods for updating beliefs during a dialogue [12]–[14], for efficient representation and reasoning with the probabilistic user model [15], and for harnessing decision rules for optimizing the choice of arguments based on the user model [16]. These developments offer a well-understood theoretical and computationally viable framework for applications such as behaviour change.

However, so far we have not considered how the user models could be acquired (i.e. how we can efficiently and effectively obtain information from users about their opinions in order to be able to construct user models). There is also the question of whether we should continue to focus on belief in arguments, or whether there are other dimensions that we could consider for evaluating arguments, such as the convincingness and appeal of the arguments.

The ability to put forward convincing arguments is vital to a persuasion system, and as observed in [6], this is not necessarily the same as putting forward believable arguments. For example, while a patient may very well believe the doctor that smoking damages his health and that it would be best for him to quit, it does not necessarily mean that he is convinced enough to actually do it. Argumentation systems also often do not consider how appealing a given argument is to the persuadee, which is perhaps surprising given the fact that presenting arguments that appeal to people is a vital aim of many product advertisements as well as of various politicians. For example, people may find arguments that are in line with their personal views to be more appealing than those that contrast them strongly. By focusing purely on the correctness of the arguments, we often do not reflect on how the persuadees are going to feel about the information we provide them. Furthermore, such a focus has led to the development of software for persuasion which, despite being correct in the dialectical argumentation sense, has been considered offensive and judgmental by its users [17].

To address these issues, we investigate empirical methods for developing the user model. We consider how we can collect opinions from users on arguments, and how we can predict opinions on a variety of arguments based on a user's opinions on some arguments. We evaluate our approach by crowd-sourcing opinions from 50 participants about 30 arguments.

Note, in this paper, our user models are restricted to the opinions they might have of arguments. Obviously, this is just one dimension of what we ultimately require for more sophisticated user models. Other dimensions of a user that could be important in persuasion include personality, biases, and emotional state of the user [18] as well as goals, commitments, motivation, and past behaviour of the user [6]. We return to these other dimensions in our discussion of related work and future work.

In the following, we explain the methods used for acquiring crowd-sourced opinions on arguments (Section II), analyze the correlations between scoring criteria (Section III), investigate the classifiers for predicting scores (Section IV-B), discuss how these classifiers can be harnessed in modelling the persuadee (Section V), review the related literature (Section VI) and discuss our contributions (Section VII).

II. Data & Methods

In this section we describe the arguments used in the experiment, which of their aspects the participants were asked to judge, the methods we used for recruiting participants, and the methods used to analyze the obtained results.

A. Arguments

For the purpose of our study, we have created a data set consisting of 30 arguments that were split into the three

primary categories – celebrity, scientific and society – each one of them with a particular format. Every category consisted of 10 arguments. The arguments have been presented in English. The topics of the arguments are medicines, recycling, electric cars, and coffee.

Celebrity arguments are of the form "Person X says Y. Therefore, you/we should Z", where X is a (fictional or real) celebrity, such as an actor or a singer, Y is a claim made by this person and Z is a possible opinion or action resulting from this claim. For example, "*Melissa Latimer, a popular health and fitness celebrity, says that coffee can disturb the natural rhythm of the body and cause sleeping issues. Therefore, you should drink less coffee and replace it with healthier options, such as green tea.*"

Scientific arguments look similar to the previous ones, but instead of using a celebrity as the "source" for a given claim, we use the studies or reports carried out by (again, fictional or real) scientific organizations. For example, "*Extensive scientific studies carried out by the Australian Government National Health and Medical Research Council show that there is no evidence that homeopathy is an effective treatment for any health condition. Therefore, we should not use it as an alternative to traditional medicine.*"

Society arguments are of the form "Y. Therefore, you/we should Z". There is no explicit evidence for the claim Y (the intention is to use "common knowledge") and the argument is meant to the reflect social pressure or the possible dangers or benefits to the society. For example, "*Vaccines are crucial in building herd immunity and preventing diseases from spreading, which is important for people with compromised immune systems. Therefore, we should receive vaccines for our wellbeing as well as for the people around us.*"

In each category, five of the arguments are pro arguments (i.e. arguments for something) and five of the arguments are con arguments (i.e. arguments against something). As this paper is not concerned with dialectical issues (i.e. with how an opinion in an argument can change with the presentation of a counterarguments), we do not necessarily give a con argument for each pro argument or vice versa.

B. Criteria for judging arguments

There are various criteria for judging an argument. Its length, formality, language, etc. can be used to estimate its convincingness [19]. Appeals to our vanity or use of flattery as deployed by salespeople can convince us to buy a given product [20]. A good argument can also be seen as one strongly grounded in facts and evidence. If we look at dialectical semantics (e.g. [4]), then the arguments we can defend, or even better, which are not attacked at all, are the ones we want.

Rather than attempt to investigate too many aspects of an argument, as a starting point we have chosen to focus on three of them - how believable, how convincing and how appealing an argument is. We chose these dimensions because they provide a seemingly diverse and insightful range of notions for evaluating an argument. Furthermore, we can easily find arguments that have a high score in one of the aforementioned dimensions but a low one in another, for example:

- "*Smoking causes numerous diseases. Therefore, you should quit.*" On its own, this is an argument that many smokers will believe. However, the number of people that will actually be convinced to quit is not significant.
- "*Education should be free for everyone independently of race, gender or religion. Therefore, we should abolish the tuition fees incurred on students by the universities.*" Many people will consider the argument very appealing. At the same time, we can acknowledge that universities need resources to function. There is a reason why the tuition fees in many of the highest ranking universities and institutes in the world force a lot of students to take up loans. Hence, as appealing as this argument may be, its convincingness may be much lower.
- "*We have a found a tumour in your brain and, if it is left untreated, you have a year left to live. You will eventually develop seizures, difficulties with speech, movement and vision and experience severe headaches. Therefore, we would advise you to undergo a surgery to remove as much of the tumour as possible and follow it up with radiotherapy.*" This is not a statement anyone wants to hear. However, as unappealing it is, we consider it to be quite convincing and it is likely that the patient will decide to undergo the treatment to extend his life.

The examples above show that these three notions are, to some degree, distinct. Therefore, it makes sense to ask how far do the differences go, whether all of these dimensions should be taken into account or, based on the correlation between them, can one of them become a proxy for another. We will address these issues in Section III. In future work, we will consider further dimensions (e.g. how plausible, or how compelling, an argument is).

C. Recruitment

In this experiment we recruited 50 participants and asked them to score the aforementioned arguments in terms of believability, convincingness, and appeal on the scale from -10 to 10. No definition was given to participants of the terms (i.e. believability, convincingness, and appeal), as we wanted to investigate empirically the diversity ways that people may score them. However, we did check that they had a reasonable understanding of the general meaning of them and saw the differences between them. For this, we presented 6 sentences and asked the participant to complete them using words from a list. For example:

- "*We went to the store just to buy a fishing rod, but the seller talked us into getting an extra fishing line and some hooks. I think he is a very ___ person*" – the answer is "convincing".
- "*My blind date turned out to be a very closed-minded and racist person. Things like these put me off. I find them ___*" – the answer is "unappealing".

- "*My 5 years old nephew said he can lift the couch over his head. I don't think he is telling the truth, which means I find all of this* ___" – the answer is "unbelievable".

The participants were not informed of the category (i.e. celebrity, scientific, or society) to which a given argument was assigned. Also, we ensured that the arguments belonging to a single category were not presented together.

The recruitment was done using Amazon Mechanical Turk (AMT for short) and the survey ran on Survey Monkey, which are both common platforms for experiments of this type. Results are available online[1]. In addition to the aforementioned test, the participants were subjected to an additional language exercise and two attention checks to ensure their skills and honesty of their work. The language exercise was comparable to the Cambridge English: First (FCE) qualification test in terms of difficulty. The attention checks were meant to disqualify participants who are too distracted or simply resort to random clicking in order to complete the survey. They were presented with two sentences (one after 10 and the other after 20 arguments) requesting them to enter particular values on the believability, convincingness and appeal scales. We ran the survey until 50 participants meeting our requirements were found. This brings us to a 66% acceptability rate.

D. Methods for prediction studies

For prediction, we use the naive Bayes classifier (a well–known method in machine learning). It is a simple approach that often performs well in comparison with more sophisticated methods. We combine it with a 5–fold cross–validation in order to ensure the quality of each classifier we train. This means that the data set is split into 5 non–overlapping parts, where 4 parts are used for creating the classifier and 1 part is used for predictions. The training and testing is then repeated 5 times in a way that every part is used for predictions only once and every part is used for creating the classifier equally many times. We thus obtain a vector of real and estimated data, which is then used for creating a confusion matrix (explained below). From it we obtain the accuracy and F1–score that are used to judge our results. This analysis has been performed using the e1071 and caret packages in R.

A confusion matrix is a table which compares the actual and predicted data. The structure of a 2x2 confusion matrix is visible in Table I. Each column represents the instances in a predicted class while each row represents the instances in an actual (i.e. real) class. The total number of occurrences of class X (other than X) in the actual data is denoted with P (N). The predictions in which X is correctly guessed as X is referred to as true positives (TP) and those in which it is misclassified as not being X are referred to as false negatives (FN). The classes that are not X but are predicted as X are called false positives (FP). Finally, if they are correctly recognized as different from X, we refer to them as true negatives (TN). A number of parameters can be calculated from these values, accuracy and F1–score in particular. Accuracy tells us how often the classifier is correct. F1–score is calculated as the harmonic mean of precision and recall. Precision is the chance that a value predicted as X is really X and recall can be seen as the chance of X being correctly predicted as X.

[1]http://www0.cs.ucl.ac.uk/staff/a.hunter/papers/empiricalappendix.zip

		Predicted		
		X	not X	
Real	X	TP	FN	= P
	not X	FP	TN	= N

TABLE I: 2x2 confusion matrix

$$\text{Accuracy} = \frac{TP+TN}{P+N} \qquad \text{Recall} = \frac{TP}{TP+FN}$$
$$\text{Precision} = \frac{TP}{TP+FP} \qquad \text{F1} = \frac{2\times Precision\times Recall}{Precision+Recall}$$

The above formulas are suitable for a 2x2 confusion matrix. In order to obtain values for larger matrices, we proceed in the following manner. The overall accuracy is the sum of the diagonal values (which represent the correct predictions) divided by the the sum of all table entries (which is the total number of predictions). Other values are calculated per class, which means that e.g. a matrix with columns and rows X, Y and Z can be transformed into three 2x2 matrices that separately focus on predictions being X or not X (resp. Y and not Y, Z and not Z). They can be either reported as such, or one can calculate their averages. In our case, the average F1–score is taken as the harmonic mean of the average precision and recall, which are calculated in the standard way.

III. Correlations between scoring criteria

For every two of the three listed attributes (i.e. believable, convincing, appealing), we have calculated the (Spearman) correlation between them for arguments belonging to the same group and for all arguments altogether. Our findings are presented in Table II.

These results show a very strong correlation between what is seen as believable and convincing in all categories. Hence, they indicate that believability is a good proxy for convincingness and therefore arguably for persuasiveness. This is an important observation as it may allow us to focus primarily on believability in user modelling.

There is also quite a strong correlation between how appealing and how believable or convincing a given argument is. However, it is lower than in the previous case. We can observe that the appeal appears to be visibly less correlated with the other criteria in the case of scientific arguments. It is possible that these arguments were seen as more formal and complicated when compared to other categories. Nevertheless, at this point we cannot say with certainty what has caused this

Category	Believable - Convincing	Convincing - Appealing	Appealing - Believable
Celebrity	0.85	0.63	0.64
Scientific	0.89	0.50	0.46
Society	0.90	0.69	0.66
Total	0.89	0.61	0.59

TABLE II: Correlations between dimensions

		Predicted								
		Appeal			Believability			Convincingness		
		Celebrity	Scientific	Society	Celebrity	Scientific	Society	Celebrity	Scientific	Society
Real	Celebrity	44	3	3	45	4	1	43	4	3
	Scientific	8	39	3	7	35	8	6	38	6
	Society	7	3	40	6	7	37	4	5	41
F1–score		0.81	0.82	0.83	0.83	0.73	0.77	0.84	0.78	0.82
Accuracy		0.82			0.78			0.81		

TABLE III: Confusion matrices for category predicting problem w.r.t. a given dimension

effect. We believe that further studies, that use more arguments and divide appeal into additional dimensions, would help us to clarify this situation.

IV. Prediction

From each participant we obtained in total 90 answers. By fixing the dimension we are interested in to any of appealing, believable or convincing, we obtain 10 scores per category for a single person. These values can be seen as a category profile of a participant. Some of the questions we could ask at this point are (1) given the profile values of a person, can we predict the category of this profile? (2) given some of the arguments from a category, how well can we predict the values for the remaining arguments? (3) how well can arguments from one category predict the values in another category? In this section we will answer these questions by training and testing naive Bayes classifiers.

A. Prediction of categories

In this subsection we focus on how well we can predict the category based on the profile values given to us by the participants. Good results in this task could indicate that the arguments inside the groups are sufficiently related in order to be able predict their topic, similarly as the weight and height of a person could be used for predicting whether we are dealing with a male or a female.

For the purpose of this task, for every dimension we create a table in which every row contains the category (celebrity, scientific or society) and the 10 answers associated with it. Therefore, every participant is described using three rows. During the cross–validation stage, we have ensured that every row of every participant is used for testing precisely once and equally many times for training. The confusion matrices comparing the real and the predicted categories are presented in Table III. In this table we can also find the F1–score associated with every class and the overall prediction accuracy. We can observe that despite the limited number of participants, the category recognition of a profile works quite well. In the vast majority of cases, if a tuple of answers is predicted as belonging to a given category, then it is in this category and vice versa. This indicates that even though the participants are not aware of the categorization of the arguments, there is a certain coherence to the categories.

B. Prediction of scores for arguments

The purpose of this part of our study is to see how well we can predict the score assigned to a given argument by a participant. In the interests of space, we will present results concerning only the believability of an argument. Note, we get very similar results for convincingness (which is unsurprising given the strong correlation between believability and convincingness), and we get slightly better results for appeal.

At this point every row in our data set used for creating the classifiers now represents the beliefs of a given participant about our 30 arguments. In order to simplify the analysis for predicting scores, we replace the original scores in the data by an interpretation. Hence, in this case a given argument can be *strongly believed* (values from 10 to 7), *believed* (6 to 3), *undecided* (2 to -2), *disbelieved* (-3 to -6) or *strongly disbelieved* (-7 to -10). However, please note that splitting the values into fewer or more classes is of course possible.

1) Approaches: We first consider predicting the exact value of an argument given other arguments from the desired category. As a starting point, we first create three classifiers for every argument, one in which the predictors are the other 9 arguments in its own category and two in which the 10 arguments belonging to the remaining categories are used. We will treat the obtained results as the baseline. However, certain arguments, despite being in the same category, might not be as strongly related as others. This may be caused by, for example, their topic, certain secondary characteristics or simply by one of them being too "noisy" due to the difficulties that the participants may have had in judging it. Therefore, what we have done is to use only some of the arguments as predictors and focus on the results that perform better than our baseline either in terms of overall accuracy or average F1–score. We call each of them a "best" classifier.

We also consider a relaxation in our predictions. For example, misclassifying a "strongly believed" argument as "believed" may, depending on the context, not be a severe error. Thus, we have allowed an error margin of 1 in our predictions, i.e. the predicted result that is at most one class away from the actual value is seen as "satisfactory". Thus, we have again looked at the behaviour of our classifiers and obtained a new baseline and the best results with an error margin 1.

2) Findings: To summarize, we have two types of classifiers per argument – baseline and best – and for each of them we consider an error margin of 0 and of 1. The accuracy and F1–scores associated with predicting an argument from a given category based on other arguments in the same category can be seen in Figure 2. Every marker on a given line represents the value obtained by a single argument belonging to the desired

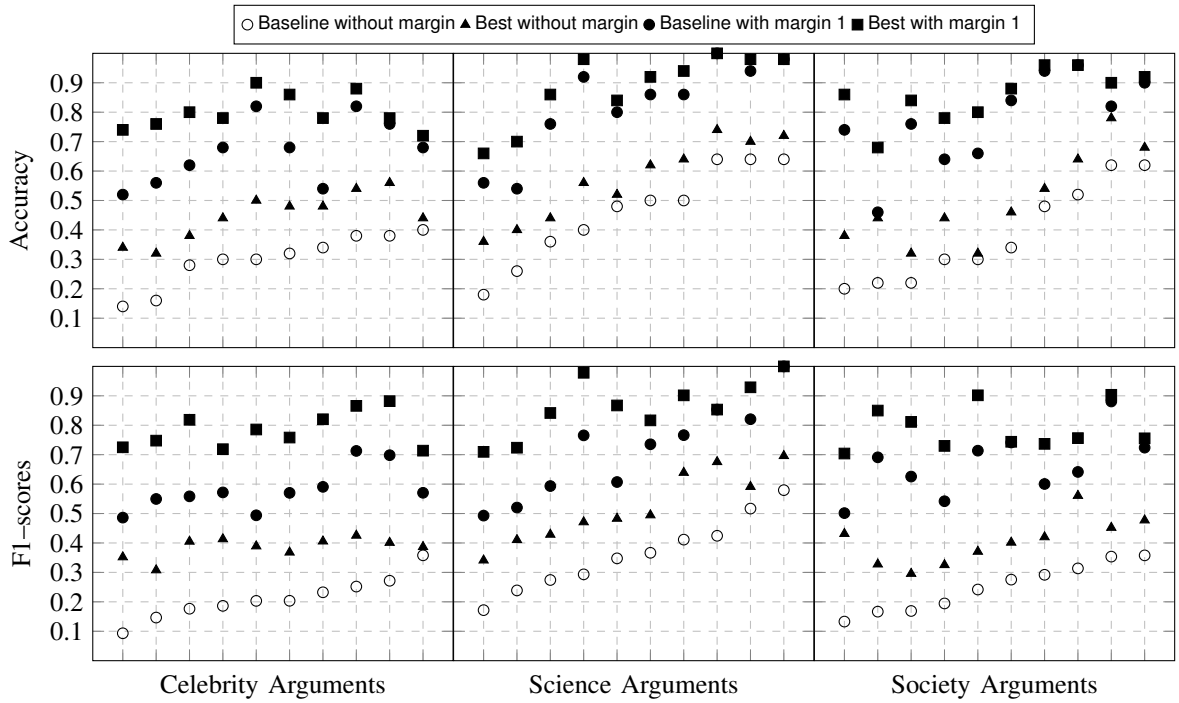

Fig. 2: Accuracy and F1–scores of predicted scores of arguments in a given category w.r.t. this category. Each point on the x–axis denotes an argument. For presentation purposes, the results have been ordered by increasing accuracy or F1–score w.r.t. the baseline without margin.

	Predicting category	Celebrity		Scientific		Society	
	Predicted category	Accuracy	F1	Accuracy	F1	Accuracy	F1
Best without margin	Celebrity	0.48	0.39	0.52	0.46	0.52	0.43
	Scientific	0.59	0.53	0.58	0.53	0.57	0.48
	Society	0.58	0.46	0.59	0.46	0.53	0.41
Best with margin 1	Celebrity	0.816	0.79	0.84	0.82	0.83	0.80
	Scientific	0.90	0.86	0.90	0.86	0.90	0.86
	Society	0.89	0.85	0.88	0.83	0.86	0.80

TABLE IV: Average accuracy and F1–scores for predicting with different categories

category. We can observe that not all arguments are equally well predicted in the baseline case. In the science and society categories we can find arguments that, even with the limited size of our sample, have an accuracy above 0.6. Unfortunately, we can also encounter arguments with accuracy lower than 0.2, which means that the chance of getting the correct score w.r.t. the baseline may be worse than if we were to use a random guess. Fortunately, in all of the cases we can find a classifier using only a subset of the predictors that performs better than the baseline. Similar observations can be made in the case of F1–score. By allowing the classifier to be off by a single class (i.e. we use error margin of 1) we obtain a baseline that performs even better than the best predictions without any margins. For certain arguments, by improving upon this baseline we can even reach accuracy of 1.

In Table IV we can find the best average accuracy and the best average F1–scores of predictions made for arguments in category X by arguments in category Y. We can observe that the scientific arguments appear to be easier to predict than others and that they tend to serve as better predictors as well. This information is particularly valuable, as obtaining and verifying such arguments appears to be less difficult than in the case of other categories. Moreover, we can expect that the credibility and general opinion of a given scientific institute may be more stable when compared to e.g. the popularity of a given celebrity. This means that the quality of such a data source should not overly decrease with time.

V. Using prediction in persuadee modelling

The methods and results in this paper show that it is possible to gather data from a set of participants on their scores (e.g. belief, convincingness, appeal) for each set of arguments, and use this data to train the classifiers using off-the-shelf software. This can prove useful in strategic argumentation, particularly when it comes to creating the user models and deciding which arguments should or should not be put forward by the system during a dialogue. If the dialogue is preceded by a profiling phase in which we find out more about a particular user, then

by knowing which arguments serve as good predictors we can limit the number of questions the system needs to ask. This may be advantageous if the dialogue concerns a sensitive topic, and the user is reluctant to reveal too much information, or the system is concerned that the user might disengage if the dialogue involves too many questions.

The predicted belief scores can be harnessed by the epistemic approach to probabilistic argumentation, where the degree to which an argument is believed is derived from a probability distribution over the subsets of arguments [8]–[10], [12]. For an argument A, $P(A) > 0.5$ represents A is believed to some degree, $P(A) = 0.5$ represents A is neither believed nor disbelieved, and $P(A) < 0.5$ represents A is disbelieved to some degree. Although the model of the user is a probability distribution over the power set of arguments in the argument graph, not over single arguments, given a tuple of answers by a participant and classifiers for argument $A_1, \ldots, A_n$ that may predict that the believability of each $A_i \in \{A_1, \ldots, A_n\}$ as k_i, it is straightforward to identify a probability distribution P such that for each $A_i, P(A_i) = k_i$. Furthermore, taking the structure of the graph into account, we may choose to assume rationality constraints (postulates) on satisfying probability distributions, and then use distance-based methods for minimally changing a probability distribution in order to satisfy the set of postulates [11].

As we discussed earlier, our framework for persuasion has so far been based on taking the user's beliefs in arguments into account when choosing moves [3], [12], [16]. However, we can harness the further dimensions of the user's opinion of convincingness and appeal by taking an aggregation of the three dimensions (e.g. a weighted average) in our decision-theoretic framework [16].

VI. Related work

Most proposals for persuasion in dialogical argumentation focus on protocols (for a review see [21]). Some strategies have been investigated (e.g. [22]–[25]) but there are relatively few proposals that formalize user modelling. A probabilistic model of the opponent has been used in a strategy allowing the selection of moves based on what it believes the other agent is aware of [26]. The history of previous dialogues has been used to predict the arguments that an opponent might put forward [27]. For modelling dialogues, a probabilistic finite state machine can represent the possible moves that each agent can make in each state of the dialogue, and this has been generalized to partially observable Markov decision processes (POMDPs) when there is uncertainty about what an opponent is aware of [28]. However, none of these proposals consider the beliefs of the opposing agent. In [29], a planning system has been used by the persuader to optimize the choice of arguments based on what premises are believed. However, there is no consideration of how the beliefs are obtained or how they updated during the dialogue.

In an empirical study, Rosenfeld and Kraus [30] (and extended in [31], [32]) undertook an experiment in order to develop a machine learning-based approach to predict the next move a participant would make in a dialogue. The machine learning models were trained on data that incorporated the sequences of arguments in a dialogue that the participants accept. Once trained, the models were able to predict the acceptance an unseen case would have. However, this work only concerned whether the participants in their studies regarded arguments as acceptable or not, and it did not consider how the participants viewed the individual arguments. Therefore, it does not allow for a model of the user to be constructed that, for instance, modelled the degree to which the user believes an argument. Hence, it does not allow for a probabilistic model to be extracted that could be used in a decision-theoretic framework for strategic argumentation (such as our decision-theoretic framework [16]) .

There are some studies of computational models of argument with participants by Rahwan *et al* [33] and Cerruti *et al* [34] that investigate reinstatement. These studies were aimed at investigating how well existing argumentation theories performed in describing user behaviour. The users were presented several argument graphs and were asked to explain how acceptable a given argument is in their opinion. The results show that in some cases, the implicit knowledge about domains can substantially affect the given acceptability levels. However, more importantly, the experiments show that the attacked argument's acceptability is lowered, but does not fall to 0, which is what would be predicted by the usual dialectical semantics for abstract argumentation. Additionally, introducing the defense for this argument raises its acceptability. However, typically it does not reach the value of 1, which is the level the usual dialectical semantics would predict. These studies lend support for using a finer grained representation of belief/disbelief, but they do not provide a framework for constructing the user models that can be directly harnessed in argumentation systems.

There are user studies that investigate the persuasiveness of arguments. Lukin *et al* [35] have shown that with some audiences, emotional arguments are more effective in persuasion than factual arguments. For this, they categorized audiences according to the OCEAN personality traits (i.e. openness to experience, extroversion, agreeableness, conscientiousness, and neuroticism). Then in a user study on the persuasiveness of healthy eating messages [36], positively framed messages (e.g. Most people believe that eating a healthy breakfast contributes to a longer lifespan) were shown to be more persuasive than negatively framed messages (e.g. Most people believe that eating an unhealthy breakfast contributes to a shorter lifespan). Furthermore, Cialdini's principles of persuasion [20] were considered (i.e. reciprocation, commitment, consensus, liking, authority, and scarcity), and it was found that arguments that appeal to authority (e.g. Studies conducted by health experts have shown that eating a healthy breakfast keeps you energized) were shown to be the most persuasive.

VII. Conclusions

In this paper, we have developed and evaluated methods for acquiring crowd-sourced opinions on arguments, and shown

how they can be used for predicting opinions on arguments. This shows how it is viable to acquire data to construct classifiers, and that these can then be deployed to substantially decrease the number of questions that need to be asked of a user. In future work, we will use the methods in this paper for training classifiers on arguments concerning two case-studies we are developing in behaviour change in healthcare.

Two main approaches to probabilistic argumentation are the epistemic approach (discussed in Section I) and the constellations approach (e.g. [37], [38]). The epistemic approach captures belief in arguments, and as discussed in Section V, can be used to represent and reason with the predicted scores. The constellations approach captures uncertainty in the structure of the graph (formalized by a probability distribution over the subgraphs). It would be interesting to extend the methods in this paper to predict probability distributions for the constellations approach.

The work in this paper is a starting point for further experiments. In particular, we would like to investigate further judging criteria. For instance, emotion in argumentation has been studied with participants in a debate where the emotional state was estimated from EEG data and automated facial expression analysis [39]. They showed, for example, that during the dialogue, the number and the strength of arguments and the relations between them could be correlated with particular emotions of the participants. Requesting the participants to state their emotional reactions to the presented arguments could help us in improving our predictions.

REFERENCES

[1] B. Fogg, "Persuasive computers: Perspectives and research directions," in *Proc. of CHI'98*. CHI, 1998, pp. 225–232.

[2] A. Hunter, "Opportunities for argument-centric persuasion in behaviour change," in *Proc. of JELIA'14*, ser. LNCS, vol. 8761. Springer, 2014, pp. 48–151.

[3] ——, "Computational persuasion with applications in behaviour change," in *Proc. of COMMA'16*, ser. FAIA, vol. 287. IOS Press, 2016, pp. 5–18.

[4] P. Dung, "On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming, and n-person games," *Artificial Intelligence*, vol. 77, pp. 321–357, 1995.

[5] R. Likert, "A technique for the measurement of attitudes," *Archives of Psychology*, vol. 140, pp. 1–55, 1932.

[6] J. Ogden, *Health Psychology*. McGraw-Hill, 2011.

[7] A. Hunter and S. Polberg, "Empirical methods for modelling persuadees in dialogical argumentation," in *Proc. of ICTAI'17*. IEEE, 2017, pp. 382–389.

[8] M. Thimm, "A probabilistic semantics for abstract argumentation," in *Proc. of ECAI'12*, ser. FAIA, vol. 242. IOS Press, 2012, pp. 750–755.

[9] A. Hunter, "A probabilistic approach to modelling uncertain logical arguments," *International Journal of Approximate Reasoning*, vol. 54, no. 1, pp. 47–81, 2013.

[10] P. Baroni, M. Giacomin, and P. Vicig, "On rationality conditions for epistemic probabilities in abstract argumentation," in *Proc. of COMMA'14*, ser. FAIA, vol. 266. IOS Press, 2014, pp. 121–132.

[11] A. Hunter and M. Thimm, "On partial information and contradictions in probabilistic abstract argumentation," in *Proc. of KR'16*. AAAI Press, 2016, pp. 53–62.

[12] A. Hunter, "Modelling the persuadee in asymmetric argumentation dialogues for persuasion," in *Proc. of IJCAI'15*. AAAI Press, 2015, pp. 3055–3061.

[13] ——, "Persuasion dialogues via restricted interfaces using probabilistic argumentation," in *Proc. of SUM'16*, ser. LNCS, vol. 9858. Springer, 2016, pp. 184–198.

[14] A. Hunter and N. Potyka, "Updating probabilistic epistemic states in persuasion dialogues," in *Proc. of ECSQARU'17*, ser. LNCS. Springer, 2017, forthcoming.

[15] E. Hadoux and A. Hunter, "Computationally viable handling of beliefs in arguments for persuasion," in *Proc. of ICTAI'16*. IEEE Press, 2016, pp. 319–326.

[16] ——, "Strategic sequences of arguments for persuasion using decision trees," in *Proc. of AAAI'17*. AAAI Press, 2017, forthcoming.

[17] H. Nguyen and J. Masthoff, "Designing persuasive dialogue systems: Using argumentation with care," in *Proc. of PERSUASIVE'08*. Springer, 2008, pp. 201–212.

[18] G. Maio and G. Haddock, *The Psychology of Attitudes and Attitude Change*. Sage, 2015.

[19] I. Habernal and I. Gurevych, "Which argument is more convincing? Analyzing and predicting convincingness of Web arguments using bidirectional LSTM," in *Proc. of ACL'16*. The Association for Computer Linguistics, 2016, pp. 1589–1599.

[20] R. Cialdini, *Influence: The Psychology of Persuasion*. HarperCollins, 1984.

[21] H. Prakken, "Formal sytems for persuasion dialogue," *Knowledge Engineering Review*, vol. 21, no. 2, pp. 163–188, 2006.

[22] M. Thimm, "Strategic argumentation in multi-agent systems," *Kunstliche Intelligenz*, vol. 28, pp. 159–168, 2014.

[23] X. Fan and F. Toni, "A general framework for sound assumption-based argumentation dialogues," *Artificial Intelligence*, vol. 216, pp. 20–54, 2014.

[24] D. Kontarinis, E. Bonzon, N. Maudet, and P. Moraitis, "Empirical evaluation of strategies for multiparty argumentative debates," in *Proc. of CLIMA'14*, ser. LNCS, vol. 8624. Springer, 2014, pp. 105–122.

[25] E. Black and A. Hunter, "Reasons and options for updating an opponent model in persuasion dialogues," in *Proc. of TAFA'15*, ser. LNCS, vol. 9524. Springer, 2015, pp. 21–39.

[26] T. Rienstra, M. Thimm, and N. Oren, "Opponent models with uncertainty for strategic argumentation," in *Proc. of IJCAI'13*. AAAI Press, 2013, pp. 332–338.

[27] C. Hadjinikolis, Y. Siantos, S. Modgil, E. Black, and P. McBurney, "Opponent modelling in persuasion dialogues." in *Proc. of IJCAI'13*. AAAI Press, 2013, pp. 164–170.

[28] E. Hadoux, A. Beynier, N. Maudet, P. Weng, and A. Hunter, "Optimization of probabilistic argumentation with Markov Decision Models," in *Proc. of IJCAI'15*. AAAI Press, 2015, pp. 2004–2010.

[29] E. Black, A. Coles, and S. Bernardini, "Automated planning of simple persuasion dialogues," in *Proc. of CLIMA'14*, ser. LNCS, vol. 8624. Springer, 2014, pp. 87–104.

[30] A. Rosenfeld and S. Kraus, "Providing arguments in discussions based on the prediction of human argumentative behavior," in *Proc. of AAAI'15*. AAAI Press, 2015, pp. 1320–1327.

[31] A. Rosenfeld and S. Kraus., "Providing arguments in discussions on the basis of the prediction of human argumentative behavior," *ACM Trans. Interactive Intelligent Systems*, vol. 6, pp. 30:1–30:33, 2016.

[32] A. Rosenfeld and S. Kraus, "Strategical argumentative agent for human persuasion," in *Proc. of ECAI'16*, ser. FAIA, vol. 285. IOS Press, 2016, pp. 320–328.

[33] I. Rahwan, M. Madakkatel, J. Bonnefon, R. Awan, and S. Abdallah, "Behavioural experiments for assessing the abstract argumentation semantics of reinstatement," *Cognitive Science*, vol. 34, no. 8, pp. 1483–1502, 2010.

[34] F. Cerutti, N. Tintarev, and N. Oren, "Formal arguments, preferences, and natural language interfaces to humans: an empirical evaluation," in *Proc. of ECAI'14*, ser. FAIA, vol. 263. IOS Press, 2014, pp. 207–212.

[35] S. Lukin, P. Anand, M. Walker, and S. Whittaker, "Argument strength is in the eye of the beholder: Audience effect in persuasion," in *Proc. of EACL'17*. ACL, 2017, pp. 742–753.

[36] R. J. Thomas, J. Masthoff, and N. Oren, "Adapting healthy eating messages to personality," in *Proc. of PERSUASIVE'17*, ser. LNCS, vol. 10171. Springer, 2017, pp. 119–132.

[37] P. Dung and P. Thang, "Towards (probabilistic) argumentation for jury-based dispute resolution," in *Proc. of COMMA'10*, ser. FAIA, vol. 216. IOS Press, 2010, pp. 171–182.

[38] H. Li, N. Oren, and T. Norman, "Probabilistic argumentation frameworks," in *Proc. of TAFA'11*, ser. LNCS, vol. 7132. Springer, 2011, pp. 1–16.

[39] S. Benlamine, M. Chaouachi, S. Villata, E. Cabrio, C. Frasson, and F. Gandon, "Emotions in argumentation: an empirical evaluation," in *Proc. of IJCAI'15*. AAAI Press, 2015, pp. 156–163.