Using Argumentative Structure to Interpret Debates in Online Deliberative Democracy and eRulemaking

JOHN LAWRENCE, Centre for Argument Technology, University of Dundee, UK
JOONSUK PARK, Department of Computer Science, Williams College, USA
KATARZYNA BUDZYNSKA, Centre for Argument Technology, University of Dundee, UK & Polish Academy of Sciences, Poland
CLAIRE CARDIE, Department of Computer Science, Cornell University, USA
BARBARA KONAT and CHRIS REED, Centre for Argument Technology, University of Dundee, UK

Governments around the world are increasingly utilising online platforms and social media to engage with, and ascertain the opinions of, their citizens. Whilst policy makers could potentially benefit from such enormous feedback from society, they first face the challenge of making sense out of the large volumes of data produced. In this article, we show how the analysis of argumentative and dialogical structures allows for the principled identification of those issues that are central, controversial, or popular in an online corpus of debates. Although areas such as controversy mining work towards identifying issues that are a source of disagreement, by looking at the deeper argumentative structure, we show that a much richer understanding can be obtained. We provide results from using a pipeline of argument-mining techniques on the debate corpus, showing that the accuracy obtained is sufficient to automatically identify those issues that are key to the discussion, attracting proportionately more support than others, and those that are divisive, attracting proportionately more conflicting viewpoints.

CCS Concepts: • **Computing methodologies** → **Discourse, dialogue and pragmatics**; *Ensemble methods*; • **Applied computing** → **E-government;**

Additional Key Words and Phrases: Argument, argumentation, corpus, dialogue, sensemaking, engagement, analytics

ACM Reference Format:
John Lawrence, Joonsuk Park, Katarzyna Budzynska, Claire Cardie, Barbara Konat, and Chris Reed. 2017. Using argumentative structure to interpret debates in online deliberative democracy and eRulemaking. ACM Trans. Internet Technol. 17, 3, Article 25 (July 2017), 22 pages.
DOI: http://dx.doi.org/10.1145/3032989

1. INTRODUCTION

In the age of vast amounts of information being communicated through the Internet, it is not surprising that political dialogue is increasingly taking place online, too. Governments around the world are increasingly utilising online platforms and social media to engage with, and ascertain the opinions of, their citizens [Howard 2001; Moon 2002]. Whilst policy makers could potentially benefit from such feedback from society, they

We acknowledge that the work reported in this article has been supported in part by EPSRC in the UK under grant EP/N014871/1, the Innovate UK under grant 101777, and the National Science Foundation in the USA under Grant 1314778.
Authors' addresses: J. Lawrence, K. Budzynska, B. Konat, and C. Reed, Centre for Argument Technology, University of Dundee, Dundee, DD1 4HN, UK; emails: {fjohn, kasia, basia, chrisg}@arg.tech; J. Park, Department of Computer Science, Williams College, Williamstown, MA 01267; email: jpark@cs.williams.edu; C. Cardie, Department of Computer Science, Cornell University, 417 Gates Hall; email: cardie@cs.cornell.edu.

ACM 1533-5399/2017/07-ART25 $15.00
DOI: http://dx.doi.org/10.1145/3032989

first face the challenge of making sense out of the large volumes of data produced. Identifying those issues that are key to the debate, those that are the most controversial, those that were successfully resolved, and those that should be further handled to achieve consensus and mutual understanding is a skilled and time-consuming task in real-life discussions.

Decision makers and government officials often do not have the time to process the data resulting from such online engagement, instead relying on a superficial summary of the points made and the strength of opinion on each side of an issue. There is a clear demand for tools and technologies that will enable policy makers to quickly and thoroughly digest the points being made and to respond accordingly. For example, Cynthia Farina, a research professor in administration of the law at Cornell University, stated that automatic identification of areas of disagreement in public debates would be extremely useful for at least two reasons. First, although commenters often take opposing positions, they rarely directly interact with one another in their disagreements. Hence, the government decision maker ends up with polar views—and little sense of which policy outcomes commenters could actually "live with." Identifying disagreement as it emerges would permit human or automatic moderation aimed at inciting the kind of discussion among commenters that could reveal mutually satisfactory compromises or at least narrow the range of dispute. Second, areas of disagreement are important focal points for summarizing and analyzing, and eventually responding to, public comments. Conflicting comments may indicate gaps or disputes about key facts, tensions between relevant values, competing views of the nature or causes of the problem, or contradictory predictions of likely remedial impacts. Moreover, disagreement signals issues on which the ultimate decision is vulnerable to challenge in court or even the legislature.

Whilst the field of controversy mining (see Section 2.1) aims to identify issues or events that attract conflicting opinions in dynamic, dialogical networks, merely identifying controversial issues falls short of the deeper understanding required by policy makers. By instead determining the argumentative and dialogical structures contained within a debate, we are able to determine not only those issues that are controversial but also those that attract agreement. Using the Argument Interchange Format (AIF) [Chesñevar et al. 2006] to represent the argumentative structure, we are able to see points of agreement and disagreement, as well as to understand why those views are held and the reasons both supporting and attacking them. Furthermore, we are able to leverage work in the growing field of argument mining to automate the processing of debates and the analysis of their argumentative structure.

Our goal here is to combine a variety of techniques, some based on general linguistic features and others on features that are specific to argumentation, to automate the task of identifying the structure of the arguments and how they interconnect in a broader discussion. Though this task is extremely demanding for current text-mining and computational linguistics techniques, our final target is not the network structure itself but rather a network structure that is sufficiently accurate to develop an *interpretative* step that gives decision makers some insight into the discussion. Here we use the simple metric of centrality and show that even with modest performance on the task of extracting the argument network, it is possible to generate rather high reliability in identifying central issues to the discussion.

In this article, we first look at related work in automatically detecting controversy and arguments (Section 2). In Section 3.1 we describe the data that we are using, taken from the RegulationRoom[1] online deliberative democracy platform. Section 3.2 moves on to look at the argumentative analysis of this data and the insights that such analysis can provide. Finally, in Section 4, we look at utilising argument-mining

[1]At http://RegulationRoom.org.

techniques to automate the analysis process, leading then into the final, interpretative step. Though only a first example of how such interpretative analytics can be developed automatically, the article's contribution is that this is the first time argument-mining techniques have been connected to such metrics that give decision makers insight into, and understanding of, complex discussions.

2. RELATED WORK

In this section, we look at related work from two areas: First, we consider the automated recognition of controversy (controversy mining), and, second, we look at the work that has been carried out in automatically recognising argumentative structures (argument mining). Although controversy mining provides an indication of where controversial issues occur, it does not offer the richer understanding obtained through study of the argumentative structure. For this reason, we have focused on those areas of controversy mining most closely related to argumentation. Similarly, for argument mining, there are a range of techniques that have been applied across different domains, and here we focus mainly on those that are most closely related to online dialogue.

2.1. Automated Recognition of Controversy

Controversy mining looks at the processes of how an issue or event attracts conflicting opinions in dynamic, dialogical networks. The clearest link between controversy and argument detection can be seen in Boltužić and Šnajder [2015], where argumentative statements are clustered based on their textual similarity, to identify prominent arguments in online debates. Controversy detection to date has largely targeted specific domains; for example, Kittur et al. [2007] looks at the cost of conflict in producing Wikipedia articles, where conflict cost is defined as "excess work in the system that does not directly lead to new article content."

Identifying controversial events in social media is considered in Popescu and Pennacchiotti [2010], where Twitter posts are used as a starting point. Specifically, when given a target entity, an event involving that entity is defined as "an activity or action with a clear, finite duration in which the target entity plays a key role." Twitter snapshots (triples consisting of a target entity (e.g., a person, or event), a specific time period, and a set of tweets about the entity that were posted within the target time frame) are then assigned a controversy score, and this score is then used to rank the snapshots.

The scope of controversy detection is broadened slightly in Choi et al. [2010], which looks at identifying controversy in news articles. In Choi et al. [2010], a controversial issue is defined as "a concept that invokes conflicting sentiments or views" and a subtopic as "a reason or factor that gives a particular sentiment or view to the issue." A method is proposed for the detection of controversial issues, based on the magnitude of sentiment information and the difference between the magnitudes for two different polarities. First, noun and verb phrases are identified as candidate issues using a mixture of sentiment models and topical information. The degree of controversy for these issues is calculated by measuring the volume of both positive and negative sentiment and the difference between them.

The role of agreement and disagreement for obtaining consensus in online discussion was explored in Rosenthal and McKeown [2015] using Internet Argument Corpus [Walker et al. 2012]. All sentences were classified as expressing agreement, disagreement, or neither. Only those pairs of posts (quote-response) were taken into account where the response immediately followed the quote. The highest reported F-score is 0.58 for agreement and 0.73 for disagreement. In our approach, this was broadened

from only adjacent pairs by using an annotation scheme that allows for marking support and conflict relations even between long distance pairs of propositions.

Another relevant work comes from the area of analysing online idealogical dialogues. In Misra et al. [2015], two concepts were introduced that are relevant to our work: the concept of central claim and the concept of argument facet. Determination of central claims was obtained by the combination of human summarization techniques on Mechanical Turk and the Pyramid method (a method of step-by-step narrowing down the number of sentences in the text). In the Pyramid method, a central claim is defined as the claim that surfaces to the top of the pyramid. The argument facet may be rephrased as the argument topic: "Just as two words can only be antonyms if they are in the same semantic field, two arguments can only be contradictory if they are about the same FACET" [Misra et al. 2015].

2.2. Automated Recognition of Arguments

Argument mining aims at developing methods and techniques for automatic extraction of arguments from texts in natural language. An argument is a complex discourse unit with boundaries easily recognisable by humans and yet hard to determine by a computer. For this reason, argument mining is often supported with rhetorical document structure, argument schemes, or dialogical relations. This area of research began to attract attention over a decade ago. Argumentative zoning [Teufel 1999; Teufel and Moens 2002] was looking at recognising argumentative discourse units from unstructured scientific articles using rhetorical structure of a document. The results varied from the highest F-score of 0.86 for the recognition of parts of articles in which an author refers to his or her own research to as low as an F-score of 0.26 for the recognition of parts in which an author presents arguments against other approaches. The authors point out that their solution is domain specific and works well for academic articles, as it relies on specifically tailored sentential features.

Automated classification of sentences as either argument or non-argument [Moens et al. 2007] on the material from discussion fora, legal judgements, newspapers, parliamentary records, and weekly magazines achieved an average accuracy of 70% using maximum entropy and multinomial Naive Bayes classifiers. In this study, the score for discussion fora (68.4%) was lower than for the newspaper articles (73.22%). The authors suggest that discussion fora contain more ambiguous arguments and are less-well-formed texts compared to the news and legal texts. Classification of sentences as argument or non-argument constitutes the first step in argument mining; however, it does not yet provide information about argumentative relations, such as reason-conclusion structure or conflict. The relations between reason and conclusion in legal texts are explored in Palau and Moens [2009]. The first step in the argument detection task was the usage of a Naive Bayes classifier to classify sentences as "argumentative" or "non-argumentative." The argumentative sentences were then classified by their argumentative function, and, finally, a set of manually crafted rules was used to determine the global argumentative structure.

Since 2014, the area of argument mining has been witnessing a rapidly increasing interest. Analysis of support and attack relations in the corpus of German argumentative microtexts [Peldszus 2014] provided a highest-achieved F-score of 0.7. Automated extraction of counter-consideration is explored in Peldszus and Stede [2015b]. A speaker may provide counter-consideration to his or her own statement in anticipation of the critique. This study provides evidence that lexical indicators (especially "but" but also "however" and "although") perform well as predictors of counter-considerations.

Further exploration of argument structure is possible with the application of argument schemes [Walton et al. 2008] to argument mining [Lawrence and Reed 2015b]. The combination of discourse indicators, topic similarity, and argument scheme

resulted in an F-score of 0.83 on the corpus of online argumentation excerpts stored in Argument Interchange Format database (AIFdb).

3. INTERPRETATION OF DEBATES THROUGH ARGUMENT NETWORKS

Democracy is founded on dialogue and debates. However, unlike in its origins in the ancient Greek poleis, modern countries are too large to ensure the direct participation in the process of law creation for every citizen. An initial solution adopted by rule-makers has been to seek for advice indirectly through social dialogue, that is, a dialogue between a government and non-governmental organisations (NGOs) that are typically set up by citizens and have non-profit status. Nowadays, governments increasingly emphasise the importance of extending this formula to direct democracy to better understand and address the real needs of society and to increase the transparency of the process of law making. To overcome the challenge of being inclusive, governments are looking for solutions making use of Internet technologies for e-participation allowing citizens to engage in the political process in a way that is easily accessible, appeals to younger generations, and allows for anonymity, which is particularly important if one wants to raise a controversial issue.

In the US, online deliberative democracy (or e-rulemaking), for example, RegulationRoom,[2] has been introduced as a multi-step process of social media outreach that federal agencies use to consult with citizens on new regulations on health and safety, finance, and other complex topics. In order to ensure public awareness and participation of new regulations, federal agencies are obliged to publish materials describing the legal basis, factual and technical support, policy rationale, and costs and benefits of a proposal. Once the agency introduces the new regulation, it has to summarise the comments it received, respond to questions and criticisms, and offer explanations where it did not implement changes. Still, user-generated feedback, despite being socially extremely important, poses a challenge of big data; for example, in the US over 200 million citizens are eligible to vote and thus can participate in RegulationRoom [Farina and Newhart 2013; Park et al. 2012].

We propose an alternative method of defining features of debates such as central claims to those discussed in Section 2.1. These can be specified using the properties of argument networks as being the main conclusion to which all other claims are leading (i.e., being on top of the graph tree). We also introduce the method of recognizing controversial and non-controversial claims by checking for conflict instances. Our solution for the determination of topic relations between propositions relies on the position on the argument tree rather than summarization or facet properties. We also demonstrate in Section 4 how argument-mining techniques, such as those described in Section 2.2, can be employed to automatically determine the argumentative structure necessary for identifying such features of a debate.

3.1. The eRulemaking Debate Corpus

Our corpus, eRulemaking_Controversy Corpus (eRCC), is composed of user comments from the RegulationRoom platform, RRP (see Table I), manually segmented into propositions. As the first step, we selected a part of the existing, but as yet unpublished, Cornell corpus, in which the US Department of Transportation was publicly consulting on the topic of *Airline Passenger Rights*. The corpus was first annotated with structures for pro-arguments labelled as Reason (see Section 3.2 for the definition) [Park and Cardie 2014]. From this corpus, we selected only those comments that had dialogical nature, that is, that attracted at least one reply. The reply structure was easily retrieved from

[2]Available at http://RegulationRoom.org.

Table I. Summary of the Language Resources for Mining Argument Networks in eRulemaking_Controversy Corpus (eRCC)

	Words	Segments	Turns	Maps	Corpus Location
Train	16,403	1,152	139	47	http://arg.tech/ercctrain
Test	7,279	505	70	23	http://arg.tech/ercctest
Total	23,682	1,657	209	70	

the data, since RRP was assigning comments with a unique ID and recording when a comment was made in response to another comment.

The resources were then transferred to the freely accessible database AIFdb [Lawrence et al. 2012],[3] which hosts multiple corpora.[4] Its key advantage is that it uses the Argument Interchange Format (AIF) [Rahwan et al. 2007], a common language for representing argument networks (graphs), that distinguishes between nodes of information (I-Nodes), instances of schemes (S-Nodes), with sub-types representing the application of rules of inference (RA-Nodes), and rules of conflict (CA-Nodes). AIFdb allows collections of nodes and edges to be grouped into nodesets corresponding to specific argumentative structures, for example, the arguments contained within a specific document. In our corpus, each dialogical thread is contained within its own nodeset.

At the second stage, the annotation was extended to identify more pro-arguments using a more fine-grained annotation scheme and to identify a new type of structure, that is, con-arguments, to account for the interactional dimension of online citizen dialogue (see Section 3.2 for the full set of labels). The addition of the category of con-arguments is necessary to be able to identify controversies, and the addition of further sub-categories of pro-arguments allows for more detailed analysis of what divides people.

The data were structured and annotated using the Online Visualisation of Arguments tool (OVA+) [Janier et al. 2014][5] and made publicly available as training and test subcorpora[6] (eRulemaking_Controversy_Train and eRulemaking_Controversy_Test; see Table I for links to each). The whole corpus contains 23,682 words, 1,657 segments (i.e., discourse units that constitute components of argument networks), 209 turns (i.e., comments users exchanged during the dialogue), and 70 maps (i.e., visualisations of argument networks, each corresponding to one thread of users' exchange). eRulemaking_Controversy is larger than the Potsdam Micro-text Corpus [Peldszus and Stede 2015a] and smaller than the more lightly structured Internet Argument Corpus [Walker et al. 2012].

3.2. Argumentative Analysis and Interpretation

Understanding the significance of, and the relation between, the points raised in a detailed online debate is not a straightforward task. Consider a conversation between three users of RegulationRoom about whether peanuts should be prohibited on planes as they may cause allergic reactions. RegulationRoom.org is an online platform for public consultancy that hosts regulation proposals from various US government agencies, allowing citizens to submit online comments.

[3]Available at http://aifdb.org.

[4]Available at http://corpora.aifdb.org/.

[5]Available at http://ova.arg-tech.org.

[6]We distinguish the test set from the training set, so the training set can be used to manually identify discourse indicators, which are used as features. The test set is not used for training the classifiers in any way.

(1) a. MALLONE: *When a food allergy is life threatening (and known to cause anaphylaxis), it considered a disability under federal laws such as Section 504 of the Rehabilitation Act of 1973 and the Americans with Disabilities Act (ADA).*
b. *In other words, people with severe peanut allergies have the right to be protected.*
c. MULDER: *No, allergies are not disabilities,*
d. *and therefore you get no special treatment under the ADA.*
e. *Federal courts have consistently ruled this way.(..)*
f. ANTANAGOGE: *Mulder's comment about the ADA is only partially true, but thoroughly exaggerated.*
g. *because there has only been one court case.*
h. *Food allergy is generally considered a disability under Section 504 and ADA.*
i. *The point Mulder exaggerates is that there is no primary legal precedent, i.e., a court opinion, saying this.(...)*
j. *The Air Carrier Access Act (ACAA) prohibits discrimination against those with disabilities by U.S. and foreign air carriers,*
k. *and Department of Transportation (DOT) regulations require airlines to accommodate travelers with disabilities.*

Such comments need to be then summarised back to a government agency (in this case, the US Department of Transportation, DOT) so decision makers can take public opinion into account in preparation of a new policy. Still, how can we effectively judge from the text itself which of the issues are the most controversial? Is it what was said towards the beginning: *In other words, people with severe peanut allergies have the right to be protected* (1-b), that attracted most reactions, or maybe the statement: *No, allergies are not disabilities* (1-c), that seems to be a common theme of what everyone is saying? This problem becomes even more challenging when we scale up to real-life conditions of numerous people contributing numerous comments. How can we make sense out of the vast amount of data? How can we effectively summarise to DOT which controversies led eventually to consensus and which ones remained unresolved? Which issues against prohibiting peanuts on planes should be taken into account and addressed by the US department of transportation?

We propose to structure a dialogue as argument networks consisting of pro-arguments (see Default Inference and Reason in Figure 1), con-arguments (Default Conflicts), comments, and the relations amongst them.[7] If controversy were conceptualised as a comment that attracts both some support and some attack (intuitively: people disagree on this issue), then it is now easy to identify in the argument network in Figure 1 that the issue (1-d) attracted the highest number of arguments pro and con, even though it is (1-b) that is a main claim of the discussion. On the other hand, if we rather understood controversy as the strength of conflict between two comments, then the conflict between (1-d) and (1-f) attracts the highest number of pro-arguments each (i.e., (1-c) for (1-d); and (1-g) and (1-i) for (1-f)). Intuitively, the more support two conflicting issues have, the higher the strength of the conflict. In this way, we can summarise to DOT that they should take into account some controversial issues identified to achieve mutual understanding between rule-makers and citizens.

Structuring the dialogue as argument networks to obtain a final eRRC dataset consisted of annotating pro-arguments (i.e., Reason, Default Inference, and argument schemes) and con-arguments (Default Conflict).

At the first stage, the data was annotated in Cornell with the initial set of pro-argument structures, that is, Reason [Park et al. 2015]. As a result, eRCC contained 125 initial pro-arguments tags (see Table II).

[7]The text is manually annotated using the software tool OVA+ http://ova.arg-tech.org/.

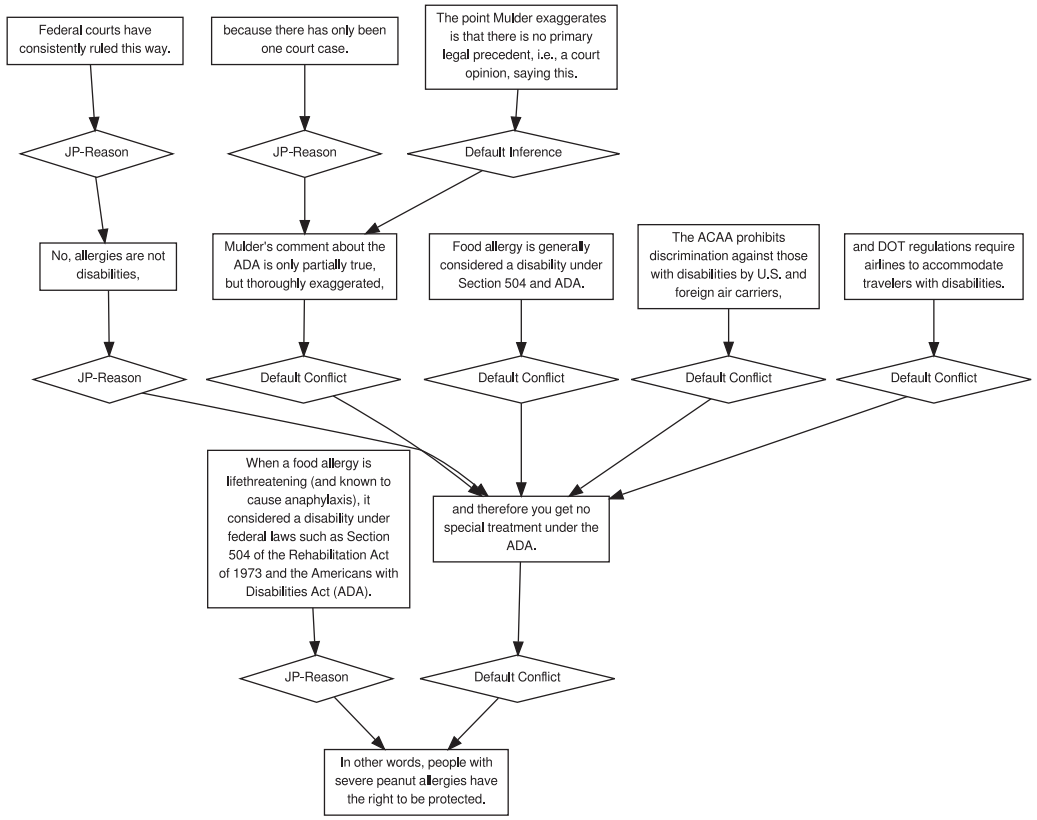

Fig. 1. Fragment of the argument network with the example of controversial issues (map #5695). Full map view available at http://arg.tech/5695.

Reason: One proposition is a Reason for another when it explains why the other proposition is true: *Peanut allergies are lethal* is a Reason for: *Peanut allergies should be banned on all flights*. Note that the validity of neither the reason nor the supported proposition affects the classification.

At the second stage, the annotation was extended by adding more pro-arguments (using more fine-grained criteria) and adding con-arguments. In the first case, we employed the theory of argument schemes [Walton et al. 2008] that are typical patterns used by people to formulate arguments. Argument schemes demonstrated to be useful in argument mining from natural language [Feng and Hirst 2011; Lawrence and Reed 2015a]. In the current corpus, three schemes for pro-arguments were annotated: argument from expert opinion, argument from example and practical reasoning. Because of the informal nature of online fora discussions, the schemes had to be slightly modified, which is marked by prefix "ER" in the scheme's name. For cases in which pro-argument was analysed, but did not follow any specific pattern we looked for, the argument structure was annotated as Default Inference. As a result, eRCC gained 463 additional pro-arguments tags in comparison with the first iteration of annotation (see Table II).

Default Inference: One proposition argues in favour of another via Default Inference when it provides information supporting the second one: *Freedom loving travelers don't*

Table II. The Total Number of Occurrences of Categories Annotated in the eRulemaking_Controversy Corpus

Corpus	#	Argument structure	#	Labels	#
eRulemaking_ Controversy	**767**	**Pro-arguments**	**670**	Reason	125
				Default Inference	419
				ERExample	38
				ERExpert Opinion	6
				ERPract Reasoning	82
		Con-arguments	**97**	Default Conflict	82
				ERAd Hominem	15

want to be told what they can and can't bring on board the aircraft to eat supports *This proposal goes too far*.

ERExpertOpinion: ER Argument from Expert Opinion was annotated when a speaker appealed to opinions provided by another person or institution that is treated as having some expertise in the area: *Food allergy occurs in 6 to 8 percent of children 4 years of age or under* is supported with ERExpertOpinion by *Prevalence information as reported by the National Institute of Allergy and Infectious Diseases*.

ERExample: ER Argument from Example was annotated when a speaker provided (usually in the form of a narration) their own or someone else's experience, situation or event, which is a particular instance of the sphere of influence of the proposed rule: *I've registered my son on a flight as peanut allergic and had the attendant try to hand him a bag of peanuts* supports with ERExample: *Flight personnel need more education on this*.

ERPracticalReasoning: In ER Practical Reasoning, the conclusion has the form of call to action, either in imperative or modal form: *A nut free zone does not work* supports with ERPracticalReasoning: *There must be a complete ban on tree nuts and peanuts on planes*.

The con-arguments were treated in a similar way. We identified one common pattern of arguing against in our corpus, Ad Hominem argument, and the rest of the con-arguments were treated as Default Conflict. As a result, eRCC obtained 81 con-arguments tags (see Table II).

Default Conflict: Two propositions are annotated as being in Conflict relation when they cannot be true at the same time. In dialogical interactions, this often does not have a form of logical opposition but of an attack: *If someone is actually that allergic, then they should stay home and not inconvenience the rest of us* is attacked by *I don't understand how not being able to eat peanuts for a few hours of your life is worth putting another life at risk*.

ERAdHominem: ER Ad Hominem Argument was annotated when one proposition was attacking not the content of what was said but the speaker ("owner") of the other propositions: *I'm a physician, epidemiologist, and mother to a four year old boy with allergies to milk, peanuts and egg* is attacked with ERAdHominem by *Methinks you have an agenda*.

For the inter-annotator agreement, a systematic sample of 10% of the corpus was extracted (by selecting every 10th argument map from the corpus) and annotated

by the second annotator. The samples were compared using the Combined Argument Similarity Score (CASS) technique ([Duthie et al. 2016], resulting in Cohen's $\kappa = \mathbf{0.92}$.[8]

The number of occurrences for each label is given in Table II. Pro-arguments (670 occurrences, 87% of the whole corpus) are significantly more frequent than con-arguments (87, 13%). In both categories, defaults are the most common (Default Inference, with 419 instances constitutes 63% of all pro-arguments and Default Conflict, with 82 instances—85% of all con-arguments). Amongst argument schemes, the most often used is practical reasoning (82 instances, 12% of all pro-arguments).

Structuring online comments as networks of pro- and con- arguments allows for classifying issues discussed at RegulationRoom on at least three main dimensions:

(1) **Centrality:** high centrality indicates issues that attract many supports and/or conflicts, typically important to highlight to DOT.
(2) **Popularity:** high popularity indicates issues that are claimed by many different users, typically useful to present to DOT for potential adoption.
(3) **Controversiality:** high controversiality indicates issues that have many pro-arguments, but also many con-arguments, typically useful to emphasise to DOT as options rejected by many users but where it might be appropriate to develop a strategy for dealing with the controversy.

Figure 2 shows how a central and controversial issue can be highlighted and reported to decision makers. The screenshot shown here is from an overview of controversial issues automatically generated from the analysed argumentative structure. An issue that has been identified as controversial is first shown, with the supporting reasons for each side of the argument shown underneath.

4. AUTOMATING THE ARGUMENTATIVE ANALYSIS PROCESS

In this section, we look at automating the argument analysis task. More specifically, we build classifiers to distinguish propositions in support relation[9] from those that are not. The task is formulated as identifying proposition pairs (ordered) in support relation from all possible ordered pairs of propositions in a given thread. By using argument-mining techniques to produce the kind of argumentative structures that we are able to obtain manually, it would be possible to give a real-time overview of the state of a particular debate, providing the kind of insights described in Section 3.2 as the debate progresses and thus allowing for interactions with the debate to resolve controversial issues, or pursue topics that are central, as they arise. Starting with manually segmented text, we then consider three techniques: First, we use topical similarity to reduce the possible search space of connected propositions; we then look at identifying discourse indicators, strong lexical cues indicating the role of a proposition in the dialogue; and, finally, we apply computational discourse analysis techniques to identify the connections between propositions.

4.1. Reducing the Search Space

Our corpus contains over 1,500 segments across 70 nodesets, corresponding to individual threads in the dialogue, resulting in over 20,000 potential connections between segments in the same nodeset. Our first step is to reduce the size of the search space. We do this using semantic similarity to determine those propositions that are discussing similar topics. This method is similar to that presented in Lawrence et al. [2014], where it is assumed first that the argument structure to be determined can be represented

[8]Both samples used for κ calculations are publicly available in AIFdb. First annotator: http://arg.tech/erkappa1; second annotator: http://arg.tech/erkappa2.
[9]A proposition *supports* another if they form a Pro-argument structure in Table II.

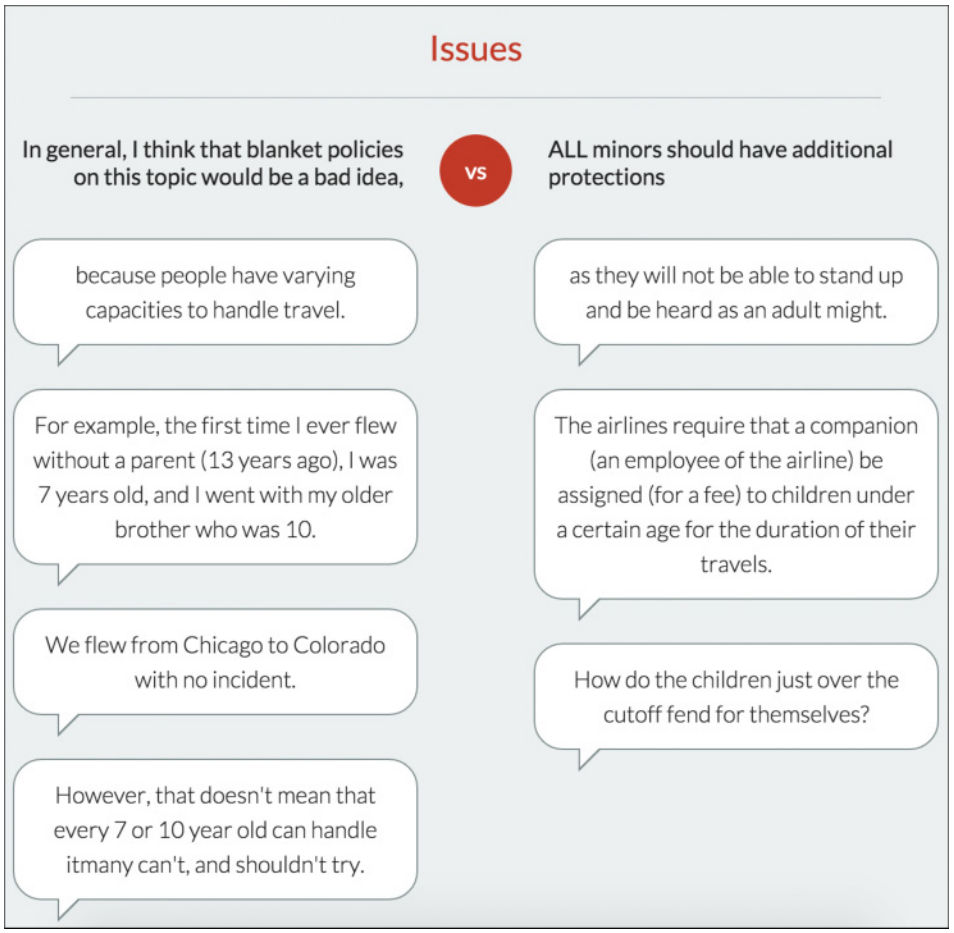

Fig. 2. Overview of a controversial issue, with support on each side, automatically generated from the analysed argumentative structure.

as a tree and second, that this tree is presented depth first. That is, the conclusion is given first and then a line of reasoning is followed supporting this conclusion. Once that line of reasoning is exhausted, the argument moves back up the tree to one of the previously made points.

Based on these assumptions, it is possible to determine connections by looking at how semantically similar each proposition is to its predecessor. If they are similar, then we assume that they are connected and the current line of reasoning is being followed. If they are not sufficiently similar, then we first consider whether we are moving back up the tree and compare the current proposition to all of those made previously and, if the most similar previous point is above a set threshold, we connect them. Finally, if the current point is not related to any of those made previously, then it is assumed that a new topic is being discussed, and the proposition is left unconnected as the root of this new argument.

We exploit the dialogical structure of our data by discounting any possible connections between propositions that are not in the same thread of the dialogue. The way that a connection is determined also gives precedence to connections between adjacent

Table III. Comparison of Different Methods for Reducing the Search Space by Determining Connectedness Using Semantic Similarity, Optimised for Maximum Recall

Method	Precision	Recall	F1
Average score	0.17	**0.92**	0.29
Maximum score	0.17	0.90	0.29
Average of top two scores	0.17	0.90	0.29
Average of top three scores	0.17	0.89	0.28
Weighted average score	0.18	0.88	0.30
word2vec	0.16	0.88	0.27
doc2vec	0.12	0.85	0.21
Sequential Baseline	*0.38*	*0.65*	*0.48*

propositions in the same comment. Two different thresholds are used, a lower threshold for sequential propositions that are more likely to be connected and a higher threshold for non-sequential propositions. In all cases, the threshold values were selected to maximise recall, whilst keeping precision at a reasonable level. This tradeoff was made as our goal is to narrow the search space, reducing the number of possible pairs as much as possible whilst losing a minimum number of connected pairs.

The first approach that we consider uses WordNet [Miller 1995] to determine the similarity between the synsets of each word in the first proposition and each word in the second. This relatedness score is inversely proportional to the number of nodes along the shortest path between the synsets. The shortest possible path occurs when the two synsets are the same, in which case the length is 1, and, thus, the maximum relatedness value is 1. We then look at the maximum of these values to pair a word in the first proposition to one in the second. From here, we then considered a range of different methods to determine whether the two propositions are connected:

(1) **Average score:** takes the sum of the scores for each pairing and divides by the total number of paired words.
(2) **Maximum score:** looks only at the pairing with the greatest score.
(3) **Average of top two scores:** takes the average of the scores for the two most similar words.
(4) **Average of top three scores:** takes the average of the scores for the three most similar words.
(5) **Weighted average score:** takes the average score for each pairing, giving a higher weight to the most similar, and then reducing this weighting as the similarity decreases.

The average precision, recall and F-score obtained using each of these possible methods is shown in Table III.

We also implemented two further methods of determining connectedness using semantic similarity. The approaches used have both been shown to perform robustly when using models trained on large external corpora [Lau and Baldwin 2016].

The first of these approaches uses word2vec [Mikolov et al. 2013], an efficient neural approach to learning high-quality embeddings for words. Due to the relatively small size of our training dataset, we used pre-trained skip-gram vectors trained on part of the Google News dataset.[10] This model contains 300-dimensional vectors for 3 million words and phrases obtained using a simple data-driven approach described in Mikolov et al. [2013].

[10]https://code.google.com/archive/p/word2vec/.

To determine similarity between propositions, we located the centroid of the word embeddings for each by averaging the word2vec vectors for the individual words in the proposition and then calculating the cosine similarity between centroids to represent the proposition similarity.

The final approach that we implemented uses a doc2vec [Le and Mikolov 2014] distributed bag of words(*dbow*) model to represent every proposition as a vector with 300 dimensions. Again, we then calculated the cosine similarity between vectors to represent the proposition similarity.

In each case, as our aim here is simply to reduce the search space, the threshold values were lowered to maximise recall and so reduce the number of possible connections whilst retaining the greatest number of those propositions that had been identified as connected in the manual analysis. These results can be seen in Table III, compared to a baseline obtained by assuming that each sequential proposition is connected. We can see that the results for each method are remarkably similar, suggesting that the limitation is not in calculating the similarity of proposition pairs but in being unable to correctly connect some pairs of propositions that are connected in the annotation but semantically differ. This issue is exactly why we have adjusted each threshold to maximise recall, and, despite the approaches tested giving overall lower accuracy than the baseline, we were able, in each case, to obtain a higher value for the recall. Although each method resulted in a similar level of accuracy, the *Average score* method performed best, and, as such, we used the pairs of connected propositions obtained by this method as input to the classifiers described in Section 4.3, reducing the number of possible connections by 17.5%.

4.2. Determining Discourse Indicators

Discourse indicators are words that serve as a clue for the argumentative function of the proposition. They can either connect two propositions (inter-proposition indicators) or constitute part of the proposition (intra-proposition indicators). Certain indicators have been listed in the literature (see Table IV for an aggregate list). To further broaden this list, we used a keyword method in certain subsets of eRCC corpus. The indicators discovered in this step were then used as features for the classifier discussed in Section 4.3.

A keyword is a word that has much higher frequency in one corpus than in other, and the keyness of a given word indicates its overuse in one corpus as compared to another corpus [Gries 2009]. The corpus for which the overuse is determined (source corpus) is compared with the reference corpus. We created 12 subcorpora of propositions holding certain argumentative function.[11] By looking at these subcorpora separately, we are able to determine those words that, for example, are more commonly found in an Attacking proposition than in a proposition that does not attack any of the others.

For each of the subcorpora, keywords were extracted using the Log Likelihood method (threshold of critical value $= 3.84$, $p < 0.05$). This allowed for the determination of the list of words overused in propositions holding certain argumentative function. From the list of obtained keywords, words that were topic specific (such as "allergy," "children," and "airplane") were removed. From the total of 12 corpora comparisons, only 6 brought relevant results (i.e., results both statistically significant and topic independent). The resulting list of keywords (presented in Table IV) indicates words specific for this type of discourse (online comments on legal regulations) that indicate propositions with

[11]Supporting, Supported, Attacking, Attacked, ERExample Prem, ERExample Concl, ERExpertOpinion Prem, ERExpertOpinion Concl, ERPracticalReason Prem, ERPracticalReason Concl, ERAdHominem Attacking, ERAdHominem Attacked.

Table IV. Overview of the Identified Discourse Cues, Both from Existing Work and Identified from the eRulemaking Training Corpus

From literature			
	List	**Indicates**	**Source - reference**
Inter	*because,therefore, after, for, since, when, assuming, so, accordingly, thus, hence, then, consequently*	Support	[Lawrence and Reed 2015b]
	however, but, though, except, not, never, no, whereas, nonetheless, yet, despite	Conflict	[Lawrence and Reed 2015b]
	as a result	Conclusion	[Webber et al. 2012]
	Reference to the first person in the covering sentence of an argument component: *I, me, my, mine, myself*	Major claim	[Stab and Gurevych 2014]
	while, whereas, whereas normally, whereas otherwise, not even, yet	complementary coordinative argumentation	[van Eemeren et al. 2007]
Intra	*cause, effect, means, end, makes that, leads to* (and other expressions that refer to causality only implicitly: *for example, cultivate, suddenly, necessarily*)	causal argument	[van Eemeren et al. 2007]
From eRCC train corpus			
	List	**Indicates**	**Source - corpus**
Intra	*argument*	indicates attacking	all attacking vs. all non-attacking
	you, your	weakly indicates attacking	all attacking vs. all non-attacking
	Negative words: *funeral, death*	weakly indicates attacking	all attacking vs. all non-attacking
	should	strongly indicates supported	all supported vs. all non-supported
	I think	indicates supported	all supported vs. all non-supported
	you	strongly indicates ERAdhominem-attacking	all ERAdh-attacking vs. all non-ERAdh-attacking
	Personal pronouns (including possessive): *him, his, he, our, my*	strongly indicates ERExample-prem	all ERExample-prem vs. all non-ERExample-prem
	Relating to expertise: *association,(s), cite, journal(s), pages, published, studies, www, http, academy, college, reported, institute,*	strongly indicates ERExpertOp-prem	all ERExpertOp-prem vs. all non-ERExpertOp-prem
	should	weakly indicates ERPractReas-concl	all ERPractReas-concl vs. all non-ERPractReas-concl

certain argumentative functions. The rationale for the choice of these words is as follows:

(1) Indicating Attacking:
   —*argument*: The users of the forum use the word "argument" in attacking, rather than in any other argumentative move, as a meta-discourse marker, in some ways announcing that they are about to attack someone's argument, as in the following: "This slippery slope argument is a false one" and "Your argument is a strawman."
   —*you, your*: These are specific for attacking moves, due to the personal engagement and AdHominem nature of many attacks, as in the following: "If you have a problem, then it is up to you to have the solution."
   —*negative words (funeral, death)*: Due to the emotional nature of the forum discussion, users refer to negative consequences and use hyperbolization to make their attack look stronger: "But some of the people on this board calling for funerals before advancing the discussion give new meaning to the Founders' fears of the tyranny of the majority."
(2) Indicating Supported
   —*should*: Due to the nature of the discussion (proposition of new legal regulations), propositions expressed in deontic modality were expressed by users and were more often used as premises (in our annotation: supported) than conclusions: "A similar problem, that should also be addressed, along with the peanut allergy problem, is the case of allowing small domestic pets in the cabin of a aircraft."
   —*I think*: This bigram is used as a hedge, lowering the level of confidence the speaker ascribes to the truth of the proposition; taken into account that in argument, asserting the truth of the conclusion cannot be stronger than asserting the truth of its weakest premises; it is not surprising that users of the forum were hedging conclusions but not premises: "I think a ban of all peanuts and nuts (or at least peanuts) would be the safest route for those with peanut allergies."

In our new approach to the indicators, we broaden the concept of lexical indicators. We assume not only connectives between propositions but also specific lexical items (unigrams, bigrams) that appear inside the phrase. It could be hard to indicate certain and not topic-specific lexical indicators or constructions for argument structure in general, but it is possible to show specific lexical features of certain argumentative schemes. For example, in ERExample speakers use *I*/*me* and action verbs and in ERExpertOpinion we can expect a Named Entity to be present. A full list of intra-proposition discourse indicators can be seen in Table IV. Some of those identified are probably genre specific (and specific for American English) but, we expect, not topic specific.

Intra-proposition discourse markers work not only for consecutive propositions but also for any propositions that are topically related (e.g., Ad Hominem attack may refer to the proposition of a person speaking many turns before).

To determine the validity of the identified indicators, we performed classification of propositions based on their presence, obtaining a precision of 0.82, and recall of 0.19, for support relations and precision of 0.73, and recall of 0.14, for attack relations. Although in both cases the precision is high, the fact that these types of indicators are often omitted means that they do not give a good indication of the argumentative structure on their own. However, when they do occur, they give a very strong indication of the role that a proposition is playing in the dialogue and, as such, provide a useful feature for the machine-learning technique discussed in the next section.

4.3. Classifying Relations Between Propositions

This component is the final step of the automation process in which propositions in support relations are identified. As previously mentioned, the task is formulated as identifying ordered propositions pairs in support relation, that is, the first proposition in the pair supports the second. The number of all possible ordered pairs of propositions is quadratic to the number of propositions in a given thread. Since the vast majority of them are not in a support relation, there is a significant imbalance in the class distribution. Thus, we only consider the proposition pairs that are classified as topically similar during search space reduction as described in Section 4.1. This is precisely why the search space reduction was optimised for recall.

Setup. We adopt a general approach in computational discourse analysis where classification algorithms, such as Support Vector Machines (SVM) and Naive Bayes, are used with various lexical and syntactic features [Park and Cardie 2012].[12] The main difference is that traditional discourse analysis in Natural Language Processing (NLP) focuses on a broader set of relations, such as contingency, comparison, expansion, and temporal, whereas only support relation is targeted in this work. Also, in previous work using Penn Discourse Treebank [Prasad et al. 2008], only adjacent text spans are considered, while we aim to deal with relations between propositions that may not be adjacent to each other. Because of this difference, the most informative features for this task are dissimilar to those for discourse analysis, though all the features have previously been employed in discourse analysis. In addition to the machine-learning approach, we also report results using a hand-coded rule-based classifier that returns true if the given pair of propositions are adjacent and contains at least one discourse marker and returns false otherwise.

Below are brief descriptions of features whose efficacy have been empirically determined in prior work,[13] along with the rationale behind them:

—**Word Pairs** is the Cartesian product of the unigrams from proposition 1 with those from proposition 2. Word pairs can potentially capture semantic support relations, for example, between "rain" and "wet." To elaborate, with enough occurrences of proposition pairs annotated as support where "rain" appears in the first and "wet" appears in the second, the model will learn that there is a support relation between "rain" and "wet." More generally, the intuition is that indicators of support relation should exist in both propositions under consideration, since we also consider propositions that are not adjacent to each other. Word pairs are an extension of unigrams to tasks involving pairs of propositions. Note that while discourse connectives, such as "because," are strong indicators of support relations, they are only applicable to proposition pairs that are adjacent.
—**First-Last-First3** is the first, last, and first three words of proposition 1 and those of 2. The goal is to capture discourse indicators or expressions that function as discourse indicators. Even when a known list of discourse indicators, such as *because, since*, and *therefore* is used as a feature, First-Last-First3 can be useful, as it also captures multiword expressions such as "as a result."
—**Verbs** is the count of pairs of verbs from proposition 1 and proposition 2 belonging to the same Levin English Verb Class [Levin 1993]; the average lengths of verb phrases as well as their Cartesian product; and, last, the part of speech of the main verb from

[12]Laplacian Smoothing was used for Naive Bayes, and SVM was training with linear kernel where the hyper-parameters were tuned through cross-validation.
[13]Word Pairs [Marcu and Echihabi 2002], First-Last-First3 [Wellner et al. 2006], Verbs [Pitler et al. 2009], and Production Rules [Lin et al. 2009].

Table V. Support vs No-Relation Binary Classification Results for Ordered Proposition Pairs: Naive Bayes and SVM Results Are Averages of 10 Rounds of Experiments with Randomly Downsampled Training Set to Balance the Class Distribution. (For SVM, This Approach Led to Better Results Than Introducing Class Weights)

Scope	Algorithm	Precision	Recall	F1	Accuracy
Global	Naive Bayes	0.02	0.94	0.05	0.16
	SVM	0.04	0.54	0.08	0.73
	Rule-based	1.00	0.42	**0.59**	**0.99**
Local	Naive Bayes	0.16	0.91	0.28	0.30
	SVM	0.24	0.49	**0.32**	**0.69**
	Rule-based	0.17	0.58	0.26	0.52

each argument. Levin Verb classes provide a means of clustering verbs according to their meanings and behaviors. Also, longer verb phrases may indicate support in the form of justification.

—**Production Rules** refers to three features denoting the use of syntactic production rules in proposition 1, proposition 2, or both. The syntactic structure of an argument can influence that of the other argument as well as its relation type. We take the smallest units of the syntactic parse trees, that is, production rules, as features to minimise the sparsity problem. A parse tree consists of applications of production rules, such as "[noun phrase] → [[determiner] [noun]]."

—**Discourse Indicators** are words that capture discourse relations among propositions, such as *because* and *therefore*. While most of them are meaningful in the cases where the propositions under consideration appear consecutively, a few of them are free from this restriction, as long as they share the same topic. See Table IV for the full list of discourse indicators.

Results. Table V summarises performances of each classifier on the test set under two different settings: *Scope* denotes whether all proposition pairs (*Global*) or only the pairs that are 2 propositions apart at most (*Local*) were used in the experiments. Both SVM and Naive Bayes classifiers were trained on the training set, whereas the rule-based classifier did not involve any training.

Both SVM and Naive Bayes perform poorly in the global scope but much better in the local scope. While the global scope is a better representation of the real scenario, in which a given proposition can support any proposition in the thread, the class imbalance makes it a challenging learning problem. The negative instances, or ordered pairs in a non-support relations, are more than 100 times the number of positive instances even after the preprocessing step. We tried to remedy the problem by introducing class weights in SVM and downsampling the negative instances to balance the training set, but the approaches were not too effective.

Table VI shows a clear difference in the set of most important features for SVM and Naive Bayes classifiers. Naive Bayes tends to attach more weight to word pair features, whereas production rules are more important for SVM.

The word pair "(peanuts, peanuts)" is correlated with a support relation and "(?,?)" with a no-support relation. The former suggests that propositions that share the same topic are more likely to be in a support relation, and the latter shows that a question is unlikely to support another question. The most important feature for SVM in the local scope is having a verb phrase consisting of a verb and adjective phrase in the supporting proposition. This could be hinting that supporting propositions often contain a detailed description.

Table VI. Most Informative Features: Features Listed in "+" and "-" Rows Are the Most Informative Features Associated with Ordered Proposition Pairs in Support and No Support Relation, Respectively. Parenthesised Features Are Word Pair Features, and Features with Arrows Are Production Rules. Last, "[p]" Means the Given Feature Appears in the Supporting Proposition (Premise), and "[c]" The Supported Proposition (Conclusion)

Scope	Algorithm		Most Informative Features
Global	Naive Bayes	+	(i,be), (a,ban), (be,should), (only,the), (not,of)
		-	(be,do), whadjp → wrb jj [c], (?,?), (a,their), last token: "?" [p,c]
	SVM	+	s → np vp . [c], (you,you), vp → vbp adjp [p], s → np vp . [p]
		-	s → np vp [c], advp → rb [c], np → nn [c], np → nns [p]
Local	Naive Bayes	+	(flight,be), (the,must), (peanuts,peanuts), (peanuts,be)
		-	(to,just), frag → sbar . [c], (that,just), adjp → jj sbar [p]
	SVM	+	vp → vb adjp [p], root → s [c], (are,you), s → np advp vp . [c]
		-	np → nns [p], sbar → in s [c], np → prp [p], vp → vbp pp [p]

Table VII. Confusion Matrix for the Classification Results of the Rule-Based Classifier (Global Scope)

		Predicted	
		Support	No-support
Actual	Support	161	116
	No-Support	0	12175

The rule-based classifier[14] performs quite well in the global scope. A quick look at the confusion matrix (Table VII) reveals that this performance was made possible by the search space reduction step—all consecutive proposition pairs that are not in support relation were filtered out. We do not see the same effect in the local scope, however, resulting in a much lower precision.

4.4. Taking an Interpretative Step

No matter how successful automatic mining of argument structure might be, the key challenge is then to provide information that allows sense to be made of the potentially very large datasets. A first example of such an interpretative step that offers end-users an insight into a debate is the notion of *centrality*. Central issues are those that play a particularly important role, and for this we can adapt eigenvector centrality (used in the Google Pagerank algorithm [Brin and Page 1998]). This measure is closer to intuitions about claim centrality in arguments than alternative measures such as the Estrada index [Estrada 2000] despite the latter's wide applicability. We have not found the Estrada index an informative measure for debate structure.

First, we build the subgraph corresponding only to vertices connected by support or conflict relationships, which we call $G_l = (V_l, E_l)$, where $V_l = \{v \in V : R(V) \in \{support, conflict\}\}$ and $\forall v_l \in V_l$, if $(v_l, v') \in E$, then $(v_l, v') \in E_l$ and if $(v', v_l) \in E$, then $(v', v_l) \in E_l$. We can then define eigencentrality over G_l as follows:

$$Central(v) =_{def} \frac{1}{\lambda} \sum_{\substack{v' \in V_l \\ \text{s.t. } (v,v') \in E_l}} Central(v'). \quad (1)$$

From the output of the classifier presented in Section 4.3, we are able to automatically generate argument maps corresponding to those in the manually annotated test corpus (an example is shown in Figure 3). These maps can then be used to compare the

[14]As previously mentioned, the rule-based classifier simply returns "true" only when a given pair of propositions are consecutive and contain one or more discourse indicator.

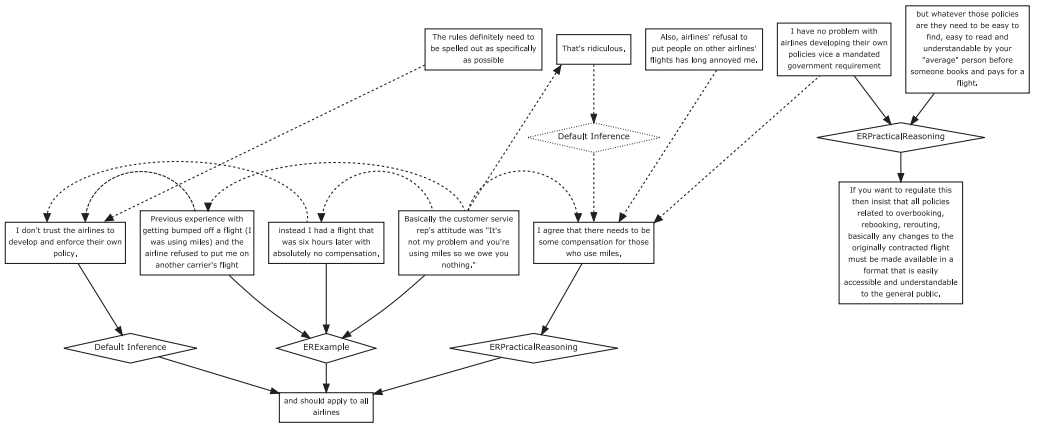

Fig. 3. Argument map comparing manual and automatically identified connections. Correctly identified connections are in bold, false positives are dashed lines, and the single false negative is represented by dotted lines.

calculated central issues for both manually and automatically annotated arguments. The issues with the highest centrality ranks for each dataset are listed below:

Top central issues from the manually annotated corpus

1. Again, the only prudent course of action is to require that distribution of peanut on airplanes be discontinued.
2. That's being a hypocrite.
3. If you have an allergy to peanuts and you know you have it then take your own precautions.
4. Please help protect people by offering people the opportunity to get peanut free flights or ban the sale and serving of nut products on the planes.
5. DOT should set maximum tarmac delay trigger.
6. Instead of conflating the possible with the inevitable, you should focus on the reality, which is that the possibility is extremely small.
7. Request peanut free services from the airlines for yourself,
8. I am utterly amazed at the ignorance displayed by some of those commenting here.

Top central issues from the automatic classification

1. Please help protect people by offering people the opportunity to get peanut free flights or ban the sale and serving of nut products on the planes.
2. Again, the only prudent course of action is to require that distribution of peanut on airplanes be discontinued.
3. If you have an allergy to peanuts and you know you have it then take your own precautions.
4. Request peanut free services from the airlines for yourself,
5. The latest research indicates that peanut allergy doubled in children from 1997 to 2002 and that number continues to increase.
6. Instead of conflating the possible with the inevitable, you should focus on the reality, which is that the possibility is extremely small.
7. Leave my peanuts alone!
8. An outright ban should be in place.

Even just superficial comparison of these lists suggests strong overlap between the highest ranked issues. This impression is borne out by more thorough analysis across the complete ranked list of 634 issues, for which the Kendall rank correlation coefficient, $\tau = 0.604$ ($p < 0.05$) [Kendall 1938]. These results suggest that although automatic identification of the argumentative structure of the text remains immensely challenging, the results obtained from the automated approach presented here are sufficient to perform the kind of analysis detailed in Section 3.2 and to provide significant insight into the nature of the debate and the issues being discussed.

5. CONCLUSION

We have shown that, despite the challenges faced in understanding and summarising the large volumes of data that can be produced from online citizen dialogue, by

analysing the argumentative structure contained within such a discussion, we are able to obtain a deeper understanding of the issues being raised than by using existing techniques such as controversy mining. Using the Argument Interchange Format to represent the argumentative structure, we are able to see not just points of agreement and disagreement but to understand why those views are held and the expression of opinions both in support and in conflict with them.

We have highlighted several possible measures that can be determined from these structures, giving a clear insight into the topic and providing policy makers with tools to understand and interpret citizen dialogues. These include areas of disagreement, areas on which people generally agree, and those areas that are central to the debate. We have selected a simple metric, centrality, to use as our exemplar and shown how even modest performance on the recovery of the argument network expressed in the discussion can yield robust results for this metric. The article has shown how a pipeline running through various computational linguistics techniques through analytical processes can be connected together. Though evaluation with users remains future work, the results in this article demonstrate for the first time that the state of the art in argument mining is already sufficient to start to offer real value to decision makers and those responsible for public policy in interpreting and gaining insight into large-scale, complex debates.

ACKNOWLEDGMENTS

We are grateful to the anonymous reviewers for their comments on earlier versions of this manuscript.

REFERENCES

Filip Boltužić and Jan Šnajder. 2015. Identifying prominent arguments in online debates using semantic textual similarity. In *Proceedings of the 2nd Workshop on Argumentation Mining*. Association for Computational Linguistics, Denver, CO, 110–115.

Sergey Brin and Larry Page. 1998. The anatomy of a large-scale hypertextual Web search engine. *Comput. Net. ISDN Syst.* 30 (1998), 107–117.

Carlos Chesñevar, Sanjay Modgil, Iyad Rahwan, Chris Reed, Guillermo Simari, Matthew South, Gerard Vreeswijk, Steven Willmott, and others. 2006. Towards an argument interchange format. *Knowl. Eng. Rev.* 21, 4 (2006), 293–316.

Yoonjung Choi, Yuchul Jung, and Sung-Hyon Myaeng. 2010. Identifying controversial issues and their subtopics in news articles. In *Intelligence and Security Informatics*. Springer, 140–153.

Rory Duthie, John Lawrence, Katarzyna Budzynska, and Chris Reed. 2016. The CASS technique for evaluating the performance of argument mining. In *Proceedings of the 3rd Workshop on Argumentation Mining*. Association for Computational Linguistics, Berlin.

Ernesto Estrada. 2000. Characterization of 3D molecular structure. *Chem. Phys. Lett.* 319, 5 (2000), 713–718.

Cynthia R. Farina and Mary J. Newhart. 2013. Rulemaking 2.0: Understanding and getting better public participation. *Cornell e-Rulemaking Initiative Publications* (2013).

Vanessa Wei Feng and Graeme Hirst. 2011. Classifying arguments by scheme. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 987–996.

Stefan Gries. 2009. Quantitative Corpus Linguistics with R. A Practical Introduction. Routledge.

Mark Howard. 2001. e-Government across the globe: How will 'e' change government. *Government Finance Review* 17, 4 (2001), 6–9.

Mathilde Janier, John Lawrence, and Chris Reed. 2014. OVA+: An argument analysis interface. In *Proceedings of the 5th International Conference on Computational Models of Argument (COMMA'14)*. 463–464.

Maurice G. Kendall. 1938. A new measure of rank correlation. *Biometrika* 30, 1/2 (1938), 81–93.

Aniket Kittur, Bongwon Suh, Bryan A. Pendleton, and Ed H. Chi. 2007. He says, she says: Conflict and coordination in Wikipedia. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 453–462.

Jey Han Lau and Timothy Baldwin. 2016. An empirical evaluation of doc2vec with practical insights into document embedding generation. In *Proceedings of the 1st Workshop on Representation Learning for NLP*. Association for Computational Linguistics, Berlin, Germany, 78–86.

John Lawrence, Floris Bex, Chris Reed, and Mark Snaith. 2012. AIFdb: Infrastructure for the argument web. In *Proceedings of the 4th International Conference on Computational Models of Argument (COMMA'12)*. 515–516.

John Lawrence and Chris Reed. 2015a. In *Argument Mining using Argumentation Scheme Structures*. (In review)

John Lawrence and Chris Reed. 2015b. Combining argument mining techniques. In *Working Notes of the 2nd Argumentation Mining Workshop (ACL'15*).

John Lawrence, Chris Reed, Colin Allen, Simon McAlister, and Andrew Ravenscroft. 2014. Mining arguments from 19th century philosophical texts using topic based modelling. In *Proceedings of the First Workshop on Argumentation Mining*. Association for Computational Linguistics, Baltimore, MD, 79–87.

Quoc V. Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *ICML*, Vol. 14. 1188–1196.

Beth Levin. 1993. *English verb classes and alternations: A preliminary investigation*. University of Chicago press.

Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. 2009. Recognizing implicit discourse relations in the penn discourse treebank. In *EMNLP*. 343–351.

Daniel Marcu and Abdessamad Echihabi. 2002. An unsupervised approach to recognizing discourse relations. In *ACL*. 368–375.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*. 3111–3119.

George A. Miller. 1995. WordNet: A lexical database for english. *Commun. ACM* 38, 11 (1995), 39–41.

Amita Misra, Pranav Anand, Jean Fox Tree, and Marilyn Walker. 2015. Using summarization to discover argument facets in online idealogical dialog. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics–Human Language Technologies (NAACL HLT'15)*.

Marie-Francine Moens, Erik Boiy, Raquel Mochales Palau, and Chris Reed. 2007. Automatic detection of arguments in legal texts. In *Proceedings of the 11th International Conference on Artificial Intelligence and Law* (2007). ACM, 225–230.

M. Jae Moon. 2002. The evolution of e-government among municipalities: Rhetoric or reality? *Publ. Admin. Rev.* 62, 4 (2002), 424–433.

Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation mining: The detection, classification and structure of arguments in text. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law* (2009). ACM, 98–107.

Joonsuk Park, Cheryl Blake, and Claire Cardie. 2015. Toward machine-assisted participation in eRulemaking: An argumentation model of evaluability. *Proceedings of the 15th International Conference on Artificial Intelligence and Law (ICAIL'15)*.

Joonsuk Park and Claire Cardie. 2012. Improving implicit discourse relation recognition through feature set optimization. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL'12)*. Association for Computational Linguistics, Stroudsburg, PA, 108–112.

Joonsuk Park and Claire Cardie. 2014. Identifying appropriate support for propositions in online user comments. In *Proceedings of the 1st Workshop on Argumentation Mining*. Association for Computational Linguistics, 29–38.

Joonsuk Park, Sally Klingel, Claire Cardie, Mary Newhart, Cynthia Farina, and Joan-Josep Vallbé. 2012. Facilitative moderation for online participation in eRulemaking. In *Proceedings of the 13th Annual International Conference on Digital Government Research*. ACM, 173–182.

Andreas Peldszus. 2014. Towards segment-based recognition of argumentation structure in short texts. In *Proceedings of the 1st Workshop on Argumentation Mining*.

Andreas Peldszus and Manfred Stede. 2015a. An annotated corpus of argumentative microtexts. In *Proceedings of the 1st European Conference on Argumentation*.

Andreas Peldszus and Manfred Stede. 2015b. Towards detecting counter-considerations in text. In *Proceedings of the 2nd Workshop on Argumentation Mining (ARG-MINING 2015)*.

Emily Pitler, Annie Louis, and Ani Nenkova. 2009. Automatic sense prediction for implicit discourse relations in text. In *ACL/AFNLP*. 683–691.

Ana-Maria Popescu and Marco Pennacchiotti. 2010. Detecting controversial events from twitter. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*. ACM, 1873–1876.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K. Joshi, and Bonnie L. Webber. 2008. The penn discourse treebank 2.0. In *LREC*. Citeseer.

I. Rahwan, F. Zablith, and C. Reed. 2007. Laying the foundations for a worldwide argument web. *Artif. Intell.* 171, 10–15 (2007), 897–921.

Sara Rosenthal and Kathleen McKeown. 2015. I couldnt agree more: The role of conversational structure in agreement and disagreement detection in online discussions. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 168.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014)*. 1501–1510.

Simone Teufel. 1999. *Argumentative Zoning: Information Extraction from Scientific Text*. Ph.D. Dissertation. University of Edinburgh.

Simone Teufel and Marc Moens. 2002. Summarizing scientific articles: Experiments with relevance and rhetorical status. *Comput. Ling.* 28, 4 (2002), 409–445.

Frans H. van Eemeren, Peter Houtlosser, and Arnolda Francisca Snoeck Henkemans. 2007. *Argumentative Indicators in Discourse: A Pragma-dialectical Study*. Vol. 12. Springer Science & Business Media.

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*. 812–817.

Douglas N. Walton, Chris Reed, and Fabrizio Macagno. 2008. *Argumentation Schemes*. Cambridge University Press.

Bonnie Webber, Markus Egg, and Valia Kordoni. 2012. Discourse structure and language technology. *Nat. Lang. Eng.* 18, 4 (2012), 437–490.

Ben Wellner, Lisa Ferro, Warren R. Greiff, and Lynette Hirschman. 2006. Reading comprehension tests for computer-based understanding evaluation. *Nat. Lang. Eng.* 12, 4 (2006), 305–334.

Received January 2016; revised November 2016; accepted December 2016