Microdebates App for Android: A tool for participating in argumentative online debates using a handheld device

Nefise Yağlıkçı
TOBB University of Economics and Technology
Ankara, Turkey
Email: ngyaglikci@gmail.com

Paolo Torroni
DISI, University of Bologna
Bologna, Italy
Email: paolo.torroni@unibo.it

***Abstract*—Microdebates App for Android is a part of a research effort aimed to propose better ways of exchanging ideas and opinions in online communities. With it, a user can argue from a handheld device, using Twitter. One can also visualize opinions of other microdebaters, explore ongoing debates, and see where the consensus is. Under the hood, Microdebates uses computational argumentation to rank opinions and drive the visualization. The result is a visual summary of the debate that takes into account semantic information such as explicit attack relations that link opinions together. To the best of our knowledge, this is the first application that brings computational argumentation to handheld devices. We describe the application and its logic, and discuss results from an empirical study.**

I. Introduction

Social media are increasingly used to support online debate and facilitate citizens' engagement in policy and decision-making [1]. Accordingly, in recent years researchers and practitioners have been proposing innovative ways of organising, presenting and extracting useful information from online discussion tools. An established research literature documents the advantages, and challenges, of making the structure and status of a dialogue or debate more visible [2]. Many proposed solutions are based on computational argumentation, a lively interdisciplinary domain that recently has gain momentum in the AI community and beyond [3]. In general, the idea is to present the user with a graph-like visual representation of elements of a debate and the connections among them.

Argument diagrams can be used to display premises and conclusions in an argument, and to show how groups of premises support conclusions that can in turn be used as premises in adjoining arguments. Araucaria,[1] for example, uses argument diagrams. The Evidence Hub [1], [4] is an argumentation-based tool to structure conversations, which puts issues, ideas, and evidence at the centre of a reflective community of practice. The Evidence Hub adopts a version of the IBIS model [5] to create argument maps which are then visualised in the form of graphs (see Figure 1). This approach is quite common [6], [7]. Other argumentation-enabled web applications are discussed in [8], [9], [6], [10], [11], [12], [13].

It would be useful to have these technologies effectively support online debates at a large scale. For example, computational argumentation could support the activities preceding deliberations, in online democracy and e-participation environments. However, a problem we see in doing so with current approaches is that graph-like structures are not necessarily the best way to represent a debate, especially when big numbers are in play. In handheld devices, for instance, with limited real estate, exploring a graph could be a compelling task. Moreover, large numbers of users could result in substantial heterogeneity. It may be unrealistic to assume that a particular argument structure is understood and correctly used by the participants in a debate. Indeed, there is limited evidence that structured argument visualization tools such as argument maps and diagrams increase understanding and engagement in online discussions.

In the present work, we take a different approach. We also aim at improving online debates and support the agreement process. However, instead of proposing new conceptual models to a prospective user, we build on largely established social media (microblogs), and we use word clouds, as opposed to diagrams, to summarise a debate.

Microblogging is a new form of communication whereby users can describe their current status in short posts distributed by instant messages, mobile phones, email or the Web [14]. A very popular platform for microblogging is Twitter, where people talk about their daily activities and seek or share information [15] by broadcasting brief textual messages (*tweets*) to their *followers* [16].

Word clouds are visual presentations of a set of words, or a subset thereof selected by some rationale, in which attributes of the text such as size, weight, or color are used to represent features, such as frequency, of the associated terms [17]. The idea is that by looking at a set of word clouds, one can form a general impression of the underlying arguments and their status in an ongoing discussion. We believe that this visualization method will be more accessible to the general public compare to other interfaces that put explicit emphasis

[1] http://araucaria.computing.dundee.ac.uk

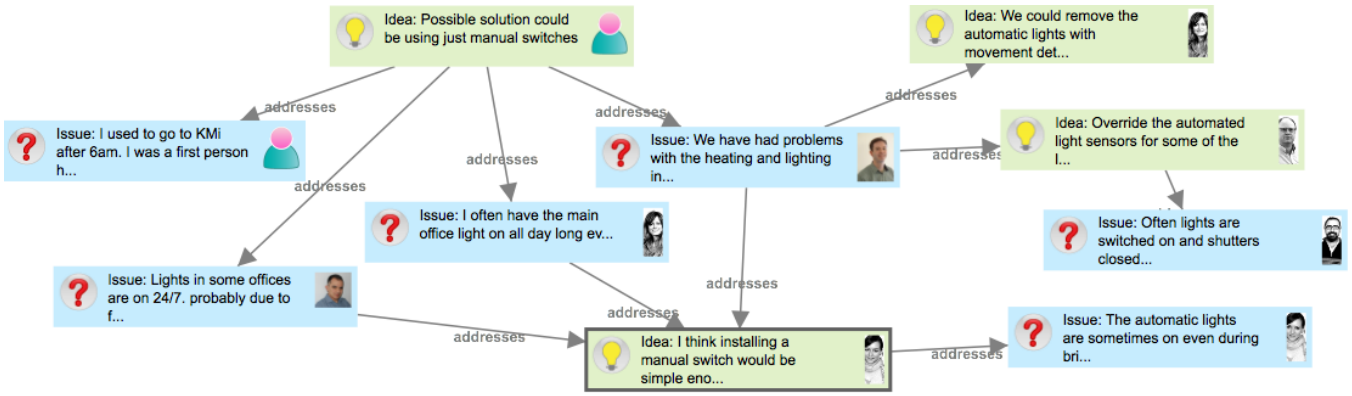

Figure 1. Network graph representation of arguments about *Lights for sensors*. Retrieved on November 22, 2013 from http://isave.evidence-hub.net/.

on a data model, which should instead remain invisible.

To this end, we devised a new method for *ranking* arguments in order to present the user with a pictorial, browsable, and *linear* view of the ongoing debate, that gives greater emphasis to more popular arguments.

We implemented an application for Android that can be used to contribute to a debate and obtain said visual summary. In this article, we present the application, its conceptual underpinning, and discuss results from our initial experimentation.

II. Microdebates

Microdebates were proposed by Gabbriellini and Torroni as a way to help organising and confronting opinions online, in an automated way [18], [19]. They consist in streams of tweets annotated with some special tags, to mark opinions and conflicts between opinions. In particular, the $$ tag (*double-cashtag*), as in $$redLooksGreat, is interpreted as (the label of) an opinion or argument *supported* by the author of the tweet, whereas the !$ tag (*bang-cashtag*), as in !$greenLooksGreat, is interpreted as (the label of) an opinion or argument *opposed* by the author of the tweet.

There is no special syntax for tweets belonging to a microdebate, other than the usual Twitter syntax which imposes a 140 character limit for a tweet, and space-free tags. However, tweets belonging to a microdebate should at least contain a discussion identifier (hashtag), and an argument identifier (double-cashtag). There are no other restrictions on the number and type of tags a tweet can/should contain. Figure 2 illustrates.

When a user broadcasts a tweet containing a double-cashtag/bang-cashtag association, a link is set between the two tags and the corresponding opinions. If another user sends out another tweet with the same association, or recasts the same tweet, that link is reinforced.

The keywords identified by double-cashtags and bang-cashtags are labels for *abstract arguments*, while the links

Figure 2. A fragment of a Twitter stream, showing a sample microdebate. Twitter organises its entries top to bottom from newest to oldest.

between such keywords represent *attack* relations as it will become clear in the next section.

III. Weighted Abstract Argumentation Frameworks

We build on Dung's abstract argumentation frameworks.

Definition 1 (AAF [20]): An Abstract Argumentation Framework (AAF) is a pair $\langle X, A \rangle$ where X is a set of *arguments*, and A is an *attack* binary relation defined on X. A set of arguments $S \subseteq X$ or "*extension*" is said to be:

- *conflict-free* iff $\nexists a, b \in X, (a, b) \in A$;
- *admissible* iff it is conflict-free and $\forall a \in S$ such that $\exists (b, a) \in A$, $\exists c \in S$ such that $\exists (c, b) \in A$ (i.e., S defends all its attackers);
- *preferred* iff it is admissible and $\nexists S' \subseteq X$ such that $S' \supset S$ and S' is admissible (i.e., S is maximal, with respect to set inclusion).

Dung's AAFs can be used to model the arguments in a microdebate and their relations. However, they do not capture accrual. Instead, we wish to reason on the magnitude of the consensus around the emerging positions. To this end, we use Weighted Abstract Argumentation Frameworks (WAAFs) instead of simple AAFs.

WAAFs were introduced by Bistarelli and Santini [21], [22] as an extension of Dung's framework. In a WAAF, attacks are labelled with a weight, indicating its relative strength, for example in terms of a probability score, or of a number of support votes. The idea is similar to that of Dunne et al's Weighted Argument Systems [23], except that WAAFs not only consider internal inconsistency, but also the balance between attacks and defense (*weighted defense*).

Bistarelli and Santini's work is based on the general notion of semiring [24] to define an arithmetics of preference values attached to arguments, and can be instantiated in different ways. For example, one can use a Weighted semiring, or a Fuzzy semiring, or other semirings. For the sake of simplicity, we will consider here an instantiation of WAAFs with *Weighted semirings*. A Weighted semiring is a tuple $\langle \mathbb{R}^+ \cup \{\infty\}, \min, +, \infty, 0 \rangle$ and it is used to define arithmetic operations on weights.

As pointed out by Dunne et al. [23], weights could represent a number of things. For instance, they could represent relative rankings of attacks, votes in support of attacks, subjective beliefs modelled using probabilities. In our work, the weight of an attack is directly proportional to the number of tweets expressing that attack.

We will now illustrate some key notions adapted from [23], [21], [22]. Literature offers a plethora of semantics for (weighted) abstract argumentation frameworks. We will focus on the *preferred* semantics.

Definition 2 (WAAF [21]): A WAAF is a triple $\langle X, A, w \rangle$ where X is a set of *arguments*, A is the *attack* binary relation defined on X, and $w : A \to \mathbb{R}^+ \cup \infty$ is a function assigning *weights* to attacks. Given $a, b \in X$, $\forall (a, b) \in A, w(a, b) = s$ means that a attacks b with "strength" $s \in \mathbb{R}^+ \cup \infty$, the latter being the domain of preference values of the Weighted semiring.

Example 1: Consider W_1 in Figure 3, from [23]. Nodes represent arguments. Edges represent attacks. Labels attached to attacks represent weights. If we ignore the weights, the argumentation framework has two preferred extensions: $S = \{a_1, a_2, a_4, a_6\}$ and $T = \{a_3, a_5, a_7, a_8\}$.

Definition 3 (Attacks for sets of arguments [21]): Given two extensions $S, T \subseteq X$ and an argument $a \in X$, we say that:

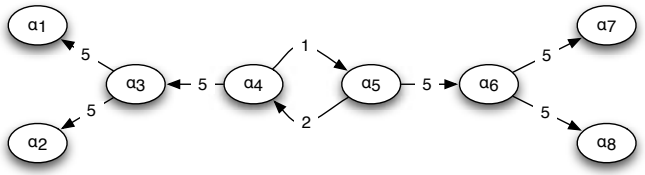

Figure 3. A sample WAAF, W_1, from [23].

- S attacks a with strength k, $w(S, a) = k$, iff $\sum_{b \in S} w(b, a) = k$;
- S attacks T with strength k, $w(S, T) = k$, iff $\sum_{a \in T} w(S, a) = k$.

Example 2: Consider again W_1 and the two preferred extensions S and T. Then,

- $w(S, a_5) = w(a_4, a_5) = 1$ (S attacks a_5 with strength 1);
- $w(S, T) = w(S, a_3) + w(S, a_5) + w(S, a_7) + w(S, a_8) = 5 + 1 + 5 + 5 = 16$ (S attacks T with strength 16).

Definition 4 (α-conflict-free extensions [21]): An extension $S \subseteq X$ is α-conflict-free iff $w(S, S) \leq \alpha$.

Example 3: In W_1, $\emptyset$ and $\{a_4, a_6\}$ are 0-conflict-free (and therefore α-conflict-free for any $\alpha \geq 0$), while $\{a_4, a_5\}$ or, say, $\{a_4, a_5, a_7\}$ are only 3-conflict free (and indeed α-conflict-free for any $\alpha \geq 3$).

Definition 5 (Weighted defense [21]): An argument b is weighted-defended by an extension S (S "*w-defends*" b) iff $\forall a \in X \setminus S$ such that $(a, b) \in A$, $w(a, b) \geq w(S, a)$.

Example 4: In W_1, $\{a_5\}$ w-defends a_5, its only element, because $w(a_4, a_5) \leq w(a_5, a_4)$ and there are no other attacks against a_5. Conversely, $\{a_5\}$ does not w-defend a_3, because there is an attack against a_3 from an argument a_4 outside of $\{a_5\}$, $a_4 \in X \setminus \{a_5\}$, whose weight is $w(a_4, a_3) = 5$, whereas the only attack from $\{a_5\}$ to a_4 is $w(a_5, a_4) = 2 < 5$.

Definition 6 (α-admissible/α-preferred extensions [21]): An α-conflict-free extension S is α-admissible iff it w-defends each of its elements. An α-admissible extension S is α-preferred if it is maximal, with respect to set inclusion.

It is worthwhile noticing that, while an α-admissible extension S is also α'-admissible for any $\alpha' \geq \alpha$, the same cannot be said of α-preferred extensions, due to the maximality requirement.

Example 5: In W_1, $\emptyset$, $\{a_5\}$, $\{a_5, a_7\}$ and $\{a_5, a_7, a_8\}$ are 0-admissible extensions, whereas $\{a_3, a_5\}$ is not α-admissible (for any α), because it does not w-defend its own element a_3. $\{a_5, a_7, a_8\}$ is not contained in any other 0-admissible extension. Therefore $\{a_5, a_7, a_8\}$ is both 0-admissible and 0-preferred. There is no other 0-preferred extension. If we are prepared to tolerate inconsistency up to

$\alpha = 3$, then $\{a_4, a_5\}$ is 3-conflict-free and has no attackers, thus it is 3-admissible. There is only one 3-preferred extension: $\{a_1, a_2, a_4, a_5, a_7, a_8\}$. Similarly, $\{a_5, a_6\}$ is the only 5-preferred extension, whereas $\{a_3, a_4, a_5, a_7, a_8\}$ and $\{a_1, a_2, a_4, a_5, a_6\}$ are both 8-preferred, being also the only such extensions. Notice that $\{a_5, a_7, a_8\}$ is 0-preferred but not 8-preferred, because it is not maximal when $\alpha = 8$.

Definitions 4 to 6 naturally extend Dung's conflict-free, admissible, and preferred semantics [20]. The concept of weighted defense is the main difference between Bistarelli and Santini's interpretation [21] and Dunne et al.'s interpretation of WAAFs [23]. In particular, according to Dunne et al.'s weighted extension of Dung's preferred semantics, i.e., the β-preferred semantics,[2] $S = \{a_1, a_2, a_4, a_6\}$ and $T = \{a_3, a_5, a_7, a_8\}$ are both 0-admissible and 0-preferred in spite of the asymmetry between $w(S, T)$ and $w(T, S)$ due to the difference between $w(a_4, a_5)$ and $w(a_5, a_4)$. In Bistarelli and Santini's interpretation, instead, $T = \{a_3, a_5, a_7, a_8\}$ is the only 0-preferred extension. This, and the availability of an efficient CP-based implementation [22], convinced us to opt for the latter as a conceptual and practical basis for our work. Other such sytems are becoming available [25].

IV. Argument ranking

If an argument a belongs to an α-preferred extension S, we know two things: (1) a can peacefully coexist with the other arguments in S, the inconsistency within S being at most α, and (2) for every tweet attacking it, there exists at least another tweet that counters the attack. So, it would seem reasonable to give arguments in S the status of "popular" argument. We could display all popular arguments with a special style that gives emphasis to them. That could be valuable output for the user.

However, we know from Example 5 that, for a given WAAF, there may be more than one α-preferred extension. That would be a problem. Clearly, we don't want to confuse the user with many possibilities. We should instead provide one single view of the ongoing debate, aggregating all viewpoints in a meaningful way.

To this end, we use extensions to define a function that produces *a unique argument ranking*, related to the argument's popularity. In particular, we distinguish between arguments belonging to *all extensions* (maximally popular), *at least one* extension (supported by some), and *none at all*.

Definition 7 (α-skeptically/α-credulously preferred): An argument $a \in X$ is α-skeptically preferred iff it belongs to all α-preferred extensions: $skep(a, \alpha)$ iff $\nexists S \subseteq X \setminus \{a\}$ s.t. S is α-preferred. It is α-credulously preferred iff it belongs to at least one α-preferred extension: $cred(a, \alpha)$ iff $\exists S \subseteq X$ s.t. $a \in S$ and S is α-preferred.

[2]The intuition behind Dunne et al.'s proposal for WAAFs is nicely explained by the "inconsistency budget" metaphor, whereby β, for budget, defines how much inconsistency we are prepared to tolerate. This tallies with Santini and Bistarelli's α value introduced in Definition 4.

For a given α, $skep(a, \alpha) \models cred(a, \alpha)$. However, $skep(a, \alpha) \not\models \exists \alpha' \neq \alpha$ such that $skep(a, \alpha')$. Finally, we will say that an argument a is non-α-preferred, $nope(a, \alpha)$, if $cred(a, \alpha)$ does not hold.

Example 6: In W_1, a_5, a_7, and a_8 are 0-skeptically preferred; a_1, a_2 and a_4 are 3-skeptically preferred; a_6 is 5-skeptically preferred, and a_3 is 8-credulously preferred.

Definition 8 (argument ranking): An argument ranking $r : X \to \mathbb{N}$ is a function defining a partial order over X.

We will represent rankings using tuples, whose elements are disjoint elements of 2^X covering X.

Example 7: A possible ranking of W_1's arguments X is $\hat{r}(X)$, which orders the arguments from most to least skeptically preferred, for varying values of α: $\hat{r}(X) = \langle\{a_5, a_7, a_8\}, \{a_1, a_2, a_4\}, \{a_6\}, \{a_3\}\rangle$. Alternatively, we could define a sequence of rankings $r_0, r_1, \ldots$, each for a fixed α, considering α-skeptically-preferred arguments more popular than α-credulously-preferred arguments, and the latter being more popular than the non-α-credulously-preferred arguments. Thus for $\alpha = 3$ we have six 3-sceptically-preferred arguments belonging to the only 3-preferred extension $\{a_1, a_2, a_4, a_5, a_7, a_8\}$; moreover, $nope(a_3, 3)$ and $nope(a_3, 6)$, thus $r_3 = \langle\{a_1, a_2, a_4, a_5, a_7, a_8\}, \emptyset, \{a_6, a_3\}\rangle$. Instead, for $\alpha = 8$, we have two 8-skeptically-preferred arguments (a_5 and a_7) and six 8-credulously preferred ones, thus $r_8 = \langle\{a_5, a_7\}, \{a_1, a_2, a_3, a_4, a_6, a_8\}, \emptyset\rangle$.

The Microdebates App uses a ranking of the latter type in the example above, for a fixed value of α.

One question that immediately arises is: How do we calibrate α? Since this ranking is not directed to automated reasoners, but to humans, this question could be satisfactorily answered only by experience with human users. We will discuss that later in this paper.

V. Microdebates for Android

The Microdebates App is distributed via Google Play.[3] A quick start guide is available at Storify.[4] Here we will focus on architectural and implementation aspects rather than on use.

The Application has a client-server architecture. The server runs a background process that manages the interaction with Twitter and maintains a MySQL database via WAMP.[5] The database contains all the information that is shown to the user on the client side: tweets, topics, word clouds, plus other data extracted from the tweets, such as attacks and weights, needed to compute the extensions (see Figure 4). Weights are determined by counting the tweets that express a given attack. The server retrieves from Twitter all new tweets about the topics listed in the database using Twitter4j,[6] a Java library for the Twitter API. The list of

[3]https://play.google.com/store/apps/details?id=it.unibo.ai.microdebates
[4]http://storify.com/paolotorroni/microdebates-for-android-for-beginners
[5]http://www.wampserver.com
[6]http://twitter4j.org

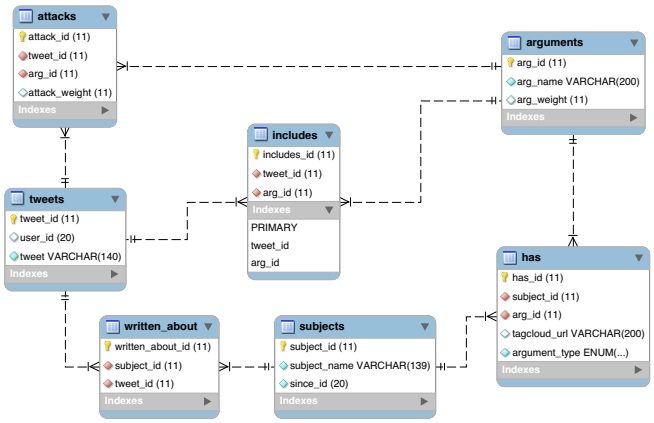

Figure 4. (Partial) ER scheme of the server-side database.

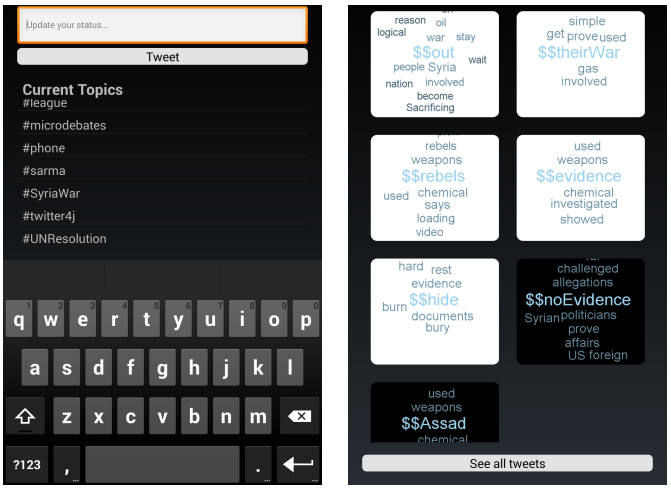

Figure 5. Microdebates activities: choose a topic, input a tweet (left); visualize a microdebate (right).

Figure 6. BestActress test case, $\alpha = 1$. On the left, all attacks weigh 1. On the right, attacks from $$Judi to $$Sandra and $$Amy weigh 2 (right).

topics is dynamic: if someone enters a new tweet from the application, introducing a new topic, the database is updated accordingly. At regular intervals, the server process looks for new tweets, and if there are any, it recomputes the extensions and updates the word clouds accordingly.

Word clouds are generated by the server application using an algorithm adapted from the one used in Wordle [26]. In particular, the server application analyses the tweets about each new topic (hashtag) to detect the language (English vs Italian) based on the occurrence of stop words. Then, for each argument in that topic (double-cashtag), it assigns a weight to each word contained in the set of tweets that support the argument. Such weight reflects the significance of the word, based on its frequency in the language and in the set of tweets. Weights determine the font size. For example, a word will be displayed in large font if it is rare in the language and frequent in the tweets. Words are then placed around the argument label (double-cashtag) in a pseudo-random fashion that avoids overlapping.

To compute the extensions, the server process invokes the ConArg methods for α-preferred extensions. ConArg also accommodates other semantics, but after a thorough analysis of the ConArg package and after nunning a number of tests, we became convinced that α-preferred extensions would be the best candidate for our application.

On the client side, the Microdebates App is actually quite simple. It does not do any reasoning. Its main purpose is to provide the user with an interface to enter new tweets and, mainly, to navigate through microdebates. The first task is trivial and again it is achieved using Twitter4j. As for the second task, the App gets all the data from our database via web services.

The App interface is straightforward. It consists of 4 activities:[7] (1) login, (2) choice of topic / tweet editing (3) microdebate visualization (4) tweet visualization. Figure 5 shows some screenshots. Interfaces to the database are included in the WAMP package.

[7]In the Android jargon, an activity is an application component that provides a screen with which users can interact in order to do something.

VI. Experiments

We first tested the application with artificially constructed WAAFs, and then conducted a study with real users. We produced 18 test microdebates. We implemented two variants of each microdebate, changing the weights on the attacks. Each variant was tested with one or more values of α. Figure 6 shows a test case (microdebate #bestActress) run with different attack weights and the same α.

This initial testing was successful and it gave expected results. It also helped us gain a clear understanding of α-preferred semantics and its behaviour in practice. We did not measure the computational effort required to obtain the semantics. Experiment have been done by others [27] with

graphs consisting of thousands of nodes (way more than we expect to have in a single debate). Indeed, the CPU time required by ConArg was negligible compare to the CPU time expended in the production of word clouds. But this is not worrisome either. We expect less recent debates to be reasonably stable, and more recent or ongoing debates to have a slowly growing number of hashtags (and thus word clouds). Moreover, producing a word cloud is not a computationally hard problem.

The study with human users was designed to answer a key question: *Does this application provide understandable, useful input to a human user; and under which circumstances?* We also wanted to gauge how much the user experience would be influenced by the system's calibration (in particular, by the value given to α), and whether having to create new cashtags would be seen as a hurdle by users not accustomed to microdebates.

The empirical results not only suggested some possible answers, but also gave valuable insights, that may in turn help a deeper understanding of the weighted argumentation semantics and their application in social contexts.

For this study, we approached ten participants in the 25-34 age group, all of them with an Android phone, a Twitter account, and a reasonable command of English. None of the participants is a native English speaker. Participants were asked to install the Microdebates App from Google Play and read the quick start guide from Storify. Then we divided all participants into two equally sized groups: Group A and Group B.

Each group was given a topic, and a 40-minute time frame, to discuss using Microdebates App. At the end of 40 minutes we gave a two-hour break. Then we gave a different topic, and an additional 40 minutes for a second microdebate. Eventually, we asked participants to answer an anonumous survey.

The topics were: *Are occupy protest movements justified?* and *Is nuclear energy justified and should it be expanded?* In the first debate, participants were allowed to create new cashtags in order to label their arguments. In the second debate, participants were given a fixed set of cashtags, each one with a brief explanation of the concepts around it. These conditions were the same for both groups. α was set to 1 for Group A's first debate (#mdoccupy) and to 3 for Group A's second debate (#mdnuke). Conversely, α was set to 3 for Group B's first debate (#mdprotest), and to 1 for Group B's second debate (#mdenergy).

In spite of equal conditions, the debates resulting from the two groups were largely different from one another (see Figures 7 and 8).

Group A: #mdoccupy consisted of 14 tweets, with 8 different arguments, 6 attack relations and a maximum weight of 3. The resulting WAAF was composed of four disconnected graphs, with three 1-skeptically preferred arguments displayed in white and five 1-credulously preferred

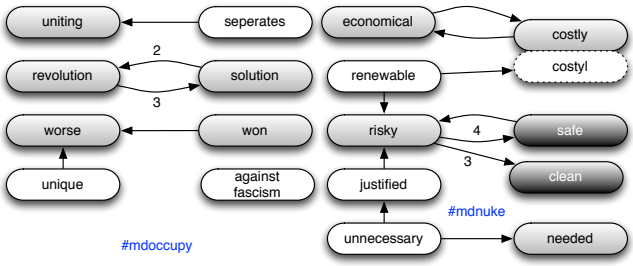

Figure 7. WAAFs from the study with Group A

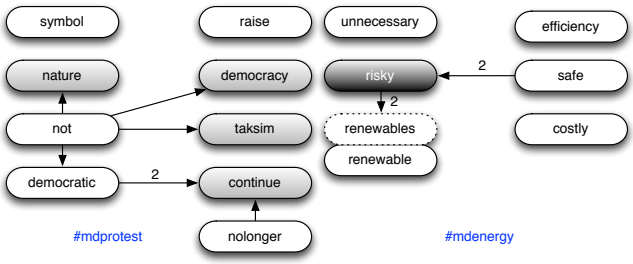

Figure 8. WAAFs from the study with Group B

argument displayed in grey. #mdnuke consisted of 21 tweets, with 9 different arguments, 10 attacks and a max weight of 4. The resulting WAAF was a fully connected graph, with two 3-skeptically preferred arguments (white), five 3-credulously preferred arguments (grey) and two losing arguments (black).

Group B #mdprotest consisted of 9 tweets, with 9 different arguments, 5 attacks, and a max weight of 2. The resulting WAAF was composed of three disconnected graphs of which one giant component and two isolated arguments, with five 3-skeptically preferred arguments (white) and four 3-credulously preferred arguments (grey). #mdenergy consisted of 7 tweets, with 6 arguments, 2 attacks, and a max weight of 2. The resulting WAAF was an extremely sparse graph, with four disconnected components, five 1-skeptically preferred arguments (white) and one 1-credulously preferred argument (black).

We observed that the structure of the debates was not visibly influenced by α, and that there was no substantial difference between debates whose cashtags were given and those with free cashtags. We can also identify two extreme situations: a fully connected graph produced by Group A, and an almost fully disconnected graph produced by Group B. From the answers to our survey, it appears that the two groups largely differed in terms of interest towards the topics. In particular, 5/5 participants of Group A and only 3/5 of Group B found the first topic interesting. Similarly, 4/5 participants of Group A and only 2/5 of Group B found the second topic interesting. This was reflected not only in the number of tweets produced (35 for Group A vs 16 for Group B), but also in the number of connections drawn between

arguments (16 vs 7). There were no differences, instead, with respect to prior experience with online social networks (Twitter or otherwise: 3/5 in both groups).

A part of the survey aimed to establish how well the microdebate helped understanding other people's opinions, how well word clouds summarized the opinions expressed in the debate, how well the colours reflected the consensus in the debate, and how useful it was to be able to express attacks between arguments.

The majority of participants declared that taglouds provide a good summary of the debate (6/10), attacks are useful (6/10), and considered the experience to be positive (7/10, with a preference for free sets of arguments as opposed to fixed sets of arguments). We did not receive any very negative feedback. Some participants declared that, after the debate, their *understanding* of the topics (2/10), and in some instances even their *opinion* (3/10), changed. Interestingly, the majority of participants from Group A found the colours to be appropriate (3/5) while noone found them misleading, whereas 2/5 participant in Group B found the colours misleading and only 1/5 found them appropriate.

There seems to be an apparent, and unsurprising, relation between a participant's interest in a topic and their contribution in the microdebate. Interestingly, we observed a direct correlation between a participant group's interest in a topic and the number of tweets and explicit attacks produced by the group, all else being equal. When at least 4/5 participants declared interest in a topic, the discussion received 14 to 21 tweets, containing 6 to 10 attacks and the connectivity of the argument network was 3. When less than 4/5 participants declared lack of interest in a topic, the discussion received 7 to 9 tweets with 2 to 5 attacks, and the connectivity was less than 3.

We can attempt some general conclusions from this experience. A more interested group seems to be likely to produce richer WAAFs (i.e., containing more information, in terms of connections and weights) and enjoy a sharper consensus. The visualization may generally perceived, by the group, to be appropriate and useful. Conversely, we can expect less interest to bring fewer tweets and sparser, shallower WAAFs. This will produces a larger grey area and amplify the effect of individual opinions and noise, thus odd results are more likely to appear, and the visualization may be perceived as misleading and not useful.

We can expect that, at least in some contexts, the network connectivity will increase with the number of tweets. This is a comforting prospect, since Microdebates are thought to be used for large debates. Another positive result is that the visualization of consensus and summary of debates are generally acceptable, and this does not seem to depend neither on α, nor directly on the number of arguments, but rather on the connectivity of the argument graph. Finally, it was interesting to see how attacks - a notion from computational argumentation theory, with a rigorous semantics - were actually used by participants unaware of the technicalities, and emerging opinions were perceived to be coherent. Indeed, 9/10 participants recognised the microdebates' potential to increase the quality of the debate.

Alongside these general observations, this study helped us identify some issues. For example, mistakes and typos, as well as multiple labelings of the same concept, can have a significant impact on the debate. Thus equipping the App's editor with syntactic/semantic analysis features could improve Microdebates significantly. Moreover, the effect of ironic comments is sometimes difficult to gauge. However, we noticed that the great majority of participants did use cashtags correctly, and even creatively.

VII. Conclusion

To the best of our knowledge, this is the first application that employs computational argumentation in a handheld device. Current approaches to summarizing and visualizing the status of an online discussion are either flat text summarization techniques that do not take into account the relations among emerging positions in a debate (e.g., a text summary or a single word cloud), or structured graphs that are difficult to browse in a small device, and that could represent a hurdle to the layperson, instead of a help.

The application is fully implemented and distributed via Google Play. Microdebates could represent a significant element of innovation in online democratic and participatory processes, especially in a pre-deliberation phase, where the objective is not to take a decision, but to let people shape options and reach a collective awareness.

The results of our experimental study are encouraging. Most significantly, user acceptance of the App's visualization of consensus and summary of debates was generally positive, and it did not visibly depend neither on α, nor on the number of arguments, the main factor influencing the result being instead the connectivity of the attack graph, which in turn was related to the degree of interest a group had in the topic under discussion.

Other experiments are under way. On 15 May 2014, together with colleagues from the University of Bologna's Political Sciences department, we organized a distributed experiment in the context of the EU presidential election live debate. There, three groups of students located in three different cities in Italy (Bologna, Siena, Trento) watched the debate online on TV, and exchanged feedback using the Microdebates App. We collected 290 tweets, which we are currently analyzing. Preliminary results show that the behaviour of users commenting on a political event is radically different from that observed in the previous situation. The conditions were different: here, new input was coming at a considerable rate (presidential candidates had to answer questions in rapid succession and in a very short time frame), while users had less opportunity to follow the TV debate and at the same time follow other tweets. User goals

and attitude were also different: we noticed a large amount of *ad hominem* arguments and ironic comments.

In general, we expect a variety of user behaviours that will depend on factors such as context, user goals, timing constraints, audience, etc. It is a part of our future work to identify and characterize application domains accordingly.

ACKNOWLEDGMENTS

This work was partially supported by EU project ePolicy, FP7-ICT-2011-7, grant agreement 288147.

We also wish to thank the Erasmus Training programme for sponsoring the first author's visit to the University of Bologna, Francesco Santini for his feedback and help with ConArg, and Aldo di Virgilio, Luca Pinto, Andrea Pedrazzani, and Daniela Giannetti for the EU presidential debate experiment.

REFERENCES

[1] A. De Liddo and S. Buckingham Shum, "Improving online deliberation with argument network visualization," in *Workshop: Digital Cities 8 at 6th C&T*, Munich, Germany, 2013.

[2] S. B. Shum, "The roots of computer supported argument visualization," in *Visualizing Argumentation*, Springer, 2003, pp. 3–24.

[3] I. Rahwan and G. R. Simari, *Argumentation in Artificial Intelligence*, 1st ed. Springer, 2009.

[4] A. De Liddo and S. Buckingham Shum, "The evidence hub: Harnessing the collective intelligence of communities to build evidence-based knowledge," in *Workshop: Large Scale Ideation and Deliberation at 6th C&T*, Munich, 2013.

[5] W. Kunz and H. W. J. Rittel, "Issues as elements of information systems," Institute of Urban & Regional Development, University of California, Working Paper 131, Jul. 1970.

[6] S. Buckingham Shum, "Cohere: Towards web 2.0 argumentation," in *Proc. 2nd COMMA*, FAIA 172. IOS, 2008, pp. 97–108.

[7] P. Baroni, M. Romano, F. Toni, M. Aurisicchio, and G. Bertanza, "An argumentation-based approach for automatic evaluation of design debates," in *Proc. CLIMA*, LNCS 8143, Springer, 2013, pp. 340–356.

[8] P. Torroni, M. Gavanelli, and F. Chesani, "Arguing on the semantic grid," in *Argumentation in Artificial Intelligence*, Springer, 2009.

[9] J. Schneider, T. Groza, and A. Passant, "A review of argumentation for the social semantic web," *Semantic Web*, vol. 4, no. 2, pp. 159–218, 01 2013.

[10] D. Cartwright and K. Atkinson, "Political engagement through tools for argumentation," in *Proc. 2nd COMMA*, FAIA 172. IOS, 2008, pp. 116–127.

[11] F. Bex, J. Lawrence, M. Snaith, and C. Reed, "Implementing the argument web," *Commun. ACM*, vol. 56, no. 10, pp. 66–73, 2013.

[12] C. I. Chesñevar, A. G. Maguitman, E. Estévez, and R. Brena, "Integrating argumentation technologies and context-based search for intelligent processing of citizens' opinion in social media," in *Proc. of ICEGOV 2012, Albany, NY*, Oct. 2012, pp. 166–170.

[13] K. Grosse, C. I. Chesñevar, and A. G. Maguitman, "An argument-based approach to mining opinions from Twitter," in *Proc. 1st AT*, CEUR-WS 918, 2012, pp. 408–422.

[14] K. M. DeVoe, "Bursts of information: Microblogging," *The Reference Librarian*, vol. 50, no. 2, pp. 212–214, 2009.

[15] A. Java, X. Song, T. Finin, and B. Tseng, "Why we twitter: understanding microblogging usage and communities," in *Proc. Web mining and social network analysis*. ACM, 2007, pp. 56–65.

[16] B. A. Huberman, D. M. Romero, and F. Wu, "Social networks that matter: Twitter under the microscope," *First Monday*, vol. 14, no. 1, January 2009.

[17] A. W. Rivadeneira, D. M. Gruen, M. J. Muller, and D. R. Millen, "Getting our head in the clouds: Toward evaluation studies of tagclouds," in *Proc. Human Factors in Computing Systems*, CHI '07. ACM, 2007, pp. 995–998.

[18] S. Gabbriellini and P. Torroni, "Large scale agreements via microdebates," in *Proc. 1st AT*, CEUR-WS 918, 2012, pp. 366–377.

[19] ——, "Microdebates: Structuring debates without a structuring tool," in *AI Commun.*, in press.

[20] P. M. Dung, "On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games," *Artif. Intell.*, vol. 77, no. 2, pp. 321–358, 1995.

[21] S. Bistarelli and F. Santini, "A common computational framework for semiring-based argumentation systems," in *Proc. ECAI*, FAIA 215. IOS, 2010, pp. 131–136.

[22] ——, "Conarg: A constraint-based computational framework for argumentation systems," in *Proc. ICTAI*. IEEE, 2011, pp. 605–612.

[23] P. E. Dunne, A. Hunter, P. McBurney, S. Parsons, and M. Wooldridge, "Inconsistency tolerance in weighted argument systems," in *Proc. AAMAS*. IFAAMAS, 2009, pp. 851–858.

[24] S. Bistarelli, U. Montanari, and F. Rossi, "Semiring-based constraint satisfaction and optimization," *J. ACM*, vol. 44, no. 2, pp. 201–236, 1997.

[25] M. Correia, J. Cruz, and J. Leite, "On the Efficient Implementation of Social Abstract Argumentation," in *Proc. ECAI*, FAIA 263. IOS, 2014, pp. 225– 230.

[26] J. Feinberg, "Wordle," in *Beautiful Visualization: Looking at Data through the Eyes of Experts*, N. I. Julie Steele, Ed. O'Reilly, Apr. 2010.

[27] S. Bistarelli, F. Rossi, and F. Santini, "A first comparison of abstract argumentation systems: A computational perspective," in *Proc. ECAI*, FAIA 271. IOS, 2014, pp. 969 – 970.