Emotions in Argumentation: an Empirical Evaluation

Sahbi Benlamine, Maher Chaouachi, Serena Villata, Elena Cabrio, Claude Frasson, Fabien Gandon

Emotions in Argumentation: an Empirical Evaluation*

Sahbi Benlamine
Univ. de Montreal, Canada
benlamim@iro.umontreal.ca

Maher Chaouachi
Univ. de Montreal, Canada
chaouacm@iro.umontreal.ca

Serena Villata
INRIA, France
serena.villata@inria.fr

Elena Cabrio
INRIA, France
elena.cabrio@inria.fr

Claude Frasson
Univ. de Montreal, Canada
frasson@iro.umontreal.ca

Fabien Gandon
INRIA, France
fabien.gandon@inria.fr

Abstract

Argumentation is often seen as a mechanism to support different forms of reasoning such that decision-making and persuasion, but all these approaches assume a purely rational behavior of the involved actors. However, humans are proved to behave differently, mixing rational and emotional attitudes to guide their actions, and it has been claimed that there exists a strong connection between the argumentation process and the emotions felt by people involved in such process. In this paper, we assess this claim by means of an experiment: during several debates people's argumentation in plain English is connected and compared to the emotions automatically detected from the participants. Our results show a correspondence between emotions and argumentation elements, e.g., when in the argumentation two opposite opinions are conflicting this is reflected in a negative way on the debaters' emotions.

1 Introduction

The Web is becoming a hybrid space where men and machines interact. In this context, detecting and managing the emotional state of a user is important to allow artificial and human actors to adapt their reactions to others' emotional states. It is also a useful indicator for community managers, moderators and editors to help them in handling the communities and the content they produce. As a typical example, Wikipedia is managed by users and bots who constantly contribute, agree, disagree, debate and update the content of the encyclopedia. In this paper, we argue that in order to apply argumentation to scenarios like e-democracy and online debate systems, designers must take emotions into account. To efficiently manage and interact with such a hybrid society, we need to improve our means to understand and link the different dimensions of the exchanges (social interactions, textual content of the messages, dialogical structures of the interactions, emotional states of the participants, etc.). Beyond the challenges individually raised by each dimension, a key problem is to link these dimensions and their analysis together with the aim to detect, for instance, a debate turning into a flame war, a content reaching an agreement, a good or bad emotion spreading in a community.

In this paper, we aim to answer the following research question: *What is the connection between the arguments proposed by the participants of a debate and their emotional status?* Such question breaks down into the following subquestions: *(1)* is the polarity of arguments and the relations among them correlated with the polarity of the detected emotions?, and *(2)* what is the relation between the kind and the amount of arguments proposed in a debate, and the mental engagement detected among the participants of the debate?

To answer these questions, we propose an empirical evaluation of the connection between argumentation and emotions. This paper describes an experiment with human participants which studies the correspondences between the arguments and their relations put forward during a debate, and the emotions detected by emotions recognition systems in the debaters. We design an experiment where 12 debates are addressed by 4 participants each. Participants argue in plain English proposing arguments, that are in positive or negative relation with the arguments proposed by the other participants. During these debates, participants are equipped with emotions detection tools, recording their emotions. We hypothesize that negative relations among the arguments correspond to negative emotions felt by the participants proposing such arguments, and vice versa for the positive relation between arguments.

A key point in our work is that, up to our knowledge, no user experiment has been carried out yet to determine what is the connection between the argumentation addressed during a debate and the emotions emerging in the participants involved in such debate. An important result is the development of a publicly available dataset capturing several debate situations, and annotating them with their argumentation structure and the emotional states automatically detected.

The paper is organized as follows. In Section 2 we describe the two main components of our framework (namely bipolar argumentation and emotions detection systems), and Section 3 describes the experimental protocol and the research hypotheses. In Section 4 we analyze our experimental results, and Section 5 compares this work with the relevant literature.

*The authors acknowledge support of the SEEMPAD associate team project (http://project.inria.fr/seempad/).

2 The Framework

In this section, we present the two main components involved in our experimental framework: *i)* bipolar argumentation theory, i.e., the formalism used to analyze the textual arguments retrieved from the debates (Section 2.1), and *ii)* the systems used to detect the degrees of attention and engagement of each participant involved in the debate as well as her facial emotions (Section 2.2).

2.1 Argumentation

Argumentation is the process of creating arguments for and against competing claims [Rahwan and Simari, 2009]. What distinguishes argumentation-based discussions from other approaches is that opinions have to be supported by the arguments that justify, or oppose, them. This permits greater flexibility than in other decision-making and communication schemes since, for instance, it makes it possible to persuade other persons to change their view of a claim by identifying information or knowledge that is not being considered, or by introducing a new relevant factor in the middle of a negotiation, or to resolve an impasse.

Argumentation is the process by which arguments are constructed and handled. Thus argumentation means that arguments are compared, evaluated in some respect and judged in order to establish whether any of them are warranted. Roughly, each argument can be defined as a set of assumptions that, together with a conclusion, is obtained by a reasoning process. Argumentation as an exchange of pieces of information and reasoning about them involves groups of actors, human or artificial. We can assume that each argument has a proponent, the person who puts forward the argument, and an audience, the person who receives the argument. In our framework, we rely on abstract bipolar argumentation [Dung, 1995; Cayrol and Lagasquie-Schiex, 2013] where we do not distinguish the internal structure of the arguments (i.e., premises, conclusion), but we consider each argument proposed by the participants in the debate as a unique element, then analyzing the relation it has with the other pieces of information put forward in the debate. In particular, in bipolar argumentation two kinds of relations among arguments are considered: the *support relation*, i.e., a positive relation between arguments, and the *attack relation*, i.e., a negative relation between arguments.

2.2 Emotions Detection

Emotions play an important role in decision making [Quartz, 2009], creative thinking, inspiration, as well as concentration and motivation. During conversations, emotions are expressed by participants according to their beliefs and viewpoints with respect to the opinions put forward by the other participants. In the argumentation process, how they feel about the others' point of view influences their reply, being it a support or an attack. To capture these different emotional reactions through automatic emotions recognition systems, cameras and/or physiological sensors (e.g., EDA[1], EEG[2], EMG[3]) can be used. Many researches propose to apply multimodal techniques, i.e. to combine different media in order to improve the automatic emotions recognition [Calvo and D'Mello, 2010; Jraidi *et al.*, 2013]. For these reasons, in our experiments we have used webcams for facial expressions analysis relying on the FACEREADER 6.0 software[4], and physiological sensors (EEG) to assess and monitor users' cognitive states in real-time [Chaouachi *et al.*, 2010].

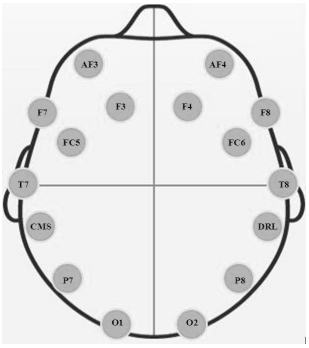

Figure 1: Emotiv Headset sensors/data channels placement.

Emotiv EPOC EEG headset. It has been used to record physiological data during the debate sessions. It contains 14 electrodes spatially organized according to International $10-20$ system[5], moist with a saline solution (contact lens cleaning solution). As shown in Figure 1, the electrodes are placed at AF3, AF4, F3, F4, F7, F8, FC5, FC6, P7, P8, T7, T8, O1, O2, and two other reference sensors are placed behind the ears. The generated data are in (μV) with a 128Hz sampling rate. The signal frequencies are between 0.2 and 60Hz.

Computing Engagement Index. The engagement index used in our study comes from the work of [Pope *et al.*, 1995] at the National Aeronautics and Space Administration (NASA). This work is based on neuroscientific research on attention and vigilance [Lubar, 1991]. It was found that the user's performance improved when this index is used as a criterion for switching between manual and automated piloting mode [Freeman *et al.*, 2000; Pope *et al.*, 1995]. Many studies showed the usefulness of integrating this index in many fields like eLearning and games to assess user's cognitive states. This index is computed from three EEG frequency bands: $\Theta(4-8Hz)$, $\alpha(8-13Hz)$ and $\beta(13-22Hz)$ as follows: $eng = \frac{\theta}{\alpha+\beta}$. This index is computed each second from the EEG signal. To reduce its fluctuation, we use a moving av-

[1]Electrodermal Activity: electrical changes measured at the surface of the skin.

[2]Electroencephalography: recording of electrical activity along the scalp.

[3]Electromyography: recording the electrical activity produced by skeletal muscles.

[4]`http://www.noldus.com/human-behavior-research/products/facereader`

[5]International $10-20$ system is an internationally recognized method to describe and apply the location of scalp electrodes in the context of an EEG test or experiment.

erage on a 40-second mobile window. Thus, the value of the index at time t corresponds to the total average of the ratios calculated on a period of 40 seconds preceding t. The extraction of the Θ, α and β frequency bands is performed by multiplying every second of the EEG signal by a Hamming window (to reduce the spectral leakage) and applying a Fast Fourier Transform (FFT). As the Emotiv headset measures 14 regions at the same time, we use a combined value of Θ, α and β frequency bands by summing their values over all the measured regions. To examine participants' engagement, we extract their minimum, average and maximum values during the debate, and we use such values to identify the range of engagement (High, Medium, Low) of every participant.

Facial expressions analysis. The primary emotion facial expressions [Ekman, 2005] from the user's emotional reactions are identified using real-time frame-by-frame analysis software FaceReader 6.0 via a webcam. FaceReader infers the emotional state by extracting and classifying in real-time 500 key points in facial muscles of the target face. These key points are provided as input to a neural network trained on a dataset of 10000 manually annotated images corresponding to these six basic emotions: *Happy*, *Sad*, *Angry*, *Surprised*, *Scared* and *Disgusted*. In addition to these emotions, the resulting file contains the *Valence*[6], and the *Arousal* of emotion as well as the probability of the *Neutral* state. In this study, to align the dominant emotion state occurring at every second, we compute the average (10 values/sec) for each column according to the camera frame rate.[7] We extract the most dominant emotion having the maximum value, pleased/unpleased valence depending on positive/negative values, and the active/inactive arousal by comparing the obtained values to 0.5.

3 The Experiment

This section details the experimental session we set up to analyze the relation between emotions and the argumentation process: we detail the protocol we have defined to guide the experimental setting (Section 3.1), and the resulting datasets (Section 3.2). Finally, we specify what are the hypotheses we aim at verifying in this experiment (Section 3.3).

3.1 Protocol

The general goal of the experimental session is to study the relation (if any) holding between the emotions detected in the participants and the argumentation flow. The idea is to associate arguments and the relations among them to the participants' mental engagement detected by the EEG headset and the facial emotions detected via the Face Emotion Recognition tool. More precisely, starting from an issue to be discussed provided by the moderators, the aim of the experiment is to collect the arguments proposed by the participants as well as the relations among them, and to associate such arguments/relations to the mental engagement states and to the facial emotions expressed by the participants. During a post-processing phase on the collected data, we synchronize the arguments and the relations put forward by the different participants at instant t with the emotional indexes we retrieved. Finally, we build the resulting bipolar argumentation graph for each debate, such that the resulting argumentation graphs are labelled with the source who has proposed each argument, and the emotional state of each participant at the time of the introduction of the argument in the discussion.

The first point to clarify in this experimental setting is the terminology. In this experiment, an *argument* is each single piece of text that is proposed by the participants of the debate. Typically, arguments have the goal to promote the opinion of the debater in the debate. Thus, an *opinion* in our setting represents the overall opinion of the debater about the issue to be debated, i.e., "Ban animal testing". The opinion is promoted in the debate through arguments, that will support (if the opinions converge) or attack (otherwise) the arguments proposed in the debate by the other participants.

The experiment involves two kinds of persons:

- *Participant*: she is expected to provide her own opinion about the issue of the debate proposed by the moderators, and to argue with the other participants in order to convince them (in case of initial disagreement) about the goodness of her viewpoint.[8]
- *Moderator*: she is expected to propose the initial issue to be discussed to the participants. In case of lack of active exchanges among the participants, the moderator is in charge of proposing pro and con arguments (with respect to the main issue) to reactivate the discussion.

The experimental setting of each debate is conceived as follows: there are 4 participants for each discussion group (each participant is placed far from the other participants, even if they are in the same room), and 2 moderators located in another room with respect to the participants. The moderators interact with the participants uniquely through the debate platform. The language used for debating is English. In order to provide an easy-to-use debate platform to the participants, without requiring from them any background knowledge, we decide to rely on a simple IRC network[9] as debate platform. The debate is anonymous and participants are visible to each others with their nicknames, e.g., *participant1*, while the moderators are visualized as *moderator1* and *moderator2*. Each participant has been provided with 1 laptop device equipped with internet access and a camera used to detect facial emotions. Moreover, each participant has been equipped with an EEG headset to detect engagement index. Each moderator used only a laptop.

The procedure we followed for each debate is:

- Participants' familiarization with the debate platform;
- The debate - participants take part into two debates each, about two different topics for a maximum of about 20 minutes each:

[6]The valence refers to the degree of pleasantness of an expressed emotion. A positive valence corresponds to an emotion with pleasant character and a negative valence to an unpleasant one.

[7]We used cameras of 10 frames/sec.

[8]Note that in this experimental scenario we do not evaluate the connection between emotions and persuasive argumentation. This analysis is left for future research.

[9]`http://webchat.freenode.net/`

  - The moderator(s) provides the debaters with the topic to be discussed;
  - The moderator(s) asks each participant to provide a general statement about his/her opinion concerning the topic;
  - Participants expose their opinion to the others;
  - Participants are asked to comment on the opinions expressed by the other participants;
  - If needed (no active debate among the participants), the moderator posts an argument and asks for comments from the participants;
- Debriefing: each participant is asked to complete a short questionnaire about his/her experience in the debate.[10]

The measured variables in the debate are: engagement (measurement tool: EEG headset), and the following emotions: Neutral, Happy, Sad, Angry, Surprised, Scared and Disgusted (measurement tool: FaceReader).

The post-processing phase of the experimental session involved *(i)* the detection of the support and attack relations among the arguments proposed in each discussion, following the methodology described in Section 3.2, and *(ii)* the synchronization of the argumentation (i.e., the arguments/relations proposed at time t) with the emotional indexes retrieved at time t using the EEG headset and FaceReader.

Participants. The experiment was distributed over 6 sessions of 4 participants each; the first session was discarded due to a technical problem while collecting data. We had a total of 20 participants (7 women, 13 men), whose age range was from 22 to 35 years. All of them were students in a North American university, and all of them had good computer skills. Since not all of them were native English speakers, the use of the Google translate service was allowed. They have all signed an ethical agreement before proceeding to the experiment.

3.2 Dataset

In this section we describe the dataset of textual arguments we have created from the debates among the participants. The dataset is composed of three main layers: *(i)* the basic annotation of the arguments proposed in each debate (i.e. the annotation in xml of the debate flow downloaded from the debate platform); *(ii)* the annotation of the relations of support and attack among the arguments; and *(iii)* starting from the basic annotation of the arguments, the annotation of each argument with the emotions felt by each participant involved in the debate.

The *basic* dataset is composed of 598 different arguments proposed by the participants in 12 different debates. The debated issues and the number of arguments for each debate are reported in Table 1. We selected the topics of the debates among the set of popular discussions addressed in online debate platforms like iDebate[11] and DebateGraph[12].

The annotation (in xml) of this dataset is as follows: we have assigned to each debate a unique numerical id, and for each argument proposed in the debate we assign an id and we annotate who was the participant putting this argument on the table, and in which time interval the argument has been proposed. An example of basic annotation is provided below:

```
<debate id="1" title="Ban_Animal_Testing">
<argument id="1" debate_id="1" participant="mod"
    time-from="19:26" time-to="19:27">Welcome to the
    first debate! The topic of the first debate is that
    animal testing should be banned.</argument>
<argument id="3" debate_id="1" participant="2"
    time-from="20:06" time-to="20:06">If we don't use
    animals in these testing, what could we use?</argument>
</debate>
```

The second level of our dataset consists in the annotation of arguments pairs with the relation holding between them, i.e., support or attack. To create the dataset, for each debate of our experiment we apply the following procedure, validated in [Cabrio and Villata, 2013]:

1. the main issue (i.e., the issue of the debate proposed by the moderator) is considered as the starting argument;
2. each opinion is extracted and considered as an argument;
3. since *attack* and *support* are binary relations, the arguments are coupled with:
   (a) the starting argument, or
   (b) other arguments in the same discussion to which the most recent argument refers (e.g., when an argument proposed by a certain user supports or attacks an argument previously expressed by another user);
4. the resulting pairs of arguments are then tagged with the appropriate relation, i.e., *attack* or *support*.

To show a step-by-step application of the procedure, let us consider the debated issue *Ban Animal Testing*. At step 1, we consider the issue of the debate proposed by the moderator as the starting argument (a):

(a) *The topic of the first debate is that animal testing should be banned.*

Then, at step 2, we extract all the users opinions concerning this issue (both pro and con), e.g., (b), (c) and (d):

(b) *I don't think the animal testing should be banned, but researchers should reduce the pain to the animal.*

(c) *I totally agree with that.*

(d) *I think that using animals for different kind of experience is the only way to test the accuracy of the method or drugs. I cannot see any difference between using animals for this kind of purpose and eating their meat.*

[10] Such material is available at `http://bit.ly/DebriefingData`

[11] `http://idebate.org/`

[12] `www.debategraph.org/`

(e) *Animals are not able to express the result of the medical treatment but humans can.*

At step 3a we couple the arguments (b) and (d) with the starting issue since they are directly linked with it, and at step 3b we couple argument (c) with argument (b), and argument (e) with argument (d) since they follow one another in the discussion. At step 4, the resulting pairs of arguments are then tagged by one annotator with the appropriate relation, i.e.: **(b)** *attacks* **(a)**, **(d)** *attacks* **(a)**, **(c)** *supports* **(b)** and **(e)** *attacks* **(d)**. For the purpose of validating our hypotheses, we decide to not annotate the supports/attacks between arguments proposed by the same participant (e.g., situations where participants are contradicting themselves). Note that this does not mean that we assumed that such situations do not arise: no restriction was imposed to the participants of the debates, so situations where a participant attacked/supported her own arguments are represented in our dataset. We just decided to not annotate such cases in the dataset of argument pairs, as it was not necessary for verifying our assumptions.

To assess the validity of the annotation task and the reliability of the obtained dataset, the same annotation task has been independently carried out also by a second annotator, so as to compute inter-annotator agreement. It has been calculated on a sample of 100 argument pairs (randomly extracted). The complete percentage agreement on the full sample amounts to 91%. The statistical measure usually used in NLP to calculate the inter-rater agreement for categorical items is Cohen's kappa coefficient [Carletta, 1996], that is generally thought to be a more robust measure than simple percent agreement calculation since κ takes into account the agreement occurring by chance. More specifically, Cohen's kappa measures the agreement between two raters who each classifies N items into C mutually exclusive categories. The equation for κ is:

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)} \quad (1)$$

where Pr(a) is the relative observed agreement among raters, and Pr(e) is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly saying each category. If the raters are in complete agreement then κ = 1. If there is no agreement among the raters other than what would be expected by chance (as defined by Pr(e)), κ = 0. For NLP tasks, the inter-annotator agreement is considered as significant when κ >0.6. Applying such formula to our data, the inter-annotator agreement results in $\kappa = 0.82$. As a rule of thumb, this is a satisfactory agreement, therefore we consider these annotated datasets as reliable (i.e., our *goldstandard* dataset where arguments are associated to participants' emotions detected by EEG/FaceReader) to be exploited during the experimental phase.

Table 1 reports on the number of arguments and pairs we extracted applying the methodology described before to all the mentioned topics. In total, our dataset contains 598 different arguments and 263 argument pairs (127 expressing the *support* relation among the involved arguments, and 136 expressing the *attack* relation among the involved arguments).

Dataset				
Topic	**#arg**	**#pair**	**#att**	**#sup**
BAN ANIMAL TESTING	49	28	18	10
GO NUCLEAR	40	24	15	9
HOUSEWIVES SHOULD BE PAID	42	18	11	7
RELIGION DOES MORE HARM THAN GOOD	46	23	11	12
ADVERTISING IS HARMFUL	71	16	6	10
BULLIES ARE LEGALLY RESPONSIBLE	71	12	3	9
DISTRIBUTE CONDOMS IN SCHOOLS	68	27	11	16
ENCOURAGE FEWER PEOPLE TO GO TO THE UNIVERSITY	55	14	7	7
FEAR GOVERNMENT POWER OVER INTERNET	41	32	18	14
BAN PARTIAL BIRTH ABORTIONS	41	26	15	11
USE RACIAL PROFILING FOR AIRPORT SECURITY	31	10	1	9
CANNABIS SHOULD BE LEGALIZED	43	33	20	13
TOTAL	**598**	**263**	**136**	**127**

Table 1: The textual dataset of the experiment.

The final dataset adds to all previously annotated information the player characteristics (gender, age and personality type), FaceReader data (dominant emotion, Valence (pleasant/unpleased) and Arousal (activated/ inactivated)), and EEG data (Mental Engagement levels)[13]. A correlation matrix has been generated to identify the correlations between arguments and emotions in the debates, and a data analysis is performed to determine the proportions of emotions for all participants. We consider the obtained dataset as representative of human debates in a non-controlled setting, and therefore we consider it as the reference dataset to carry out our empirical study.

An example, from the debate about the topic "Religion does more harm than good" where arguments are annotated with emotions (i.e., the third layer of the annotation of the textual arguments we retrieved), is as follows:

```
<argument id="30" debate_id="4" participant="4"
time-from="20:43" time-to="20:43"
emotion_p1="neutral" emotion_p2="neutral"
emotion_p3="neutral" emotion_p4="neutral">
Indeed but there exist some advocates of the devil
like Bernard Levi who is decomposing arabic countries.
</argument>
<argument id="31" debate_id="4" participant="1"
time-from="20:43" time-to="20:43"
emotion_p1="angry" emotion_p2="neutral"
emotion_p3="angry" emotion_p4="disgusted">
I don't totally agree with you Participant2: science
and religion don't explain each other, they tend to
explain the world but in two different ways.
</argument>
<argument id="32" debate_id="4" participant="3"
time-from="20:44" time-to="20:44"
emotion_p1="angry" emotion_p2="happy"
emotion_p3="surprised" emotion_p4="angry">
Participant4: for recent wars ok but what about wars
happened 3 or 4 centuries ago?
</argument>
```

3.3 Hypotheses

This experiment aims to verify the link between the emotions detected on the participants of the debate, and the arguments and their relations proposed in the debate. Our hypotheses therefore revolve around the assumption that the participants'

[13]The datasets of textual arguments are available at `http://bit.ly/TextualArgumentsDataset`.

emotions arise out of the arguments they propose in the debate:

H1 : There are some emotional and behavioral trends that can be extracted from a set of debates.

H2 : The number and the strength of arguments, attacks and supports exchanged between the debaters are correlated with particular emotions throughout the debates.

H3 : The number of expressed arguments is connected to the degree of mental engagement and social interactions.

4 Results

In order to verify these hypotheses, we first computed the mean percentage of appearance of each basic emotion across the 20 participants. Results show (with 95% confidence interval) that the most frequent emotion expressed by participants was *anger*, with a mean appearance frequency ranging from 8.15% to 15.6% of the times. The second most frequent emotion was another negative emotion, namely *disgust*, which was present 7.52% to 14.8% of the times. The overall appearance frequency of other emotions was very low. For example, the frequency of appearance of happiness was below 1%. Even if this result might be surprising at a first glance, this trend can be justified by a phenomenon called *negativity effect* [Rozin and Royzman, 2001]. This means that negative emotions have generally more impact on a person's behavior and cognition than positive ones. So, negative emotions like anger and disgust have a tendency to last in time more than positive emotions like happiness.

With regard to the mental engagement, participants show in general a high level of attention and vigilance in 70.2% to 87.7% of the times. This high level of engagement is also correlated with appearance of anger (r=0.306), where r refers to the Pearson product-moment correlation coefficient. It is a standard measure of the linear correlation between two variables X and Y, giving a value between $[1, -1]$, where 1 is a total positive correlation, 0 means no correlation, and -1 is a total negative correlation. This trend confirms that, in such context, participants may be thwarted by the other participants' arguments or attacks, thus the level of engagement tends to be high as more attention is allowed to evaluate the other arguments or to formulate rebutting or offensive arguments. Thus, our experiments confirm behavioral trends as expected by the first hypothesis.

Figure 2 shows an evolution of the first participant's emotions at the beginning of the first debate. The most significant lines of emotions are surprise and disgust (respectively, the line with squares and the line with circles). The participant is initially surprised by the discussion (and so mentally engaged) and then, after the debate starts, this surprise switches suddenly into disgust, due to the impact of the rejection of one of her arguments; the bottom line with circles grows and replaces the surprise as the participant is now actively engaged in an opposed argument (thus confirming our hypothesis 2). Finally, the participant is calming down. In this line, Figure 3 highlights that we have a strong correlation (r= 0.83) in Session 2 showing that the number of attacks provided in the debate increased linearly with the manifestation of more disgust emotion.

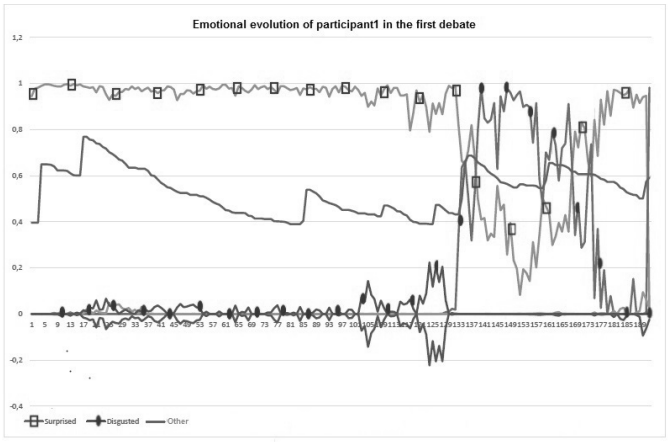

Figure 2: Emotional evolution of participant 1 in debate 1.

	NB ARG	ATTACK	SUPPORT
Pleasant	0,0962	0,1328	-0,0332
Unpleasant	-0,0962	-0,1328	0,0332
High ENG	-0,0718	**-0,6705**	0,2459
LowENG	-0,2448	0,2115	-0,1063
Neutral	0,0378	0,6173	-0,1138
Disgusted	-0,0580	**-0,4367**	-0,3621
Scared	0,1396	-0,0952	**0,5755**
Angry	-0,1018	-0,4386	0,0582

Figure 3: Correlation table for Session 2 (debated topics: *Advertising is harmful* and *Bullies are legally responsible*.

In the second part of our study, we were interested in analyzing how emotions correlate with the number of attacks, supports and arguments. We have generated a correlation matrix to identify the existent correlations between arguments and emotions in debates. Main results show that the number of arguments tends to decrease linearly with manifestations of sadness (r=-0.25). So when the participants start to feel unpleasant emotions, such as sadness, the number of arguments decreases, showing a less positive social behavior[14] and a tendency to retreat into herself. This negative correlation between the number of arguments and sadness even reaches very high level in certain debates (i.e. a mean correlation r= -0.70 is registered in the two debates of the second session). Another negative linear relationship is registered with regard to the number of attacks and the anger expressed by the participant (r=0.22). Participants who tend to attack the others in the debate are less angry than those whose number of attacks is smaller. Figure 4 shows the correlation table for Session 3. The analysis of the results we obtained shows the occurrence of strong correlations between emotions and attacks / media / number of arguments in some discussions, but not in others. This is an interesting index to investigate in future work.

Figure 5 shows the most significant correlations we de-

[14]By positive social behavior, we mean that a participant aims at sharing her arguments with the other participants. This attitude is mitigated if unpleasant emotions start to be felt by the participant.

	NB ARG	ATTACK	SUPPORT
Pleasant	**0,7067**	-0,3383	-0,3800
Unpleasant	-0,7067	0,3383	0,3800
High ENG	**-0,6903**	-0,3699	-0,1117
LowENG	-0,1705	**0,5337**	-0,0615
Neutral	0,8887	-0,0895	-0,3739
Disgusted	0,1017	**0,8379**	**0,5227**
Scared	0,2606	-0,4132	**-0,7107**
Angry	**-0,7384**	-0,5072	-0,0937

Figure 4: Correlation table for Session 3 (debated topics: *Distribute condoms at schools* and *Encourage fewer people to go to the university*).

	NB ARG	ATTACK	SUPPORT
Pleasant	**0,1534**	0,0134	-0,0493
Unpleasant	-0,1534	-0,0134	0,0493
High ENG	-0,0246	-0,0437	**0,3185**
LowENG	**0,2054**	0,1147	0,1592
Neutral	0,0505	0,1221	-0,2542
Disgusted	-0,0177	-0,0240	**0,2996**
Scared	-0,0278	0,0297	-0,2358
Angry	0,0344	**-0,2206**	0,0782

Figure 5: General correlation table of the results.

tected. For instance, the number of supports provided in the debate increased linearly with the manifestation of high levels of mental engagement (r=0.31). This trend is more pronounced when the debate does not trigger controversies and conflicts between the participants. For example, in the debate *Encourage fewer people to go to university*, all the participants shared the same opinion (against the main issue as formulated by the moderator) and engaged to support each other's arguments. The correlation between the number of supports and the engagement was very high (r=0.80) in this debate. The number of attacks is more related to low engagement. The moderator can provide more supporting arguments to balance participants' engagement, and if the attacks are increasing, that means participants tend to disengage. The experiments show that participants maintaining high levels of vigilance are the most participative in the debate and resulted in a more positive social behavior (thus confirming our hypothesis 3).

5 Related Work

[Cerutti *et al.*, 2014] propose an empirical experiment with humans in the argumentation theory area. However, the goal of this work is different from our one, emotions are not considered and their aim is to show a correspondence between the acceptability of arguments by human subjects and the acceptability prescribed by the formal theory in argumentation. [Rahwan *et al.*, 2010] study whether the meaning assigned to the notion of *reinstatement* in abstract argumentation theory is perceived in the same way by humans. They propose to the participants of the experiment a number of natural language texts where reinstatement holds, and then asked them to evaluate the arguments. Also in this case, the purpose of the work differs from our one, and emotions are not considered at all.

Emotions are considered, instead, by [Nawwab *et al.*, 2010] that propose to couple the model of emotions introduced by [Ortony *et al.*, 1988] in an argumentation-based decision making scenario. They show how emotions, e.g., gratitude and displeasure, impact on the practical reasoning mechanisms. A similar work has been proposed by [Dalibón *et al.*, 2012] where emotions are exploited by agents to produce a line of reasoning according to the evolution of its own emotional state. Finally, [Lloyd-Kelly and Wyner, 2011] propose emotional argumentation schemes to capture forms of reasoning involving emotions. All these works differ from our approach since they do not address an empirical evaluation of their models, and emotions are not detected from humans.

Several works in philosophy and linguistics have studied the link between emotions and natural argumentation, like [Carofiglio and de Rosis, 2003; Gilbert, 1995; Walton, 1992]. These works analyze the connection of emotions and the different kind of argumentation that can be addressed. The difference with our approach is that they do not verify their theories empirically, on actual emotions extracted from people involved in an argumentation task. A particularly interesting case is that of the connection between persuasive argumentation and emotions, studied for instance by [DeSteno *et al.*, 2004].

6 Conclusions

In this paper, we presented an investigation into the link between the argumentation people address when they debate with each other, and the emotions they feel during these debates. We conducted an experiment aimed at verifying our hypotheses about the correlation between the positive/negative emotions emerging when positive/negative relations among the arguments are put forward in the debate. The results suggest that there exist clear trends that can be extracted from emotions analysis. Moreover, we also provide the first open dataset and gold standard to compare and analyze emotion detection in an argumentation session.

Several lines of research have to be considered as future work. First, we intend to study the link between emotions and persuasive argumentation. This issue has already been tackled in a number of works in the literature (e.g., [DeSteno *et al.*, 2004]), but no empirical evaluation has been addressed yet. Second, we aim to study how emotion persistence influence the attitude of the debates: this kind of experiment has to be repeated a number of times in order to verify whether positive/negative emotions before the debate influence the new interactions. Third, we plan to add a further step, namely to study how sentiment analysis techniques are able to automatically detect the polarity of the arguments proposed by the debaters, and how they are correlated with the detected emotions. Moreover, we plan to study emotions propagation among the debaters, and to verify whether the emotion can be seen as a predictor of the solidity of an argument, e.g., if I write an argument when I am angry I may make wrong judgments.

References

[Cabrio and Villata, 2013] Elena Cabrio and Serena Villata. A natural language bipolar argumentation approach to support users in online debate interactions. *Argument & Computation*, 4(3):209–230, 2013.

[Calvo and D'Mello, 2010] Rafael A. Calvo and Sidney D'Mello. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing*, 1(1):18–37, 2010.

[Carletta, 1996] Jean Carletta. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2):249–254, 1996.

[Carofiglio and de Rosis, 2003] Valeria Carofiglio and F. de Rosis. Combining logical with emotional reasoning in natural argumentation. In Conati C, Hudlicka E, and editors Lisetti C, editors, *9th International Conference on User Modeling. Workshop Proceedings*, page 915, 2003.

[Cayrol and Lagasquie-Schiex, 2013] Claudette Cayrol and Marie-Christine Lagasquie-Schiex. Bipolarity in argumentation graphs: Towards a better understanding. *Int. J. Approx. Reasoning*, 54(7):876–899, 2013.

[Cerutti *et al.*, 2014] Federico Cerutti, Nava Tintarev, and Nir Oren. Formal arguments, preferences, and natural language interfaces to humans: an empirical evaluation. In *ECAI 2014 - 21st European Conference on Artificial Intelligence*, pages 207–212, 2014.

[Chaouachi *et al.*, 2010] Maher Chaouachi, Pierre Chalfoun, Imène Jraidi, and Claude Frasson. Affect and mental engagement: towards adaptability for intelligent systems. In *Proceedings of the 23rd International FLAIRS Conference, Daytona Beach, FL.*, 2010.

[Dalibón *et al.*, 2012] Santiago Emanuel Fulladoza Dalibón, Diego César Martinez, and Guillermo Ricardo Simari. Emotion-directed argument awareness for autonomous agent reasoning. *Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial*, 15(50):30–45, 2012.

[DeSteno *et al.*, 2004] David DeSteno, Duane T. Wegener, Richard E. Petty, Derek D. Rucker, and Julia Braverman. Discrete emotions and persuasion: The role of emotion-induced expectancies. *Journal of Personality and Social Psychology*, 86:4356, 2004.

[Dung, 1995] Phan Minh Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artif. Intell.*, 77(2):321–358, 1995.

[Ekman, 2005] Paul Ekman. *Basic Emotions*, pages 45–60. John Wiley & Sons Ltd, 2005.

[Freeman *et al.*, 2000] Frederick G. Freeman, Peter J. Mikulka, Mark W. Scerbo, Lawrence J. Prinzel, and Keith Clouatre. Evaluation of a psychophysiologically controlled adaptive automation system, using performance on a tracking task. *Applied Psychophysiology and Biofeedback*, 25(2):103–115, 2000.

[Gilbert, 1995] Michael A. Gilbert. Emotional argumentation, or, why do argumentation theorists argue with their mates? In F.H. van Eemeren, R. Grootendorst, J.A. Blair, and C.A. Willard, editors, *Proceedings of the Third ISSA Conference on Argumentation*, volume II, 1995.

[Jraidi *et al.*, 2013] Imène Jraidi, Maher Chaouachi, and Claude Frasson. A dynamic multimodal approach for assessing learner's interaction experience. In *Proceedings of the 15th ACM on International Conference on multimodal interaction*, pages 271–278. ACM, 2013.

[Lloyd-Kelly and Wyner, 2011] Martyn Lloyd-Kelly and Adam Wyner. Arguing about emotion. In *Advances in User Modeling - UMAP 2011 Workshops*, pages 355–367, 2011.

[Lubar, 1991] Joel F. Lubar. Discourse on the development of eeg diagnostics and biofeedback for attention-deficit/hyperactivity disorders. *Biofeedback and Self-regulation*, 16(3):201–225, 1991.

[Nawwab *et al.*, 2010] Fahd Saud Nawwab, Paul E. Dunne, and Trevor J. M. Bench-Capon. Exploring the role of emotions in rational decision making. In *Computational Models of Argument: Proceedings of COMMA 2010*, pages 367–378, 2010.

[Ortony *et al.*, 1988] A. Ortony, G. Clore, and A. Collins. *The Cognitive Structure of Emotions*. Cambridge University Press, Cambridge, 1988.

[Pope *et al.*, 1995] Alan T. Pope, Edward H. Bogart, and Debbie S. Bartolome. Biocybernetic system evaluates indices of operator engagement in automated task. *Biological psychology*, 40(1):187–195, 1995.

[Quartz, 2009] Steven R. Quartz. Reason, emotion and decision-making: risk and reward computation with feeling. *Trends in cognitive sciences*, 13:209–215, 2009.

[Rahwan and Simari, 2009] Ihad Rahwan and Guillermo Simari, editors. *Argumentation in Artificial Intelligence*. Springer, 2009.

[Rahwan *et al.*, 2010] Iyad Rahwan, Mohammed Iqbal Madakkatel, Jean-François Bonnefon, Ruqiyabi Naz Awan, and Sherief Abdallah. Behavioral experiments for assessing the abstract argumentation semantics of reinstatement. *Cognitive Science*, 34(8):1483–1502, 2010.

[Rozin and Royzman, 2001] Paul Rozin and Edward B. Royzman. Negativity bias, negativity dominance, and contagion. *Personality and social psychology review*, 5(4):296–320, 2001.

[Walton, 1992] Douglas Walton. *The Place of Emotion in Argument*. Pennsylvania State University Press, University Park, 1992.