Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)

Argument Mining from Speech: Detecting Claims in Political Debates

Marco Lippi
DISI, University of Bologna, Italy
marco.lippi3@unibo.it

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

Abstract

The automatic extraction of arguments from text, also known as argument mining, has recently become a hot topic in artificial intelligence. Current research has only focused on linguistic analysis. However, in many domains where communication may be also vocal or visual, paralinguistic features too may contribute to the transmission of the message that arguments intend to convey. For example, in political debates a crucial role is played by speech. The research question we address in this work is whether in such domains one can improve claim detection for argument mining, by employing features from text and speech in combination. To explore this hypothesis, we develop a machine learning classifier and train it on an original dataset based on the 2015 UK political elections debate.

Introduction

Context

The studies conducted in the late Sixties under Merhabian's lead, summarized in (Mehrabian 1981), became extremely popular because they were the first attempt to quantify the impact of nonverbal communication. In particular, his team run a number of experiments designed to measure the relative importance of verbal and nonverbal messages when a communicator is talking about their feelings and attitudes. While his famous equation *Total Liking = 7% Verbal Liking + 38% Vocal Liking + 55% Facial Liking* has often been and still is subject to misquotes and unwarranted generalizations, the fact remains that non-verbal elements are particularly important for communicating feelings and attitude. In particular, voice quality is an essential part of prosody, it serves as a strong indicator of the affective states of the speaker, and is perhaps the most strongly recognised feature of paralinguistic speech, albeit subconsciously (Campbell 2007).

The context of the present work is not psychology but argumentation, and in particular argument mining. This is a booming area in artificial intelligence, where the combined efforts of experts in machine learning, computational linguistics and argumentation are producing innovative methods for detecting and processing argumentative elements—such as claims and premises—in natural language, with promising results (Levy et al. 2014; Lippi and Torroni 2015b). All the focus so far has been on written text: to the best of our knowledge, no attempt has ever been made at detecting arguments from spoken language. However, thanks also to breakthrough advancements in deep learning, speech recognition technology has reached commercial-level maturity and has become truly pervasive, due also to services such as Siri, Google Now and Cortana. It seems sensible then to consider speech as a natural—albeit noisy[1]—input for argument detection systems, on par with text.

Motivation

The hypothesis that motivates our study is that vocal features of speech can improve argument mining. This hypothesis is inspired by results in nonverbal communication but also by recent studies on the connections between the argumentation process and the emotions felt by the people involved in such process. Benlamine et al. (2015) show, through an experiment with human participants, that there is indeed a correlation between number and strength of arguments exchanged between debaters, and emotions such as anger. At the same time, we know that emotional states can be recognized from speech with reasonable accuracy (El Ayadi, Kamel, and Karray 2011). All this leads us to believe that a correlation between vocal features of speech and arguments should be investigated.

We elected political debate as a domain of choice. We based our study on an original corpus constructed from material available online from the 7-party leader's debate of April 2, 2015 which preceded the last UK general elections.[2] Our choice of that particular event was not only motivated by the availability of input data, but also by the good quality of the speech audio signal (the debate was moderated and subject to strict rules) and the potential outreach to areas such as political communication and media studies. Indeed, televised debates are becoming increasingly popular also thanks to the impact they have upon first-time voters (Coleman 2011). Political discourse on the other hand has been a long standing domain of interest for argumentation

[1] According to Apple, currently Siri has a 5% word error rate. Source: VentureBeat's coverage of WDC 2015, San Francisco, June 2015. http://venturebeat.com

[2] https://www.youtube.com/watch?v=7Sv2AOQBd_s

studies (Fairclough and Fairclough 2012).

Objectives

The main objectives of this paper are threefold. First, we want to present the design and implementation of the first system that can detect claims from speech. Second, we want to discuss the results of our study on the use of audio features for text-based argument mining. These results go in the direction of corroborating our starting hypothesis: vocal features of speech can indeed improve argument mining. Finally, we wish to present a third valuable result: the development of a publicly available dataset, the first of its kind, which can be used to further research on argument mining from speech.

Organization

We first introduce argumentation mining to the non expert reader. We then discuss design, methods and architecture of a system for claim detection from speech. Next, we describe the dataset and motivate the choices we made for its creation. The final sections are devoted to experimental evaluation and discussion of results and future directions.

Argumentation Mining

The ever growing availability of data, together with tremendous advances in computational linguistics and machine learning, created fertile ground for the rise of a new area of research called *argumentation* (or *argument*) *mining*. The main goal of argumentation mining is to automatically extract arguments from generic textual corpora, in order to provide structured data for computational models of arguments and reasoning engines.

Argument models

Argumentation literature is rich with argument representation models; however, the most widely used model in argumentation mining is a simple *claim/premise* model. This model identifies an argument with a subjective, possibly controversial statement (claim) supported by other text segments (premises), which may describe facts grounded in reality, statistics, results of studies, expert opinions, etc. Other influential argument models are Toulmin's model (1958), IBIS (Kunz and Rittel 1970) and Freeman's (1991). However, their representational fit for practical use in diverse genres is a matter of discussion (Newman and Marshall 1991; Habernal, Eckle-Kohler, and Gurevych 2014).

Corpora

Argumentation mining started with seminal work in the legal domain (Teufel 1999; Mochales Palau and Moens 2011) but it boomed only in the last few years thanks to advances in enabling technologies such as machine learning and natural language processing, and especially to the availability of carefully annotated corpora. Some of the most significant ones to date are the AIFdb hosted by the Centre for Argument Technology at the University of Dundee,[3] the NoDE benchmark data base (Cabrio and Villata 2014) which gathers arguments collected from Debatepedia,[4] ProCon,[5] and other debate-oriented resources, two datasets consisting in 285 LiveJournal blogposts and 51 Wikipedia discussion forums developed by Rosenthal and McKeown (2012) for extracting opinionated claims, the persuasive essays corpus developed by Habernal et al. (2014), and finally the largest available dataset for claim detection, developed at IBM Research from a document set composed of 547 Wikipedia articles (Rinott et al. 2015). The IBM corpus contains 6,984 argument elements categorized into *context-dependent* claims and evidence facts (the "premises"), each relevant to one of 58 given topics. For its development, the IBM team identified an ontology whereby a *context-dependent claim* is "a general concise statement which directly supports or contests the topic," and *context-dependent evidence* is "a text segment which directly supports the claim in the context of a given topic." Aharoni et al. (2014) also define a taxonomy of evidence, where the three types of evidence are: *Study* (results of a quantitative analysis of data given as numbers or as conclusions), *Expert* (testimony by a person/group/committee/organization with some known expertise in or authority on the topic), and *Anecdotal* (a description of specific event instances or concrete examples). We built on this taxonomy to define a model for arguments used in televised political debates.

Inter-annotator agreement

One aspect of these corpora that deserves a brief discussion is the reliability of annotations. That is usually measured with a Kappa test that returns an inter-annotator agreement $\kappa = \frac{p_o - p_e}{1 - p_e}$, where p_o is the proportion of units in which the annotators agreed, and p_e is the proportion of units for which agreement is expected by chance (Cohen 1960). Alongside Cohen's κ there are other measures, such as Krippendorf's α_U (Krippendorff 2003). Habernal et al. (2014) offer a thorough study on argument annotation and show that, even with a simple claim-premise argument model, inter-annotator agreement is easily non-substantial or moderate. For example they report for (Rosenthal and McKeown 2012)'s corpus a κ of 0.50-0.57, and for an annotation study of their own based on 80 documents an α_U in the range 34.6-42.5. Most available datasets do not report any indication of agreement. In any case, the moderate level of agreement reflects a well-known issue in argumentation mining: argument analysis is seldom uncontroversial, and deciding how to annotate text is often matter of discussion even for human experts. That is one of the reasons why argumentation mining and its satellite sub-tasks, such as claim detection, are such challenging tasks.

Methods used in argumentation mining

Mainstream approaches to claim detection in textual documents employ methods at the intersection of machine learning and natural language processing. Most of them address the claim mining task as a sentence classification problem,

[3] http://corpora.aifdb.org

[4] http://www.debatepedia.com

[5] http://www.procon.org

where the goal is to predict whether a sentence contains a claim. To this aim, a machine learning classifier is usually employed, which takes a representation of the sentence (for example, in the form of some set of features) and produces a binary value that predicts whether the sentence is classified as positive (containing a claim) or not. Sometimes a score can be the output of the classifier, representing the confidence of the prediction. Among the feature sets used so far to represent sentences within this task, we can mention classic approaches from the field of text categorization, such as bag-of-words, part-of-speech tags, and their bigrams and trigrams (Mochales Palau and Moens 2011; Stab and Gurevych 2014), as well as highly sophisticated features such as sentiment and subjectivity scores or information from ontologies and thesauri (Levy et al. 2014). Among the methods used in argumentation mining we mention structured kernel machines for evidence/claim detection (Rooney, Wang, and Browne 2012; Lippi and Torroni 2015b), conditional random fields to segment the boundaries of each argument component (Goudas et al. 2014; Sardianos et al. 2015; Park, Katiyar, and Yang 2015), recursive neural networks to build representation of words (Sardianos et al. 2015), context-free grammar to predict relations between argumentative entities (Mochales Palau and Moens 2011), binary SVM classifiers to predict links in a claim/premise model (Stab and Gurevych 2014), naïve Bayes classifiers for evidence detection and for the prediction of links between evidence and claims (Biran and Rambow 2011), textual entailment to analyze inter-argument relations (Cabrio and Villata 2012), and many more. A brief survey of machine learning technologies used in argumentation mining applications is offered by Lippi and Torroni (2015a).

Using speech

We now discuss the unique challenges and opportunities posed by claim mining from speech, and the way we address this task.

Challenges and opportunities

Claim detection from text is by itself an extremely complex task. Clearly, moving to the domain of spoken language poses a number of additional challenges. This is because in the first place one needs to employ a speech recognition system, in order to automatically translate the audio signal into textual documents. Even using state-of-the-art systems, this step is bound to introduce errors. These may range from misinterpreting single words, to wrongly recognizing whole portions of sentences, to even producing a syntactically incorrect output, which could significantly distort the original meaning of the phrase. Such errors will most likely decrease the effectiveness of some existing approaches to claim detection, and they may seriously hinder usability of the methods that are not meant to be used for constructing features on noisy inputs. This is the case, for example, with methods based on the constituency parse trees of the sentences. An effect of noise could be that in some cases the parse tree may not exist, or it may be completely misleading. Even approaches that use external classifiers, as in the case of sentiment or subjectivity estimators, are likely to be affected by similar issues.

Yet, in the case of spoken language input data, additional useful information comes from the audio signal itself. Within this context, the task of automatically detecting claims in a spoken sentence can be seen as an instance of semantic information extraction from acoustic sources. Such a task shares similarities with spoken language understanding (Tur and De Mori 2011), sound analysis and classification (Istrate et al. 2006), and especially emotion recognition (El Ayadi, Kamel, and Karray 2011). Recently, the research fields of audio classification and retrieval, speech recognition, acoustic modeling, have seen significant advancements, due to the great impact of deep learning (e.g., see (LeCun, Bengio, and Hinton 2015) and references therein). Nevertheless, most of deep learning works have focused on classic tasks such as speaker identification, speaker gender recognition, music genre and artist classification (Lee et al. 2009). Some steps have also been made recently towards spoken language understanding, notably for the task of slot-filling (Mesnil et al. 2013), but not yet for problems related to semantic information extraction from audio.

Speech technologies for claim detection

The work by El Ayadi et al. (2011) presents an extensive review on the existing approaches to emotion recognition in speech, with a detailed study of different kinds of features that have been used within this task.

In this work we employ *Mel-frequency cepstral coefficients* (MFCCs), a classic feature set representing the spectral components of the audio signal. MFCCs have been used, with success, for a variety of tasks in audio classification and retrieval (Tzanetakis and Cook 2002; Guo and Li 2003; Grosse et al. 2007), among which also emotion recognition (Casale et al. 2008). The computation of MFCCs starts from the absolute value of the Short-Time Fourier Transform of the audio signal, from which a bank of filters is extracted on a Mel-frequency scale (i.e., with bands equally spaced). Then, the Discrete Cosine Transform of the log-energy of each filter is computed, and the first N coefficients of the DCT represent the MFCCs of the input signal. Clearly, input signals of different lengths are represented by MFCC vectors of different sizes. In order to have fixed-size feature vectors, typically some statistics of the original MFCC vectors are computed, such as the minimum and maximum values, the average and the standard deviation. Any machine learning classifier can thus be trained to recognize sentences containing claims, by taking as input text- and/or audio-based features.

System architecture

Our claim detection system thus implements a pipeline like the one depicted in Figure 1. The audio sequence is first fed as input to a speech recognition system, in order to extract

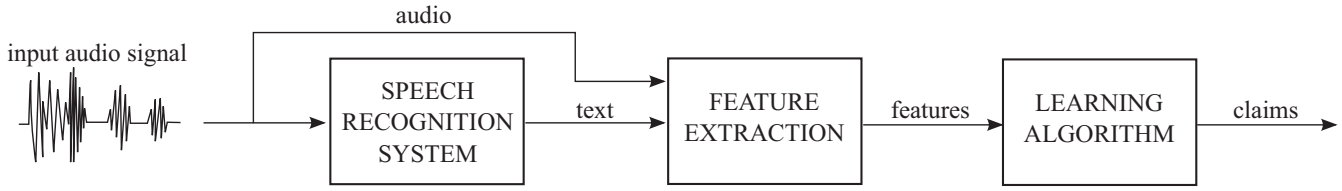

Figure 1: Pipeline of our claim detection system from audio.

the text.[6] Then, a feature extraction module processes both the recognized text and the acoustic signal, so as to produce the features to be exploited by a machine learning classifier. A model is trained using a supervised dataset (see next section). The model can be then used to detect claims in new examples.

Dataset

Dataset creation was a nontrivial and crucial process, which we describe in this section.

Collection of resources and annotation process

We selected the two-hour debate aired by Sky News on April 2, 2015 with all seven candidates to UK Prime Ministerial elections,[7] and we extracted the audio sequences for three of them: David Cameron, Nick Clegg, and Ed Miliband. We pre-processed the data so as to cut the audio sequences into frames representing self-contained sentences. We also cut off those portions of the audio signals where there was some overlapping between different speakers, which sometimes happens during the debate sessions between the candidates. We thus collected 386 audio samples: 122 for David Cameron, 104 for Nick Clegg, and 160 for Ed Miliband. The samples are, clearly, all of different length.

After collecting the raw data, a transcript was produced which represents our *ground truth*. The transcript includes 9,666 words in total: 3,469 spoken by Cameron, 2,849 by Clegg, and 3,348 by Miliband. An annotation process was then carried out, in order to label each audio fragment as containing a claim (C) or not (N). The annotation process was performed by two annotators who have a good expertise of annotation styles for argument mining. An initial ontology of argumentative elements was created to guide the annotators. The annotators then discussed their individual analyses in order to provide a single, stable analysis they both agree with. The discussions also helped to enhance the ontology, which became stable after some iterations.

Ontology

The final ontology is an adaptation of the one proposed by Aharoni et al. (2014) for the construction of the IBM corpus which was introduced earlier. It defines two types of argumentative elements: evidence and claim.

A *claim* is the conclusion of an argument, and it is represented by a statement which asserts a sort of opinion, or the thesis that is the result of a reasoning process. For example, the sentence "when you have an economy with out-of-control debt, out-of-control welfare, out-of-control spending, young people suffer the most" is a claim, whereas "let me answer Terry very directly about the things we can really do to make a difference with our NHS" is not, since it does not assert any thesis or opinion.[8]

Evidence is a segment of text that directly supports a claim. For example, the sentence "[The NHS] needs 8 billion pounds by the end of the next Parliament, that is what a man called Simon Stevens, who runs NHS England, is independent of politics, has said it needs" is evidence, while "we've been talking about the difficult decisions we had to make to turn the economy around" is neither evidence, nor claim.

Building on (Aharoni et al. 2014), we further specialized claims and evidence into sub-categories, according to their type. We extended their definition of evidence types—which had not been devised for annotating political debates in the first place, but for Wikipedia pages—and introduced a fourth type of evidence, while at the same time we defined a simple taxonomy of claims. Although the model we used for claim detection is unaware of these types—we restricted our study to plain claim detection, as it is done, for instance, by Levy et al. (2014)—defining this ontology was key to achieving an acceptable inter-annotator agreement (see next paragraph).

Evidence can be of the following types: **Study**: reports basic facts grounded on reality, mathematical truths, historical events, or results of quantitative/statistical analyses of data; **Expert**: reports testimony by a person/group/committee/organization with some known expertise in or authority on the topic; **Anecdotal**: describes specific events/instances (evidence from anecdotes) or concrete examples; and **Facevalue**: promises, confessions, mind reading, things that are presented as true simply because they are uttered, without providing any objective support.

Anecdotal and *Expert* are as in (Aharoni et al. 2014). An example of *Expert* evidence is the sentence we reported above, citing Stevens as an expert independent source. We extended *Study* in order to include, for instance, facts grounded in reality or reported events. To illustrate, "last year we had the fastest growing economy of any of the major Western countries," or "we went ahead and invested in

[6] In a fully automated pipeline, the first stage should include a sentence segmentation tool (Liu et al. 2006). We do not cover this aspect in the present work.

[7] https://www.youtube.com/watch?v=7Sv2AOQBd_s

[8] All the examples in this paper are taken from the 7-party debate. The entire corpus with annotated claims can be downloaded from http://argumentationmining.disi.unibo.it

our NHS as part of a balance plan for our country" fall in this category. We had to introduce a fourth category, because very frequently the arguments used by party leaders in this debate used some form of opponent mind-reading or reference to future events, in an effort to substantiate their claims, as in "we will have common sense spending reductions [so outside key areas like education and health spending will fall]," or "[here's what Ed Miliband isn't telling you because] he doesn't support any of the spending reductions and efficiencies we've had to make. He wants to make a very big cut. He wants to put up taxes and cut your pay going into your monthly payslip at the end of the month and taking your money out, because he thinks he can spend that money better than you." We named this fourth type of evidence *Face-value*, because the one thing these segments of text have in common is that they require an act of faith in the speaker.[9]

Our taxonomy for claims defines three types of claims, which frequently appeared in the debate: **Epistemic**, i.e., about knowledge or beliefs: what is true or believed to be true; **Practical**, i.e., about actions: what should be done, what are the consequences of actions, what alternatives are available, etc.; and **Moral**, i.e., about values or preferences: what matters or should be preferred.

To illustrate: "our survival rate for cancer that used to be some of the worse in Europe now actually is one of the best in Europe, we are changing the NHS and we are improving it" contains an *Epistemic* claim about something believed to be true ("we are changing the NHS and we are improving it") supported by a *Study*. Instead, "what is the alternative to making reductions on welfare? it is putting up taxes," "it's key that we keep a strong economy in order to fund a strong NHS," and "cuts will have to come, but we can do it in a balanced way, we can do it in a fair way" are all *Practical* claims, since they are about available options and actions to take. Finally, examples of *Moral* claims are "I don't want Rebecca, I don't want my own kids, I don't want any of our children to pay the price for this generation's mistake" or "this is the most important national institution and national public service that we have."

Agreement

We run the Kappa test after the ontology reached a stable state but before the discussion to decide the final labeling. The test yielded the following results: for Cameron κ was equal to 0.52, for Clegg 0.57, for Miliband 0.47, and the combined κ was equal to 0.53, thus an overall "fair to good" agreement was reached, not unlike the agreement reached in several other corpora for argumentation mining (Habernal, Eckle-Kohler, and Gurevych 2014).

[9]One could of course run a finer-grained argumentative analysis of the debate, for instance to identify various fallacies such as the "straw man" in the last example. However, that would be beyond the scope of this work, since the main point of this taxonomy was to define precise guidelines for labeling claims, and it turned out to be important to define broad classes of evidence, to some extent. Moreover, more sophisticated models would most likely reduce inter-annotator agreement without necessarily increasing claim detection performance.

Table 1: Experimental results of the UK 2015 Political Election corpus. For each of the three candidates we report the F_1 measure, macro-averaged on a 10-fold cross validation.

	Cameron	Clegg	Miliband
Random baseline	46.9	44.3	30.4
GroundTruth	55.2	50.6	58.7
GroundTruth+Audio	61.2	59.0	62.5
GoogleSpeech	48.2	50.5	31.3
GoogleSpeech+Audio	52.6	52.5	29.3

Experiments

Methodology

We run experiments on our dataset by employing several different settings. As for the speech recognition software, we employed the Google Speech APIs.[10] We also used the manually built transcript as a ground truth for the speech recognition data, in order to assess the impact of a perfect speech recognition system within our context. We thus compared the ground truth and the recognized text, and from both data we extracted bag-of-words and bag-of-bigrams representations for original terms, part-of-speech tags, and lemmas. For the audio signals, we used the RastaMat library[11] to compute the first 16 MFCCs and extract their minimum, maximum, average and standard deviation. We then employed them in combination with the features extracted from the text to constitute a single feature vector. As for the learning system, we employed Support Vector Machines and exploited a 10-fold cross validation on the dataset of each candidate. We employed both linear and rbf kernel, with parameters (C for linear kernel, C and γ for rbf) chosen by an inner 5-fold cross validation on the training set.

Discussion of results

Table 1 summarizes the main results obtained by all the considered models. We report F_1 as a standard performance measurement in imbalanced binary classification problems, that is, where the positive class contains much fewer examples than the negative. The information introduced by the acoustic features always improves the performance of the system (around 5% F_1 improvement), except for one candidate (Miliband), but only when the speech recognition system's output is used in place of the ground truth. In that case, in fact, the performance of the speech recognizer is particularly weak, as the features extracted from the recognized text only achieve an F_1 equal to 31.3%.

To illustrate the behaviour of the claim detection system in one concrete example, the following fragment: "we've created 2 million jobs, let's create a job for everyone who wants and needs one" (Cameron) contains evidence ("we've

[10]Using Google Speech APIs in batch mode is allowed only for brief audio sequences, about 10-15 seconds long. We thus had to split longer spoken sequences into smaller samples, and perform recognition on each of them independently from one another. This process could yield sub-optimal results with respect to the available on-line Google Speech system. We employed the en-UK language.

[11]http://labrosa.ee.columbia.edu/matlab/rastamat

created 2 million jobs") but no claim, according to our ontology. The output of the speech recognizer is "who created 2 million jobs that create a job for everyone who wants and needs one." Based on the sole text, the sentence is wrongly classified as containing a claim, both when using the ground truth and when using the speech recognizer. It is instead correctly classified as not containing a claim with the help of the audio features, again, both from the ground truth and from Google Speech's output.

Conclusion and Future Work

We started from a few simple observations: voice quality is a strongly recognized feature of paralinguistic speech, and a strong indicator of the affective states of the speaker; experimental studies highlighted a correlation between argumentation process and emotions felt by those involved in the process; there are tools to accurately recognize emotions from voice audio features. We then hypothesized that vocal features of speech can improve argument mining. We focused on televised political debates because that is indeed an argumentative setting with emotions involved, and also because of their outreach potential to other areas. In order to set up an experiment that could shed light on our hypothesis, we selected one particular debate and built an original dataset around it. We then built a claim detection system that reads an audio sequence and produces a text document where claims are labeled.

The experimental results corroborate our hypothesis. In particular, the performance of our system on ground truth and audio together is significantly higher that the performance on ground truth alone. This shows that, at least in settings with limited noise and with good speakers, the voice signal does have features that can be used to improve claim detection. Our results also show that the outcome strongly depends on the speaker. This is not surprising, and it tallies with real-life experience. Different speakers in general have different skills in using vocal cues such as articulation, sonority, and tempo, and they also have different persuasive power. Indeed, research shows that vocal cues can directly affect the clarity, credibility, and receptivity of a speaker's message (Reid 2013). A surprisingly good performance instead is noted in claim detection using Google Speech+Audio, which sometimes even exceeds the performance based on the ground truth alone (see Clegg's column in Table 1). This shows that speech recognition technology can be used to implement a fully-fledged claim detection system from speech, and that features in the voice signal are enough, sometimes, to make up for the noise introduced by the speech recognizer. In cases where the noise is too high instead the same audio features do not have any corrective effect (see Miliband), probably because the output of the speech recognition stage is so corrupted that it becomes meaningless: no significant features are extracted from the text. In fact, in that case the results are statistically indistinguishable from a random baseline.

Televised election debates such as the 7-party debate are becoming increasingly popular. They are also the focus of considerable research.[12] In the future, we plan to study the applicability of claim detection technology to new interactive tools for instant audience feedback, such as, for instance, the Debate Replay (Plüss and De Liddo 2015). Another application domain is social media analytics, where a great deal of communication is multi-modal, but current argument mining systems are restricted to text only. Finally, there is also a clear potential for improvement in the system we constructed for this experiment: while we did try a number of configurations and voice features before we decided to focus on MFCC vectors because they were yielding the best results—sufficient, anyway, to corroborate our hypothesis—there are still many other configurations which we did not explore yet. Let the 7-party debate dataset we release be an incentive for others to take on the challenge and start on this promising new line of research.

Acknowledgments

This work was partially supported by the ePolicy EU project FP7-ICT-2011-7, grant agreement 288147. Possible inaccuracies of information are under the responsibility of the project team. The text reflects solely the views of its authors. The European Commission is not liable for any use that may be made of the information contained in this paper.

References

Aharoni, E.; Polnarov, A.; Lavee, T.; Hershcovich, D.; Levy, R.; Rinott, R.; Gutfreund, D.; and Slonim, N. 2014. A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. In *Proc. 1st Workshop on Argumentation Mining*, 64–68. ACL.

Benlamine, S.; Chaouachi, M.; Villata, S.; Cabrio, E.; Frasson, C.; and Gandon, F. 2015. Emotions in argumentation: an empirical evaluation. In *Proc. 24th IJCAI*, 156–163. AAAI Press.

Biran, O., and Rambow, O. 2011. Identifying justifications in written dialogs by classifying text as argumentative. *Int. J. Semantic Computing* 5(4):363–381.

Cabrio, E., and Villata, S. 2012. Combining textual entailment and argumentation theory for supporting online debates interactions. In *Proc. 50th ACL*, 208–212. Jeju, Korea: ACL.

Cabrio, E., and Villata, S. 2014. NoDE: A benchmark of natural language arguments. In *Proc. COMMA 2014*, volume 266 of *Frontiers in Artificial Intelligence and Applications*, 449–450. IOS Press.

Campbell, N. 2007. On the use of nonverbal speech sounds in human communication. In *Verbal and Nonverbal Communication Behaviours, LNCS 4775*, 117–128. Springer.

Casale, S.; Russo, A.; Scebba, G.; and Serrano, S. 2008. Speech emotion classification using machine learning algorithms. In *2008 IEEE Int. Conf. on Semantic Computing*, 158–165. IEEE.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psych. Measurement* 20:37–46.

[12] See, e.g., the ESPRC EDV project, http://edv-project.net

Coleman, S., ed. 2011. *Leaders in the Living Room The Prime Ministerial Debates of 2010: Evidence, Evaluation and Some Recommendations*. RISJ Challenges. Reuters Institute for the Study of Journalism.

El Ayadi, M.; Kamel, M. S.; and Karray, F. 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. *Patt. Recognition* 44(3):572–587.

Fairclough, I., and Fairclough, N. 2012. *Political Discourse Analysis: A Method for Advanced Students*. Routledge.

Freeman, J. B. 1991. *Dialectics and the macrostructure of arguments: A theory of argument structure*, volume 10. Walter de Gruyter.

Goudas, T.; Louizos, C.; Petasis, G.; and Karkaletsis, V. 2014. Argument extraction from news, blogs, and social media. In Likas, A.; Blekas, K.; and Kalles, D., eds., *Artificial Intelligence: Methods and Applications, LNCS 8445*, 287–299. Springer.

Grosse, R. B.; Raina, R.; Kwong, H.; and Ng, A. Y. 2007. Shift-invariance sparse coding for audio classification. In *UAI 2007, Vancouver*, 149–158. AUAI Press.

Guo, G., and Li, S. Z. 2003. Content-based audio classification and retrieval by support vector machines. *Trans. Neur. Netw.* 14(1):209–215.

Habernal, I.; Eckle-Kohler, J.; and Gurevych, I. 2014. Argumentation mining on the web from information seeking perspective. In *Frontiers and Connections between Argumentation Theory and Natural Language Processing*, volume 1341 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Istrate, D.; Castelli, E.; Vacher, M.; Besacier, L.; and Serignat, J.-F. 2006. Information extraction from sound for medical telemonitoring. *IEEE Transactions on Information Technology in Biomedicine* 10(2):264–274.

Krippendorff, K. H. 2003. *Content Analysis: An Introduction to Its Methodology*. Sage Publications, Inc, 2nd edition.

Kunz, W., and Rittel, H. W. 1970. *Issues as elements of information systems*, volume 131. Berkeley, California, US: Institute of Urban and Regional Development, University of California.

LeCun, Y.; Bengio, Y.; and Hinton, G. 2015. Deep learning. *Nature* 531:436–444.

Lee, H.; Pham, P.; Largman, Y.; and Ng, A. Y. 2009. Unsupervised feature learning for audio classification using convolutional deep belief networks. In *Advances in neural information processing systems*, 1096–1104.

Levy, R.; Bilu, Y.; Hershcovich, D.; Aharoni, E.; and Slonim, N. 2014. Context dependent claim detection. In *COLING 2014, Dublin, Ireland*, 1489–1500. ACL.

Lippi, M., and Torroni, P. 2015a. Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology,* in press.

Lippi, M., and Torroni, P. 2015b. Context-independent claim detection for argument mining. In *Proc. 24th IJCAI*, 185–191. AAAI Press.

Liu, Y.; Shriberg, E.; Stolcke, A.; Hillard, D.; Ostendorf, M.; and Harper, M. 2006. Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. *IEEE Transactions on Audio, Speech, and Language Processing* 14(5):1526–1540.

Mehrabian, A. 1981. *Silent Messages: Implicit Communication of Emotions and Attitudes*. Belmont, CA: Wadsworth.

Mesnil, G.; He, X.; Deng, L.; and Bengio, Y. 2013. Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding. In *INTERSPEECH*, 3771–3775.

Mochales Palau, R., and Moens, M.-F. 2011. Argumentation mining. *Artificial Intelligence and Law* 19(1):1–22.

Newman, S. E., and Marshall, C. C. 1991. Pushing Toulmin too far: Learning from an argument representation scheme. Technical report, Xerox Palo Alto Research Center, 3333 Coyote Hill Road, Palo Alto, CA 94034.

Park, J.; Katiyar, A.; and Yang, B. 2015. Conditional random fields for identifying appropriate types of support for propositions in online user comments. In *Proceedings of the Second Workshop on Argumentation Mining*. ACL.

Plüss, B., and De Liddo, A. 2015. Engaging citizens with televised election debates through online interactive replays. In *Proc. ACM International Conference on Interactive Experiences for TV and Online Video, TVX*, 179–184. ACM.

Reid, M. B. 2013. Paralinguistic cue and their effect on leader credibility. Technical report, School of Professional Studies, Gonzaga University.

Rinott, R.; Khapra, M.; Alzate, C.; Dankin, L.; Aharoni, E.; and Slonim, N. 2015. Show me your evidence – an automatic method for context dependent evidence detection. In *Proc. 2015 EMNLP*, 440–450. ACL.

Rooney, N.; Wang, H.; and Browne, F. 2012. Applying kernel methods to argumentation mining. In *FLAIRS, Marco Island, Florida*. AAAI Press.

Rosenthal, S., and McKeown, K. 2012. Detecting opinionated claims in online discussions. In *6th IEEE Int. Conf. on Semantic Computing*, 30–37. IEEE Computer Society.

Sardianos, C.; Katakis, I. M.; Petasis, G.; and Karkaletsis, V. 2015. Argument extraction from news. In *Proceedings of the Second Workshop on Argumentation Mining*, 56–66. ACL.

Stab, C., and Gurevych, I. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proc. 2014 EMNLP*, 46–56. ACL.

Teufel, S. 1999. Argumentative zoning. *PhD Thesis, University of Edinburgh*.

Toulmin, S. E. 1958. *The Uses of Argument*. Cambridge University Press.

Tur, G., and De Mori, R. 2011. *Spoken language understanding: Systems for extracting semantic information from speech*. John Wiley & Sons.

Tzanetakis, G., and Cook, P. 2002. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing* 10(5):293–302.