DART: a Dataset of Arguments and their Relations on Twitter

Tom Bosc, Elena Cabrio, Serena Villata

DART: a Dataset of Arguments and their Relations on Twitter

Tom Bosc, Elena Cabrio, Serena Villata
INRIA Sophia Antipolis, University of Nice Sophia Antipolis, CNRS - I3S
930 route des Colles - Bât. Les Templiers, 06903 Sophia Antipolis CEDEX, France
tom.bosc@inria.fr, {cabrio, villata}@i3s.unice.fr

Abstract

The problem of understanding the stream of messages exchanged on social media such as Facebook and Twitter is becoming a major challenge for automated systems. The tremendous amount of data exchanged on these platforms as well as the specific form of language adopted by social media users constitute a new challenging context for existing argument mining techniques. In this paper, we describe a resource of natural language arguments called DART (Dataset of Arguments and their Relations on Twitter) where the complete argument mining pipeline over Twitter messages is considered: (i) we identify which tweets can be considered as arguments and which cannot, and (ii) we identify what is the relation, i.e., support or attack, linking such tweets to each other.
Keywords: Argument mining, social media, textual entailment

1. Introduction

Understanding and interpreting streams of messages exchanged on social networks such as Facebook and Twitter in an automated way is becoming a major challenge in Natural Language Processing (NLP). This tremendous amount of data provides a huge amount of information to be mined, in particular for monitoring opinions about controversial issues in political scenarios and to manage reputation analysis of brands advertising. In this scenario, users express their viewpoints by means of textual messages posted online. In this paper, we present a dataset to support the development of frameworks tackling the argument mining pipeline over Twitter data, as highlighted in (Gabbriellini and Torroni, 2012). More precisely, in this paper, we answer the following research questions:

- How to distinguish from a stream of tweets those textual exchanges that are arguments and those that are not?
- How to identify the relationship, i.e., attack or support, among two (or more) tweets to study the evolution of a certain discussion?

To answer these questions we rely on argumentation theory (Rahwan and Simari, 2009), and more precisely, on the argument mining pipeline (Lippi and Torroni, 2016). Arguments are supposed to support, contradict, explain statements, and are used to support decision making. Abstract argumentation theory provides a framework that allows to detect winning arguments from a set of arguments and the relations among them. Applying the argument mining pipeline to Twitter means addressing the following issues: *(i)* the set of tweets to be analyzed has to be identified, i.e., a selection of potential tweet-arguments that express an idea or an opinion on a given topic, *(ii)* the identified tweet-arguments are coupled with each other forming meaningful pairs, and *(iii)* the relations, i.e., support or attack, among the tweet-arguments of each pair have to be identified. Finally, identifying and linking arguments together allows for a synthetic view of the ongoing discussion.
Differently from texts extracted from traditional sources, such as newspapers, novels or legal texts, messages from Twitter are squeezed, noisy and often unstructured. More specifically, the challenge in identifying tweet-arguments tackles the following points: *(i)* the 140-characters limit forces users to express their ideas very succinctly; *(ii)* the quality of the language in Twitter is deteriorated, contains a lot of variants in spelling, mistakes or spare characters; *(iii)* Twitter's API filters tweets on hashtags but cannot retrieve all the replies to these tweets if they do not contain the same hashtags.
To address these issues, taking into account the peculiarities of Twitter as data source, we propose a methodology to build DART (Dataset of Arguments and their Relations on Twitter) to detect tweet-arguments from a stream of tweets, and to establish the relations between them. Although there exist already several datasets for specific tasks on Twitter, as we discuss in Section 4., a dataset focused on identifying argumentation structures on Twitter has never been proposed, up to our knowledge, despite the increasing number of argument mining techniques proposed in the last years.
The paper is organized as follows. Section 2. describes the methodology we adopted for the creation of the dataset for argument mining on Twitter, and Section 3. provides a description of the DART dataset. Section 4. compares the related literature with the proposed resource. Finally, some conclusions are drawn.

2. Methodology

In this work, we propose a methodology for the creation of a dataset to be used for the automatic detection of arguments in Twitter, and for the automatic assignment of relations among such tweet-arguments (namely *support* vs *attack* (Cabrio and Villata, 2013)).

2.1. Step 1: arguments annotation

Given a set of tweets on the same topic, the first step consists in annotating those tweets that can be considered as *arguments*. The classical structure of an argument is a (set of) premise(s) supporting a conclusion, but in human linguistic interactions some of these parts may be left implicit. In Twitter, this situation is often taken to the extreme due to the characters constraints. Writing guidelines for the tweet-arguments annotation task that will lead to an unambiguous

annotation is therefore far from trivial. Note that the argument annotation task is carried out on a single tweet and not on subparts of it (this aspect can be explored in a future work).

A text containing an opinion is considered as an argument. For example, in the following tweet the opinion of the author is clearly expressed in the second sentence (i.e., *I won't be running out to get one*):

> *RT @mariofraioli: What will #AppleWatch mean for runners? I can't speak for everyone, but I won't be running out to get one. Will you? http://t.co/xBpj0HWKPW*

We consider as arguments also claims expressed as questions (either rhetorical questions, attempts to persuade, containing sarcasm or irony), as in the following example:

> *RT @GrnEyedMandy: What next Republicans? You going to send North Korea a love letter too? #47Traitors*

or:

> *Perhaps Apple can start an organ harvesting program. Because I only need one kidney, right? #iPadPro #AppleTV #AppleWatch*

Tweets containing factual information are annotated as arguments, given that they can be considered as premises or conclusions. For example:

> *RT @HeathWallace: You can already buy a fake #AppleWatch in China http://t.co/WpHEDqYuUC via @cnnnews @mr_gadget http://t.co/WhcMKuMWcd*

Defining the amount of world knowledge needed to determine whether a text is a fact or an opinion when it contains unknown acronyms and abbreviations can be pretty tough. Consider the following tweet:

> *RT @SaysSheToday: The Dixie Chicks were attacked just for using 1A right to say they were ashamed of GWB. They didn't commit treason like the #47Senators*

where the mentioned entities *The Dixie Chicks*, *GWB*, and *1A right* are strictly linked to the US politics, and hardly interpreted by people out of the US politics matters. In this case, annotators are asked to suppose that the mentioned entities exist, and focus on the phrasing of the tweets.

However, if tweets contain pronouns only (preventing the understanding of the text), we consider such tweets as not "self-contained" , and thus non arguments. It can be the case of replies, as in the following example, in which the pronoun *he* is not referenced anywhere in the tweet.

> *@FakeGhostPirate @GameOfThrones He is the one true King after all ;)*

For tweets containing an advertisement to push into visiting a web page, if an opinion or factual information is also present, then the tweet is considered as an argument, otherwise it is not. Consider the following example:

> *RT @NewAppleDevice: Apple's smartwatch can be a games platform and here's why http://t.co/uIMGDyw08I*

It contains factual information that can be understood even without visiting the link. On the contrary, the following tweet is not an argument, given that it does not convey an independent message while excluding the link:

> *For all #business students discussing #AppleWatch this morning. Give it a test drive thanks to @UsVsTh3m: http://t.co/x2bGc9j1Gl.*

To reach a consensus on the task definition and on the guidelines, three expert annotators have annotated a batch of 100 tweets (see Section 3.1. for more details) to compute the inter annotator agreement (Krippendorff's $\alpha = 0.74$).

2.2. Step 2: pairs creation

The tweets annotated as arguments in the previous step are then paired to allow for the identification of the relationship between them (see Section 2.3.).

Our first idea was to reconstruct the structure of a discussion (as done in (Cabrio and Villata, 2013)), but we then realize that on Twitter most of the time users express themselves about a certain topic, without specifically replying to other users, therefore the dialogue reconstruction would become too artificial. Therefore we decided to turn to a different strategy for the pairs creation: first of all, identical or almost identical tweet-arguments are pruned to avoid redundancy; secondly, arguments discussing about the same topic (or the same aspect of it) are grouped together, and pairs are created within such groups. For example, in a debate about politics, arguments can be expressed about, e.g., social issues, economy, justice, therefore only relations among arguments pertaining to the same subtopic are relevant. A naive approach would have consisted in creating pairs randomly among all the tweets, but this approach would yield to mostly unrelated pairs. This is why we decided to group tweets by semantic similarity so that random pairs inside each groups are probably more related. The underlying assumption here is that semantically similar tweets will probably share argumentative relations.

In order to group in an automated way tweets which are about the same topic, we firstly tested topic modelling approaches. The idea is that certain *sub-topics* would emerge and give us the right amount of similarity that is desirable. The major problem that we faced is the difficulty of finding in an automated way meaningful sub-topics. We tested both Latent Dirichlet Allocation (Blei et al., 2003), and more powerful models such as Correlated Topic Models (Blei and Lafferty, 2006), but the interpretability of the clusters did not improve (Chang et al., 2009).

For that reason, for the dataset creation, we turned the clustering problem into a classification problem. We manually created categories for each topic (three annotators have independently proposed a list of categories per topic, that have been subsequently discussed to reach a consensus), and then the tweet-arguments have been annotated by the same annotators according to the category they belong to.

2.3. Step 3: arguments linking

As introduced before, pairs of tweet-arguments are annotated with the relations connecting them, either a positive relation (i.e., a *support* relation in abstract bipolar argumentation (Cayrol and Lagasquie-Schiex, 2005)) a negative relation (i.e., an *attack* relation in abstract argumentation (Dung, 1995)), or an *unrelated* relation. In addition, following the same guidelines proposed by (Cabrio and Villata, 2013), pairs are also annotated according to the Recognizing Textual Entailment (RTE) framework, i.e., pairs linked by a support relation as *entailment/non-entailment*, and pairs linked by an attack relation as *contradiction/non-contradiction*.
Consider the following example, where Tweet-A supports Tweet-B, and there is also an entailment relation among them (i.e., Tweet-A entails Tweet-B).

> Tweet-A: *The letter #47Traitors sent to Iran is one of the most plainly stupid things a group of senators has ever done. http://t.co/oEJFlJeXjy*
>
> Tweet-B: *Republicans Admit: That Iran Letter Was a Dumb Idea http://t.co/Edj57f4nE8. You think?? #47Traitors*

On the contrary, in the following example, we have that Tweet-C attacks Tweet-A, but does not contradict it.

> Tweet-C:*#47Traitors is a joke. Given the definition of treason, it would be on the Obama administration if Iran developed a nuclear bomb.*

However, after a first annotation round to test the guidelines provided in (Cabrio and Villata, 2013), we realized that a few additional instructions should be added with the aim to consider the specificity of the Twitter scenario. The instructions we introduced are as follows:

- if both Tweet-A and Tweet-B in a pair are factual tweets, and they are related to the same issue, the pair must be annotated as *support*, as in:

  > Tweet-A: *.@AirStripmHealth + #AppleWatch provides HIPPA compliant capabilities for physicians, mothers, babies, and more #AppleEvent*
  >
  > Tweet-B: *accessible heart rate monitors and opinions on that #iWatch #apple #accessibility #ios* `https://t.co/ySYM8dk0Pf` *via @audioBoom*

- if both Tweet-A and Tweet-B in a pair are opinion tweets, and they are related to the same issue, the pair must be annotated as *support*, as in:

  > Tweet-A: *Think of how much other stuff you can buy with the money you spend on an #AppleWatch*
  >
  > Tweet-B: *#AppleWatch Tempting, but not convinced. #appletv Yes. #iPhone6sPlus No plan to upgrade #iPadPro little high price, wait & watch*

- if Tweet-B is a factual tweet, and Tweet-A is an opinion on the same issue, the pair must be annotated as *support*, as in:

  > Tweet-A: *Wow. Your vitals on your iwatch. That's bonkers. #AppleEvent*
  >
  > Tweet-B: *accessible heart rate monitors and opinions on that #iWatch #apple #accessibility #ios* `https://t.co/ySYM8dk0Pf` *via @audioBoom*

- if Tweet-A is a factual tweet, and Tweet-B expresses someone's wishes to buy the product or an opinion about it, the pair must be annotated as *unknown*, as in:

  > Tweet-A: *Mom can listen to baby's heart rate with #AppleWatch #airstrip*
  >
  > Tweet-B:*Wow!!! Look at what the #Ap, pleWatch can do for #doctors that's amazing! Seeing their vitals? I just got chills! In a good way #AppleEvent*

3. DART

In the following, we provide some more details and statistics on the application of the presented methodology for the creation and annotation of DART.

3.1. Data

To extract tweets discussing a given topic, Twitter makes an API available[1] that filters the tweets according to the presence of a (set of) given hashtag(s), related to the target topic. We selected four topics currently heavily commented and discussed on Twitter, spanning over different domains, e.g., politics, events, and products release. More precisely, we selected the following four topics:

- (politics) the letter to Iran written by 47 senators on 10/03/2015 (e.g., #47Traitors, #IranLetter)
- (politics) the referendum in Greece for or against Greece leaving European Union on 10/07/2015 (e.g., #Grexit, #GreeceCrisis)
- (product release) the release of Apple iWatch on 10/03/2015 (e.g., #AppleWatch, #iWatch)
- (product release) the airing of Episode 4 (Season 5) of the series Game of Thrones on 4/05/2015 (e.g., #GameOfThrones, #GoT)

Moreover, tweets are deduplicated as much as possible so that our data does not contain a tweet and all its retweet occurrences. For each of the aforementioned topics, 1000 tweets relatively homogeneous through time are selected, leading to a total of 4000 tweets.

[1] `https://dev.twitter.com/overview/api`

Topic	# arg	# not arg	# tot
47 Traitors	768	214	982
Grexit	746	241	987
Apple Watch	623	352	975
Games of Thrones	565	374	939
TOTAL	2702	1181	3883

Table 1: Statistics on Step 1: tweet-arguments annotation

3.2. Step 1: arguments annotation

To carry out the annotation process of this first step extensively, three European students have been hired (from Luxembourg, Italy and Germany), that do not have any background knowledge in linguistics. They have been trained on the task guidelines defined in Section 2.1., and each of them carries out the full annotation task.

In addition to the tags *argument* vs *non-argument*, annotators are offered with the possibility to tag a tweet as *unknown*, to avoid highly uncertain annotations. Given the 4000 tweets on the four topics mentioned above, Table 1 reports on the number of tweets annotated as arguments. As can be seen, 67% of the collected tweets can be considered as tweet-arguments. The percentage of tweet-arguments varies among the topics, and in particular it is lower for the Games of Thrones messages, due to the fact that most of them are advertisement on the coming episode of the series (tweets that are not considered as arguments).

In the reconciliation phase among the three students annotators to build the final dataset, we chose the label that was annotated by at least 2 annotators out of 3 (majority voting mechanism). If all the annotators disagree or if more than one annotator labels the tweet as unknown, then such tweet is discarded. We compute the inter-annotator agreement between the expert annotators and the reconciled student annotations on 250 tweets of the first batch, resulting in $\alpha_{47traitors}$ = 0.81 (Krippendorff's α handles missing values, the label "unknown" in our case).

3.3. Step 2. Pairs creation

Once identified tweet-arguments in the stream of tweets, we combined them into pairs with the aim to predict the relations between them, as explained in Section 2.2..

To make the dataset bigger, we extracted a new set of 20000 tweets following another announcement made by Apple regarding its watch (on 9/03/2016), since we noticed that this follow-up debate was richer in terms of argumentative tweets, as some users already owned an Apple watch. We applied a classifier trained on the dataset described in Section 3.2. to detect arguments (we ignore tweets classified as unknown). We used a 3-fold cross-validation (we alternately train the model on the tweets of the first two topics, and leave the third topic out as a validation set) with randomized hyperparameter search (Bergstra and Bengio, 2012). Because the classes are unbalanced, and the balance is not necessarily the same across all datasets, the training phase weights the errors inversely proportional to class frequencies. Then, tweets have been tokenized with Twokenize[2], and annotated with their PoS applying Stanford POS tagger. POS tags are then used as features, as well as bigrams of tags. As a baseline model, we train a logistic regression model on these features only, then we tested also augmented features as normalized tokens and bigrams of tokens, and this effectively improves over the baseline. The best model (Logistic regression, L2-penalized with $\lambda = 100$) is obtained by using all the features and retraining on the 3 folds. It yields to an F1-score of 0.78 over the test set (the Apple Watch set of $\sim$ 1000 tweets), that can be considered as satisfactory.

We applied such classifier to the new set of 20000 tweets, and we selected the 2200 tweets for which the probability of being tweet-arguments was the highest (to be comparable to the manual annotation).

Finally, as explained in Section 2.2., three expert annotators created a set of categories generic enough to be applied to other product announcements, i.e., *features (F)*, *price (P)*, *look (L)*, *buying announcements (B)*, *advertisement (A)*, *predictions on the future of the product (S)*, *news (N)*, and *others (O)*). Moreover, the category *features* has been subdivided in the following more fine-grained and product-dependent categories: *health*, *innovation*, and *battery*.

On the 2200 tweet-arguments, we have discarded 19 misclassified tweets (manually detected by the annotators as non arguments, while annotating the categories). The remaining 2181 tweet-arguments on the Apple Watch release have then been classified in the above mentioned categories (see Table 2). Annotators were allowed to tag tweets with more than one category, if suitable, as shown in the following example:

> *The gold Apple Watch Edition Will start at $10,000* `http://t.co/NUl7gIXLkC\ nVerge2015` *#AppleWatch #Bahrain #Watch #Wearables*

annotated both with the categories *price* and *news*.

	O	A	B	F	L	N	P	S
#	720	175	370	619	205	65	189	112

Table 2: Statistics on # tweets per category

3.4. Step 3. Arguments linking

Given the categories of tweet-arguments returned in Step 2, pairs are randomly created between tweets belonging to the same category. Within a pair, tweets are randomly assigned as Tweet-A and Tweet-B.

Concerning the annotation of such pairs, the annotators followed the guidelines defined in Section 2.3. Two expert annotators annotated $\sim$600 pairs of tweet-arguments in each categories *look*, *price*, *health*, and a batch of 100 tweets on the category *predictions* (Table 3 reports on the obtained dataset). As it could be expected, most of the pairs are tagged as *unknown*, meaning that they are unrelated (mostly because they talk about different subtopics of the same issue). Inter-annotator agreement has been calculated on 99 pairs (33 pairs randomly extracted from each of the three first topics), resulting in Krippendorff α = 0.67.

[2] `http://www.cs.cmu.edu/~ark/TweetNLP/`

	Support	Attack	Unknown	Total
# in look	72	30	498	600
# in price	134	44	412	590
# in health	222	31	348	601
# in predictions	18	17	65	100
# TOTAL	446	122	1323	1891

Table 3: Statistics on relations among tweet-arguments

We realize that the annotation of the relations on pairs of tweets was more difficult than expected. As an example, consider the following pair:

> Tweet-A: *Can't believe the designers of #AppleWatch didn't present a better shaped watch. It's still too clunky looking & could've been more sleek.*
>
> Tweet-B:*@APPLEOFFICIAL amazing product updates. Apple TV looks great. BUT! Please make a bigger iWatch! Not buying it until it's way bigger.*

On the one hand, the viewpoints emerging from the tweets agree in that the watch is not properly sized. On the other hand, they disagree since one user finds it too big, and the other one too small, which are opposite viewpoints.
To overcome this problem, an additional annotation round could be carried out to highlight partial support, as for the partial entailment relation (Levy et al., 2013). This additional annotation is left as future work.

4. Related work

The first stage of the argument mining pipeline is to detect arguments within the input texts, while the second stage consists in predicting what are the relations holding between the arguments identified in the first stage. This is an extremely complex task, as it involves high-level knowledge representation and reasoning issues, and, for this reason, existing approaches assume simplifying hypotheses, like the fact that evidence is always associated with a claim (Aharoni et al., 2014). To tackle these two challenging tasks, high-quality annotated corpora are needed, like for instance those proposed in (Reed and Rowe, 2004; Palau and Moens, 2011; Levy et al., 2014; Aharoni et al., 2014; Stab and Gurevych, 2014; Cabrio and Villata, 2014; Habernal et al., 2014), to be used as a training set for any kind of aforementioned prediction. None of the above mentioned corpora deals with Twitter data.
Argumentation is applied to Twitter data by (Grosse et al., 2015) who extract a particular version of textual arguments they called "opinions" based on incrementally generated queries. Their goal is to detect conflicting elements in an opinion tree to avoid potentially inconsistent information. Both the goal and the adopted methodology is different from the one we present in this paper, and consequently the related resources differ.
An exhaustive state of the art about argument mining resources is available in (Lippi and Torroni, 2016).

5. Conclusions and ongoing work

In this paper, we have proposed a methodology to build a dataset to support the development of frameworks addressing the argument mining pipeline. We first define the guidelines to detect tweet-arguments among a stream of tweets about a certain topic, e.g., the Apple Watch release; second, we couple the identified arguments with each other to form pairs, and finally, we provide a methodology to identify which kind of relation holds between the arguments composing a pair, i.e., support or attack. The presented methodology is then exploited to built the DART resource (available by request from the authors of the paper).
Several research lines are investigated as future work. First, we plan to identify further kinds of relations between the arguments. For instance, we are currently identifying which of the supports are also entailments, and which of the attacks are contradictions as well, in line with the work of (Cabrio and Villata, 2013). Second, the identification of evidences and claims is an open challenge never addressed over Twitter data. Third, we are finalizing the resource considering the three annotation phases for all the topics we considered, not only for the Apple Watch one. Finally, we are working on the definition of an argument mining framework, trained over the DART resource, able to identify the tweet-arguments from a strem of tweets, and to predict which relation, i.e., support or attack, holds between two tweet-arguments.

6. Acknowledgements

The work carried out in this paper is funded by the CARNOT project, where Inria collaborated with the start-up VigiGlobe in Sophia Antipolis, France.

7. Bibliographical 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 *Proceedings of the First Workshop on Argumentation Mining*, pages 64–68. Association for Computational Linguistics.

Bergstra, J. and Bengio, Y. (2012). Random search for hyper-parameter optimization. *J. Mach. Learn. Res.*, 13(1):281–305, February.

Blei, D. M. and Lafferty, J. D. (2006). Correlated topic models. In *In Proceedings of the 23rd International Conference on Machine Learning*, pages 113–120. MIT Press.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March.

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

Cabrio, E. and Villata, S. (2014). Node: A benchmark of natural language arguments. In Simon Parsons, et al., editors, *Computational Models of Argument - Proceedings*

of COMMA 2014, Atholl Palace Hotel, Scottish Highlands, UK, September 9-12, 2014, volume 266 of *Frontiers in Artificial Intelligence and Applications*, pages 449–450. IOS Press.

Cayrol, C. and Lagasquie-Schiex, M.-C. (2005). On the acceptability of arguments in bipolar argumentation frameworks. In *Symbolic and quantitative approaches to reasoning with uncertainty*, pages 378–389. Springer Berlin Heidelberg.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., and Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*, pages 288–296.

Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial intelligence*, 77(2):321–357.

Gabbriellini, S. and Torroni, P. (2012). Large scale agreements via microdebates. In *Proceedings of the First International Conference on Agreement Technologies, AT 2012, Dubrovnik, Croatia, October 15-16, 2012*, pages 366–377.

Grosse, K., González, M. P., Chesñevar, C. I., and Maguitman, A. G. (2015). Integrating argumentation and sentiment analysis for mining opinions from twitter. *AI Commun.*, 28(3):387–401.

Habernal, I., Eckle-Kohler, J., and Gurevych, I. (2014). Argumentation mining on the web from information seeking perspective. In Elena Cabrio, et al., editors, *Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing, Forlì-Cesena, Italy, July 21-25, 2014.*, volume 1341 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Jan Hajic et al., editors. (2014). *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*. ACL.

Levy, O., Zesch, T., Dagan, I., and Gurevych, I. (2013). Recognizing partial textual entailment. In *ACL (2)*, pages 451–455. The Association for Computer Linguistics.

Levy, R., Bilu, Y., Hershcovich, D., Aharoni, E., and Slonim, N. (2014). Context dependent claim detection. In Hajic and Tsujii (Hajic and Tsujii, 2014), pages 1489–1500.

Lippi, M. and Torroni, P. (2016). Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology*.

Palau, R. M. and Moens, M. (2011). Argumentation mining. *Artif. Intell. Law*, 19(1):1–22.

Rahwan, I. and Simari, G. R. (2009). *Argumentation in Artificial Intelligence*. Springer Publishing Company, Incorporated, 1st edition.

Reed, C. and Rowe, G. (2004). Araucaria: Software for argument analysis, diagramming and representation. *International Journal on Artificial Intelligence Tools*, 13(4):983.

Stab, C. and Gurevych, I. (2014). Annotating argument components and relations in persuasive essays. In Hajic and Tsujii (Hajic and Tsujii, 2014), pages 1501–1510.