Argument Mining on Twitter: Arguments, Facts and Sources

Mihai Dusmanu and **Elena Cabrio** and **Serena Villata**
Université Côte dAzur, CNRS, Inria, I3S, France

Abstract

Social media collect and spread on the Web personal opinions, facts, fake news and all kind of information users may be interested in. Applying argument mining methods to such heterogeneous data sources is a challenging open research issue, in particular considering the peculiarities of the language used to write textual messages on social media. In addition, new issues emerge when dealing with arguments posted on such platforms, such as the need to make a distinction between personal opinions and actual facts, and to detect the source disseminating information about such facts to allow for provenance verification. In this paper, we apply supervised classification to identify arguments on Twitter, and we present two new tasks for argument mining, namely facts recognition and source identification. We study the feasibility of the approaches proposed to address these tasks on a set of tweets related to the Grexit and Brexit news topics.

1 Introduction

Argument mining aims at automatically extracting natural language arguments and their relations from a variety of textual corpora, with the final goal of providing machine-processable structured data for computational models of arguments and reasoning engines (Peldszus and Stede, 2013; Lippi and Torroni, 2016). Several approaches have been proposed so far to tackle the two main tasks identified in the field: *i) arguments extraction*, i.e., to detect arguments within the input natural language texts and the further detection of their boundaries, and *ii) relations prediction*, i.e., to predict what are the relations holding between the arguments identified in the first task[1]. Social media platforms like Twitter[2] and newspapers blogs allow users to post their own viewpoints on a certain topic, or to disseminate news read on newspapers. Being these texts short, without standard spelling and with specific conventions (e.g., hashtags, emoticons), they represent an open challenge for standard argument mining approaches (Snajder, 2017). The nature and peculiarity of social media data rise also the need of defining new tasks in the argument mining domain (Addawood and Bashir, 2016; Llewellyn et al., 2014).

In this paper, we tackle the first standard task in argument mining, addressing the research question: *how to mine arguments from Twitter?* Going a step further, we address also the following subquestions that arise in the context of social media: *i)* how to distinguish factual arguments from opinions? *ii)* how to automatically detect the source of factual arguments? To answer these questions, we extend and annotate a dataset of tweets extracted from the streams about the Grexit and the Brexit news. To address the first task of argument detection, we apply supervised classification to separate *argument-tweets* from non-argumentative ones. By considering only argument-tweets, in the second step we apply again a supervised classifier to recognize tweets reporting factual information from those containing opinions only. Finally, we detect, for all those arguments recognized as factual in the previous step, what is the source of such information (e.g., the CNN), relying on the type of the Named Entities recognized in the tweets. The last two steps represent new tasks in the argument

[1] We refer the reader interested in more details on argument mining to (Peldszus and Stede, 2013; Lippi and Torroni, 2016) as survey papers, and to the proceedings of the Argument Mining workshop series (`https://argmining2017.wordpress.com/`).

[2] `www.twitter.com`

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2317–2322
Copenhagen, Denmark, September 7–11, 2017. ©2017 Association for Computational Linguistics

mining research field, of particular importance in social media applications.

2 Mining arguments on Twitter

In this section, we describe the approaches we have developed to address the following tasks: *i)* Argument detection, *ii)* Factual vs opinion classification, and *iii)* Source identification, on social media data. Our experimental setting - whose goal is to investigate the tasks' feasibility on such peculiar data - considers a dataset of tweets related to the political debates on whether or not Great Britain and Greece had to leave the European Union (i.e. #Brexit and #Grexit threads in Twitter).

2.1 Experimental setting

Dataset.[3] The only available resource of annotated tweets for argument mining is DART (Bosc et al., 2016a). From the highly heterogeneous topics contained in such resource (i.e. the letter to Iran written by 47 U.S. senators; the referendum for or against Greece leaving the EU; the release of Apple iWatch; the airing of the 4th episode of the 5th season of the TV series Game of Thrones), and considering the fact that tweets discussing a political topic generally have a more developed argumentative structure than tweets commenting on a product release, we decided to select for our experiments the subset of the DART dataset on the thread *#Grexit* (987 tweets). Then, following the same methodology described in (Bosc et al., 2016a), we have extended such dataset collecting 900 tweets from the thread on *#Brexit*. From the original thread, we filtered away retweets, accounts with a bot probability >0.5 (Davis et al., 2016), and almost identical tweets (Jaccard distance, empirically evaluated threshold). Given that tweets in DART are already annotated for task 1 (argument/non-argument, see Section 2.2), two annotators carried out the same task on the newly extracted data. Moreover, the same annotators annotated both datasets (Grexit/Brexit) for the other two tasks of our experiments, i.e. *i)* given the argument tweets, annotation of tweets as either containing factual information or opinions (see Section 2.3), and *ii)* given factual argument tweets, annotate their source when explicitly cited (see Section 2.4). Tables 1, 2 and 3 contain statistical information on the datasets.

Inter annotator agreement (IAA) (Carletta, 1996) between the two annotators has been calculated for the three annotation tasks, resulting in κ=0.767 on the first task (calculated on 100 tweets), κ=0.727 on the second task (on 80 tweets), and Dice=0.84 (Dice, 1945)[4] on the third task (on the whole dataset). More specifically, to compute IAA, we sampled the data applying the same strategy: for the first task, we randomly selected 10% of the tweets of the Grexit dataset (our training set); for task 2, again we randomly selected 10% of the tweets annotated as argument in the previous annotation step; for task 3, given the small size of the dataset, both annotators annotated the whole corpus.

dataset	# argument	# non-arg	total
Brexit	713	187	900
Grexit	746	241	987
total	1459	428	1887

Table 1: Dataset for task 1: argument detection

dataset	# factual arg.	# opinion	total
Brexit	138	575	713
Grexit	230	516	746
total	368	1091	1459

Table 2: Dataset for task 2: factual arguments vs opinions classification

dataset	# arg. with source cit.	# arg. without source cit.	total
Brexit	40	98	138
Grexit	79	151	230
total	119	249	368

Table 3: Dataset for task 3: source identification

Classification algorithms. We tested Logistic Regression (LR) and Random Forest (RF) classification algorithms, relying on the *scikit-learn* tool suite[5]. For the learning methods, we have used a Grid Search (exhaustive) through a set of predefined hyper-parameters to find the best performing ones (the goal of our work is not to optimize

[3] Annotated data are available upon request to the authors.

[4] Dice is used instead of κ to account for partial agreement on the set of sources detected in the tweets.

[5] http://scikit-learn.org/

the classification performance but to provide a preliminary investigation on new tasks in argument mining over Twitter data). We extract argument-level features from the dataset of tweets (following (Wang and Cardie, 2014)), that we group into the following categories:

- *Lexical (L):* unigram, bigram, WordNet verb synsets;
- *Twitter-specific (T):* punctuation, emoticons;
- *Syntactic/Semantic (S):* we have two versions of dependency relations as features, one being the original form, the other generalizing a word to its POS tag in turn. We also use the syntactic tree of the tweets as feature. We apply the Stanford parser (Manning et al., 2014) to obtain parse trees and dependency relations;
- *Sentiment (SE):* we extract the sentiment from the tweets with the Alchemy API[6], the sentiment analysis feature of IBM's Semantic Text Analysis API. It returns a polarity label (positive, negative or neutral) and a polarity score between -1 (totally negative) and 1 (totally positive).

As baselines we consider both LR and RF algorithms with a set of basic features (i.e., lexical).

2.2 Task 1: Argument detection

The task consists in classifying a tweet as being an argument or not. We consider as arguments all those text snippets providing a portion of a standard argument structure, i.e., opinions under the form of claims, facts mirroring the data in the Toulmin model of argument (Toulmin, 2003), or persuasive claims, following the definition of argument tweet provided in (Bosc et al., 2016a,b). Our dataset contains 746 argument tweets and 241 non-argument tweets for Grexit (that we use as training set), and 713 argument tweets and 187 non-argument tweets for Brexit (the test set). Below we report an example of argument tweet (a), and of a non-argument tweet (b).

(a) *Junker asks "who does he think I am". I suspect elected PM Tsipras thinks Junker is an unelected Eurocrat. #justsaying #democracy #grexit*

(b) *#USAvJPN #independenceday #JustinBieberBestIdol Macri #ConEsteFrioYo happy 4th of july #Grefenderum Wireless Festival*

We cast the argument detection task as a binary classification task, and we apply the supervised algorithms described in Section 2.1. Table 4 reports on the obtained results with the different configurations, while Table 5 reports on the results obtained by the best configuration, i.e., LR + All features, per each category.

Approach	Precision	Recall	F1
RF+L	0.76	0.69	0.71
LR+L	0.76	0.71	0.73
LR+all features	0.80	0.77	**0.78**

Table 4: Results obtained on the test set for the argument detection task (L=lexical features)

Category	P	R	F1	#arguments per category
non-arg	0.46	0.60	0.52	187
arg	0.89	0.82	0.85	713
avg/total	0.80	0.77	**0.78**	900

Table 5: Results obtained by the best model on each category of the test set for the argument detection task

Most of the miss-classified tweets are either ironical, e.g.:

If #Greece had a euro for every time someone mentioned #Grexit and #Greferendum they would probably have enough for a bailout. #GreekCrisis

that was wrongly classified as argument, or contain reported speech, e.g.:

Jeremy Warner: Unintentionally, the Greeks have done themselves a favour. Soon, they will be out of the euro `http://t.co/YmqXi361Gj` *#Grexit*

that was wrongly classified as non argument. Our results are comparable to those reported in (Bosc et al., 2016b) (they trained a supervised classifier on the tweets of all topics in the DART dataset but the iWatch, used as test set). Better performances obtained in our setting are most likely due to a better feature selection, and to the fact that in our case

[6] https://www.ibm.com/watson/alchemy-api.html

the topics in the training and test sets are more homogeneous.

2.3 Task 2: Factual vs opinion classification

This task consists in classifying argument-tweets as containing factual information or being opinion-based (Park et al., 2015). Our interest focuses in particular on factual argument-tweets, as we are interested then in the automated identification of their sources. This would allow then to rank factual tweet-arguments depending on the reliability or expertise of their source for subsequent tasks as fact checking. Given the huge amount of work in the literature devoted to opinion extraction, we do not address any further analysis on opinion-based arguments here, referring the interested reader to (Liu, 2012).

An argument is annotated as *factual* if it contains a piece of information which can be proved to be true (see example (a) below), or if it contains "reported speech" (see example (b) below). All the other argument tweets are considered as "opinion" (see example (c) below).

(a) *72% of people who identified as "English" supported #Brexit (while no majority among those identifying as "British")* `https://t.co/MuUXqncUBe`

(b) *#Hollande urges #UK to start #Brexit talks as soon as possible.* `https://t.co/d12TV8JqYD`.

(c) *Trump is going to sell us back to England. #Brexit #RNCinCLE*

Our dataset contains 230 factual argument tweets and 516 opinion argument tweets for Grexit (training set), and 138 factual argument tweets and 575 opinion argument tweets for Brexit (test set).

To address the task of factual vs opinion arguments classification, we apply the supervised classification algorithms described in Section 2.1. Tweets from Grexit dataset are used as training set, and those from Brexit dataset as test set. Table 6 reports on the obtained results, while Table 7 reports on the results obtained by the best configuration, i.e. LR + All features, per each category.

Most of the miss-classified tweets contain reported opinions/reported speech and are wrongly classified by the algorithm as opinion - such behaviour could be expected given that sentiment features play a major role in these cases, e.g.,

Approach	Precision	Recall	F1
RF+L	0.75	0.68	0.71
LR+L	0.75	0.75	0.75
LR+all features	0.81	0.79	**0.80**

Table 6: Results obtained on the test set for the factual vs opinion argument classification task (L=lexical features)

Category	P	R	F1	#arguments per category
fact	0.49	0.50	0.50	138
opinion	0.88	0.87	0.88	575
avg/total	0.81	0.79	**0.80**	713

Table 7: Results obtained by the best model on each category of the test set for the factual vs opinion argument classification task

Thomas Piketty accuses Germany of forgetting history as it lectures Greece `http://t.co/B0UqPn0i6T` *#grexit*

Again, the other main reason for miss-classification is sarcasm/irony contained in the tweets, e.g.,

So for Tsipras, no vote means back to the table, for Varoufakis, meant Grexit?

that was wrongly classified as fact.

2.4 Task 3: Source identification

Since factual arguments (as defined above) are generally reported by news agencies and individuals, the third task we address - and that can be of a value in the context of social media - is the recognition of the information source that disseminates the news reported in a tweet (when explicitly mentioned). For instance, in:

The Guardian: Greek crisis: European leaders scramble for response to referendum no vote. `http://t.co/cUNiyLGfg3`

the source of information is The Guardian newspaper. Such annotation is useful to rank factual tweet-arguments depending on the reliability or expertise of their source in news summarization or fact-checking applications, for example.

Our dataset contains 79 factual argument tweets where the source is explicitly cited for Grexit (training set), and 40 factual argument tweets where the source is explicitly cited for Brexit (test set). Given the small size of the available annotated dataset, to address this task we implemented a simple string matching algorithm that relies on a gazetteer containing a set of Twitter usernames and hashtags extracted from the training data, and a list of very common news agencies (e.g. BBC, CNN, CNBC). If no matches are found, the algorithm extracts the NEs from the tweets through (Nooralahzadeh et al., 2016)'s system, and applies the following two heuristics: *i)* if a NE is of type `dbo:Organisation` or `dbo:Person`, it considers such NE as the source; *ii)* it searches in the abstract of the DBpedia[7] page linked to that NE if the words "news", "newspaper" or "magazine" appear (if found, such entity is considered as the source). In the example above, the following NEs have been detected in the tweet: "The Guardian" (linked to the DBpedia resource `http://dbpedia.org/page/The_Guardian`) and "Greek crisis" (linked to `http://dbpedia.org/page/Greek_government-debt_crisis`). Applying the mentioned heuristics, the first NE is considered as the source. Table 8 reports on the obtained results. As baseline, we use a method that considers all the NEs detected in the tweet as sources.

Approach	Precision	Recall	F1
Baseline	0.26	0.48	0.33
Matching+heurist.	0.69	0.64	**0.67**

Table 8: Results obtained on the test set for the source identification task

Most of the errors of the algorithm are due to information sources not recognized as NEs (in particular, when the source is a Twitter user), or NEs that are linked to the wrong DBpedia page. However, in order to draw more interesting conclusions on the most suitable methods to address this task, we would need the increase the size of the dataset.

3 Discussion and Future work

This paper investigated argument mining tasks on Twitter data. The main contribution is twofold: first, we propose one of the very few approaches of argument mining on Twitter, and second, we propose and evaluate two new tasks for argument mining, i.e., facts recognition and source identification. These tasks are particularly relevant when applied to social media data, in line with the open popular challenges of fact-checking and source verification to which these results contribute.

The issue of argument detection on Twitter has already been addressed in the literature. Bosc et al. (2016a,b) address a binary classification task (argument-tweet vs. non argument), as first step of their pipeline. Goudas et al. (2015) experiments machine learning techniques over a dataset in Greek extracted from social media. They first detect argumentative sentences, and second identify premises and claims. However, none of them is neither interested in distinguishing facts from opinions nor to identify the arguments' sources. An argumentation-based approach is applied to Twitter data to extract opinions in (Grosse et al., 2015), with the aim of detecting conflicting elements in an opinion tree to avoid potentially inconsistent information. Both the goal and the adopted methodology are different from ours.

Being it a work in progress, several open issues have to be considered as future research. Among them, we are currently extending the dataset of annotated tweets both in terms of annotated tweets per topic, and in terms of addressed topics (e.g. Brexit after the referendum, Trump), in order to have more instances of facts and sources. On such extended dataset, we plan to run experiments using the three modules of the system as a pipeline.

Moreover, we plan to extend our pipeline by considering also the links provided in the tweets to verify their sources, i.e., if a tweet claims to report information from the CNN but the link actually redirects towards an advertisement website the source is not indubitably the CNN.

References

Aseel Addawood and Masooda Bashir. 2016. "what is your evidence?" A study of controversial topics on social media. In *Proceedings of the Third Workshop on Argument Mining, hosted by the 54th Annual Meeting of the Association for Computational Linguistics, ArgMining@ACL 2016, August 12, Berlin, Germany*. http://aclweb.org/anthology/W/W16/W16-2801.pdf.

Tom Bosc, Elena Cabrio, and Serena Villata. 2016a. DART: a dataset of arguments and their relations on twitter. In Nicoletta Calzolari, Khalid

[7] `http://www.dbpedia.org`

Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016.*. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2016/summaries/611.html.

Tom Bosc, Elena Cabrio, and Serena Villata. 2016b. Tweeties squabbling: Positive and negative results in applying argument mining on social media. In Pietro Baroni, Thomas F. Gordon, Tatjana Scheffler, and Manfred Stede, editors, *Computational Models of Argument - Proceedings of COMMA 2016, Potsdam, Germany, 12-16 September, 2016.*. IOS Press, volume 287 of *Frontiers in Artificial Intelligence and Applications*, pages 21–32. https://doi.org/10.3233/978-1-61499-686-6-21.

Jean Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Comput. Linguist.* 22(2):249–254. http://dl.acm.org/citation.cfm?id=230386.230390.

Clayton Allen Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. 2016. Botornot: A system to evaluate social bots. In *Proceedings of the 25th International Conference Companion on World Wide Web*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, WWW '16 Companion, pages 273–274. https://doi.org/10.1145/2872518.2889302.

L. R. Dice. 1945. Measures of the amount of ecologic association between species. *Journal of Ecology* 26:297–302.

Theodosios Goudas, Christos Louizos, Georgios Petasis, and Vangelis Karkaletsis. 2015. Argument extraction from news, blogs, and the social web. *International Journal on Artificial Intelligence Tools* 24(5). https://doi.org/10.1142/S0218213015400242.

Kathrin Grosse, María Paula González, Carlos Iván Chesñevar, and Ana Gabriela Maguitman. 2015. Integrating argumentation and sentiment analysis for mining opinions from twitter. *AI Commun.* 28(3):387–401. https://doi.org/10.3233/AIC-140627.

Marco Lippi and Paolo Torroni. 2016. Argumentation mining: State of the art and emerging trends. *ACM Trans. Internet Techn.* 16(2):10:1–10:25. http://doi.acm.org/10.1145/2850417.

B. Liu. 2012. *Sentiment Analysis and Opinion Mining*. Synthesis digital library of engineering and computer science. Morgan & Claypool. https://books.google.fr/books?id=Gt8g72e6MuEC.

Clare Llewellyn, Claire Grover, Jon Oberlander, and Ewan Klein. 2014. Re-using an argument corpus to aid in the curation of social media collections. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014.*. pages 462–468. http://www.lrec-conf.org/proceedings/lrec2014/summaries/845.html.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*. pages 55–60. http://www.aclweb.org/anthology/P/P14/P14-5010.

Farhad Nooralahzadeh, Cédric Lopez, Elena Cabrio, Fabien L. Gandon, and Frédérique Segond. 2016. Adapting semantic spreading activation to entity linking in text. In *Natural Language Processing and Information Systems - 21st International Conference on Applications of Natural Language to Information Systems, NLDB 2016, Salford, UK, June 22-24, 2016, Proceedings*. pages 74–90. http://dx.doi.org/10.1007/978-3-319-41754-7_7.

Joonsuk Park, Cheryl Blake, and Claire Cardie. 2015. Toward machine-assisted participation in erulemaking: an argumentation model of evaluability. In *Proceedings of the 15th International Conference on Artificial Intelligence and Law, ICAIL 2015, San Diego, CA, USA, June 8-12, 2015*. pages 206–210. https://doi.org/10.1145/2746090.2746118.

Andreas Peldszus and Manfred Stede. 2013. From argument diagrams to argumentation mining in texts: A survey. *IJCINI* 7(1):1–31. https://doi.org/10.4018/jcini.2013010101.

Jan Snajder. 2017. Social media argumentation mining: The quest for deliberateness in raucousness. *CoRR* abs/1701.00168. http://arxiv.org/abs/1701.00168.

S.E. Toulmin. 2003. *The Uses of Argument*. Cambridge University Press. https://books.google.fr/books?id=53whAwAAQBAJ.

Lu Wang and Claire Cardie. 2014. Improving agreement and disagreement identification in online discussions with A socially-tuned sentiment lexicon. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, WASSA@ACL 2014, June 27, 2014, Baltimore, Maryland, USA*. pages 97–106. http://aclweb.org/anthology/W/W14/W14-2617.pdf.