Detecting deceptive reviews using Argument Mining

Oana Cocarascu, Francesca Toni
Imperial College London

Abstract

The unstoppable rise of social networks and the web is facing a serious challenge: identifying the truthfulness of online opinions and reviews. In this paper we use Argument Mining to extract Argumentation Frameworks (AFs) from reviews and explore whether the use of these AFs can improve the performance of machine learning techniques in detecting deceptive behaviour, resulting from users lying in order to mislead readers. The AFs represent how arguments from reviews relate to arguments from other reviews as well as to arguments about the goodness of the items being reviewed.

Detecting deceptive reviews

Nowadays the decision of purchasing a specific product or service is often based on online reviews. However, the authenticity and truthfulness of these reviews is not guaranteed and content communities, review and news websites are susceptible to deceptive content. Different deception strategies exist: falsification (contradictions/lies), exaggeration (superlative information), omission (hiding information) and misleading information (irrelevant information/topic changes) (Appling, Briscoe, and Hutto 2015). It is known that deceptive reviews cannot be easily identified manually (Ott et al. 2011).

Most work on detecting deceptive reviews uses machine learning techniques and features extracted by Natural Language Processing (NLP) (e.g. see (Crawford et al. 2015)). We propose new features, obtained through (forms of) Argument Mining, and experiment with their use by several machine learning techniques in two domains.

Argument Mining is a relatively new research area which involves, for instance, the automatic detection of arguments in text, of arguments components, as well as of relations between arguments (e.g. see (Palau and Moens 2011; Peldszus and Stede 2013; Lippi and Torroni 2015) for overviews). Our approach to detecting deceptive reviews mines Argumentation Frameworks (AFs) as understood in AI, and in particular *Abstract Argumentation Frameworks* (AAFs) (Dung 1995) and *Bipolar Argumentation Frameworks* (BAFs) (Cayrol and Lagasquie-Schiex 2005). These AFs represent dialectical (attack for AAFs and attack/support for BAFs) relationships between arguments, with arguments seen simply as abstract entities, and are equipped with semantics/algorithms to compute acceptability (Dung 1995; Cayrol and Lagasquie-Schiex 2005) or dialectical strength (Rago et al. 2016) of arguments, given the relationships amongst them. In our approach, the strengths of arguments in the AFs we mine contribute new *argumentative features* for standard machine learning classifiers.

We use two methods for Argument Mining. The first method uses sentiment analysis to construct an AAF from a set of reviews whereas the second method uses relation-based Argumentation Mining (Carstens and Toni 2015b) alongside sentiment analysis to mine a BAF from a set of reviews. The second method associates arguments to (noun-level) topics in reviews, whereas in the first method arguments are topic-independent.

Our new argumentative features are calculated using the strength of arguments in AFs to capture the impact of each review on determining how good/bad the item (product or service) is with respect to all reviews about that item. Thus, these argumentation features can be seen as adding a semantic layer to the analysis of deceptive behaviour in reviews on top of the syntactic analysis given by standard NLP when using machine learning techniques. Our approach can also be seen as integrating argumentation and machine learning, in the spirit, for instance, of (Možina et al. 2008; Gao and Toni 2014; Carstens and Toni 2015a; Carstens 2016), but in a different context (deception detection) and using a novel methodology (argumentative features).

In order to test the usefulness of our novel argumentative features to determine deceptive reviews, we experiment with various machine learning classifiers, using the gold standard consisting of hotel reviews of 20 Chicago hotels (Ott, Cardie, and Hancock 2013) and restaurant reviews (Li et al. 2014). We show experimentally that, for a number of classifiers, using argumentative features yields no change or better results in classifier performance. In the case of the AAF-based argumentative features, we obtain an improvement of 1.5% accuracy for the hotel dataset and 2.25% for the restaurant dataset when compared to the baseline. In the case of the BAF-based argumentative features, we obtain an improvement of 3.5% accuracy for the hotel domain and 4% for the restaurant domain when compared to the baseline. In the experiments, to determine both AAF- and BAF-based argumentative features, we use an off-the-shelf sentiment analysis classifier. To determine BAF-based argumentative features, we train a relation-based Argument Mining classifier, achieving 96.19% F_1 for determining support/attack/neither relationships between sentences.

References

Appling, D. S.; Briscoe, E. J.; and Hutto, C. J. 2015. Discriminative models for predicting deception strategies. In *Proceedings of the 24th International Conference on World Wide Web*, 947–952. ACM.

Carstens, L., and Toni, F. 2015a. Improving out-of-domain sentiment polarity classification using argumentation. In *IEEE International Conference on Data Mining Workshop*, 1294–1301.

Carstens, L., and Toni, F. 2015b. Towards relation based argumentation mining. In *Proceedings of the 2nd Workshop on Argumentation Mining*, 29–34. Association for Computational Linguistics.

Carstens, L. 2016. *Using Argumentation to improve classification in Natural Language problems*. Ph.D. Dissertation, Imperial College London.

Cayrol, C., and Lagasquie-Schiex, M.-C. 2005. On the acceptability of arguments in bipolar argumentation frameworks. In *Proceedings of the 8th European conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, 378–389. Springer Berlin Heidelberg.

Crawford, M.; Khoshgoftaar, T.; Prusa, J.; Richter, A.; and Al Najada, H. 2015. Survey of review spam detection using machine learning techniques. *Journal of Big Data* 2(1):1–24.

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.

Gao, Y., and Toni, F. 2014. Argumentation accelerated reinforcement learning for cooperative multi-agent systems. In *ECAI*, 333–338. IOS PRESS.

Li, J.; Ott, M.; Cardie, C.; and Hovy, E. H. 2014. Towards a general rule for identifying deceptive opinion spam. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL*, 1566–1576. Association for Computer Linguistics.

Lippi, M., and Torroni, P. 2015. Argument Mining: A Machine Learning Perspective. In *The 2015 International Workshop on Theory and Applications of Formal Argument*, 163–176. Springer International Publishing.

Možina, M.; Guid, M.; Krivec, J.; Sadikov, A.; and Bratko, I. 2008. Fighting knowledge acquisition bottleneck with argument based machine learning. In *Proceedings of the 2008 Conference on ECAI 2008: 18th European Conference on Artificial Intelligence*, 234–238. IOS Press.

Ott, M.; Choi, Y.; Cardie, C.; and Hancock, J. T. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 309–319. Association for Computational Linguistics.

Ott, M.; Cardie, C.; and Hancock, J. T. 2013. Negative deceptive opinion spam. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

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

Peldszus, A., and Stede, M. 2013. From argument diagrams to argumentation mining in texts: A survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)* 7(1):1–31.

Rago, A.; Toni, F.; Aurisicchio, M.; and Baroni, P. 2016. Discontinuity-free decision support with quantitative argumentation debates. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Fifteenth International Conference, KR*, 63–73. AAAI Press.