What works and what does not: Classifier and feature analysis for argument mining

Ahmet Aker, Alfred Sliwa, Yuan Ma, Ruishen Liu
Niravkumar Borad, Seyedeh Fatemeh Ziyaei, Mina Ghbadi
University of Duisburg-Essen
a.aker@is.inf.uni-due.de
alfred.sliwa.92, yuan.ma, ruishen.liu, niravkumar.borad
seyedeh.ziyaei, mina.ghobadi@stud.uni-due.de

Abstract

This paper offers a comparative analysis of the performance of different supervised machine learning methods and feature sets on argument mining tasks. Specifically, we address the tasks of extracting argumentative segments from texts and predicting the structure between those segments. Eight classifiers and different combinations of six feature types reported in previous work are evaluated. The results indicate that overall best performing features are the structural ones. Although the performance of classifiers varies depending on the feature combinations and corpora used for training and testing, Random Forest seems to be among the best performing classifiers. These results build a basis for further development of argument mining techniques and can guide an implementation of argument mining into different applications such as argument based search.

1 Introduction

Argument mining refers to the automatic extraction of arguments from natural texts. An argument consists of a claim (also referred to as the conclusion of the argument) and several pieces of evidence called premises that support or reject the claim (Lippi and Torroni, 2016).

As a research area argument mining has seen a rapid progress in the last three-to-five years (Lippi and Torroni, 2015). Current studies report methods for argument mining in legal documents (Moens et al., 2007; Reed et al., 2008), persuasive essays (Nguyen and Litman, 2015; Stab and Gurevych, 2014b), Wikipedia articles (Levy et al., 2014), user comments (Park and Cardie, 2014), online products (Wyner et al., 2012), social media (Goudas et al., 2014) and news articles (Sardianos et al., 2015).

Argument mining is a process that involves the following steps, each of which is a research area in itself addressed by several studies: *identifying argumentative segments in text* (Moens et al., 2007; Wyner et al., 2012; Park and Cardie, 2014; Goudas et al., 2014; Levy et al., 2014; Lippi and Torroni, 2015; Swanson et al., 2015; Sardianos et al., 2015; Lawrence et al., 2014), *clustering recurring arguments* (Boltužić and Šnajder, 2015; Misra et al., 2015), *classification of premises as supporting (pro) or rejecting (contra)* (Stab and Gurevych, 2014b; Nguyen and Litman, 2015), *determining argument structure* (Cabrio and Villata, 2012; Lawrence et al., 2014; Ghosh et al., 2014) and *mapping arguments into pre-defined argument schemas* (Feng and Hirst, 2011).

In terms of methods all these studies rely on supervised machine learning. Among the different classification approaches applied Support Vector Machines, Naïve Bayes and Logistic Regression are the most common ones. Also different feature types have been investigated for the different steps of the argument mining task. Among the features types the prominent ones are *structural*, *lexical*, *syntactic*, *indicators* and *contextual* features as summarized by Stab and Gurevych (2014b).

Given this variety of work on argument mining time is ripe for an extensive comparative analysis of the performance of different machine learning techniques on different argument mining tasks using different data sets. Such an analysis should serve as a basis for further development of argument mining techniques and also inform those who want to implement argument mining components into other applications.

In this paper we offer such a comparative analysis of machine learning methods and features with respect to two argument mining tasks: (1) identifying argumentative segments in text, i.e. the classification of textual units (usually sentences) into claims, premises or none and (2) the prediction of argument structure, i.e. connecting claims and premises. We re-implement a rich set of features reported by related work and evaluate eight different classification systems. We perform our investigation on two different well-known corpora: (1) the persuasive essays corpus reported by Stab and Gurevych (2016) and (2) the Wikipedia claim and premise data reported by Aharoni et al. (2014).

2 Experimental Settings

2.1 Data

We investigate the feature and classifier performances on two corpora. The first corpus consists of over 400 persuasive essays where arguments are annotated as claim, premise or major claim (Stab and Gurevych, 2016). For our purposes we consider each major claim as a claim to keep the argumentation model as simple as possible and ensure comparability between data sets. The second corpus consists of over 300 Wikipedia articles in which arguments are annotated as either Context Dependent Claim (CDC) or Context Dependent Evidence (CDE) in the context of a given topic (Aharoni et al., 2014).

2.2 Features

We evaluate several feature types proposed in previous work (Stab and Gurevych, 2014b): *Structural features* consider statistics about tokens and punctuation. *Lexical features* capture information on unigram frequency, as well as salient verbs and adverbs. *Syntactic features* incorporate occurrences of frequent POS-Sequences. *Indicators introduce* a list of argumentative keywords. *Contextual features* take into account structural and lexical features of surrounding sentences. In terms of data preprocessing we performed lemmatization before feature extraction step but left out removing stopwords as they are relevant for determining arguments. For instance stopwords like because, therefore, etc. are indeed good indicators for argumentative text.

Each feature set is scaled to a range between 0 and 1 and normalized by tf-idf. Furthermore, we also investigated *word embeddings* as an additional feature type by using the pre-trained Google News corpus consisting of 3 million 300-dimension English word vectors [1].

2.3 Tasks

2.3.1 Detection of Argumentative Sentences

The first classification task involves identifying argumentative sentences in natural texts. This is considered as a three-class classification task, where sentences are classified as claim, premise and none. The gold standard data contains texts annotated either as premise or claim. To determine the non-argumentative sentences, which are necessary for developing a classifier to distinguish between positive and negative examples, we include sentences for which there is no annotations.

2.3.2 Prediction of Argumentative Structures

The second classification task aims to identify the relationship between claims and premises. This task is treated as a binary classification task: a claim and a premise can be in a linked or unlinked relation. All annotated pairs of claims and premises are taken as linked examples. To determine the unlinked examples we take a subset of both annotated premises and claims and calculate the cross product of these two sets.[2] The selection of negative pairs is a randomized process where repetition of single arguments are possible but not as a complete pair.

2.4 Classifiers

We investigate 8 classifiers, some of which have been used by previous studies (LinearSVC, Logistic Regression, Random Forest, Multinominal Naïve Bayes (MNB)) and some of which we implement for the first time for the above tasks: Nearest Neighbor, AdaBoosted Decision Tree (AdaBoost), Gaussian Naïve Bayes (GNB) and Convolutional Neural Networks (CNNs)). Each classifier, except the CNN, has been trained and tested on each possible combination of the six feature types.

3 Results

For each corpus we performed stratified 10-fold cross validation. The results are reported using macro F1-score.

[1] https://code.google.com/archive/p/word2vec/

[2] All linked pairs are discarded from this set.

Feature Type Combination	MNB	LinearSVC	Log. Regr.	Random Forest	AdaBoost	Near. Neigh.	GNB
Structural	.62/.6	.69/.65	.68/.65	.76/.64	.58/.64	.76/.61	.51/.58
Lexical	.41/.37	.53/.37	.53/.37	.48/.51	.39/.5	.42/.48	.48/.37
Indicators	.28/.41	.29/.44	.28/.44	.3/.47	.27/.44	.29/.42	.26/.4
Syntactic	.23/.37	.23/.37	.23/.37	.29/.37	.23/.37	.34/.37	.39/.3
Contextual	.23	.48	.48	.47	.48	.47	.48
Word Embeddings	.23/.37	.51/.45	.36/.37	.42/.42	.48/.45	.45/.44	.48/.48
All	.65/.55	**.81**/.59	.79/.62	.75/.5	.76/.58	.71/.56	.63/.43
All without Embeddings	.64/.55	.76/.63	.76/.63	.78/.65	.76/**.66**	.71/.57	.62/.43
All without Contextual	.64	.79	.76	.72	.58	.7	.63
All without Syntactic	.64/.55	.8/.59	.78/.62	.75/.51	.76/.58	.72/.56	.63/.43
All without Indicators	.64/.57	.8/.6	.78/.64	.75/.5	.76/.58	.73/.62	.7/.52
All without Lexical	.61/.55	.8/.59	.77/.62	.76/.5	.76/.58	.73/.57	.56/.43
All without Structural	.39/.43	.65/.47	.61/.45	.55/.46	.6/.49	.47/.53	.39/.41

Table 1: F1-scores of 7 classifiers for different feature combinations for the persuasive essay corpus. The results are shown as X/Y where X refers to the score for the task of detecting argumentative sentences and Y refers to the score for argument structure prediction task.

3.1 Results for Persuasive Essays

In the corpus of persuasive essays we have 3832 premise examples, 2256 claim examples and 1317 non-argumentative examples for the sentence detection task. For structure prediction task we obtained 3117 positive examples for support relations between premises and claims and 2200 negative examples for non-supporting relations.

The classification results are reported in Table 1. CNN results for both corpora are presented in Section 3.3.

For the task of argumentative sentence detection the best overall result on persuasive essays is achieved by combining all six feature sets yielding an F1-score of 81% achieved by the Linear SVC classifier. The structural features achieve the best results among the single feature types. Similar results have been also reported in (Stab and Gurevych, 2014a) for a smaller corpus of 90 persuasive essays. Also in the leave-one-out setting removing the structural features leads to the largest loss in performance. Lexical features are the next most useful feature for separating argumentative sentences from non-argumentative ones. Syntactic features are found to be least useful for this task. The performance of the classifiers based on these features only is low and removing them from a set of features does not lead to a substantial reduction in performance.

For the task of predicting the argument structure the best overall results (66%) are achieved by AdaBoost classifier based on all features without word embeddings. Table 1 indicates that the structural features are again the best performing feature set among the single ones achieving an F1-score of 65% in combination with Logistic Regression and LinearSVC. This single structural feature set even outperforms combined feature sets (excluding the ALL without Word Embeddings feature) showing that inclusion of the other feature types, in particular word embeddings lead only to noise. The other feature types all perform substantially worse than the structural feature type and their overall performance is similar.

Due to the great performance of the structural feature we computed significance test between this feature (took the best results) and all the other single features with their best performance. Results of the significance test are shown in the first two rows (after the table heading) of Table 3.

3.2 Results on Wikipedia Data

For the Wikipedia corpus we extracted 2858 premise and claim examples and 1200 non-argumentative examples for sentence detection classification task.[3] For structure prediction classification task we obtained 1232 positive examples for support relations between premises and claims and 1200 negative examples for non-supporting relations. The negative relational instances are those that bear wrong pairings. The results for the Wikipedia corpus are shown in Table 2.

Table 2 reveals that for argumentative sentence detection the structural features again achieve the best results among the single feature types and

[3] We randomly selected 1200 non-argumentative examples that were not annotated. We admit that these negative examples can still have argumentative sentences because the Wikipedia corpus contains only topic dependent claims and premises. Any claim or premise not topic related was not annotated.

Feature Type Combination	MNB	LinearSVC	Log. Regr.	Random Forest	AdaBoost	Near. Neigh.	GNB
Structural	.80/.52	.90/.54	.85/.55	.94/.55	.92/.55	.92/.56	.84/.36
Lexical	.73/.53	.81/.52	.80/.52	.85/.52	.75/.52	.66/.47	.64/.53
Indicators	.38/.47	.52/.47	.52/.47	.58/.50	.53/.54	.29/.44	.33/.36
Syntactical	.20/.33	.33/.33	.33/.33	.45/.33	.44/.33	.43/.33	.41/.33
Contextual	0.18	0.27	0.27	0.74	0.27	0.64	0.31
Word Embeddings	.20/.52	.72/.53	.64/.54	.85/.47	.76/.48	.68/.53	.61/.53
All	.92/.52	.94/.57	.93/.59	.95/.48	.92/.53	.84/.56	.88/.43
All without Embeddings	.92/.49	.93/.54	.93/.53	**.96**/.57	.93/.55	.83/.54	.85/.37
All without Contextual	0.92	0.94	0.93	0.95	0.92	0.84	0.88
All without Syntactic	.92/.53	.94/.58	.93/**.60**	.95/.5	.92/.54	.83/.55	.88/.43
All without Indicators	.92/.53	.94/.55	.93/.57	.94/.48	.92/.48	.85/.6	.9/.55
All without Lexical	.83/.49	.94/.56	.91/.58	.94/.51	.93/.51	.84/.56	.87/.47
All without Structural	.77/.5	.87/.53	.84/.53	.88/.49	.82/.51	.73/.55	.66/.47

Table 2: F1-scores of different classifiers on different feature type combinations for the Wikipedia corpus. The results are shown as X/Y where X refers to score for the task of detecting argumentative sentences and Y refers to the score for predicting argumentative structure.

Feature	Str.	Lex.	Ind.	Syn.	Con.	Emb.
Arg.	-	Y	Y	Y	Y	Y
Str.	-	Y	Y	Y	-	Y
Arg.	-	Y	Y	Y	Y	Y
Str.	-	N	Y	Y	-	N

Table 3: Significance using using Student's t-test between the structural features and the others for the essay (first 2 rows) and the Wikipedia corpus (last 2 rows). When conducting multiple analyses on the same dependent variable, the chance of achieving a significant result by pure chance increases. To correct for this we did a Bonferroni correction. Results are reported after this correction. In the cells Y means yes and N means no-significance.

lead to largest loss in performance when removed from the set of all features. The best scoring classifier is Random Forest, which based on structural features achieves an F1-score of 94%. The best overall result is achieved by random Forest classifier by combining five feature sets without word embeddings. The F1 score in this setting is 96%. As in the persuasive essay corpus, the arguments in Wikipedia corpus are also best identified using structural features. The lexical feature type gains the next best evaluation results in both single and leave-one-out feature settings. Syntactic features do not have a substantial influence in separating argumentative from non-argumentative sentences, which was also observed within the persuasive essay corpus. Overall, the scores for Wikipedia are substantially higher than those obtained for the essay corpus.

For the structure prediction task on the Wikipedia corpus Table 2 indicates that structural feature proved best feature type for argument structure prediction, achieving an F1-score of 56% in Nearest Neighbors classifier. The performance of syntactic features is the lowest, while lexical and word embedding feature types perform in general comparably to the structural features. Best results are achieved when word embeddings, lexical, indicators and structural feature types are combined leading to an F1-score of 60% in combination with Logistic Regression classifier.

Similar to the essay corpus we computed the significance test between the structural feature set with the other single feature sets. The results are shown in the last two rows of Table 3.

3.3 Results with CNN

Finally, for the purpose of detecting argumentative pieces of text as well as structure prediction we have adopted the Convolutional Neural Network (CNN) architecture described by Kim (2014), who applied it to the task of sentiment analysis. Apart from changing the inputs from sentimental sentences to argumentative pieces of text, we kept the original architecture, as well as all settings used for training as described by Kim (2014).

Table 4 shows the results of our adopted CNN classifier for both corpora. We can see the CNN has a good performance in argumentative sentence detection, it achieves an F1-score of 74% for the persuasive essay corpus and an F1-score of 75% for the Wikipedia data.[4] In terms of structure pre-

[4] Note that in case of the CNN we do not distinguish between claim, premise but rather argumentative or non-argumentative. We tried to run CNN to perform the claim, premise and none class classification however, the results

diction it leads to an F1-score of 73% for the persuasive essay corpus and 52% for the Wikipedia corpus.

Data Source	argumentative or not	structure
Essays-CNN	0.74	0.73
Wikipedia-CNN	0.75	0.52

Table 4: F1-scores of CNN on both persuasive essay and Wikipedia corpora

4 Conclusion

In this paper we presented a comparative analysis of supervised classification methods for two argument mining tasks. Specifically, we investigated six feature types proposed by previous work implemented in 8 classifiers, some of which have been proposed before and some of which were new. We addressed two argument mining tasks: (1) the detection of argumentative pieces of text and (2) predicting the structure between claims and premises. We performed our analysis on two different corpora: persuasive essays and Wikipedia articles. The most robust result in our analysis was the contribution of structural features. For both corpora and both tasks, these features were consistently the most relevant ones. Likewise, syntactic features were not useful in any of the experimental settings. The classifier performance varied across features and corpora and we did not get a robust result for one classifier consistently outperforming others. However, Random Forest classifier showed best results on the Wikipedia Corpus and results comparable to the best ones for the essays corpus. In our future work we plan to expand our investigation by including other corpora to test on as well as Recurrent Neural Networks. Also note for the final version of the paper we plan to include an extensive error analysis which we omit now due to space limitations.

Acknowledgements

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) under grant No. GRK 2167, Research Training Group "User-Centred Social Media".

were substantially lower than what is reported in Table 4.

References

Ehud Aharoni, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfreund, and Noam Slonim. 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.

Filip Boltužić and Jan Šnajder. 2015. Identifying prominent arguments in online debates using semantic textual similarity. In *2nd Workshop on Argumentation Mining (ArgMining 2015)*.

Elena Cabrio and Serena Villata. 2012. Combining textual entailment and argumentation theory for supporting online debates interactions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*. Association for Computational Linguistics, pages 208–212.

Vanessa Wei Feng and Graeme Hirst. 2011. Classifying arguments by scheme. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, pages 987–996.

Debanjan Ghosh, Smaranda Muresan, Nina Wacholder, Mark Aakhus, and Matthew Mitsui. 2014. Analyzing argumentative discourse units in online interactions. In *Proceedings of the First Workshop on Argumentation Mining*. pages 39–48.

Theodosis Goudas, Christos Louizos, Georgios Petasis, and Vangelis Karkaletsis. 2014. Argument extraction from news, blogs, and social media. In *Hellenic Conference on Artificial Intelligence*. Springer, pages 287–299.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882* .

John Lawrence, Chris Reed, Colin Allen, Simon McAlister, Andrew Ravenscroft, and David Bourget. 2014. Mining arguments from 19th century philosophical texts using topic based modelling. In *Proceedings of the First Workshop on Argumentation Mining*. Citeseer, pages 79–87.

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context dependent claim detection .

Marco Lippi and Paolo Torroni. 2015. Context-independent claim detection for argument mining. In *Proceedings of the Twenty-Fourth International Conference on Artificial Intelligence*. pages 185–191.

Marco Lippi and Paolo Torroni. 2016. Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology (TOIT)* 16(2):10.

Amita Misra, Pranav Anand, JEF Tree, and MA Walker. 2015. Using summarization to discover argument facets in online idealogical dialog. In *NAACL HLT*. pages 430–440.

Marie-Francine Moens, Erik Boiy, Raquel Mochales Palau, and Chris Reed. 2007. Automatic detection of arguments in legal texts. In *Proceedings of the 11th international conference on Artificial intelligence and law*. ACM, pages 225–230.

Huy V Nguyen and Diane J Litman. 2015. Extracting argument and domain words for identifying argument components in texts. In *Proceedings of the 2nd Workshop on Argumentation Mining*. pages 22–28.

Joonsuk Park and Claire Cardie. 2014. Identifying appropriate support for propositions in online user comments. In *Proceedings of the First Workshop on Argumentation Mining*. pages 29–38.

Chris Reed, Raquel Mochales Palau, Glenn Rowe, and Marie-Francine Moens. 2008. Language resources for studying argument. In *Proceedings of the 6th conference on language resources and evaluation-LREC 2008*. ELRA, pages 91–100.

Christos Sardianos, Ioannis Manousos Katakis, Georgios Petasis, and Vangelis Karkaletsis. 2015. Argument extraction from news. *NAACL HLT 2015* page 56.

Christian Stab and Iryna Gurevych. 2014a. Annotating argument components and relations in persuasive essays. In *COLING*. pages 1501–1510.

Christian Stab and Iryna Gurevych. 2014b. Identifying argumentative discourse structures in persuasive essays. In *EMNLP*. pages 46–56.

Christian Stab and Iryna Gurevych. 2016. Parsing argumentation structures in persuasive essays. *arXiv preprint arXiv:1604.07370* .

Reid Swanson, Brian Ecker, and Marilyn Walker. 2015. Argument mining: Extracting arguments from online dialogue. In *Proceedings of 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2015)*. pages 217–227.

Adam Wyner, Jodi Schneider, Katie Atkinson, and Trevor JM Bench-Capon. 2012. Semi-automated argumentative analysis of online product reviews. *COMMA* 245:43–50.