On Selecting Training Corpora for Cross-Domain Claim Detection

Robin Schaefer and **René Knaebel** and **Manfred Stede**
Applied Computational Linguistics
University of Potsdam
14476 Potsdam, Germany
{robin.schaefer|rene.knaebel|stede}@uni-potsdam.de

Abstract

Identifying claims in text is a crucial first step in argument mining. In this paper, we investigate factors for the composition of training corpora to improve cross-domain claim detection. To this end, we use four recent argumentation corpora annotated with claims and submit them to several experimental scenarios. Our results indicate that the "ideal" composition of training corpora is characterized by a large corpus size, homogeneous claim proportions, and less formal text domains.

1 Introduction

In the last decade, argument mining (AM) has grown into a fruitful area of research (Stede and Schneider, 2018; Lawrence and Reed, 2020). While early studies tended to focus on the annotation and detection of argument components in edited text domains (Levy et al., 2014), more recently the field progressed in different new directions. This includes intensified work on social media texts such as Twitter, e.g. by Schaefer and Stede (2022), argument quality assessment (Wachsmuth et al., 2017) and the identification of argumentation strategies (Al-Khatib et al., 2017).

Arguments consist of several components, and their identification is traditionally split into several subtasks, such as detecting argumentative text segments, specifying their function, and finding the relations among them. Given that a claim is the central component of an argument, claim detection often constitutes a crucial part of an AM pipeline.

In this paper, we combine work in claim detection with recent advances in learning contextualized word representations in order to study cross-domain claim detection on the following set of recent English argumentation corpora: Change My View (CMV) posts (Hidey et al., 2017), persuasive essays (Stab and Gurevych, 2017), micro texts (Peldszus and Stede, 2015) and political US debates (Haddadan et al., 2019). We selected them for achieving variation in genre or register, formality level, and topic. In principle, these dimensions should be distinguished, but for present purposes we do not study them separately, and thus we follow the common practice to use "domain" as an unspecific cover term. Ultimately, we are interested in investigating the "ideal" composition of a training corpus for detecing claims in new domains or corpora.

The paper is structured as follows. In Section 2, we present relevant related work. In Section 3, we describe the used corpora, and we outline our methods in Section 4. We present our results in Section 5 and provide a discussion in Section 6.

2 Related Work

Early work on claim detection was presented by Levy et al. (2014), who introduced the concept of context-dependent claims for finding claims that are relevant for a particular predetermined topic and context. Based on this idea, Lippi and Torroni (2015) proposed an approach to more general topic-independent claim detection, where the context of the argumentation was not given to the detection model as input.

Haddadan et al. (2019) focus on political debates and approach argument detection with the two subtasks of identifying argumentative sentences and subsequent classification of claims and premises. They report 0.84 F1 and 0.67 F1 scores for both tasks, respectively. Our work differs from their study by only focusing on claims and classifying them directly, i.e., against "all other" material.

Stab and Gurevych (2017) propose models for argument role classification with mostly handcrafted features. Later, in their work on a large heterogeneous corpus of argumentive sentences, Stab et al. (2018) develop an LSTM cell that incorporates topic information in the process of sentence-level claim detection. They demonstrate the beneficial effect of this additional information of about 0.05

Proceedings of the 9th Workshop on Argument Mining, co-located with the 29th International Conference on Computational Linguistics, pages 181–186
October 17, 2022.

Corpus	Domain	Type	#Docs	#Sentences	#Claims
CMV	web	monologue	107	3966	1356 (34%)
Essay	student essays	monologue	402	6743	2108 (31%)
Micro	various	monologue	112	451	112 (25%)
USDEB	politics	dialogue (spoken)	42	38309	14418 (38%)

Table 1: Overview of studied corpora.

F1 score compared to LSTM cells without topic information. Reimers et al. (2019) build on top of previously proposed recurrent architectures and successfully examine the positive influence of different contextualized word embeddings on the task of classifying argument components.

Daxenberger et al. (2017) investigate cross-domain claim identification in order to shed light on differences and similarities in claim conceptualizations across domains. They utilize linguistic feature-based and neural approaches (with and without pre-trained word embeddings). Their study is a direct precursor of our work—we use some more recent data, and in addition, incorporate recent contextualized word embeddings that serve as input for our recurrent neural network classifier. For claim detection, their best feature-free models report 0.62 F1 and 0.67 F1 for persuasive essays and micro texts, respectively.

3 Data

For determining factors influencing claim detection, we chose four English argumentation corpora of varying register (monologue and dialogue), formality level (written text and transcribed speeches), and topics. See Table 1 for statistics.

To facilitate the task, we also ensured that our corpora have less variety in claim proportions than those used by Daxenberger et al. (2017). All our corpora contain further annotations, e.g., premises, but we only use claim annotations in this study.

CMV. Hidey et al. (2017) annotate claims, premises and semantic types of argument components on the Change My View corpus from Tan et al. (2016), reporting an IAA of 0.63 for claims. We segment this user-generated data into 3966 sentences. 34% sentences contain a claim.

Essay. The corpus of argumentative essays (Stab and Gurevych, 2017) consists of 402 persuasive essays annotated for three argument components (major claim, claim, and premise) and their relations (support and attack). Annotators achieved IAA scores of 0.88 and 0.64 for major claims and claims, respectively. All argument components are annotated on clause level. We combine *major claim* and *claim* into one single claim class. After sentence splitting, we obtain 6743 units, 31% of which contain a claim.

Micro. The argumentative microtext corpus (Peldszus and Stede, 2015) was developed in a controlled setting where participants created short texts containing a single argument. Annotators then built a complete argumentation graph per text, the agreement was 0.83. Texts were originally written in German and then professionally translated to English. We work on this version; it consists of 451 sentences, 25% of which contain a claim.

USDEB. The USElecDeb60To16 corpus of Haddadan et al. (2019) is a collection of transcripts of political TV debates between 1960 and 2016. Annotators labeled argumentativeness of sentences and sentences containing argument components, i.e., claim and premise. They achieved an IAA for component annotation of 0.40, which indicates the challenge for analyzing spoken language of this kind. This is the only corpus in our set where the number of claims exceeds those of premises. After sentence splitting, we obtain 38309 sentences. 38% contain a claim.

To account for potential positional effects, we calculated percentages of claim positions by dividing a sentence into three equal parts on a token basis: beginning, middle and ending. A claim could potentially occur in individual parts or the combination of beginning and middle, middle and ending or all three parts. The percentages show that for the vast majority of sentences containing a claim, the claim occurs in all three parts. Only in 1%-4% of sentences do the claims occur in two parts. See Table 2 for details.

4 Method

For preprocessing, we perform tokenization and sentence segmentation with the Trankit toolkit (Nguyen et al., 2021). Following Daxenberger et al. (2017), we label a sentence as a claim if any token

Corpus	B & M	M & E	B & M & E
CMV	3%	2%	26%
Essay	2%	4%	25%
Micro	2%	1%	20%
USDEB	3%	3%	30%

Table 2: Claim position percentages of combined sentence parts (B=Beginning; M=Middle, E=Ending). The percentages refer to full corpus size.

within the sentence is part of a claim. However, note that this may lead to some imprecision in classification, as sentences with a claim may contain additional premises or non-argumentative parts. To study this potential issue we additionally experimented with *elementary discourse units* (EDU) (Mann and Thompson, 1988) replacing sentences as the unit of classification. For EDU identification, we use an end-to-end neural segmentation approach proposed by Wang et al. (2018) that works on already-split sentences. We adopt the previously described mapping for sentences and label individual EDUs containing at least one token referring to a claim as positive training instances. In this step, a single claim might be split into two separate discourse units, which increases the number of training instances. In general, classifying EDUs instead of full sentences is more precise, since the proportion of positive labels within a positively labeled instance is higher than on the sentence level.

We conduct one in-domain and four cross-domain experiments in order to identify promising scenarios for claim detection:

1. Train and test models on single corpora (in-domain; S1).
2. Train and test models on the union of all four corpora (S2).
3. Utilize the same test sets as in S2 but train only on three corpora, which allows us to identify the effects of removing individual corpora (S3).
4. Adopt a leave-one-out approach by training across three corpora and testing on the remaining one (S4).
5. Allow for pair comparisons by training on individual corpora and testing on a different one (S5).

We apply 10-fold cross-validation and compute the average model performance in all experiments.

		Claim Class			Macro
		F1	P	R	F1
S1)	CMV	0.72	0.74	0.69	0.79
	Essay	0.67	0.70	0.64	0.76
	Micro	0.73	0.82	0.69	0.82
	USDEB	0.73	0.75	0.71	0.78
S2)	All Corpora	0.72	0.72	0.72	0.78
S3)	No CMV	0.71	0.71	0.71	0.77
	No Essay	0.71	0.70	0.71	0.77
	No Micro	0.72	0.71	0.73	0.78
	No USDEB	0.49	0.72	0.37	0.65
S4)	CMV	0.46	0.56	0.39	0.62
	Essay	0.55	0.56	0.55	0.67
	Micro	0.59	0.52	0.68	0.71
	USDEB	0.37	0.75	0.25	0.58

Table 3: Results for experiments (except S5). *In-Domain* (S1): Training, validating and testing within a single domain. *Cross-Domain*: S2) Training/validating and testing on union of all corpora; S3) Training/validating with all except the mentioned corpus and testing with the same 4-corpora sets as in S2; S4) Training/validating with three corpora and testing with the mentioned corpus (leave-one-out).

For S1 (in-domain) and S2, we reserve 10% of the data for validation and testing, respectively. In S3, however, the validation set is <10% while the test set is larger given that we use the same test sets for S3 as for S2 while removing individual corpora from the training and validation sets. In S4 (leave-one-out) and S5 (pair comparison), 20% of the training corpora are used for validation while the whole respective testing corpus is used for testing.

Our classification pipeline was implemented using the FLAIR framework (Akbik et al., 2019), which offers a simple interface for training BERT-related models (Devlin et al., 2019), among others. We use a simple recurrent neural network on top of context-sensitive embeddings extracted using RoBERTa (Liu et al., 2019). In particular, we make use of *roberta-argument* (Stab et al., 2018), which was pre-trained on roughly 25,000 sentences annotated for +/- argumentative. The last hidden state is finally processed by a linear layer with softmax activation. The full neural network, including the pre-trained RoBERTa embeddings, is updated during training. In addition, we trained models using the classic *base-cased* BERT model (Devlin et al., 2019) for comparison.

Pair (Train - Test)	Claim Class F1	P	R	Macro F1	Pair (Train - Test)	Claim Class F1	P	R	Macro F1
CMV - Essay	0.55	0.49	0.63	0.65	Micro - CMV	0.13	0.24	0.09	0.43
CMV - Micro	0.45	0.38	0.57	0.60	Micro - Essay	0.11	0.50	0.06	0.46
CMV - USDEB	0.47	0.66	0.37	0.62	Micro - USDEB	0.30	0.36	0.28	0.46
Essay - CMV	0.21	0.66	0.13	0.51	USDEB - CMV	0.54	0.53	0.55	0.64
Essay - Micro	0.33	0.61	0.24	0.60	USDEB - Essay	0.57	0.52	0.65	0.67
Essay - USDEB	0.22	0.86	0.13	0.50	USDEB - Micro	0.57	0.44	0.82	0.67

Table 4: Results for corpus pair experiments (S5). Models were trained and validated on the first corpus and tested on the second corpus.

5 Results

All results presented in this section are produced with the RoBERTa architecture trained on sentence units. We conducted additional experiments with BERT models and with EDUs, which on the whole lead to worse results. For EDUs, this is especially the case for the claim class, which we are particularly interested in. We will discuss this briefly in Section 6. In the following, we report macro F1 scores and F1, precision, and recall for the claim class. See Table 3 for result of S1-S4 and Table 4 for results of S5.

5.1 In-Domain

S1 shows good results for all corpora. The best macro F1 score was achieved for the Micro corpus (0.82). However, the less formal CMV and USDEB still come relatively close. F1 scores for the claim class are considerably lower, which is to be expected, as it is the smaller class for all corpora.

5.2 Cross-Domain

Models trained and tested across all four corpora (S2) yield results comparable to S1. Removing the CMV, Essay, or Micro corpus from the training set while still testing on all corpora (S3) does not influence results. However, removing the USDEB corpus reduces the recall of the claim class, which leads to a drop in F1.

Leave-one-out experiments (S4) show mixed results. Best results were achieved when the Micro corpus was not part of the training set (macro F1: 0.71; claim F1: 0.59). Testing on the Essay corpus also works comparatively well. Results obtained from removing the USDEB corpus from the training set, however, are low (macro F1: 0.58; class F1: 0.37).

S5 (pair comparison; Table 4) shows substantial variance, especially with respect to the claim class results. Models trained on USDEB yield the most robust results with comparatively little variance in F1 scores. Models trained on CMV show the best results when tested with the Essay corpus. In comparison, Essay and Micro perform worse as training corpora. While models trained on the Essay corpus yield the best results when tested with the Micro corpus, all pairs show low results for the claim class (F1: 0.21-0.33). The lowest results occur for the Micro corpus with F1 scores of 0.11-0.30 for the claim class.

6 Discussion & Conclusion

As noted above, our BERT and EDU results cannot compete with the sentence-level RoBERTa results. We surmise that RoBERTa may have outperformed BERT as it was pre-trained on an argument detection task; likewise, since it was trained on sentences, EDU performance may be lower.

While being a potentially interesting factor, we argue that the claim position in a sentence does not substantially affect our results. Statistics on claim position show that claims in the vast majority of claim sentences occur in the beginning, middle, and ending, i.e. they cover more than 66% of tokens in a sentence. Only in 1%-4% of sentences of a given corpus, claims merely occur in the beginning and middle or middle and ending, i.e. they cover a span of 34%-66% tokens in a given sentence. Of course, this does not mean that position cannot have an effect in general, and justifies more research in the future.

In contrast, our results suggest that different factors influence the choice of a suitable corpus for training claim detection models. First, corpus size seems to play a crucial role. This is especially the

case when several corpora are combined for training. Removing individual corpora from the training set while testing on all corpora (S3) shows that only the removal of the largest corpus (USDEB) has a profound effect on the results. Our leave-one-out experiments (S4) confirm this finding, as models trained on all corpora except USDEB obtain worse results than models trained in other leave-one-out scenarios. Also, training on USDEB in a corpus pair scenario (S5) consistently yields good results, indicating that a large training size has a beneficial effect, while training on the small Micro corpus yields the worst results.

Second, although claim proportions vary less in our corpora than in those used by Daxenberger et al. (2017), differences in claim proportions may still have an effect. For instance, while USDEB is the largest corpus in our set, it also contains the highest proportion of claims, which may render it difficult for models trained on corpora with lower claim proportions to sufficiently capture the class distribution in USDEB. Still, it appears that size effects outweigh claim proportion effects given that claim detection results improve when the Micro corpus is left out for training, which is the corpus that is both the smallest and the one with the lowest claim proportion.

Third, our results suggest that domain plays a role. Recall that the Essay and Micro corpora represent relatively "edited" text types, while CMV and USDEB contain web data and oral debates, which can be described as less formal. This may affect the way argumentation takes place. Our corpus pair experiments show that models trained on the less formal CMV and USDEB yield better results than models trained on Essay and Micro. Note that corpus size does not explain this pattern given that the CMV corpus is smaller than the Essay corpus.

Conclusion. In this paper we present several experiments to investigate cross-domain claim detection. Our results indicate that corpus size, differences in claim proportions, and content domain influence the composition of an effective training corpus. We argue that a large training set size, homogeneous claim proportions, and less formal language improve the results, and we plan to investigate this in further experiments that examine the broad notion of "domain" more closely and consider factors like monologue/dialogue or formality level as separate dimensions. Also, we plan to extend the work to premise detection and thus move closer to "full" arguments. Finally, we are interested in investigating the effect of claim position in units larger than sentences, for instance by using sequence labeling techniques.

Acknowledgements

This research has been supported by the German Research Foundation (DFG) with grant number 455911521, project “LARGA” in SPP “RATIO”. We would like to thank the anonymous reviewers for their valuable feedback.

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, Minneapolis, Minnesota. Association for Computational Linguistics.

Khalid Al-Khatib, Henning Wachsmuth, Matthias Hagen, and Benno Stein. 2017. Patterns of argumentation strategies across topics. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1351–1357, Copenhagen, Denmark. Association for Computational Linguistics.

Johannes Daxenberger, Steffen Eger, Ivan Habernal, Christian Stab, and Iryna Gurevych. 2017. What is the essence of a claim? cross-domain claim identification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2055–2066, Copenhagen, Denmark. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Shohreh Haddadan, Elena Cabrio, and Serena Villata. 2019. Yes, we can! mining arguments in 50 years of US presidential campaign debates. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4684–4690, Florence, Italy. Association for Computational Linguistics.

Christopher Hidey, Elena Musi, Alyssa Hwang, Smaranda Muresan, and Kathy McKeown. 2017. Analyzing the semantic types of claims and premises in an online persuasive forum. In *Proceedings of the 4th Workshop on Argument Mining*, pages 11–21,

Copenhagen, Denmark. Association for Computational Linguistics.

John Lawrence and Chris Reed. 2020. Argument Mining: A Survey. *Computational Linguistics*, 45(4):765–818.

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context dependent claim detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1489–1500, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Marco Lippi and Paolo Torroni. 2015. Context-independent claim detection for argument mining. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, page 185–191. AAAI Press.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text & Talk*, 8:243 – 281.

Minh Van Nguyen, Viet Dac Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen. 2021. Trankit: A lightweight transformer-based toolkit for multilingual natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 80–90, Online. Association for Computational Linguistics.

Andreas Peldszus and Manfred Stede. 2015. An annotated corpus of argumentative microtexts. In *Argumentation and Reasoned Action: Proceedings of the 1st European Conference on Argumentation*, volume 2, pages 801—-816, Lisbon. College Publications, London.

Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2019. Classification and clustering of arguments with contextualized word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 567–578, Florence, Italy. Association for Computational Linguistics.

Robin Schaefer and Manfred Stede. 2022. GerCCT: An annotated corpus for mining arguments in german tweets on climate change. In *Proceedings of the Language Resources and Evaluation Conference*, pages 6121–6130, Marseille, France. European Language Resources Association.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. 2018. Cross-topic argument mining from heterogeneous sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3664–3674, Brussels, Belgium. Association for Computational Linguistics.

Manfred Stede and Jodi Schneider. 2018. *Argumentation Mining*, volume 40 of *Synthesis Lectures in Human Language Technology*. Morgan & Claypool.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016*, pages 613–624. ACM.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187, Valencia, Spain. Association for Computational Linguistics.

Yizhong Wang, Sujian Li, and Jingfeng Yang. 2018. Toward fast and accurate neural discourse segmentation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 962–967, Brussels, Belgium. Association for Computational Linguistics.