IMHO Fine-Tuning Improves Claim Detection

Tuhin Chakrabarty
Dept. Of Computer Science
Columbia University
tc2896@columbia.edu

Christopher Hidey
Dept. Of Computer Science
Columbia University
ch3085@columbia.edu

Kathleen McKeown
Dept. Of Computer Science
Columbia University
kathy@cs.columbia.edu

Abstract

Claims are the central component of an argument. Detecting claims across different domains or data sets can often be challenging due to their varying conceptualization. We propose to alleviate this problem by fine tuning a language model using a Reddit corpus of 5.5 million opinionated claims. These claims are self-labeled by their authors using the internet acronyms IMO/IMHO (in my (humble) opinion). Empirical results show that using this approach improves the state of art performance across four benchmark argumentation data sets by an average of 4 absolute F1 points in claim detection. As these data sets include diverse domains such as social media and student essays this improvement demonstrates the robustness of fine-tuning on this novel corpus.

1 Introduction

Toulmin's influential work on argumentation (2003) introduced a claim as an *assertion that deserves our attention*. More recent work describes a claim as *a statement that is in dispute and that we are trying to support with reasons* (Govier, 2010). While some traits of claims are defined by their context, such as that claims usually need some support to make up a 'complete' argument (e.g., premises, evidence, or justifications), the exact definition of a claim may vary depending on the domain, register, or task. Daxenberger et al. (2017) try to solve the problem of claim conceptualization by training models across one data set and testing on others, but their cross-domain claim detection experiments mostly led to decreased results over in-domain experiments.

To demonstrate that some properties of claims are shared across domains, we create a diverse and rich corpus mined from Reddit and evaluate on held out datasets from different sources. We use Universal Language Model Fine-Tuning (ULMFiT) (Howard and Ruder, 2018), which pre-trains a language model (LM) on a large general-domain corpus and fine-tunes it on our Reddit corpus before training a final classifier to identify claims on various data sets.

We make the following contributions:

- We release a dataset of 5.5 million opinionated claims from Reddit,[1] which we hope will be useful for computational argumentation.
- We show transfer learning helps in the detection of claims with varying definitions and conceptualizations across data sets from diverse domains such as social media and student essays.
- Empirical results show that using the Reddit corpus for language model fine-tuning improves the state-of-the-art performance across four benchmark argumentation data sets by an average of 4 absolute F1 points in claim detection.

2 Related Work

Argumentation mining (AM) is a research field within the growing area of computational argumentation. The tasks pursued within this field are highly challenging and include segmenting argumentative and non-argumentative text units, parsing argument structures, and recognizing argumentative components such as claims- the main focus of this work. On the modeling side, Stab and Gurevych (2017) and Persing and Ng (2016) used pipeline approaches for AM, combining parts of the pipeline using integer linear programming (ILP). Eger et al. (2017) proposed state-of-art sequence tagging neural end-to-end models for AM. Schulz et al. (2018) used multi-task learning (MTL) to identify argumentative components,

[1] https://bitbucket.org/tuhinch/imho-naacl2019

Proceedings of NAACL-HLT 2019, pages 558–563
Minneapolis, Minnesota, June 2 - June 7, 2019. ©2019 Association for Computational Linguistics

challenging assumptions that conceptualizations across AM data sets are divergent and that MTL is difficult for semantic or higher-level tasks.

Rosenthal and McKeown (2012) were among the first to conduct cross-domain experiments for claim detection. However they focused on relatively similar data sets like blog articles from LiveJournal and Wikipedia discussions. Al-Khatib et al. (2016), on the other hand, wanted to identify argumentative sentences through cross-domain experiments. Their goal was, however, to improve argumentation mining via distant supervision rather than detecting differences in the notions of a claim. Daxenberger et al. (2017) showed that while the divergent conceptualization of claims in different data sets is indeed harmful to cross-domain classification, there are shared properties on the lexical level as well as system configurations that can help to overcome these gaps. To this end they carried out experiments using models with engineered features and deep learning to identify claims in a cross-domain fashion.

Pre-trained language models have been recently used to achieve state-of-the-art results on a wide range of NLP tasks (e.g., sequence labeling and sentence classification). Some of the recent works that have employed pre-trained language models include ULMFiT (Howard and Ruder, 2018), ELMo (Peters et al., 2018), GLoMo (Yang et al., 2018), BERT (Devlin et al., 2019) and OpenAI transformer (Radford et al., 2018). While these models have demonstrated success on a variety of tasks, they have yet to be widely used in argumentation mining.

3 Data

As the goal of our experiments is to develop models that generalize across domains, we collect a large, diverse dataset from social media and fine-tune and evaluate on held out data sets.

3.1 Self-labeled Opinion Data Collection

In order to obtain a data set representative of claims, we need a method of automatic data collection that introduces minimal linguistic bias. We thus mine comments containing the acronyms IMO (in my opinion) or IMHO (in my humble opinion) from the social media site Reddit. IM(H)O is a commonly used acronym[2] with the only purpose of identifying one's own comment as a personal opinion. We provide some examples[3] below:

> That's virtually the same as neglect right there **IMHO**.
>
> **IMO**, Lakers are in big trouble next couple years

To use these examples for pre-training, we need only to remove the acronym (and any resulting unnecessary punctuation).

We collect Reddit comments from December 2008 to August 2017 through the pushshift.io API, resulting in 5,569,962 comments. We perform sentence and word tokenization using Spacy. We then extract only the sentence containing IMO or IMHO and discarded the surrounding text. We refer to the resulting collection of comments as the **IMHO** dataset.

3.2 Labeled Claim Data

The IMHO dataset contains no negative examples, only labeled opinions. Furthermore, opinions in this dataset may be only a claim or both a claim and a premise. As our goal is to identify claims, we thus consider four data sets from argumentation mining. As argumentation appears in both monologue and dialogue data, we choose two datasets created from student essays and two from social media. Peldszus and Stede (2016) created a corpus of German microtexts (**MT**) of controlled linguistic and rhetorical complexity. Each document includes a single argument and does not exceed five argumentative components. This corpus was translated to English, which we use for our experiments. The persuasive essay (**PE**) corpus (Stab and Gurevych, 2017) includes 402 student essays. The scheme comprises major claims, claims, and premises at the clause level. This corpus has been used extensively in the argumentation mining community. The corpus from Habernal and Gurevych (2017) includes user-generated web discourse (**WD**) such as blog posts, or user comments annotated with claims and premises as well as backings, rebuttals and refutations. Finally, Hidey et al. (2017) propose a two-tiered annotation scheme to label claims and premises and their semantic types in an online persuasive forum (**CMV**) using a sample of 78 threads from the subreddit Change My View, with the long-term goal

[2] https://reddit.zendesk.com/hc/en-us/articles/205173295-What-do-all-these-acronyms-mean-

[3] Examples have been modified to protect user privacy

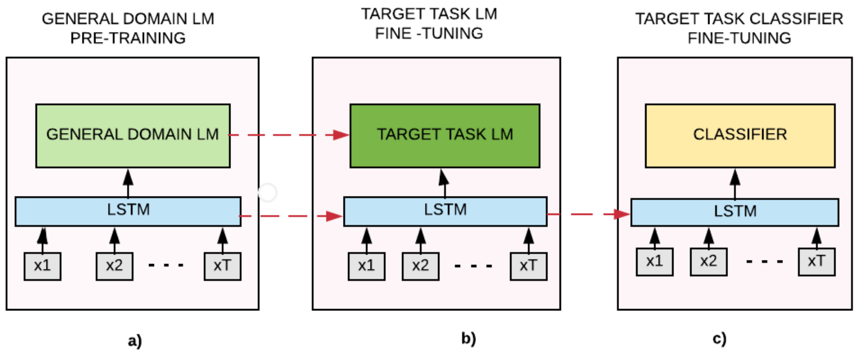

Figure 1: Schematic of ULMFiT, showing three stages. The dashed arrows indicate that the parameters from the previous stage were used to initialize the next stage.

	#Claims	#Sentences	%Claims
MT	112	449	24.94
PE	2108	7116	29.62
WD	211	3899	5.41
CMV	1206	3541	34.0

Table 1: Table showing number of claims and total number of sentences in the data sets along with the percentage of claims in them

of understanding what makes a message persuasive. As with Daxenberger et al. (2017), we model claim detection at the sentence level, as this is the only way to make all data sets compatible to each other. Table 1 gives an overview of the data.

4 Model

As the IMHO dataset is only self-labeled with claim data but does not contain non-claims, we need a method of incorporating this dataset into a claim detection model. We thus use a language model fine-tuning approach, which requires only data similar to the task of interest.

The Universal Language Model Fine-Tuning method (ULMFiT) (Howard and Ruder, 2018) consists of the following steps: a) General-domain LM pre-training b) Task-specific LM fine-tuning and c) Task-specific classifier fine-tuning. In step (a), the language model is trained on Wikitext-103 (Merity et al., 2017) consisting of 28,595 preprocessed Wikipedia articles and 103 million words capturing general properties of language. Step (b) fine-tunes the LM on task-specific data, as no matter how diverse the general-domain data used for pre-training is, the data of the target task will likely come from a different distribution. In step (c), a classifier is then trained on the target task, fine-tuning the pre-trained LM but with an additional layer for class prediction. The models all use a stacked LSTM to represent each sentence. For stages (a) and (b), the output of the LSTM is used to make a prediction of the next token and the parameters from stage (a) are used to initialize stage (b). For stage (c), the model is initialized with the same LSTM but with a new classifier layer given the output of the LSTM.

This process is illustrated in Figure 1. We refer the reader to Howard and Ruder (2018) for further details.

In our work, we maintain steps (a) and (c) but modify step (b) so that we fine-tune the language model on our IMHO dataset rather than the task-specific data. The goal of ULMFiT is to allow training on small datasets of only a few hundred examples, but our experiments will show that fine-tuning the language model on opinionated claims improves over only task-specific LM fine-tuning.

5 Results and Experiments

Table 2 show the results on the four data sets. We compare to two baselines. The numbers in the **CNN** column are taken directly from the results of the deep learning experiments mentioned in the work of Daxenberger et al. (2017). Their deep learning experiments consisted of 4 different models: a) bidirectional LSTM b) LSTM c) CNN initialized with random word embeddings and d) CNN initialized with word2vec. In their experiments for MT and PE, a CNN initialized with random word embeddings gave the best results and for WD a CNN with word2vec gave the best results. As CMV is a new data set we experimented with all four models and obtained the best result using a CNN with random initialization. The **Task-Specific LM Fine-Tuning** column

	Metric	CNN		Task-Specific LM Fine-Tuning		IMHO LM Fine-Tuning	
		Claim	Macro	Claim	Macro	Claim	Macro
WD	P	50.0	72.5	50.0	72.5	**54.0**	**75.9**
	R	20.4	59.2	20.0	59.8	**24.0**	**61.7**
	F	28.9	62.6	28.5	62.7	**33.3**	**65.2**
MT	P	66.5	79.0	66.2	78.5	**71.0**	**80.9**
	R	68.2	78.5	68.0	77.8	**71.8**	**81.4**
	F	67.3	78.6	67.0	78.1	**71.2**	**81.1**
PE	P	60.9	73.2	62.3	73.2	**62.6**	**74.4**
	R	61.2	74.0	65.8	75.1	**66.0**	**75.0**
	F	61.1	73.6	64.0	74.1	**64.3**	**74.8**
CMV	P	54.0	65.1	55.0	68.0	**55.7**	**69.5**
	R	53.0	62.5	59.0	65.0	**60.0**	**65.3**
	F	53.5	63.8	57.0	66.4	**57.8**	**67.3**

Table 2: Table showing the results on four data sets. Each cell contains the Precision (P), Recall (R) and F-score (F) for Claims as well as the Macro Precision, Recall and F-score for the binary classification.

contains the results obtained by fine-tuning the language model on each respective dataset while the **IMHO LM Fine-Tuning** column contains the results from fine-tuning the language model on IMHO. As in previous work, we report both Claim F1 and Macro F1.

The experiments were carried out in a 10-fold cross-validation setup with fixed splits into training and test data and the F1 scores are averaged over each of the folds. Each model was run 10 times to account for variance and the results reported in the table are an average of 10 runs. We use the same hyper-parameters as Howard and Ruder (2018) except for a batch size of 32 for MT and 64 for the remaining data sets. The learning rate for classifier fine-tuning is set to 0.0001. We train our classifier for 5 epochs on each data set.

We obtain statistically significant results ($p < 0.05$ using Chi Squared Test) over all CNN models trained only on the task-specific datasets. We also find that for all models, IMHO LM Fine-Tuning even performs better than Task-Specific LM Fine-Tuning, and is significantly better for the MT and WD datasets (which both contain very few claims). For the MT and WD datasets, Task-Specific LM Fine-Tuning actually performs worse than the CNN models.

6 Qualitative Analysis

To understand how using the IMHO dataset improved over the CNN and Task-Specific Fine-Tuning settings, we show examples that were incorrectly classified by the two baseline models but correctly classified by the IMHO Fine-Tuning. We retrieve the most similar example in the IMHO dataset to these misclassified samples according to TF-IDF over unigrams and bigrams. Table 3 presents the examples labeled by their dataset and the corresponding IMHO example. We find that the IMHO dataset contains n-grams indicative of claims, e.g. *can be very rewarding*, *should be taken off the market*, and *should intervene*, demonstrating that the IMHO LM Fine-Tuning learns representations of claims based on discriminatory phrases. In fact, the CMV example is almost an exact paraphrase of the IMHO example, differing only in the phrase *anecdotal evidence* compared to *my anecdotal experience*. At the same time, we find that many of the topics in these datasets occur in the IMHO dataset as well, such as *public schooling* and *licence fees*, suggesting that the language model learns a bias towards topics as well.

While empirical results indicate that IMHO Fine-Tuning helps in claim detection, we also investigated whether the language model introduces any bias towards types of claims. To this end, we also evaluated examples classified incorrectly by the model. Table 4 shows sentences that are predicted to be opinionated claims by our model but are actually non-claims. We note that a portion of these misclassified examples were premises used to back a claim which could be classified correctly given additional context. For instance, the second example from the MT data set in the table backs

Dataset	Sentence
MT	If there must be rent increases , there should also be a cap to avoid nasty surprises
MT	Video games namely FIFA in my case , can fascinate young people for hours more intensively and emotionally than any sport in the world !
PE	Last but not the least , using public transportation is much safer than using private transportation
PE	In a positive point of view , when people without jobs have hand phones that have access to the internet , they will be able to browse the net for more job opportunities
CMV	Cheating is evidence , that *something* must be wrong

Table 4: Sentences which are actually non-claims but predicted as claims by IMHO Fine-Tuning

Dataset	Sentence
WD	I send my daughter to public school but if I could afford to I would definitely send her to a nearby private school and not have to deal with lots of the problems in public schools.
IMHO	There is no telling that a private school will be better than public, that 's a parents choice, I pulled my kid from private school and went to public school that choice was made because the school we had access to was new and he excellent ratings and it was superior to the private school.
MT	That's why they should be taken off the market, unless they're unbreakable .
IMHO	Should be taken off the market.
MT	The Tv/Radio licence fee can only be required of all citizens/households equally.
IMHO	Radio 4 and Radio 6 music are pretty much worth the licence fee.
MT	Since, however, in Russia besides gas and oil only propaganda and corruption rule, the EU should intervene right away.
IMHO	Neither Russia or the EU should intervene in this case
CMV	Other than anecdotal evidence, I haven't seen anything to support this claim.
IMHO	I have personally seen no evidence to support this claim, but that's just my anecdotal experience .
PE	However, flourishing tourism in a place can be very rewarding in terms of local economy.
IMHO	It can be very rewarding.

Table 3: Sentences from each dataset and their nearest neighbor in the IMHO dataset

the claim *It would be fair to make them into an Olympic event* while the first example from the PE data set backs the claim *There is no reason that governments should hesitate to invest in public transportation, a healthy, safe and economical way of transporting*. While discriminatory phrases like *should* or *must be* and comparative statements like *much safer than* or *more ... than any* are often indicative of claims, the lack of context may lead to incorrect classifications. Language modeling with additional context sentences or jointly modeling context (e.g. by predicting relations between claims and premises) may address these errors.

7 Conclusion

We have collected a large dataset of over 5 million self-labeled opinionated claims and validated their utility on a variety of claim detection domains. Second, we demonstrate that by fine-tuning the language model on our IMHO dataset rather than each individual claim dataset, we obtain statistically significant improvement over previous state-of-the-art performance on each of these datasets on claim detection. Finally, our empirical results and error analysis show that there are features indicative of claims that transfer across data-sets.

In the future, we plan to expand this work beyond single sentences as the data-set for LM Fine-Tuning used in our experiments consists of sentences containing IM(H)O without additional context. We plan to experiment with modeling the context sentences from Reddit as well by using models such as BERT (Devlin et al., 2019), which perform well on pair classification tasks, as the fine-tuning step rather than ULMFiT. As BERT pre-training includes a next-sentence prediction task, we expect this model to be effective for modeling argumentative context and to be beneficial for predicting premise or justifications for these claims and the relations between these argumentative components.

Acknowledgements

The authors thank Johannes Daxenberger for sharing the sentence level data from their previous experiments and Smaranda Muresan for her insightful discussions on improving the contributions of the paper as well as thoughtful ideas on how to conduct a good error analysis. The authors also thank Olivia Winn for suggesting the paper title and the anonymous reviewers for helpful comments.

References

Khalid Al-Khatib, Henning Wachsmuth, Matthias Hagen, Jonas Kohler, and Benno Stein. 2016. Domain mining of argumentative text through distant supervision. In *15th Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1395–1404.

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.

Jacob Devlin, Ming-Wei Chang, Lee Kenton, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 17th Annual Meeting of the North American Association for Computational Linguistics*.

Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. 2017. Neural end-to-end learning for computational argumentation mining. In *In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics.*, pages 11–22.

Trudy Govier. 2010. *A Practical Study of Argument*. 7th edition Cengage Learning, Wadsworth.

Ivan Habernal and Iryna Gurevych. 2017. Argumentation mining in user-generated web discourse. In *Computational Linguistics, 43(1)*, pages 125–179.

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

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, pages 328–339.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. In proceedings of the international conference on learning representations.

Andreas Peldszus and Manfred Stede. 2016. An annotated corpus of argumentative microtexts. In *In Argumentation and Reasoned Action: Proceedings of the 1st European Conference on Argumentation*, pages 810–815.

Isaac Persing and Vincent Ng. 2016. End-to-end argumentation mining in student essays. In *In Proceedings of the 15th Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.*, pages 1384–1394.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Sara Rosenthal and Kathy McKeown. 2012. Detecting opinionated claims in online discussion. In *Inproceedings of the Sixth IEEE International Conference on Semantic Computing (IEEE ICSC2012) Special Session on Semantics and Sociolinguistics in Social Medi*.

Claudia Schulz, Steffen Eger, Johannes Daxenberger, Tobias Kahse, and Iryna Gurevych. 2018. Multi-task learning for argumentation mining in low-resource settings. In *Proceedings of NAACL-HLT 2018*, pages 35–41.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. In *Computational Linguistics, pages in press, preprint available at arXiv:1604.07370.*

Stephen E. Toulmin. 2003. *The Uses of Argument*. Cambridge University Press, New York.

Zhilin Yang, Jake Zhao, Bhuwan Dhingra, Kaiming He, William W. Cohen, Ruslan Salakhutdinov, and Yann LeCun. 2018. Glomo: Unsupervisedly learned relational graphs as transferable. In *arXiv:1806.05662*.