DeInStance: Creating and Evaluating a German Corpus for Fine-Grained Inferred Stance Detection

Anne Göhring, Manfred Klenner, Sophia Conrad
Department of Computational Linguistics
University of Zurich, Switzerland
goehring,klenner,conrad@cl.uzh.ch

Abstract

We introduce **deInStance**, a corpus of 1000 politicians' answers in German (**de**) containing sentences labeled with explicitly expressed and **in**ferred **stance**s - pro and con relations - by 3 annotators. They achieved an acceptable inter-rater agreement given the inherent subjective nature of the task. A first baseline, a fine-tuned BERT-based token classifier, achieved F_1-scores of around 70% . Our focus is on the difficult subclass of sentences comprising only non-polar words, but still with an (implicit) pro or con perspective of the writer.

1 Introduction

When people are asked about their position with regard to a certain topic, they typically answer by elaborating an argumentation in favor or against this topic. Argument mining is concerned with the structure of such arguments and the classification of each part. This happens at the clause level: a clause might be a claim, the support of a claim, etc. But what about the entities and events within the different parts of a clause? What can the reader infer from the writer's perspective on these different subtopics contained in the author's text? In this work, we present a new resource together with a first neural sequence labeling model of such inferred fine-grained stances. The goal is to find all those entities (called targets) in a text that the writer approves (*pro*) or disapproves (*con*), explicitly or implicitly. These targets might be aspects of the overall topic of the text, but also any entity mentioned in the text towards which the writer seems to bear a positive or negative attitude. Among these non-aspect targets are entities reflecting the writer's moral convictions, political views, and all sorts of other preferences.

2 Corpus Annotation

As a starting point, we took the German part of the freely available xstance[1] corpus. The original corpus contains politicians' stances consisting of an explicit position (from strongly/weakly against to weakly/strongly in favor) together with a comment as answer to given questions from different topics. We annotated a subcorpus[2] of 1000 answers where each word receives a label: *pro* (in favour of), *con* (against) or *none* (neutral). A *pro* relation indicates that the writer approves (i.e. is in favor of) the denoted entity or event; correspondingly for *con*.[3]

2.1 Objective of the Annotation

In order to clarify the annotation task and to show the differences to aspect-based sentiment analysis, take the following question: *Do you support the introduction of minimum wage for employees?* One answer is: *Another unhelpful blanket proposal from the mothballs of socialism, which would further weaken our country's competitiveness*. Socialism (text author is against it) and the competitiveness (author is in favour) are somehow related to the question, but they are not aspects in the sense of aspect-based sentiment analysis. Aspects are strongly correlated categories of an item (e.g. the price of a product). Another writer puts it in the following way: *This would be the breaking of a promise*. The author is against such a *break of a promise*, which again is not an aspect (of *minimum wage*). This characterization of our setting shows that we cannot reduce the task to a mere aspect-based sentiment analysis.

[1] https://github.com/ZurichNLP/xstance (Vamvas and Sennrich, 2020)

[2] The data is available on request.

[3] We only labeled the dependency heads of the corresponding target phrases with pro's and con's, all other words are *none*.

2.2 Annotation Guidelines

Our annotation guidelines are brief. Annotate those *pro/con* relations of the text author that (1) explicitly state his/her stance or (2) implicitly bears or must bear towards the entities mentioned in his/her comment. It is crucial to be aware that the borderline between these cases sometimes is fuzzy. When does an opinion starts to become stated explicitly? We thus decided not to annotate the implicit/explicit distinction. We just annotated the writer's attitude : *pro* or *con*.

Guideline (1) is plain. Given *I have welcomed the liberalizations that have been implemented*, there is a *pro* relation of the writer towards *liberalizations*. There are a number of linguistic indicators for an explicit assertion of stance:

1. a personal statement (first person pronoun) with a verb of (dis)approval: *I approve it*
2. predicative statements: *Liberalization is good*
3. modal constructions: *Liberalization should be carried out*
4. verb-based inference schemata: *It prevents a solution to other problems*

(1) most explicitly states that the writer is in favor of *it* and the positive evaluation in (2) immediately gives rise to a *pro* relation of the writer towards liberalization. (3) expresses the need to have liberalization. This again points out that he/she is in favor of liberalization. In (4), *prevent* casts a *con* relation between its logical subject (*it*) and the theme role (*solution*). A contra relation towards a positively connotated theme indicates a negative subject and suggests that the writer stands in *con* relation towards it (here *it*).

The second annotation objective is concerned with relations that are not directly asserted or stated by the linguistic means from above, but either must hold as a kind of presupposition or do hold because they follow from some conventional pragmatic reasoning. Take the following examples:

1. *After liberalization the employees are paid even less*
2. *The quality of education should not depend on the income*
3. *This is what the constitution says*

The pragmatically used particle *even* in (1) together with world knowledge (less pay is bad) indicates that the writer regarded it as negative, if the employees got less money. This, in turn, means that she/he must be (maybe only in a situation-specific way) in favor of the employees - not a particular subset of but the group of employees in general. He/She cares about their situation. Also, she/he is against the mentioned liberalization, which is not explicitly stated but inferred.

(2) is a response to the following question: Should the government be more committed to equal educational opportunities? Only if he/she is in favor of education, the answer can be understood as an approval: education must be one of his/her values. However, there is no *pro* nor *con* relation with respect to *income*.

The question underlying the answer in (3) is: Should the government increase its support for non-profit housing construction? The comment (just this sentence) is an example of an implicature trigger. We cannot give the whole implicature chain, but in principle it goes like this: The constitution is in favor of it, I, the writer, cite this authority and it thus is an authority of mine and I hereby indicate that I am in favor of it as well. Thus, the writer can be understood as being in favor of the constitution, this is the annotation goal here.

Such attitudes depend on the subjective understanding and reasoning of the annotators. However, it is a worthy goal to not only be able to identify the writer's directly stated stance, but also to fix her obligations, values, preferences that become visible in what is semantically/pragmatically implied.

2.3 Annotation Results

The 1000 comments containing 32,274 tokens in 2183 sentences were manually labeled by 3 trained raters. We performed independent harmonization at various progress points, each annotator checking the differences between the others' annotation and their own, adjusting it if needed.

As a simple concrete example, the sentence *'In the long term, Switzerland belongs to the EU.'* is labeled by all three annotators as:

Langfristig	*gehört*	*die*	*Schweiz*	*zur*	***EU***
none	none	none	none	none	**pro**

Although annotators A1 and A3 tend to label more tokens (around 12.5%) than A2 (10.5%), our annotations are sparse. The proportion of *pro* and *con* labels is approx 70-75% and 25-30%, respectively (see Table 1). This imbalance probably deteriorates the results for the *con* label.

To evaluate the reliability of our annotations, we calculate Cohen's kappa for the agreement and

annotator	pro	con	none
A1	2986	1132	28156
A2	2412	974	28888
A3	2870	1141	28263

Table 1: Label distributions for each annotator (A1-A3). Tokens are either labeled as *pro* (in favor of) or as *con* (against), or they are not (none).

Krippendorff's alpha for the disagreement between the different raters. On the whole corpus, the inter-rater reliability measured by Krippendorff's alpha is above the acceptability threshold of 0.667. The pairwise kappa coefficients show a higher agreement between annotators A1 and A3 (0.8578); annotators A1 and A2 disagree most (0.7229).

3 Experiments

Attention-based models are the current architecture of choice for many natural language processing tasks. For training the stance labeling models, we used the self-attentional transformer (Vaswani et al., 2017) implementation provided by HuggingFace (Wolf et al., 2020): the class BertForTokenClassification is defined as a token classification model on top of a language model, i.e. a linear classification layer on top of the tokens' hidden state output. We chose the pretrained German BERT model from DBMDZ[4] to train our models.

3.1 Configurations

The experiment settings vary for the datasets used, but the model parameters are fix throughout the runs (see Appendix A). On the data configuration side, we take each rater's labeled dataset separately and mix these annotations in various ways:

- Major: majority label per token
- Inter: intersection label (same or *none*)
- Concat: concatenation of all annotations

The setting *Major* means that we took those annotations that two or all raters have tagged, whereas in *Inter* only those are taken that all raters have selected. To simulate a weighted average, we also simply concatenated the labeled data from the three annotators to form one larger *Concat* set. We trained models also with the individual annotations (models M1-M3) in order to see whether the annotations are reasonable (i.e. reproducible).

[4]https://huggingface.co/dbmdz/bert-base-german-cased

3.2 Results

All our models achieve modest though reasonable F_1-scores given the challenging task. To mitigate the anecdotal character of a single evaluation, we randomly shuffle the annotated comments into 10 different dataset splits, and run the training and evaluation on each split (cross-validation). For instance, given the annotations of annotator $A1$, we trained a model (called M1) on a train set split, used it to predict labels for the test set split and evaluated this with respect to the annotations of $A1$ for that test set split (see Table 2).

model	acc	F_1	pro		con	
			prec	rec	prec	rec
M1	93.2	69.2	70.0	71.0	67.8	63.7
M2	93.6	66.4	66.3	68.5	65.5	63.4
M3	93.4	69.7	70.0	71.0	68.5	66.2
Major	93.7	**70.4**	70.5	**73.0**	68.3	66.3
Inter	**94.4**	64.2	64.5	67.2	63.1	58.1
Concat	93.4	**70.4**	**71.3**	71.1	**70.0**	**67.3**
C_{fair}	93.1	68.1	68.3	71.7	64.8	62.5

Table 2: Accuracy, F_1, precision, and recall results of the different models: models for individual annotators (M1-M3), majority (Major), intersection (Inter), and concatenation (Concat and C_{fair}).

On average, these baseline models attain an overall accuracy of 93-94%, achieve better precision and recall for *pro* than for *con* labels, from the lowest *con* recall of 58% to the highest *pro* precision of 71%. The high accuracy is due to the high number of (word) instances of the none class (i.e. a word that is neither *pro* nor *con*). There is no clear best setting, but *Major* is better reproducible with respect to F_1 than *Inter*. It is therefore a good choice for a gold standard generation strategy in our case.[5]

All these results are evaluated within each data configuration, e.g. the intersection model on the intersection test data. This does not allow for a direct comparison of the models. We thus run cross-configuration evaluations, where we created a single test set from the annotations of A1 and evaluated with respect to it (see Table 3). For instance, a model trained on the majority (*Major*) data applied to this test set has a accuracy of 92.9% (second line of the table).

[5]Note that, for a fair comparison with the other settings, the concatenation of the same data annotated by 3 different raters, i.e. the fact that training data is tripled is compensated at training time by the number of epochs divided by 3 (C_{fair}).

model	acc	F_1	prec	rec
M1	93.2	69.1	70.7	67.5
Major	92.9	67.9	70.0	65.9
Inter	92.6	62.7	**77.2**	52.7
Concat	**93.3**	69.0	74.9	64.0
C_{fair}	93.1	**69.5**	68.7	**70.4**

Table 3: Cross-configuration results

The comparison with the manual annotations of A2 and A3 (not with the predictions of their models, M2 and M3!) represents the upper-bound of "human models": the resp. F_1 scores are 73.5% and 86.7%. The accuracy and F_1 scores of A1's model *M1* (i.e. an intra-configuration evaluation) come close to human performance, but the gap is substantial: 69.1% versus 73.5% and 86.7% (both *Concat* models contain part of A1 and performs on par with *M1*). So either *Concat* or *Major* are the natural choice for producing the final gold standard.

3.3 Discussion

About 20% of the annotated sentences do not contain any explicit polar words, according to a per se limited lexical resource[6], of course. Are these "non-polar" sentences harder for a model to tag than the "polar" sentences?

Splitting the non-polar sentences from the polar ones, we trained a polar model on A1's annotations and evaluated it once on the non-polar, i.e. exclusive subset from the same annotator (see Table 4). Comparing these results to the individual intra-configuration results for M1-M3 shown in Table 2, we can observe similar tendencies for *pro* and *con* labeling quality levels. Although further evaluations are needed to confirm these preliminary results, this could indicate that baseline BERT models can bridge the gaps remaining in polar lexicons.

label	F_1	prec	rec
pro	0.68	0.71	0.66
con	0.66	0.66	0.67

Table 4: F_1, precision and recall of A1's polar model P1 evaluated on the non-polar subset

Apart from some cases where such non-polar words are just (polarity) lexicon gaps, there are some challenging examples of sarcasm and underlying world knowledge. For example, the words *Umwelt* (environment) and *Landschaft* (landscape) have no explicit polarity, though they may have a positive connotation, but the author of the following sarcastic comment *'Umwelt und Landschaft kann man nur einmal kaputt machen.'* (*'Environment and landscape can be destroyed only once.'*) reveals a *pro* position towards both terms. As a further example, consider the word *Atomkraftwerk* (nuclear power plant) and its two different labels (pro, none) in the following sentence: *Darum ist es sicherer wenn die Schweiz eigene Atomkraftwerke**pro** besitzt als Strom aus ausländischen Atomkraftwerken**none** zu beziehen.* ('That is why it is safer for Switzerland to have its own nuclear power plants than to buy electricity from foreign nuclear power plants.')

4 Related Work

As far as we know, there is no prior work on fine-grained stances in German texts.

Luo et al. (2020) analyse the opinions in the highly topical and controversial debate of climate change. Their BERT-based classifier achieves 75% accuracy for the stance detection of global warming. The main differences to our work concern the language, the granularity of the labeled units, and the number, i.e. diversity of topics. While they label whole English sentences with stance on one topic, we detect all possible targets at token-level in German politicians' comments on various issues.

Allaway and McKeown (2020) specify a connotation lexicon that includes the cultural and emotional perspectives of the writer. Although many words do have a context independent connotation, in our texts a word often switches its polarity depending on the context.

5 Conclusion

In texts expressing stance, we not only find explicitly communicated opinions that comprise a person's overall opinion towards the target, but also his/her implicitly given preferences and values which establish common ground for the reader's understanding of the argumentation. We have introduced **deInStance**, a corpus on such a fine-grained level and carried out experiments with a baseline BERT model showing a reasonable performance. Predicting fine-grained stance could be beneficial for overall stance detection, but it also could be used to get closer to an author's personal profile.

[6]We use the Polart lexicon (Klenner et al., 2009) available from the IGGSA webpage.

References

Emily Allaway and Kathleen R. McKeown. 2020. A unified feature representation for lexical connotations. *CoRR*, abs/2006.00635.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Manfred Klenner, Angela Fahrni, and Stefanos Petrakis. 2009. PolArt: A robust tool for sentiment analysis. In *Proceedings of the 17th Nordic Conference of Computational Linguistics (NODALIDA 2009)*, pages 235–238, Odense, Denmark. Northern European Association for Language Technology (NEALT).

Yiwei Luo, Dallas Card, and Dan Jurafsky. 2020. DeSMOG: Detecting stance in media on global warming. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3296–3315, Online. Association for Computational Linguistics.

Jannis Vamvas and Rico Sennrich. 2020. X-Stance: A multilingual multi-target dataset for stance detection. In *Proceedings of the 5th Swiss Text Analytics Conference (SwissText) & 16th Conference on Natural Language Processing (KONVENS)*, Zurich, Switzerland.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

A Appendices

Appendix A. Model settings

Regarding the model settings, we fine-tune the pretrained cased German model from DBMDZ while training our token classifier, i.e. all the weights are updated, not only the classifier's weights. We train models on a single GPU (NVIDIA GeForce GTX TITAN X) for 3 epochs without early stopping. We use the Adam optimizer (Kingma and Ba, 2015) with default epsilon=1e-08. We use a learning rate of 5e-5 for all experiments with a training batch size of 32, with no gradient accumulation. We set the random seed to 1, the maximum sequence length to 256. As the number of examples varies between 1720 and 1765 training sentences in the different data splits, the optimization process runs through 162 to 168 steps.[7]

[7] multiplied by 3 for the concatenation configuration: 5160 to 5295 training sentences, processed in 486 to 498 optimization steps.