Political claim identification and categorization in a multilingual setting: First experiments

Urs Zaberer and **Sebastian Padó** and **Gabriella Lapesa**
Institut für Maschinelle Sprachverarbeitung
University of Stuttgart, Germany
{urs.zaberer,sebastian.pado,gabriella.lapesa}@ims.uni-stuttgart.de

Abstract

The identification and classification of political claims is an important step in the analysis of political newspaper reports; however, resources for this task are few and far between. This paper explores different strategies for the cross-lingual projection of political claims analysis. We conduct experiments on a German dataset, DebateNet2.0, covering the policy debate sparked by the 2015 refugee crisis. Our evaluation involves two tasks (claim identification and categorization), three languages (German, English, and French) and two methods (machine translation – the best method in our experiments – and multilingual embeddings).

1 Introduction

The identification of political claims in news is a core step in the analysis of policy debates. *Discourse networks*, whose nodes correspond to claims and the actors who advance them, provide a rich source of information on phenomena such as formation of coalitions (who agrees with whom), shift in salience due to external events (e.g., migration waves making the issues of refugee accommodation more central in a debate), emergence of leadership, and polarization of a discourse (Leifeld and Haunss, 2012; Koopmans and Statham, 1999; Hajer, 1993).

Political claims are defined as demands, proposals or criticism that are *supported* or *opposed* by an *actor* (a person or a group of persons). Political claims generally form a call to action: they refer to something that should (or should not) be done in a policy domain (e.g., assigning empty flats to refugees). Thus, political claims are related to, but add a new perspective on, the Argument Mining question of what claims are, and what are the best strategies for modeling them across domains (Daxenberger et al., 2017; Schaefer et al., 2022).

The potential and challenges of the NLP support to political claim analysis have been thoroughly explored in the recent years in a monolingual setting (Chen et al., 2020; Dayanik et al., 2022); however, there are very few resources available in multilingual or crosslingual settings. Thus, there is little work on the comparison of policy debates in different countries, either completely automatic, or semi-automatic (supporting the inductive development of annotation guidelines in a new language).

This paper reports on cross-lingual pilot experiments on two tasks (claim identification and categorization), comparing two well known approaches to cross-lingual transfer in NLP in general, and argument mining in particular: machine translation and multilingual embeddings (Eger et al., 2018; Toledo-Ronen et al., 2020). We first work with a reference dataset for the German migration policy debate (Blokker et al., 2023), and on its projection to English and French, before moving on to a newly annotated English test set on the same topic. Machine Translation turns out to be the best cross-lingual projection strategy.

2 Experimental Setting

2.1 Tasks

This work focusses on two constituent tasks of political claim analysis (Padó et al., 2019). Our first task is **claim identification**, performed as a binary classification task at the sentence level. Our second task is **claim categorization**, phrased as a multi-label classification task at the sentence level.[1]

2.2 Data

We carry out two experiments. In the first one, we use a German corpus, DebateNet, which we automatically translate into English and French: this represents a cross-lingual transfer within the same media outlet. In the second experiment, we transfer our DebateNet models to an original English dataset based on the *Guardian* newspaper.

[1] For our evaluation in the claim categorization task, we consider all claims in the manually annotated gold standard.

Proceedings of the 19th Conference on Natural Language Processing (KONVENS 2023), pages 219–228
September 18–22, 2023.

DebateNet 2.0. Blokker et al. (2023) is a dataset[2] targeting the German public debate on migration policies in the context of the 2015 so-called 'refugee crisis'. It is based on 700 articles from the German quality newspaper *die Tageszeitung (taz)* with a total of 16402 sentences.

Political claims are annotated as textual spans, and each claim span is associated with at least one of 110 categories drawn from a theory-based codebook (annotation guidelines). Around 15% of sentences are annotated to contain a claim span. In total, the dataset contains 3442 claim spans corresponding to 4417 claim labels (i.e., each claim span is associated with an average of 1.3 claim categories). Annotations are first proposed by pairs of students of political science, with an inter-coder reliability is $\kappa = 0.59$ (Padó et al., 2019), and then accepted, rejected or merged by domain experts. We randomly split DebateNet into a training, development, and test set with a ratio of 80:10:10.

Crucially for our experiments, the 110 fine-grained categories are organized into 8 top-level categories which encode general domains of the migration policy field. In the claim categorization experiments in this paper we focus on the 8 top-level categories. Table 5 in the Appendix shows them with the percentage of claims annotated for each category and illustrative examples.

Guardian test set To compare German news translated into English to actual UK news, we collected an English-language test set of 36 articles from the British quality newspaper Guardian, extracted from the World News section and published in 2015. To make our test set as compatible as possible with *DebateNet2.0*, we look at the five months most represented in *DebateNet2.0* and within each month sample from articles written in the seven-day spans with the highest frequency of articles in *DebateNet2.0*. Articles were further filtered by keywords (*migrant, refugee, asylum, Germany, Syria, Afghanistan* and their morphological and syntactic variants) and by the mention of the most salient political actors (politician and parties).

The Guardian test set was manually annotated by a native speaker, a MSc-level student in Computational Linguistics, based on the *DebateNet2.0* guidelines. Claims were identified and assigned to one of the 8 top-level categories described in the previous section. Across the 36 articles with 1347 sentences, the test set contains 82 claim spans which correspond to 101 claim categories (mean of 1.2 categories per span).[3] Refer to Table 5 in the Appendix for the distribution of claim categories.

2.3 Methods

2.3.1 Projection methods

With the German DebateNet2.0 as our starting point, and the goal of testing the feasibility of cross-lingual projection to English and French (as target languages), we compare the two most established projection methods (Eger et al., 2018; Toledo-Ronen et al., 2020): machine translation (to make the modeling task monolingual) and multilingual embeddings (to let the model bridge the language gap implicitly). This yields three experimental conditions:

Translate-train: We machine-translate the German training data into the target languages and fine-tune a monolingual target-language model on it, to be evaluated on the target-language test data.[4]

Translate-test: We machine-translate the test data into German (as described above) and apply a monolingual German model fine-tuned on the original German data to it. For the DebateNet experiments in Section 3.1, we can only simulate this setting, as we do not have genuine foreign-language test data. We simulate it with a back-translation: first, we machine-translate the German DebateNet test set into the target language (EN/FR); then we translate the simulated EN/FR test sets back into German. It is only on the Guardian test set (Section 4) that we can fully evaluate our models in the translate-test configuration.

Multilingual: We employ multilingual embeddings, fine-tune them on the original German data, and apply the resulting classifier on the target language test data, exploiting the model's internal alignment of the source and target languages.

For both claim identification and classification, we re-implement standard Transformer-based models from the literature (Dayanik and Padó, 2020). We use BERT as well as its German, French and multilingual versions. Details on the classifier setups for both tasks follow below.

[2] http://hdl.handle.net/11022/ 1007-0000-0007-DB07-B

[3] 30 claims, albeit identified by our annotator, could not be classified in any categories of the codebook.

[4] We uses the DeepL translator via its web interface on a free trial of the "advanced" plan as of August 2022.

2.3.2 Claim identification

Translate-train: For English, we select the uncased model (`bert-base-uncased`) based on its performance on the development set, and we set learning rate to 5e-5 and warm-up steps to 30. The same configuration is used for the German monolingual baseline. For French, we select the base version of CamemBERT, `camembert-base`, with a learning rate of 4e-5 with 30 warm-up steps.

Translate-test: we employ a German BERT model, `bert-base-german-cased`, fine-tuned on the original German dataset. The hyperparameters are the same as for English translate-train.

Multilingual: Based on performance on the development set, we select the cased variant of the multilingual BERT from the Huggingface transformer library, `bert-base-multilingual-cased`. Training this model requires a lower learning rate of 2.5e-5 and correspondingly more epochs.

2.3.3 Claim categorization

Translate-train: For the English model, we assess both the cased and uncased versions. Since the uncased one (`bert-base-uncased`) again performs slightly better, we select it and use a learning rate of 5e-5. Experiments on the corresponding development sets establishes 25 warm-up steps as a reasonable choice for all configurations in Task 2. The French model – the same as for the claim identification task – requires a learning rate of 4e-5.

Translate-test: We employ `bert-base-german-cased` with a learning rate of 4e-5. The same model is also used for the monolingual German baseline model.

Multilingual: Based on performance on the development set, we select `bert-base-multilingual-uncased` with a low learning rate of 3e-5 and correspondingly more epochs.

3 Experiment 1: Within-outlet cross-lingual transfer

3.1 Claim Identification on DebateNet

The left-hand side of Table 1 shows results for the first main experiment, comparing the translate-train, translate-test, and the multilingual embedding approaches to claim identification to a monolingual baseline.[5] For comparison, we also run the translate-train and translate-test approaches on the multilingual model (multilingual:en:de and multilingual:de:de-en). The language labels de-en and de-fr stand for German data translated into EN or FR and back-translated into German.

Setup	Train	Test	Id	Cat
BL (mono)	de	de	56.2	**70.5**
Translate-train	en	en	**57.3**	67.8
Translate-train	fr	fr	**57.4**	69.7
Translate-test	de	de-en	55.8	69.5
Translate-test	de	de-fr	58.3	69.8
Multilingual	de	en	45.8	50.3
Multilingual	de	fr	51.1	51.0
Multilingual†	de	de-en	52.0	60.0
Multilingual†	en	de	55.4	64.1

Table 1: DebateNet test set results: F1 scores (positive class for claim identification (ID), macro average for claim categorization (Cat)). BL (mono): monolingual baseline.

The main contrast of this set of experiments is the one between the translate-train approach and the multilingual embeddings approach with respect to their performance on the target languages (EN/FR). For both target languages, the translate-train approach outperforms the monolingual baseline and the multilingual embedding approach. We ascribe this (small) performance gain to the higher quality of the embeddings available for the target languages: The monolingual English model, `bert-base`, is trained on a much larger corpus (English Wikipedia and BookCorpus) than `bert-base-german`, which is only trained on the significantly smaller German Wikipedia. The French model's training corpus is also over ten times larger than the German one. This also means the translation process, albeit not perfect, has not degraded the claim "signal" in the training data.

This point is also supported by the results for the "simulated" translate-test approach, which (cf. Section 2.3) can be considered a test of translation quality. Since the performance is in line with the monolingual baseline (de-en) or even slightly superior to it (de-fr)[6], the claim signal is preserved

[5] Unless indicated by a dagger †, reported values for all conditions are the averages of two runs to reduce variance.

[6] The exact reason for the improved performance in the de-fr setup is to be further investigated. Given that we consider the translate-test setup in DebateNet as a translation quality check, the result is not highlighted in bold even if higher than

	Target: yes	Target: no
Predicted: yes	71	39
Predicted: no	75	822

Table 2: Claim identification (DebateNet) confusion matrix of the best model for English (translate-train)

through the back-translation process.

In contrast, the multilingual embeddings perform poorly, below the monolingual baseline. The bottom part of Table 1 shows additional experiments we carried out to better understand this result. We find that a monolingual setup with multilingual embeddings (DE-DE) still performs below the monolingual baseline, but the performance gap is narrower than for the cross-lingual setups (DE-EN and DE-FR). Reverting the direction of the mapping, contrasting the performance of English-German (55.6) vs. German-English (45.8), again speaks in favor of the German representations being the weak point – the training data for the English-German multilingual embeddings setup is the same as that of the translate-train approach.

The confusion matrix for the best cross-lingual model for English (translate-train), Table 2, shows many fewer false negatives than false positives (i.e., a high precision). Regarding application to the (semi-)automatic extraction of discourse networks, this outcome is complementary to the high-recall approach applied by Haunss et al. (2020) to the German annotation in DebateNet, but lends itself to high-precision human-in-the-loop approaches like the one proposed by Ein-Dor et al. (2019) for argument mining.

Error Analysis. The misclassified instances provide some more insight into the model. For instance, we might expect the word “fordern” (“demand”, “call for”) to frequently appear in claims and therefore lead the model to make a positive prediction. Indeed, in the misclassified instances of the German-French translate-test model, forms of the word “fordern” or “Forderung” are 13 times more likely to be FP than FN even though there are almost twice as many FNs. We can therefore conclude that this word influences the model in the expected way. We bolster these observations with more formal methods: using saliency-based analysis (Simonyan et al., 2014) we can assign each token a relevance for the model’s prediction. The results partially confirm this: the token “fordert” gets scores above 0.9 throughout. However, other forms, like the infinitive, receive lower scores, presumably because the 3rd person singular is more highly associated with concrete claiming situations.

translate-train.

Saliency scores are highly correlated between models and between languages. E.g., the sentence “Der bayerische Ministerpräsident Horst Seehofer begrüßte die Pläne” and its corresponding English version ‘Bavaria’s prime minister Horst Seehofer welcomed the plans.’, are both labeled as claims. In both cases, the highest saliency is assigned to “Pläne”/”plans”. A systematic comparison of scores among models is however complicated by the differences in tokenizations among embedding models. Alternatively, we can compare instances misclassified by different models. Here, we observe large overlap. On one test run, the multilingual German-French model misclassified 122 out of 1007 test instances, while the monolingual English model misclassified 120 instances. These instances have an overlap of 58% (random assignment, should result in 12% overlap). This suggests that the models struggle with the same instances. A first qualitative inspection at such "difficult" instances has ruled out the impact of proper names, length of sentences as well as the type of involved actors; further analysis in this direction is required.

3.2 Claim Categorization on DebateNet

The right-hand side of Table 1 shows the results for the claim categorization task (F1 macro over all classes; Tables 6–9 in the Appendix provide per-category results). Unsurprisingly, this fine-grained task is more challenging for cross-lingual transfer. None of the experimental configurations beats the monolingual baseline. As in claim id, translate-train outperforms multilingual embeddings.

Error analysis. Inspection of sentences shows that many misclassifications arise from misleading local lexical material in the sentences. For example, “Die SPD findet dies könnte die Integration unterstützen“ (“The Social Democratic party believes this could support integration”) includes the word ’integration’ which is a strong cue for the claim category ’integration’, which the model predicts. However, the correct category is ’residency’, as becomes clear from the broader context of the article. Another example is: "Die sollen ja auch in der Gesellschaft ankommen" (“They must arrive

Setup	Train	Test	Id	Cat
translate-train	en	en	**25.5**	51.0
translate-test	de	de-en	20.6	**53.4**
multilingual	de	en	20.0	39.0

Table 3: Guardian test set results for claim identification (Id, F1 of positive class) and claim categorization (Cat, macro F1)

in society after all”), with misleading cue ’society’ indicating claim category ’society’ and gold category ’integration’. A saliency analysis, as before, confirmed this pattern: the “red herring” cues consistently receive the highest saliency scores in the sentences. Notably, the error pattern persists in the case of literal translations, but disppears when the translation changes the wording (‘mit Sicherheit’ – “with security/certainty” $\rightarrow$ ‘certainly’).

4 Experiment 2: Cross-outlet cross-lingual transfer

Results on the Guardian test set are shown in Table 3. For claim identification, the translate-train approach outperforms the other approaches, confirming the trend seen on the DebateNet data. For claim categorization, translate-test outperforms translate-train and multilingual embeddings. Both of these results are in line with our findings in Exp. 1.

For both tasks, we see a substantial decrease of performance on the Guardian data (-30 points for claim identification, -15 points for claim categorization). Since our previous experiment also used English data, this difference cannot be due to cross-lingual differences, but rather to differences between the two outlets, taz and the Guardian. Indeed, we see that a British newspaper is likely to report differently on German domestic affairs than a German newspaper, which leads to differences in claim form and substance: They tend to focus on the internationally most visible actors and report claims on a more coarse-grained level. They also overreport the claim categories most relevant for the British readership: claims migration control account for 22% of all claims in DebateNet but for 34% in the Guardian. In contrast, domestic (German) residency issues make up 14% of the DebateNet claims but only 2% of the Guardian claims. See Table 5 in the Appendix for a detailed breakdown and example claims.

	Target: yes	Target: no
Predicted: yes	29	147
Predicted: no	83	1088

Table 4: Claim identification (Guardian): confusion matrix of the best model for English (translate-train)

Thus, even if the Guardian claims might be structurally easier to recognize, the cross-outlet differences in claim distribution make transferring model representations from DebateNet to the Guardian hard. The confusion matrix for claim identification in Table 4 shows a low-precision scenario, in contrast to the high precision of the cross-lingual within-DebateNet setup.

It is interesting to note that claim identification suffers much more (-30 points) than claim categorization (-15 points), indicating that the model of claim topics survives the transfer to another outlet better than the model of what constitutes a claim.

5 Conclusion

This paper explores different strategies for the cross-lingual projection of political claims analysis from German into English and French. Our experiments establish the potential of machine translation for both claim identification and categorization, setting the stage for further investigations on the factors affecting projection performance and on the applicability of cross-lingual transfer for similar analyses. Multilingual embeddings yielded worse results, in line with previous analyses arguing that they attempt to solve a harder (since more open-ended) task than Machine Translation (Pires et al., 2019; Barnes and Klinger, 2019). We find that the language is not the only relevant dimension, though: in fact, the differences in presentation between German and British articles on German affairs go substantially beyond the language gap (Vu et al., 2019).

Acknowledgements

This study was partially funded by Deutsche Forschungsgemeinschaft (DFG) through MARDY (Modeling Argumentation Dynamics) within SPP RATIO and by Bundesministerium für Bildung und Forschung (BMBF) through E-DELIB (Powering up e-deliberation: towards AI-supported moderation). We are grateful to Brandon Sorensen, who annotated the Guardian test set.

Limitations

Our main experiment was limited to German, English, and French, three typologically very similar languages. Generalization to more distant languages is presumably harder, but was outside the scope of our study. Our Guardian test set is very small (albeit not significantly smaller than out-of-domain gold sets often gathered for validation purposes), and annotating it was challenging due to the need to apply a codebook developed for the German debate to an English source. We are currently working on improving the size and quality of our test set.

While our experiments are reassuring as regards translation quality, we cannot exclude that translation biases may have been introduced in the data. We are also aware that DeepL is not the only option for automatic translation; evaluating different translation methods, however, falls outside the scope of this work.

Ethical Considerations

At the level of datasets and annotations, we employed an existing dataset (DebateNet2.0). Our own annotation contribution (the Guardian test set) was based on publicly available data; moreover, the annotation task was carried out following best practices. The Guardian test set is available upon request.

At the modeling level, we use previously defined models that are publicly available; in this sense, our contribution does not raise new ethical questions (e.g. in terms of misuse potential). To the contrary, our focus is on understanding how these models transfer across languages and what biases can potentially arise in this transfer, as shown by our focus on error analysis.

References

Jeremy Barnes and Roman Klinger. 2019. Embedding projection for targeted cross-lingual sentiment: Model comparisons and a real-world study. *Journal of Artificial Intelligence Research*, 66:691–742.

Nico Blokker, Andre Blessing, Erenay Dayanik, Jonas Kuhn, Sebastian Padó, and Gabriella Lapesa. 2023. Between welcome culture and border fence: The European refugee crisis in German newspaper reports. *Language Resources and Evaluation*, 57:121–153.

Wei-Fan Chen, Khalid Al Khatib, Benno Stein, and Henning Wachsmuth. 2020. Detecting media bias in news articles using Gaussian bias distributions. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4290–4300, Online. 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.

Erenay Dayanik, Andre Blessing, Nico Blokker, Sebastian Haunss, Jonas Kuhn, Gabriella Lapesa, and Sebastian Pado. 2022. Improving neural political statement classification with class hierarchical information. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2367–2382, Dublin, Ireland. Association for Computational Linguistics.

Erenay Dayanik and Sebastian Padó. 2020. Masking actor information leads to fairer political claims detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4385–4391, Online. Association for Computational Linguistics.

Steffen Eger, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2018. Cross-lingual argumentation mining: Machine translation (and a bit of projection) is all you need! In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 831–844, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Liat Ein-Dor, Eyal Shnarch, Lena Dankin, Alon Halfon, Benjamin Sznajder, Ariel Gera, Carlos Alzate, Martin Gleize, Leshem Choshen, Yufang Hou, Yonatan Bilu, Ranit Aharonov, and Noam Slonim. 2019. Corpus wide argument mining - a working solution. In *AAAI Conference on Artificial Intelligence*.

Maarten A Hajer. 1993. Discourse Coalitions and the Institutionalization of Practice: The Case of Acid Rain in Britain. In *The Argumentative Turn in Policy Analysis and Planning*, pages 43–76. Duke University Press.

Sebastian Haunss, Jonas Kuhn, Sebastian Pado, Andre Blessing, Nico Blokker, Erenay Dayanik, and Gabriella Lapesa. 2020. Integrating manual and automatic annotation for the creation of discourse network data sets. *Politics and Governance*, 8(2).

Ruud Koopmans and Paul Statham. 1999. Political Claims Analysis: Integrating Protest Event And Political Discourse Approaches. *Mobilization*, 4(2):203–221.

Philip Leifeld and Sebastian Haunss. 2012. Political Discourse Networks and the Conflict over Software Patents in Europe. *European Journal of Political Research*, 51(3):382–409.

Sebastian Padó, Andre Blessing, Nico Blokker, Erenay Dayanik, Sebastian Haunss, and Jonas Kuhn. 2019. Who sides with whom? towards computational construction of discourse networks for political debates. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2841–2847, Florence, Italy. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Robin Schaefer, René Knaebel, and Manfred Stede. 2022. On selecting training corpora for cross-domain claim detection. In *Proceedings of the 9th Workshop on Argument Mining*, pages 181–186, Online and in Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings*.

Orith Toledo-Ronen, Matan Orbach, Yonatan Bilu, Artem Spector, and Noam Slonim. 2020. Multilingual argument mining: Datasets and analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 303–317, Online. Association for Computational Linguistics.

Hong Tien Vu, Yuchen Liu, and Duc Vinh Tran. 2019. Nationalizing a global phenomenon: A study of how the press in 45 countries and territories portrays climate change. *Global Environmental Change*, 58:101942.

A Appendix

A.1 Datasets: quantitative details and comparison

Class	Label	%DN	%G	Examples
C1	Controlling Migration	22	34	*DN*: A fixed resettlement programme is needed, with binding annual admission quotas. *G*: Angela Merkel stressed the need for a fairer distribution of refugees across the EU
C2	Residency	14	2	*DN*: These urgent procedures shall be carried out in special reception facilities. *G*: We have to find suitable accommodation for all of them.
C3	Integration	9	3	*DN*: The CDU insists on an integration obligation for migrants. *G*: Michael Fuchs called on the government to set up language courses and to send job centre employees to assess newcomers
C4	Domestic Security	3	8	*DN*: The head of the police union, RainerWendt, has called for a "ban mile around refugee shelters". *G*: We should not hand over our streets to hollow rallying cries
C5	Foreign Policy	16	11	*DN*: The current problems with the refugees must nevertheless be solved at European and international level, she said. *G*: Tomas de Maizière said pressure should be applied to rejectionist nations such as Hungary, Slowakia and the Czech Republic.
C6	Economy + Labour Market	3	7	*DN*: A condition for waiving such proof, however, must be that collective bargaining conditions or a minimum wage apply in order to prevent dirty competition to the detriment of all employees. *G*: Folkerts-Landau said the influx of refugees has the potential not just to invigorate our economy but to protect prosperity for the future generations
C7	Society	17	21	*DN*: And Reinhard Marx, chairman of the Catholic Bishops' Conference, criticized the strict separation between war refugees and economic refugees. *G*: As chancellor, I come to the defense of Muslims, most of whom are upright, constitutionally loyal citizens
C8	Procedures	15	14	*DN*: The federal government is planning a new law to speed up asylum procedures. *G*: Gerd Mueller called on Tuesday for the EU to appoint a European Refugees commissioner and said it had to treat the problem with more urgency

Table 5: Claim categories: class, labels, distribution (percentage of total claims), and example claim in DebateNet2.0 (DN) (manually translated into English) and Guardian test set (G).

A.2 Per-category Results

Class	#instances in test	Precision	Recall	F1 score
C1 (Controlling Migration)	35	0.67	0.83	0.74
C2 (Residency)	2	0.66	0.74	0.70
C3 (Integration)	3	0.66	0.60	0.63
C4 (Domestic Security)	8	0.50	0.44	0.47
C5 (Foreign policy)	11	0.87	0.76	0.81
C6 (Economy)	7	0.88	0.50	0.64
C7 (Society)	21	0.70	0.67	0.69
C8 (Procedures)	14	0.75	0.70	0.72
micro avg		0.71	0.71	0.71
macro avg		0.71	0.66	0.67

Table 6: Claim categorization: precision, recall and F1 values for the different classes, translate-train French

Class	#instances in test	Precision	Recall	F1 score
C1 (Controlling Migration)	35	0.66	0.74	0.70
C2 (Residency)	2	0.68	0.70	0.69
C3 (Integration)	3	0.72	0.51	0.60
C4 (Domestic Security)	8	0.40	0.33	0.36
C5 (Foreign policy)	11	0.85	0.65	0.73
C6 (Economy)	7	0.80	0.57	0.67
C7 (Society)	21	0.77	0.56	0.65
C8 (Procedures)	14	0.76	0.58	0.66
micro avg		0.71	0.62	0.67
macro avg		0.70	0.58	0.63

Table 7: Claim categorization: precision, recall and F1 values for the different classes, translate-train English

Class	#instances in test	Precision	Recall	F1 score
C1 (Controlling Migration)	35	0.76	0.71	0.73
C2 (Residency)	2	0.76	0.69	0.72
C3 (Integration)	3	0.72	0.58	0.64
C4 (Domestic Security)	8	0.40	0.33	0.36
C5 (Foreign policy)	11	0.86	0.65	0.74
C6 (Economy)	7	0.83	0.36	0.50
C7 (Society)	21	0.86	0.56	0.68
C8 (Procedures)	14	0.73	0.61	0.66
micro avg		0.76	0.62	0.68
macro avg		0.74	0.56	0.63

Table 8: Claim categorization: precision, recall and F1 values for the different classes, German baseline

Class	#instances in test	Precision	Recall	F1 score
C1 (Controlling Migration)	35	0.74	0.78	0.76
C2 (Residency)	2	0.69	0.84	0.76
C3 (Integration)	3	0.72	0.62	0.67
C4 (Domestic Security)	8	0.48	0.61	0.54
C5 (Foreign policy)	11	0.81	0.81	0.81
C6 (Economy)	7	0.70	0.50	0.58
C7 (Society)	21	0.72	0.66	0.69
C8 (Procedures)	14	0.70	0.68	0.69
micro avg		0.72	0.73	0.72
macro avg		0.70	0.69	0.69

Table 9: Claim categorization: precision, recall and F1 values for the different classes. Model: best cross-lingual model (translate-test)

Class	Precision	Recall	F1 score
C1 (Controlling Migration)	0.66	0.66	0.66
C2 (Residency)	0.25	0.50	0.33
C3 (Integration)	0.50	0.67	0.57
C4 (Domestic Security)	1.00	0.25	0.40
C5 (Foreign policy)	0.45	0.82	0.58
C6 (Economy)	0.50	0.29	0.36
C7 (Society)	0.76	0.76	0.76
C8 (Procedures)	0.57	0.29	0.38
micro avg	0.61	0.58	0.60
macro avg	0.59	0.53	0.51

Table 10: Claim categorization: precision, recall and F1 values for the different classes on Guardian dataset. Model: translate-test