

Identifying Stance-Bearing Keywords in Public Debates with Instruction-Tuned Language Models

Milena Belosevic^{ID}

¹ Faculty of Linguistics and Literary Studies, Bielefeld University, Bielefeld, Germany

Abstract

The paper presents a computational approach to discourse analysis that identifies stance-bearing keywords: evaluative lexical units embedded in arguments that signal support for or opposition to a debated issue. Based on a dataset of 4035 arguments from the migration debate in Germany between 2015 and 2017, we explore how language models can be instruction-tuned to extract such lexical units. Human-annotated keywords serve as the basis for training and evaluation, providing a gold standard for model comparison. We apply supervised fine-tuning on BübleLM, a German-only language model, and compare it with both a three-shot prompted version and two established baselines: a named entity recognition (NER)-tuned German BERT model and a similarity-based keyword extractor (KeyBERT with RoBERTa embeddings). The findings show that the instruction-tuned BübleLM aligns more closely with human annotations than the baselines and its three-shot prompted variant. This suggests that domain-specific tuning can capture evaluative cues central to discourse-specific argumentation. The study contributes to ongoing efforts in computational humanities to combine machine learning with discourse-sensitive human annotation of politically salient language.

Keywords: language models, stance detection, instruction-tuning, human annotation, politically salient language, discourse analysis, argumentation, keyword extraction

1 Introduction

Understanding how social reality is shaped through language requires close attention to the rhetorical strategies employed in public debates, particularly in arguments that express a stance on controversial issues. However, identifying whether a speaker supports or opposes a given stance requires contextual knowledge about the topic in question.

This paper proposes a computational approach to identifying stance-bearing keywords, context-sensitive lexical units (i.e., spans, contiguous sequences of words in a text) that convey either a positive or negative position towards a particular controversial topic expressed by the argument. Stance-bearing keywords are an important part of discourse, defined as a collection of texts on a specific topic [15]. They influence public opinion, provoking either approval or rejection. For example, in the following argument in favor of migration: “It is necessary to abandon sweeping prejudices against the ‘foreigners’ and to be aware of the exceptional situation that the workers have to cope with in our northern regions”. (example from [16]), *abandon sweeping prejudices against the “foreigners”* and *be aware of the exceptional situation* verbalize the stance in favor of refugees, whereas other parts of the argument (*in our northern regions*) may be contextually neutral. Stance-bearing keywords include both “flag words” that signal solidarity (e.g., *Willkommenskultur* / “welcome culture” used to promote a humane treatment of refugees)

Milena Belosevic^{ID}. “Identifying Stance-Bearing Keywords in Public Debates with Instruction-Tuned Language Models.” In: *Computational Humanities Research 2025*, ed. by Taylor Arnold, Margherita Fantoli, and Ruben Ros. Vol. 3. Anthology of Computers and the Humanities. 2025, 405–435. <https://doi.org/10.63744/NskPfU7etX83>.

© 2025 by the authors. Licensed under Creative Commons Attribution 4.0 International (CC BY 4.0).

and “stigma words” (e.g., *Sozialtourismus* / “welfare tourism”, implying that migrants exploit the social welfare system) that delegitimize opposing views [9]. Identifying such keywords can support the automated discovery of stance in large corpora, reduce annotation costs, and provide new insights into identifying arguments in public language use.

Although recent studies in discourse analysis have focused on identifying arguments with machine learning methods [10], the task of extracting evaluative components within arguments, particularly in German political discourse, has not been systematically investigated. We address this gap by fine-tuning BübleLM [5], a German-only language model, on a corpus of over 4,000 manually annotated arguments about migration in Germany (2015–2017). Our approach reflects an understanding of arguments as recurring discursive patterns that do not always have to be verbalized in the same way, but that appear in many texts as recurring, similar ways of plausibly connecting facts [15].

Our work builds on recent advances in argument mining [13] dealing with the role of external knowledge [7] and key phrases [4] in identifying argumentative relations. However, our work shifts the focus from relational structures to argumentative units and their evaluative function. The paper also extends beyond traditional named entity recognition (NER) approaches, which often fail to capture context-dependent expressions that require domain-specific knowledge. In doing so, we contribute to the ongoing tendency in the computational humanities to combine machine learning with discourse-aware analysis of socially and politically salient language.

2 Methodology

2.1 Dataset and human annotation

The dataset comprises 1,772 newspaper articles in German published between April 2015 (shortly before Germany admitted refugees) and April 2017 (ending before the 2017 federal election campaign began), drawn from leading German outlets across the political spectrum: the conservative Frankfurter Allgemeine Zeitung, the liberal Süddeutsche Zeitung, the left-wing Tageszeitung, and weekly news magazines Der Spiegel and Die Zeit. From this corpus, 4,035 arguments related to the reception of refugees were previously identified and annotated for stance (supporting or opposing migration) by [1]. Building on this argument-level annotation, we conducted a second annotation layer.

Two trained annotators were tasked with extracting between one and six stance-bearing keywords per argument. The keywords should capture the central features of stance for or against refugees and represent emotionally loaded spans (“flag” words or “stigma” words, see Section 1). Keywords could be single words, multi-word phrases, subordinate clauses, or a maximum of one full sentence (see Appendix A.1). The annotators were not informed of the stance label during annotation to minimize bias. The guidelines also defined core terms (argumentation, discourse, and stance-bearing keywords) and emphasized the need to incorporate extra-linguistic knowledge.

Annotator 1 identified an average of 3.42 stance-bearing spans per argument, while annotator 2 annotated 2.99 on average. Subsequently, a third annotator created the gold-standard set by selecting one of the two sets in full, without merging or modifying spans (since span-by-span adjudication of over 4000 examples would have been prohibitively time-consuming).¹

Inter-rater reliability was assessed at the span level prior to adjudication using both exact and soft character-based overlap (Table 1).

The modest inter-annotator agreement at the exact span level ($F_1 = 22.97\%$) improved considerably with soft overlap (partial credit for partially overlapping spans), reaching $F_1 = 49.48\%$, which indicates the difficulty of achieving precise alignment in discourse-sensitive tasks. Annotator 1 aligned more closely with the adjudicated spans ($F_1 = 90.81\%$) than annotator 2 ($F_1 =$

¹ Gold-standard keyword set: <https://github.com/milenabelosevic/Stance-Bearing-Keywords->.

Comparison	Precision	Recall	F1
Annotator 1 vs. Annotator 2 (exact)	23.95%	24.62%	22.97%
Annotator 1 vs. Annotator 2 (soft)	52.47%	52.73%	49.48%
Annotator 1 vs. Gold (soft)	89.27%	94.27%	90.81%
Annotator 2 vs. Gold (soft)	58.14%	63.08%	58.33%

Table 1: Span-level agreement between annotators and with the adjudicated gold set. Soft match: $\geq 60\%$ character overlap.

58.33%), reflecting the adjudicator’s preference for annotator 1 in 60.7% of cases. Annotated and adjudicated examples are shown in Appendix Table 4.

2.2 Main model and baselines

BübleLM, a German decoder-only language model [5], was instruction-tuned for the task of extracting discourse-relevant keywords from public arguments about migration. BübleLM was originally pre-trained on 3.5 billion tokens from the contemporary web content, legislative documents, news data, and Wikipedia sources. We applied QLoRA (quantized low-rank adaptation) [6], a memory-efficient fine-tuning technique, using 4-bit quantization and low-rank adapters. The dataset was split into training (80%), validation (10%), and test (10%) sets. The model was trained for three epochs to generate semicolon-separated keywords from argumentative input. It was guided by a task-specific prompt derived from annotation guidelines (see Appendix A.3). At inference time, the instruction-tuned BübleLM was also evaluated in a three-shot prompting setup (Appendix A.4) to assess its ability to generalize from a limited context. Both training and validation loss showed early convergence and no signs of overfitting after step 400, suggesting that the model learned to recognize recurring keywords associated with stance in the refugee debate. Detailed training metrics are provided in Appendix Table 5.

Although models fine-tuned for NER can reliably label surface-level entities, they often struggle to capture the contextual nuances needed for discourse-level tasks. To illustrate these limitations, we include a German BERT model (deepset/gbert-base, [2]) as a baseline. This model has been widely used in German NLP tasks and is well-suited for token-level sequence labeling tasks such as NER. We fine-tuned the pretrained German BERT model using the same dataset split described above and applied the standard BIO tagging scheme, where each token is labeled as the beginning (B), inside (I), or outside (O) of a keyphrase. Fine-tuning was performed using QLoRA to reduce memory usage and computational costs. Predicted BIO tags were grouped into full keyword spans. While this method is technically robust, it is not optimized for identifying context-sensitive expressions, making it a useful contrast to the instruction-tuned model.

To establish a stronger baseline, we employed the off-the-shelf method without any fine-tuning or training on our data, specifically embedding-based keyphrase extraction with KeyBERT [8]. KeyBERT is an unsupervised method that identifies potential keywords from the input by comparing them to the overall meaning of the input text, using vector-based similarity. A pre-trained multilingual sentence transformer (*T-Systems-onsite/cross-en-de-roberta-sentence-transformer*) was used to support German-language inputs. As an unsupervised baseline, KeyBERT was applied only to the test set to ensure a fair comparison with the fine-tuned model.

For BübleLM and KeyBERT, we applied a post-processing filter that retained only predicted spans appearing verbatim in the input text. This ensured a fair comparison and prevented the evaluation from rewarding predictions that were not grounded in the input. For GBERT, no filtering

was needed because it predicts spans using BIO labels directly from the input tokens. Human-annotated spans served as the gold/reference spans.

3 Results

3.1 Model evaluation

For both the main model and the baselines, we report soft span-level F1 score, precision, and recall as the primary evaluation metrics. These metrics allow for partial matches between the predicted and gold spans, for example, giving partial credit if the model predicts *integration* when the gold span is *integration of refugees*. This reflects the fact that stance-bearing keywords within arguments are often semantically close but not identical in wording. The low exact span matching across all models ($\leq 4.2\%$, Appendix Table 6) confirms the need for more flexible evaluation when modeling discourse-specific meaning.

We allow BübleLM to predict multiple spans per argument without enforcing disjointness. This means that overlapping or nested spans (e.g., *Integration* and *Integration von Flüchtlingen*) are retained and scored individually. Such nested spans occurred in 47.77% of BübleLM predictions, and 88.37% contained at least some form of overlapping behavior. They were evaluated *as-is*, without removing overlapping predictions. This approach reflects the model’s raw output behavior.

For the NER-tuned baseline (German BERT), we additionally report both token-level metrics, as standard in NER, and soft span-level scores for comparison with the instruction-tuned model.

Although human-annotated spans are strict (i.e., not rephrased) substrings of the input, instruction-tuned BübleLM may generate semantically related but non-identical spans. For unsupervised methods like KeyBERT, which extract spans directly from the input, semantic similarity may still occur when the selected span partially captures the meaning of the gold reference, even if token overlap is limited (e.g., gold span: *recognition of degrees* vs. extracted span: *degrees obtained in home countries*). To capture this behavior, we compute embedding-based similarity between predicted and gold spans for instruction-tuned BübleLM and KeyBERT. This measure complements the main evaluation metrics by revealing whether incorrect predictions are still semantically meaningful. We do not apply this metric to the NER-tuned GBERT model, as its predictions adhere to strict BIO labeling and do not generate free-form text spans that would benefit from semantic similarity evaluation.

Besides the main evaluation metrics, we also analyzed the models’ extraction behavior using complementary measures: overlap coverage (span recall), which measures how many gold spans are at least partially matched ($\geq 50\%$ overlap), and span count and length, which captures how many keywords the model predicts and how long they tend to be. These metrics were calculated for all models. Although the GBERT base follows strict BIO tagging, span-based evaluation ensures comparability with KeyBERT and BübleLM.

We first report the macro-averaged soft span-level precision, recall, and F1 scores for all models in Table 2.

Model	Soft Precision	Soft Recall	Soft F1
Instruction-tuned BübleLM	52.70%	65.45%	53.97%
3-shot Prompting (Instr-Tuned BübleLM)	49.77%	54.96%	47.71%
KeyBERT (baseline)	32.86%	38.01%	31.31%
GBERT NER (BIO-tagged, baseline)	11.15%	5.06%	6.06%

Table 2: Macro-averaged soft span-level precision, recall, and F1 across models.

The instruction-tuned BübleLM achieved the best performance (F1: 53.97%), outperforming both its 3-shot variant (47.71%) and the baselines (KeyBERT 31.91%, GBERT NER 6.06%). Although three-shot prompting requires no additional training, instruction-tuning led to better task alignment. KeyBERT lags behind language-model-based approaches, especially in terms of recall. The NER-tuned GBERT model demonstrated limited alignment with gold annotations, achieving an F1 score of only 6.06%. This result implies that sequence labeling alone is not well-suited for capturing discourse-relevant spans. The findings support prior research showing that language models outperform traditional methods due to their capability to capture semantic nuance in context [11].

The instruction-tuned BübleLM achieved the highest semantic similarity (58.43%), followed by its three-shot variant (54.09%). KeyBERT, despite relying entirely on embedding similarity, scored lower (48.60%). This suggests that instruction tuning improves not only the formal similarity between model output and gold spans but also leads to predictions that are semantically more faithful to the gold spans. We did not compute semantic similarity for the NER-tuned GBERT model, as its BIO-labeled output reflects token-level predictions rather than semantically interpretable spans.

To further analyze model behavior, we computed span count and length statistics (Table 3).

Model	# Spans	Len.	Gold Cov.	Pred. Cov.
Gold (Ref.)	3.25	4.67	–	–
Instr.-tuned BübleLM	6.20	4.85	56.67	44.62
3-shot Prompted BübleLM	6.52	4.06	48.93	37.69
KeyBERT (Baseline)	1.00	43.89	11.39	–
GBERT NER (BIO-tagged, baseline)	0.28	2.29	6.12	73.91

Table 3: Span-level statistics: average number and length of spans, and overlap-based coverage metrics ($\geq 50\%$ token overlap). Coverage is shown both from the gold span perspective (recall-like) and the predicted span perspective (precision-like).

Human-annotated gold spans averaged 3.25 spans per example, with each span approximately 4.67 tokens long. The instruction-tuned BübleLM produced nearly twice as many spans (6.20), closely matching human span length (4.85 tokens). Its 3-shot variant overgenerated slightly more spans (6.52) and tended toward shorter spans (4.06 tokens). This higher span count reflects BübleLM’s tendency to produce overlapping or nested spans. Since we retained all spans without filtering, the model’s output was evaluated as is. In contrast, KeyBERT returned only one span per example on average, often very long (43.89 tokens), indicating that its similarity-based selection often included entire sentences or clauses. GBERT predicted very few spans (0.28 per example), typically short (2.29 tokens).

In terms of gold span coverage, BübleLM outperformed all models (56.67%), followed by the 3-shot variant (48.93%), with KeyBERT (11.39%) and GBERT (6.12%) far behind. BübleLM also achieved the highest precision-style coverage, suggesting that its predictions were both accurate and well-targeted. GBERT’s predicted span coverage was 73.91%, suggesting that although it rarely predicts spans, they tend to be well-aligned when it does.

3.2 Error analysis

We conducted a quantitative error analysis of model predictions, distinguishing between partial matches and complete misses (Figure 1).

The instruction-tuned BübleLM model produced the most reliable predictions, with 97.52%

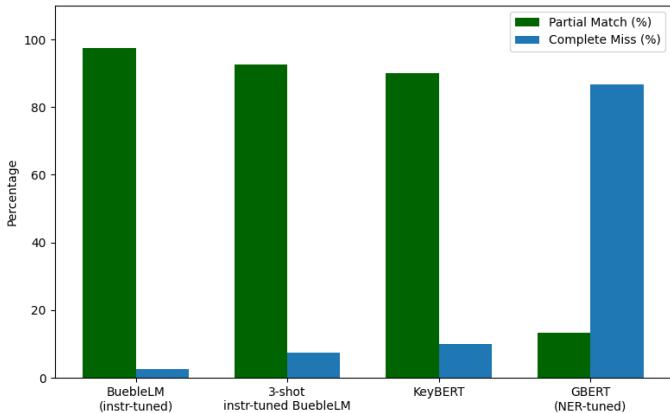


Figure 1: Quantitative error distribution across models. Partial matches indicate any token-level overlap between predicted and gold spans. Complete misses indicate no overlap.

showing at least partial overlap with the gold-standard spans. The three-shot variant also performed well (92.57%), followed closely by KeyBERT (90.10%). In contrast, the GBERT baseline yielded a very high complete miss rate (86.63%), indicating that its token-level labeling approach is ineffective in identifying discourse-specific keywords.

To illustrate the models’ behavior, we compared examples from three categories: the 10 best partial matches (with soft F1 scores close to 1), the 10 worst partial matches (soft F1 scores near 0), and 10 complete misses (soft F1 scores equal to 0). It is striking that the best partial predictions in all models miss one or two tokens compared to the gold span or have one or two additional tokens not included in the gold span. For example, given the gold span: “Einwanderung zu einem Verlustgeschäft werden könnte”; “Deutschland zieht nicht gerade die am besten ausgebildeten Migranten an” (“Immigration could become a losing proposition”; “Germany does not exactly attract the best-educated migrants”), the instruction-tuned BübleLM correctly captured the core phrase “nicht gerade die am besten ausgebildeten Migranten” but omitted the first clause. KeyBERT, by contrast, selected “nach Europa fliehen, laufen vor den Extremisten weg” (“flee to Europe, run away from the extremists”) instead of the gold span “laufen vor den Extremisten weg” (“run away from the extremists”).

The three-shot BübleLM often predicted semantically equivalent expressions, such as “ein paar schwarze Schafe” (“a few black sheep”), even if exact span boundaries differed from the gold span “darunter eben auch ein paar schwarze Schafe” (“among them also a few black sheep”). Notably, all models captured idiomatic expressions well, such as “sich einen Schalter umlegen müssen” (“having to flip a mental switch”) (KeyBERT) or “schwarze Schafe” (“black sheep”) (three-shot BübleLM) (see Appendix Section 3.2).

Worst partial matches (soft F1 score near 0) include cases where the argumentation comprises a stance against refugees but actually argues in favor of refugees. The models usually misinterpret the spans that indicate criticism of such positions or vice versa. For example, instead of identifying the negative stance in “Frauen verniedlicht und respektlos zu behandeln” (“to treat women in a belittling and disrespectful way”), the three-shot BübleLM incorrectly predicts “Frauen werden respektvoll behandelt” (“women are treated respectfully”), reversing the intended message. Similarly, the instruction-tuned BübleLM extracts “Wir schaffen das” (“We can do it”) from the argument against refugee reception. KeyBERT often selects irrelevant spans, such as “sehr laut geworden” (“became very loud”), which is not the argumentative core. In these cases, the models fail to capture the precise stance-bearing aspects that human annotators identified as central.

Complete misses represent severe boundary misalignments or the prediction of unrelated spans

that share little or no semantic overlap with the gold annotations. For instance, the three-shot BübleLM predicted “Profiling; Übersetzer; sich gut in der Materie auskennen” (“profiling; translator; to be well-versed in the subject matter”) in a context discussing skill mismatches in the labor market, missing the more relevant gold span “das Problem; viele Laufbahnen weniger streng reglementiert” (“the problem; many career paths are less strictly regulated”). The instruction-tuned BübleLM generated “missachtete Verlierer; fatale Weise Würde zu erlangen” (“neglected losers; to gain dignity in a fatal way”), instead of the gold span “soll man zugreifen; bevor es andere tun” (“one should act; before others do”), thereby shifting the argumentative focus. KeyBERT often selected spans irrelevant to the argumentation, such as “Sarrazin identifiziert zwar tatsächliche Probleme ja Deutschland kann” (“Sarrazin identifies actual problems yes Germany can”) instead of the more meaningful gold spans such as “die Schwierigkeit”, “ihre Familien dauerhaft zu integrieren”, and “nicht mal eben die Lücke füllen, die der demografische Wandel reißt” (“the difficulty”, “to permanently integrate their families”, “not simply fill the gap caused by demographic change”).

To better understand the nature of prediction errors, we categorized each example into one of six error types. Figure 2 visualizes the normalized frequency of each error type per model.

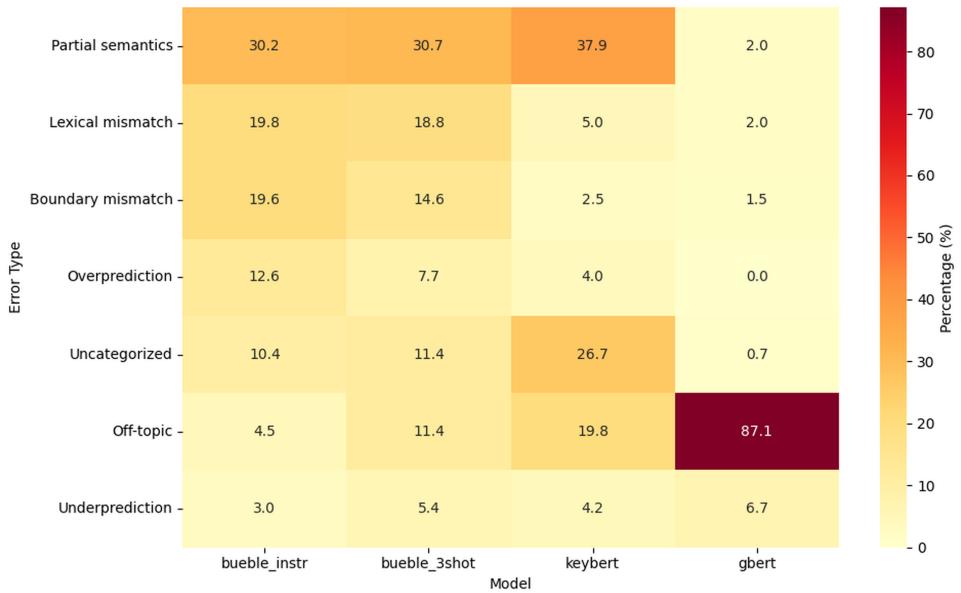


Figure 2: Relative distribution of error types per model. Each column is normalized to sum to 100%, allowing for the comparison of typical error patterns across models regardless of differences in the total number of predictions.

KeyBERT exhibited a strong tendency toward partial semantic matches (37.9%) and uncategorized errors (26.7%), whereas off-topic predictions were prevalent in GBERT (87.1%), indicating frequent semantic drift or irrelevant predictions. In contrast, the instruction-tuned BübleLM and its three-shot variant exhibited more balanced error profiles, with higher rates of lexical mismatches and boundary errors, suggesting that they effectively captured relevant content but struggled with precise phrasing or span limits. These results emphasize that the instruction-tuned model approximates meaning more effectively, while baselines are more prone to producing contextually irrelevant outputs.

4 Conclusions and future work

The paper proposed a computational approach to discourse analysis, focusing on stance-bearing lexical units that express a positive or negative position within arguments in the German migration

discourse. Among the tested approaches, the instruction-tuned BübleLM most closely aligned with human annotations, showing the importance of adapting language models through domain-specific instruction. The model can be used as a tool to support human annotation in qualitative discourse analysis (e.g., in the historic discourse semantic approach [15]). Discourse-level stance extraction enables us to explore not only which attitudes are dominant at a particular point but also how they are legitimized through language over time.

However, in addition to the model limitations discussed in Section 3.2, language models may inherit biases from their training data, which can affect the quality of extraction and raise ethical concerns. Finally, potential annotator bias remains, as the gold spans were fully adopted from a single annotator per instance. Generalizability beyond German newspaper data, as well as transfer to other languages, should be evaluated in future work.

To assess whether BübleLM’s performance can be improved, we are currently testing other German-only (LLäMlein [12]) and multilingual models (Teuken-7b-instruct [14], EuroLLM-Instruct [3]) and exploring alternative prompting strategies and out-of-domain generalization.

References

- [1] Belosevic, Milena. *Vertrauen und Misstrauen in der Flüchtlingsdebatte 2015-2017*. Berlin: Buske, 2022. DOI: 10.46771/978-3-96769-198-6.
- [2] Chan, Branden, Schweter, Stefan, and Möller, Timo. “German’s Next Language Model”. In: *Proceedings of the 28th International Conference on Computational Linguistics*, ed. by Donia Scott, Nuria Bel, and Chengqing Zong. Barcelona, Spain (Online): International Committee on Computational Linguistics, 2020, pp. 6788–6796. DOI: 10.18653/v1/2020.coling-main.598. URL: <https://aclanthology.org/2020.coling-main.598>.
- [3] Colombo, Pierre, Noiry, Nathan, Irurozki, Ekhine, and Clemenccon, Stephan. “What are the best systems? New perspectives on NLP Benchmarking”. 2022. DOI: 10.48550/arXiv.2202.03799.
- [4] Das, Nilmadhab, Saradhi, Vijaya V, and Anand, Ashish. “On the Role of Key Phrases in Argument Mining”. In: *Findings of the Association for Computational Linguistics: NAACL 2025*, ed. by Luis Chiruzzo, Alan Ritter, and Lu Wang. Albuquerque, New Mexico: Association for Computational Linguistics, 2025, pp. 5759–5772. DOI: 10.18653/v1/2025.findings-naacl.319.
- [5] Delobelle, Pieter, Akbik, Alan, et al. “BübleLM: A small German LM”. In: (2024).
- [6] Dettmers, Tim, Pagnoni, Artidoro, Holtzman, Ari, and Zettlemoyer, Luke. “QLORA: efficient finetuning of quantized LLMs”. In: *Proceedings of the 37th International Conference on Neural Information Processing Systems*. NIPS ’23. New Orleans, LA, USA: Curran Associates Inc., 2023. DOI: 10.5555/3666122.3666563.
- [7] Gemechu, Debela and Reed, Chris. “External Knowledge-Driven Argument Mining: Leveraging Attention-Enhanced Multi-Network Models”. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, ed. by Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen. Miami, Florida, USA: Association for Computational Linguistics, 2024, pp. 3688–3709. DOI: 10.18653/v1/2024.emnlp-main.216.
- [8] Grootendorst, Maarten. “KeyBERT: Minimal keyword extraction with BERT.” Version v0.3.0. 2020. DOI: 10.5281/zenodo.4461265. URL: <https://doi.org/10.5281/zenodo.4461265>.

- [9] Jung, Matthias and Wengeler, Martin. “Wörter - Argumente - Diskurse Was die Öffentlichkeit bewegt und was die Linguistik dazu sagen kann”. In: *Sprache - Sprachwissenschaft -Öffentlichkeit*, ed. by Gerhard Stickel. Berlin, Boston: De Gruyter, 1999, pp. 143–171. DOI: 10.1515/9783110622645-012.
- [10] Kiemes, Carina, Müller, Marcus, and Wengeler, Martin. “The Analysis of Argumentation Topoi: A Qualitative Approach Goes to Large Corpora”. In: *Nota Bene* 2, no. 1 (2025), ed. by Steven Schoonjans. Special issue: *The Quality of Quantity, the Quantity of Quality*, pp. 102–126.
- [11] Mansour, Nacef Ben, Rahimi, Hamed, and Alrahabi, Motasem. “How Well Do Large Language Models Extract Keywords? A Systematic Evaluation on Scientific Corpora”. In: *Proceedings of the 1st Workshop on AI and Scientific Discovery: Directions and Opportunities*, ed. by Peter Jansen, Bhavana Dalvi Mishra, Harsh Trivedi, Bodhisattwa Prasad Majumder, Tom Hope, Tushar Khot, Doug Downey, and Eric Horvitz. Albuquerque, New Mexico, USA: Association for Computational Linguistics, 2025, pp. 13–21. DOI: 10.18653/v1/2025aisd-main.2.
- [12] Pfister, Jan, Wunderle, Julia, and Hotho, Andreas. “LLÄMmlein: Transparent, Compact and Competitive German-Only Language Models from Scratch”. 2025. DOI: 10.48550/arXiv.2411.11171.
- [13] Stede, Manfred and Schneider, Jodi. *Argumentation Mining*. Cham: Springer, 2019.
- [14] Thellmann, Klaudia et al. “Towards Multilingual LLM Evaluation for European Languages”. 2024. DOI: 10.48550/arXiv.2410.08928.
- [15] Wengeler, Martin. “Patterns of argumentation and the heterogeneity of social knowledge”. In: *Journal of Language and Politics* 14, no. 5 (Dec. 2015), pp. 689–711. DOI: 10.1075/jlp.14.5.04wen.
- [16] Wengeler, Martin. *Topos und Diskurs: Begründung einer argumentationsanalytischen Methode und ihre Anwendung auf den Migrationsdiskurs (1960-1985)*. Vol. 244. de Gruyter, 2012.

A Appendix

A.1 Annotation Guidelines

Below are the annotation guidelines for human annotators used to produce a gold keyword set (translated from German).

Annotation goal. The following texts contain arguments for or against refugees, drawn from German media debate on refugee reception (2015–2017). Your task is to identify and extract **1 to 6 keywords** per argument that best support the position taken.

Keywords may consist of single words (e.g., *integration*), phrases (e.g., *the integration of refugees*), clauses (e.g., *that they came to Germany out of necessity*), or a maximum of one full sentence (but not paragraphs).

Keywords must be **representative of the argument**, capturing its evaluative stance. For example:

Willy Brandt's outlook is grim. The West, the chancellor prophesied last Friday at the sidelines of the EC summit in Copenhagen, faces “its greatest challenge since the Great Depression.”

(SPIEGEL, 17.12.1973, via <https://diskursmonitor.de/glossar/topos/>)

Good keywords: *grim; the greatest challenge since the Great Depression*

Poor keyword: *at the sidelines of the EC summit in Copenhagen*

Annotation instructions:

- Extract well-formed, meaningful spans; avoid fragments (e.g., *the greatest*).
- Prefer contentful parts of arguments (e.g., *grim over is grim*).
- Do not add or paraphrase content. Extract spans exactly as they appear in the text.
- Include preceding articles (e.g., *the challenge*).
- Separate multiple keywords with semicolons.
- Avoid subphrases within longer spans unless they carry distinct argumentative meaning.
- If spans overlap, extract the longer variant that more effectively expresses the stance.
- Use your extralinguistic knowledge of the refugee debate to interpret meaning.

Definitions

- **Stance-bearing keyword:** A word, phrase, or clause that expresses a positive or negative stance toward refugees. These may include *stigma terms* (e.g., *burden, fear, foreign infiltration*) or *flag terms* (e.g., *humanity, successful integration*).
- **Argumentation:** Recurrent discourse patterns used to plausibly connect facts, reflecting typical reasoning by specific social groups in historical context.
- **Discourse:** The totality of factors that shape meaning in context—textual, situational, action-related, and extralinguistic.

A.2 Comparison annotation

Table 4: Comparison of adjudicated keywords and annotator selections exemplified on two arguments. English translations are shown in italics below the German entries.

Text (excerpt)	Adjudicated Key-words	Annotator 1	Annotator 2
<p>Ich kenne es nur zu gut. Bei Einbruch der Dunkelheit konnte ich in den Ländern, in denen ich gelebt habe, nicht mehr auf der Straße sein. Da werden so manche Gutmenschen hier noch Aha-Erlebnisse haben. Und wie reagiert die politische Führung auf die Eskalation? Hilflos.</p> <p><i>I know it only too well. After dark, in the countries where I lived, I could no longer be out on the streets. Some do-gooders here will still have their aha moments. And how does the political leadership react to the escalation? Helpless.</i></p>	<p>Gutmenschen; Aha-Erlebnisse; Eskalation <i>do-gooders; aha moments; escalation</i></p>	<p>Aha-Erlebnisse; Eskalation <i>aha moments; escalation</i></p>	<p>Eskalation; Aha-Erlebnisse; Gutmenschen <i>escalation; aha moments; do-gooders</i></p>

Continued on next page

Table 4 – continued from previous page

Text (excerpt)	Adjudicated Key-words	Annotator 1	Annotator 2
<p>Nein, aber es gibt Unsicherheiten, weil wir Fluchtursachen, Fluchtwiege und Fluchtbereitschaft nur relativ einschätzen können. Als die Russen im syrischen Aleppo Bomben warfen, haben sich bis zu 100 000 Menschen in Bewegung gesetzt. Das lässt sich nicht vorhersehen. Anfang 2015 hätte auch niemand damit gerechnet, dass allein aus Hessen mehr als 10 000 Rückführungen in die Heimatländer vollzogen würden.</p> <p><i>No, but there are uncertainties because we can only relatively assess the causes of flight, escape routes, and the willingness to flee. When the Russians dropped bombs on Aleppo, up to 100,000 people began to move. That cannot be predicted. At the beginning of 2015, no one would have expected that more than 10,000 deportations from Hesse alone would be carried out.</i></p>	<p>Unsicherheiten, weil wir Fluchtursachen; Fluchtwiege und Fluchtbereitschaft nur relativ ein- schätzen können; lässt sich nicht vorhersehen <i>uncertainties be- cause we can only assess causes of flight, escape routes and will- ingness to flee rel- atively; cannot be predicted</i></p>	<p>lässt sich nicht vorhersehen; Unsicherheiten; Fluchtursachen <i>cannot be pre- dicted; uncer- tainties; causes of flight</i></p>	<p>lässt sich nicht vorhersehen; Unsicher- heiten, weil wir Fluchtursachen; Fluchtwiege und Fluchtbereitschaft nur relativ ein- schätzen können <i>cannot be pre- dicted; uncertain- ties because we can only assess causes of flight, escape routes and willingness to flee relatively</i></p>

A.3 Prompting instruction-tuned BübleLM

To adapt BübleLM to the task of discourse-relevant keyword extraction, we instruction-tuned the model using a custom prompt format. Each training instance consisted of a single argumentative text and an output string containing one to six semicolon-separated keyword spans. The prompt included an explicit instruction derived from the annotation guidelines used for human annotation.

Aufgabe: Extrahiere ein bis sechs Schlüsselbegriffe oder -ausdrücke aus dem folgenden Text, die die argumentative Perspektive am besten repräsentieren. Die Schlüsselbegriffe sollen positive oder negative Einstellungen gegenüber Flüchtlingen versprachlichen. Sie können aus Stigmawörtern (z. B. „Belastung“, „Angst“, „Überfremdung“) oder Fahnenwörtern (z.B. „Humanität“, „gelungene Integration“, „wenn sie Teil dieser Gesellschaft werden“) bestehen.

Ein Schlüsselbegriff ist ein bedeutungstragendes Wort, eine Phrase, ein Satzteil oder ein ganzer Satz (jedoch kein Absatz). Wähle wohlgeformte, zusammenhängende Ausdrücke, wie sie im Text vorkommen, ohne sie zu paraphrasieren oder zu verkürzen. Artikel am Anfang sollen einbezogen werden. Subphrasen innerhalb längerer Ausdrücke sind nur zu extrahieren, wenn sie eine eigene argumentative Bedeutung tragen. Bevorzuge längere Varianten, wenn diese eine Einstellung ausdrücken. Die Begriffe sollen durch Semikola getrennt werden. Die Auswahl erfordert auch außersprachliches Wissen über die Flüchtlingsdebatte.

English translation:

Task: Extract one to six key terms or expressions from the following text that best represent the argumentative perspective. These key elements should express positive or negative attitudes toward refugees. They may include stigma terms (e.g., “burden,” “fear,” “overforeignization”) or virtue terms (e.g., “humanity,” “successful integration,” “if they become part of this society”).

A keyword is a meaningful word, phrase, sentence fragment, or a single full sentence, but not a paragraph. Select well-formed, contiguous expressions exactly as they appear in the text, without paraphrasing or shortening them. Include any preceding articles. Subphrases within longer expressions should only be extracted if they carry a distinct argumentative meaning. Prefer longer variants if they express an attitude. Terms should be separated by semicolons. Selection also requires extra-linguistic knowledge about the refugee debate.

A.4 3-shot prompt on top of instruction-tuned BübleLM

We used three manually selected examples from the training set to enable the model to "see" domain-specific patterns. The examples were provided as a full input text + corresponding gold keyword spans. They vary in stance towards refugees (positive, negative, and ambiguous)

Extrahiere die wichtigsten Schlüsselbegriffe, phrasen oder einzelne Sätze (aber keinen ganzen Absatz), die die Einstellung gegenüber Migration/den Flüchtlingen am besten versprachlichen aus der folgenden Argumentation. Gib sie als Liste an, getrennt durch Semikola.

'Extract the most important keywords, phrases or single sentences (but no paragraphs) that best express the stance toward migration/the refugees from the following argument. Provide them as a list, separated by semicolons.'

Beispiel 1 /Example 1 *stance: neutral/both positive and negative, not shown to model:*

Text: Allerdings geht es der BA auch nicht anders als den übrigen staatlichen Stellen sie wissen nur wenig über die Menschen, die da gekommen sind. Alter, Geschlecht, Nationalität, das schon falls die Angaben denn stimmen. Aber ansonsten? Deshalb geht eine schnelle Eingreiftruppe in Flüchtlingsunterkünfte, um sich mehr Informationen zu holen: welche Sprachen die Flüchtlinge sprechen, von welchen Jobs sie träumen, was sie dafür mitbringen. Early Intervention heißt das Modellprojekt, das die BA seit 2014 in mittlerweile neun Städten testet und das von Januar 2016 an bundesweit Standard wird.

'However, the Federal Employment Agency (BA) is no different from other government authorities — they know very little about the people who have arrived. Age, gender, nationality — that much, if the information is accurate. But beyond that? That's why a rapid response team is sent into refugee shelters to gather more information: which languages the refugees speak, what jobs they aspire to, and what qualifications they possess. This model project, called "Early Intervention," has been tested in nine cities by the BA since 2014 and will become standard nationwide starting in January 2016.'

Schlüsselbegriffe: wissen nur wenig über die Menschen, die da gekommen sind; sich mehr Informationen zu holen

Keywords: know very little about the people who have arrived; gather more information

Beispiel 2: / Example 2: *stance: positive, not shown to the model*

Text: Die vielen Flüchtlinge, die ich in den letzten Monaten kennengelernt habe, sind mehr als willens und bereit, sich zu integrieren und ganz schnell Teil unserer Gesellschaft zu werden, sagte Dreyer weiter.

'The many refugees I have gotten to know in recent months are more than willing and ready to integrate and to become part of our society very quickly, Dreyer continued.'

Schlüsselbegriffe: willens und bereit sich zu integrieren; ganz schnell Teil unserer Gesellschaft zu werden

Keywords: more than willing and ready to integrate; become part of our society very quickly

Beispiel 3: /Example 3 *stance: negative, not shown to the model*

Text: Eine Frau berichtet von Tätern, die sie für Nordafrikaner hielt, bei einem anderen mutmaßlichen Täter könne laut Abendblatt und Mopo Migrationshintergrund nicht ausgeschlossen werden.

'A woman reports about perpetrators she believed to be North Africans; in the case of another suspected perpetrator, a migrant background cannot be ruled out according to Abendblatt and Mopo.'

Schlüsselbegriffe: Tätern, die sie für Nordafrikaner hielt; Migrationshintergrund nicht ausgeschlossen

Keywords: perpetrators she believed to be North Africans; a migrant background cannot be ruled out

Text: [neuer Text]

Schlüsselbegriffe:

A.5 Training diagnostics for instruction-tuned BübleLM

Metric	Value
Average throughput	4.67 samples/sec
Training steps per second	0.585
Final training loss	5.82
Initial training loss	5.91
Training loss at step 1200	5.74
Validation loss stabilized after	Step 400
Minimum validation loss (step 1200)	5.8326
Overfitting observed	No

Table 5: Training diagnostics for instruction-tuned BübleLM

A.6 Exact match results

Model	Exact Match (%)	95% CI
BübleLM (3-shot)	0.00	(0.00, 0.00)
BübleLM (instruction-tuned)	0.25	(0.00, 0.74)
KeyBERT (Baseline)	0.00	(0.00, 0.00)
GBERT (NER-tuned, Baseline)	4.21	(2.23, 5.96)

Table 6: Exact match accuracy with 95% confidence intervals, computed via bootstrap resampling (n=1000). A prediction is counted as correct only if all predicted spans exactly match the gold spans in form and count.

A.7 Error analysis

Table 8: Three high-scoring “partial match” examples from the instruction-tuned BübleLM model. Each predicted span partially overlaps with at least one gold span. English translations are shown in italics below each German entry.

Text (excerpt)	Gold Keywords	Model Prediction	P	R	F1
Das Ifo-Institut in München ist dagegen skeptisch und warnt, dass Einwanderung zu einem Verlustgeschäft werden könnte. Deutschland zieht nicht gerade die am besten ausgebildeten Migranten an, urteilt Ifo-Chef Hans-Werner Sinn. <i>The Ifo Institute in Munich is skeptical and warns that immigration could become a losing proposition. Germany does not exactly attract the best-educated migrants, says Ifo head Hans-Werner Sinn.</i>	Einwanderung zu einem Verlustgeschäft werden könnte; Deutschland zieht nicht gerade die am besten ausgebildeten Migranten an <i>immigration could become a losing proposition; Germany does not attract highly educated migrants</i>	Deutschland zieht nicht gerade die am besten ausgebildeten Migranten an; Einwanderung zu einem Verlustgeschäft <i>Germany doesn't attract educated migrants; immigration as a loss</i>	1.0	1.0	0.999
Wie viele Flüchtlinge hat Haver & Boecker denn schon eingestellt? Einen Syrer als Praktikanten. Einen? Warum nicht zehn? So viel mehr Arbeit haben wir nicht. Wir hätten auch gar nicht die richtigen Bewerber. Das ist ein Problem. <i>How many refugees has Haver & Boecker hired? One Syrian intern. One? Why not ten? We don't have that much more work. We also don't really have the right applicants. That's a problem.</i>	gar nicht die richtigen Bewerber; ein Problem <i>not the right applicants; a problem</i>	gar nicht die richtigen Bewerber; ein Problem; nicht die richtigen Bewerber <i>not the right applicants; a problem</i>	1.0	1.0	0.999

Text (excerpt)	Gold Keywords	Model Prediction	P	R	F1
Eine Anwohnerin sprach von ihrer Angst, abends auf die Straße zu gehen und ihren Kindern das Radfahren zu erlauben. <i>A local resident spoke of her fear of going out in the evening and letting her children ride their bikes.</i>	Angst; abends auf die Straße zu gehen; Kindern das Radfahren zu erlauben <i>fear; going out in the evening; letting children cycle</i>	Angst; abends auf die Straße zu gehen; Kindern das Radfahren zu erlauben <i>fear; going out in the evening; letting children cycle</i>	1.0	1.0	0.999

Table 9: Three “complete miss” examples from the 3-shot instruction-tuned BübleLM model. None of the predicted spans overlap with the gold keywords. Soft precision, recall, and F1 are all 0.0. English translations are shown in italics.

Text (excerpt)	Gold Keywords	Model Prediction	P	R	F1
Das Problem ist, die Fähigkeiten für den deutschen Arbeitsmarkt zu bewerten. In Deutschland gibt es für die meisten Berufe eine standardisierte Ausbildung. In Syrien, Irak oder Iran sind viele Laufbahnen weniger streng reguliert. Das ist das eine. Das andere ist die Verständigung: Eigentlich bräuchte man für das Profiling einen Übersetzer, der sich gut in der Materie auskennt. Aber dafür gibt es keine Mittel. <i>The problem is evaluating skills for the German job market. Most professions have standardized training in Germany. In Syria, Iraq, or Iran, careers are less regulated. The other issue is communication: you'd need a translator who knows the field. But there's no funding.</i>	das Problem; viele Laufbahnen weniger streng reguliert <i>the problem; less regulated careers</i>	Profiling; Übersetzer; keine Mittel; Verständigung <i>profiling; translator; no funding; communication</i>	0.0	0.0	0.0

Text (excerpt)	Gold Keywords	Model Prediction	P	R	F1
Natürlich gab es auch in Waldkirch Bedenken. Den ganzen Sommer über gab es deutschlandweit Meldungen von sexueller Belästigung in Freibädern. Die rechte Facebookseite XY-Einzelfall listet 160 Vorfälle in diesem Jahr, von üblen Sex-Attacken war in der Bild-Zeitung die Rede. <i>There were also concerns in Waldkirch. Nationwide reports of sexual harassment in pools appeared all summer. The right-wing Facebook page XY-Einzelfall listed 160 cases; the Bild newspaper spoke of vile sex attacks.</i>	Bedenken; Meldungen von sexueller Belästigung in Freibädern <i>concerns; reports of harassment in pools</i>	160 Vorfälle; üble Sex-Attacken <i>160 incidents; vile sex attacks</i>	0.0	0.0	0.0

Text (excerpt)	Gold Keywords	Model Prediction	P	R	F1
<p>In diesen Situationen hilft der Gebrauch des Verstandes oft ungemein weiter: Bei Flüchtlingen handelt es sich um Menschen wie Du und ich, mit allen Stärken und Schwächen, Vorlieben und Abneigungen, sie hatten – im Gegensatz zu uns Europäern – lediglich das Pech, an einem lebensfeindlichen Ort geboren zu werden. Will sagen, auch unter Flüchtlingen passieren unschöne Dinge, genau wie unter unseresgleichen. Man sollte zugunsten der Neuankommelinge in Anrechnung bringen, dass sie sich in einer absoluten Ausnahmesituation befinden und die angesichts dieses Umstandes auftretenden Überreaktionen sind bewundernswert niedrig.</p> <p><i>Refugees are people like you and me, with strengths and weaknesses. They were unlucky to be born in hostile places. One should recognize that they are in an exceptional situation. Their overreactions are admirably rare.</i></p>	<p>Menschen wie Du und ich; Stärken und Schwächen; an einem lebensfeindlichen Ort geboren; absolute Ausnahmesituation</p> <p><i>people like you and me; born in hostile places; exceptional situation</i></p>	<p>unschöne Dinge; bewundernswert niedrig; Umstand</p> <p><i>unpleasant things; admirably rare; circumstance</i></p>	0.0	0.0	0.0

Table 10: Three best “partial match” examples from the 3-shot instruction-tuned BübleLM model. All predicted spans partially overlap with gold spans. English translations are shown in italics.

Text (excerpt)	Gold Keywords	Model Prediction	P	R	F1
Aus der Praxis ist zu hören, die Motivation sei zumeist nicht das Problem. Der Deutsche Gewerkschaftsbund verweist darauf, dass der Integrationswille erheblich größer sei als das Angebot der Bundesregierung. <i>From practice we hear that motivation is usually not the problem. The German Trade Union Confederation points out that the willingness to integrate is far greater than the government’s offer.</i>	Motivation sei zumeist nicht das Problem; Integrationswille erheblich größer sei als das Angebot der Bundesregierung <i>motivation not the problem; willingness to integrate > gov. support</i>	Integrationswille; Motivation sei zumeist nicht das Problem; erheblich größer sei ... <i>willingness to integrate; motivation not the problem; far greater than offer</i>	1.0	0.94	0.97

Text (excerpt)	Gold Keywords	Model Prediction	P	R	F1
<p>22 Jahre später gibt es eine andere Ausgangssituation. Zum einen, weil das Asylrecht ja bereits so sehr ausgehöhlt ist, dass eine weitere Verschärfung kaum denkbar scheint. Zum anderen aber auch, weil das Land nicht mehr durch die Nachwhehen der Wiedervereinigung verunsichert ist, sondern – auch weil es von der Finanzkrise in anderen EU-Staaten profitiert – wirtschaftlich gut dasteht. Wir können es uns leisten, die Flüchtlinge mit Wohlwollen aufzunehmen.</p> <p><i>22 years later, the situation has changed. Asylum law is already so weakened that further tightening seems unthinkable. And the country is no longer insecure after reunification, but instead—thanks to the financial crisis in other EU states—economically stable. We can afford to welcome refugees benevolently.</i></p>	wirtschaftlich gut dasteht; können es uns leisten; die Flüchtlinge mit Wohlwollen aufzunehmen <i>economically stable; can afford; welcome refugees</i>	wirtschaftlich gut dasteht; können es uns leisten; die Flüchtlinge mit Wohlwollen aufzunehmen <i>economically stable; can afford; welcome refugees</i>	0.94	1.0	0.97

Text (excerpt)	Gold Keywords	Model Prediction	P	R	F1
<p>Das Ministerium nennt diese aber in einer Reihe mit normalen Kriminellen, Kriegsverbrechern und Einzelpersonen extremistischer Gesinnung. Mit anderen Worten: Wenn so viele Menschen kommen, dann sind darunter eben auch ein paar schwarze Schafe.</p> <p><i>The ministry lists them alongside criminals, war criminals, and extremists. In other words: with so many people arriving, a few black sheep are among them.</i></p>	<p>darunter eben auch ein paar schwarze Schafe <i>a few black sheep among them</i></p>	<p>ein paar schwarze Schafe; so viele Menschen kommen; darunter eben auch ... <i>a few black sheep; so many people come</i></p>	0.90	1.0	0.95

Table 11: Three “complete miss” examples from the GBERT NER-tuned model. All predicted spans yielded soft precision, recall, and F1 scores of 0.0. English translations are shown in italics.

Text (excerpt)	P	R	F1
<p>Das Problem ist, die Fähigkeiten für den deutschen Arbeitsmarkt zu bewerten. In Deutschland gibt es für die meisten Berufe eine standardisierte Ausbildung. In Syrien, Irak oder Iran sind viele Laufbahnen weniger streng reguliert. Das ist das eine. Das andere ist die Verständigung: Eigentlich bräuchte man für das Profiling einen Übersetzer, der sich gut in der Materie auskennt. Aber dafür gibt es keine Mittel.</p> <p><i>The problem is evaluating skills for the German job market. Most jobs in Germany require formal training. In Syria, Iraq, or Iran, many careers are less regulated. Another issue is communication: you'd need a translator familiar with the topic. But there's no funding.</i></p>	0.0	0.0	0.0

Text (excerpt)	P	R	F1
<p>Dass zwei Männer, die als Flüchtlinge ins Land kamen, für die Taten von Würzburg und Ansbach verantwortlich seien, verhöhne nicht nur das Land, das sie aufgenommen habe. Es verhöhnt die Helfer, die Ehrenamtlichen, die sich so sehr um die Flüchtlinge kümmern. Und es verhöhnt die vielen anderen Flüchtlinge, die wirklich Hilfe vor Gewalt und Krieg bei uns suchen, sagte Merkel.</p> <p><i>That two men who came as refugees were responsible for the Würzburg and Ansbach attacks mocks not only the country that received them. It mocks the volunteers who care for refugees—and the many others who truly seek protection from war and violence, said Merkel.</i></p>	0.0	0.0	0.0
<p>Zunächst muss man strikt unterscheiden zwischen Schutzbedürftigen und nicht Schutzbedürftigen, sonst lösen wir einen Sogeffekt aus, den wir nicht bewältigen können – und wir verlieren die Zustimmung der Bevölkerung. Deswegen dränge ich darauf, dass wir Asylbewerber aus den Westbalkanländern klar sagen, dass sie keine Chance haben, in Deutschland als politisch Verfolgte anerkannt zu werden.</p> <p><i>We must distinguish strictly between those in need of protection and those who are not—otherwise we risk an uncontrollable pull effect and lose public support. That's why I insist that asylum seekers from the Western Balkans must be told clearly they have no chance of being recognized as politically persecuted in Germany.</i></p>	0.0	0.0	0.0

Table 12: Three best-matching examples from the GBERT NER-tuned model. These predictions show high soft precision, recall, and F1 values. English translations are shown in italics.

Text (excerpt)	P	R	F1
Die Studentenstadt im Dreiländereck hat allerdings seit Jahren Probleme mit unbegleiteten minderjährigen Ausländern. Im Frühjahr 2014 führte das erstmals zu größeren politischen Debatten: Lange vor den Ereignissen auf der Kölner Domplatte hatte eine Gruppe von 40 minderjährigen Flüchtlingen aus Nordafrika Anwohner und Passanten am Stühlinger Kirchplatz regelrecht terrorisiert. Von Januar bis Mitte Mai 2014 zählte die Polizei allein an diesem Platz 232 Straftaten. <i>The student city at the tri-border region has had problems with unaccompanied minor foreigners for years. In spring 2014, this led to major political debates: long before the events at Cologne Cathedral Square, a group of 40 minor refugees from North Africa harassed residents and passers-by at Stühlinger Kirchplatz. From January to mid-May 2014, police recorded 232 crimes at that square alone.</i>	1.0	1.0	1.0
Ifo-Forscher um Michele Battisti mutmaßen, dass Einwanderung Unternehmer dazu verleitet, mehr Arbeitssplätze anzubieten, von denen auch Einheimische profitieren. Der Grund liegt darin, dass Zuwanderer oft weniger scharf verhandeln und so weniger verdienen als ihre einheimischen Kollegen. <i>Ifo researchers around Michele Battisti suggest that immigration encourages employers to offer more jobs, which also benefit locals. The reason is that immigrants often negotiate less aggressively and thus earn less than their local colleagues.</i>	1.0	0.80	0.89
“Zwölf Jahre hat Ibrahim im Tourismus gearbeitet und träumt nun von einem Job in Deutschland. Hier bekomme ich mehr Eindrücke und Informationen als in meinem Heim, in dem ich nur dahinvegetiere, sagt er. Dieses Projekt soll auch eine Brücke zwischen den Einheimischen und den Flüchtlingen herstellen. Es soll Verständnis füreinander schaffen, ergänzt Eben.” <i>“Ibrahim worked in tourism for twelve years and now dreams of a job in Germany. Here I get more impressions and information than in my home, where I just vegetate, he says. This project is also meant to build a bridge between locals and refugees. It aims to create mutual understanding, adds Eben.”</i>	0.79	1.0	0.88

Table 13: Three “complete miss” examples from KeyBERT. The predicted keywords do not overlap with the gold annotations. English translations are shown in italics.

Text (excerpt)	Gold Keywords	Predicted Key-words	P	R	F1
Müller: Unsere Syrer hatten Angst in den Massenunterkünften vor einigen Landsleuten, die sie dem Daesch zuordneten. Und sie selbst sind nicht radikalisierbar? Die sind aus solcher Not, mit solchen Sehnsüchten hierhergekommen, so warmherzig, haben so viel auf sich genommen, aber wenn es nun nicht klappt mit Arbeiten, Geldverdienen, Integration, sondern Perspektivlosigkeit zur Realität wird, dann lege ich meine Hand für keinen ins Feuer. <i>Müller: Our Syrians were afraid in mass shelters of some compatriots they linked to Daesh. They came here from hardship, with longing, warm-hearted, endured a lot. But if integration fails and hopelessness becomes reality, I won't vouch for anyone.</i>	nicht klappt mit Arbeiten Geldverdienen Integration; Perspektivlosigkeit zur Realität wird	syrer hatten angst in den massenunterkünften vor einigen;			
müller unsere syrer hatten angst in den massenunterkünften	0.0	0.0	0.0	0.0	0.0

Text (excerpt)	Gold Keywords	Predicted Key-words	P R F1
<p>Wie 1990 erwartet uns eine Herausforderung, die Generationen beschäftigen wird. Doch anders als damals soll nun zusammenwachsen, was bisher nicht zusammen gehörte. Ost- und Westdeutsche hatten ja dieselbe Sprache, blickten auf dieselbe nationale Kultur und Geschichte zurück. Ost- und Westdeutsche standen selbst in Zeiten der Mauer durch Kirchengemeinden, Verwandte oder Freunde in direktem Kontakt miteinander und wussten über die Medien voneinander Bescheid. Wie viel größere Distanzen dagegen sind zu überwinden in einem Land, das zum Einwanderungsland geworden ist. Zu diesem Land gehören heute Menschen verschiedener Herkunftsländer, Religionen, Hautfarben, Kulturen – Menschen, die vor Jahrzehnten eingewandert sind, und zunehmend auch jene, die augenblicklich und in Zukunft kommen, hier leben wollen und auch eine Bleibeperspektive haben.</p> <p><i>Just like 1990, we face a generational challenge. But now we must unite what was never united before. East and West Germans shared culture and stayed in contact. Now, in a diverse immigration country, much greater distances must be overcome.</i></p>	Herausforderung, die Generationen beschäftigen wird; größere Distanzen; überwinden; Menschen verschiedener Herkunftsländer, Religionen, Hautfarben, Kulturen	zusammen gehörte ost und west- deutsche hatten ja dieselbe sprache;	

Text (excerpt)	Gold Keywords	Predicted Key-words	P R F1
nationale kultur und geschichte zurück ost und westdeutsche	0.0	0.0	0.0
<p>Vorrangig ist es jetzt, die Menschen so schnell wie möglich arbeitsfähig zu machen. Das kann die Bundesagentur für Arbeit nicht allein. Daher ist es gut und lobenswert, dass sich viele Unternehmen – oft auf Wunsch der Mitarbeiter – Gedanken machen, was sie tun können – von Sprachkursen über Lebenshilfeangebote (Wie schreibe ich eine Bewerbung?) bis hin zu Praktikantenstellen. Der Gesetzgeber sollte sich überlegen, ob er in sein Vorhaben, die Wartezeit von Asylanten bis zur Aufnahme einer Festanstellung zu verkürzen, nicht auch die Teilzeitarbeit einbezieht. Sie könnte vielen Flüchtlingen den Einstieg ins Erwerbsleben erleichtern. Niemand kann sagen, welche Maßnahme was bringt. Sicher ist nur, dass alle denkbaren Möglichkeiten ergriffen und umgesetzt werden müssen, wenn die Integration Erfolg haben soll.</p> <p><i>The priority is to make people employable quickly. Agencies can't do it alone. Companies help with language courses and internships. Lawmakers may allow part-time work earlier. Refugees could thus ease into the labor market. All options must be implemented for integration to succeed.</i></p>	<p>den Einstieg ins Erwerbsleben erleichtern; alle denkbaren Möglichkeiten ergriffen und umgesetzt werden müssen</p>	<p>wartezeit von asylanten bis zur aufnahme;</p>	

Text (excerpt)	Gold Keywords	Predicted Key-words	P	R	F1
menschen so schnell wie möglich arbeitsfähig	0.0	0.0		0.0	

Table 14: Three best “partial match” examples from KeyBERT. Predicted keywords partially overlap with gold keywords. English translations of text and keywords are shown in italics.

Text (excerpt)	Gold Keywords	Predicted Key-words	P	R	F1
<p>Wie würden Sie die Angst vor Kriminalität entkräften? Wir hatten einen Polizeigewerkschafter zum Vortrag geladen. Er hat mit Zahlen belegt, dass Flüchtlinge in keiner Weise krimineller sind als Deutsche. <i>How would you counter the fear of crime? We invited a police unionist to give a talk. He showed with data that refugees are in no way more criminal than Germans.</i></p> <p>wie würden sie die angst vor kriminalität entkräften <i>refugees in no way more criminal than Germans; how would you counter the fear of crime</i></p>	<p>Flüchtlinge in keiner Weise krimineller sind als Deutsche <i>Refugees are in no way more criminal than Germans</i></p>	<p>flüchtlinge in keiner weise krimineller sind als deutsche;</p>	0.78	1.00	0.88

Text (excerpt)	Gold Keywords	Predicted Key-words	P R F1
<p>Auch Flüchtlinge mit eingeschränktem Aufenthaltrecht sollten nach Ansicht des Chefs der Arbeitsagenturen, Frank-Jürgen Weise, unter bestimmten Voraussetzungen in Deutschland bleiben dürfen. Wenn jemand nicht kriminell ist, integriert in Sprache und Arbeit ist, dann spricht aus meiner Arbeitsmarktsicht sehr viel dafür, dass so jemand bleiben darf, sagte er der Deutschen Presse-Agentur.</p> <p><i>Even refugees with restricted residence rights should, under certain conditions, be allowed to stay in Germany, says Frank-Jürgen Weise.</i></p> <p><i>If someone is not criminal and is integrated into work and language, there is strong support for letting them stay.</i></p> <p>in deutschland bleiben dürfen wenn jemand nicht kriminell</p> <p><i>stay in Germany under certain conditions; stay in Germany if not criminal</i></p>	<p>unter bestimmten Voraussetzungen in Deutschland bleiben dürfen</p> <p><i>be allowed to stay in Germany under certain conditions</i></p>	<p>unter bestimmten voraussetzungen in deutschland bleiben dürfen;</p>	<p>0.78</p> <p>1.00</p> <p>0.88</p>

Text (excerpt)	Gold Keywords	Predicted Key-words	P	R	F1
<p>Wie die Kölner Polizei in der vergangenen Woche bereits in einem Bericht für den Innenausschuss des Landtags ausführte, seien Täter aus nordafrikanischen Ländern vor allem im Bereich Taschendiebstahl aktiv. Mittlerweile sind 40 Prozent der ermittelten Taschendiebe aus Nordafrika, sagte Günther Korn, Leiter des Kölner Taschendiebstahls-Kommissariats. Die signifikante Entwicklung dieser Tätergruppe verdeutlichte er mit folgenden Zahlen: In den Jahren 2010 und 2011 habe es in Köln 15 polizeibekannte nordafrikanische Straftäter gegeben. 2014 gab es dann bereits 1.000.</p> <p><i>According to a recent report by Cologne police, perpetrators from North African countries are especially active in pickpocketing. By 2014, 40 percent of identified pickpockets in Cologne were from North Africa.</i></p>	<p>Täter aus nordafrikanischen Ländern vor allem im Bereich Taschendiebstahl aktiv;</p>				
<p>40 Prozent der ermittelten Taschendiebe aus Nordafrika</p> <p><i>Offenders from North Africa active in pickpocketing; 40% of pickpockets from North Africa</i></p> <p>nordafrikanischen ländern vor allem im bereich taschendiebstahl aktiv</p> <p><i>percent of identified pickpockets from North Africa; North Africans active in pickpocketing</i></p>	<p>prozent der ermittelten taschendiebe aus nordafrika;</p>	0.81	0.90	0.85	

Text (excerpt)	Gold Keywords	Model Prediction
<p>Zeit: Fürchten Sie die islamische Einwanderung? Heinsohn: Im Gegen teil. Sowie Muslime auf dem Radar der internationalen Talentscouts auftauchen, soll man zugreifen, bevor es andere tun. Der Islamismus in Deutschland entspringt selten einer religiösen Aufwallung. Es geht um missachtete Verlierer, die sich einer Ideologie der Überlegenheit verschreiben, um auf fatale Weise Würde zu erlangen.</p> <p><i>Zeit: Are you afraid of Islamic immigration? Heinsohn: On the contrary. As soon as Muslims appear on talent scouts' radar, one should act before others do. Islamism rarely stems from religion; it's about losers seeking dignity through superiority.</i></p>	<p>soll man zugreifen; bevor es andere tun <i>one should act; before others do</i></p>	<p>missachtete Verlierer; fatale Weise Würde zu erlangen; Islamismus in Deutschland entspringt selten ...; sich einer Ideologie verschreiben <i>neglected losers; fatal dignity; Islamism not religious; ideological superiority</i></p>
<p>Wenn von der Terrorgefahr geredet wird, die von jungen, wütenden Männern ausgeht, dann ist Deutschland zurzeit leider auch von seinen eigenen Einwohnern bedroht, und es wäre dringend an der Zeit, zu fragen, welche massiven politischen Fehler, welche sozial- und wirtschaftspolitische Wurschtigkeit dazu führen konnte, dass in französischen Banlieues und in deutschen Kleinstädten so viele junge Männer so werden konnten, wie sie geworden sind.</p> <p><i>Germany is also threatened by its own citizens. What massive political failures led young men in suburbs to radicalize?</i></p>	<p>Terrorgefahr; wütenden Männern; eigenen Einwohnern bedroht <i>terror threat; angry men; own citizens</i></p>	<p>massiven politischen Fehler; sozial- und wirtschaftspolitische Wurschtigkeit; junge Männer wurden wie sie sind <i>political failures; economic indifference; became what they are</i></p>
<p>Es dauerte ein wenig, aber dann kam bei Merkels Politikollegen wie bei ihren WählerInnen die Angst hoch. Die elementare Angst in fast allen Menschen vor Fremden und vor großen Veränderungen. Wobei: So groß ist die Veränderung gar nicht.</p> <p><i>Fear rose in Merkel's camp — the fear of foreigners and big changes. But the change isn't even that big.</i></p>	<p>So groß ist die Veränderung gar nicht <i>change isn't that big</i></p>	<p>Angst; Angst; Angst <i>fear; fear; fear</i></p>

Table 7: Three worst “complete miss” examples from the instruction-tuned BübleLM model: All predicted spans scored 0.0 on soft precision, recall, and F1. English translations are shown in italics below each German entry.