# Twente-BMS-NLP at PerspectiveArg 2024: Combining Bi-Encoder and Cross-Encoder for Argument Retrieval

**Leixin Zhang**
University of Twente
l.zhang-5@utwente.nl

**Daniel Braun**
University of Twente
d.braun@utwente.nl

## Abstract

The paper describes our system for the Perspective Argument Retrieval Shared Task. The shared task consists of three scenarios in which relevant political arguments have to be retrieved based on queries (Scenario 1). In Scenario 2 explicit socio-cultural properties are provided and in Scenario 3 implicit socio-cultural properties within the arguments have to be used. We combined a Bi-Encoder and a Cross-Encoder to retrieve relevant arguments for each query. For the third scenario, we extracted linguistic features to predict socio-demographic labels as a separate task. However, the socio-demographic match task proved challenging due to the constraints of argument lengths and genres. The described system won both tracks (relevance and diversity) of the shared task.

## 1 Introduction

The perspective argument retrieval shared task (Falk et al., 2024) addresses the challenge of incorporating socio-cultural factors into argument retrieval. It is based on the x-stance dataset (Vamvas and Sennrich, 2020) and includes three scenarios: baseline, explicit, and implicit. Queries in the baseline scenario are provided without socio-demographic requirements on extracted arguments. In the explicit and implicit scenarios, one socio-demographic feature is provided along with the query as an additional requirement to retrieve arguments that match the label, (e.g.{gender: male}). An extracted argument is considered a true candidate only if it is relevant to the query text and matches the socio-demographic label provided with the query in explicit and implicit scenarios.

The difference between the explicit and implicit scenarios is that in the explicit scenario, the socio-demographic information of argument authors is available in the corpus, whereas in the implicit scenario it is not. This means that in the implicit scenario, systems must predict or match the socio-demographic features in addition to determining relevance to the query.

In our system, we combine bi-encoder and cross-encoder models to retrieve relevant arguments. Additionally, we predict socio-demographic features from argument texts in the implicit scenario, using sentence embeddings, n-gram of part-of-speech (POS) tags and stop words, and token length distributions as input features. The system performed best in both tracks of the shared task, relevance and diversity.

## 2 Related Work

Previous studies have employed several strategies to model query and argument sentences before they can be matched subsequently: Simple statistical features like token n-grams or part-of-speech (POS) n-grams (Clement and Sharp, 2003), TF-IDF (Ramos et al., 2003), or word2vec (Sardianos et al., 2015; Jang and Kwon, 2023). Word embeddings with mean pooling or other techniques to generate sentence embeddings of fixed lengths (Devlin et al., 2018; Liu et al., 2019), sentence embedding models such as Sentence-BERT (Reimers and Gurevych, 2019) or LaBSE (Feng et al., 2020), and ensemble approaches combining models from the aforementioned categories (Zhang and Çöltekin, 2024; Reimers et al., 2019).

Sentence embedding models typically use a bi-encoder architecture, such as a Siamese Neural Network. The relevance of two sentences is often measured with the cosine similarity of their embeddings. Bi-encoders are more suitable for symmetric searches where two sentences are interchangeable (Muennighoff, 2022). Asymmetric scenarios like answering a specific question often benefit more from cross-encoder models. Thakur et al. (2020) pointed out the challenge in training bi-encoders to represent two asymmetric sentences within a shared vector space. Beyond encoder architectures,



*Proceedings of the 11th Workshop on Argument Mining*, pages 164–168

Muennighoff (2022) proposed a method that employs decoders for sentence embeddings and semantic search.

### 2.1 Socio-Demographic Features

One goal of the the perspective argument retrieval shared task is to retrieve a set of diverse with regard to the socio-demographic features of the argument providers, such as age and political stances. Most works that attempt to explicitly extract such features require long texts as input. In age and gender prediction, for example, the majority of studies work with texts with 250 words or more (Peersman et al., 2011). By contrast, for the dataset in this shared task, the average length of each argument is less than 30 words (25.86 excluding punctuation tokens). Some studies also work with shorter texts: Peersman et al. (2011) analyzed chat messages averaging 12.2 tokens each and achieved an accuracy of 88.8% for age prediction. Zhang and Zhang (2010) studied blog posts averaging 15 tokens per segment, achieving an accuracy of 72.10% for gender prediction. When shorter texts are used, these are often collected from social media, which may include more personal styles such as abbreviations or emoticons. This may simplify the task compared to the dataset of the shared task, which consists of political arguments from a dedicated platform, less likely to contain such explicit clues.

## 3 Dataset Analysis

In the training and development datasets, we observed that the queries and candidates are structured as follows: the same query is provided in three languages: German, French, and Italian. Though as separate query entries, each language version shares the same list of candidate arguments. Options for addressing this cross-lingual setting include using or fine-tuning cross-lingual sentence embedding models or translating different languages into one.

Additionally, we noted that an argument candidate only appears under one query (or the same query of three language versions). This suggests that the dataset might have been created from an existing set of query arguments (in the X-stance dataset), with arguments randomized and compiled into a mixed arguments corpus, rather than through annotations of argument relevance for each query. This setup could pose challenges for traditional semantic search tasks: if an argument could validly answer two different queries, it is still tied to only the query with which it was originally associated, and appearing under the other query would be considered a false retrieval, even if it might be correct.

A potential strategy to address this issue is to evaluate each candidate's relevance across all queries, assigning it to only one query. However, this approach has a drawback: if an argument is incorrectly assigned to one query, it precludes the possibility of it being correctly assigned to another query. Ultimately, we adopted a hybrid strategy. For queries in the development set, we only considered candidates that had not appeared under training queries. Similarly, for each test set, we only considered candidates that had not appeared in either the training or development queries, which helps narrow down the argument pools and potentially enhance retrieval accuracy. This strategy was implemented across all test sets and scenarios. Nevertheless, we also provide unfiltered results in this paper for broader comparison with the baselines.

## 4 System Design

### 4.1 Baseline Scenario

As discussed in Section 2, previous work has shown that bi-encoders generally perform less well than cross-encoders for asymmetric retrieval tasks. However, cross-encoders have a disadvantage in terms of computing complexity. If there are $M$ samples in the query set and $N$ samples in the corpus, the model needs to be run $M \times N$ times, compared to $M + N$ for bi-encoders. In our system, we employed a strategy to combine both: we used a bi-encoder (`paraphrase-multilingual-mpnet-base-v2`) to retrieve the top 1000 argument candidates and then used a cross-encoder model (`ms-marco-MiniLM-L-12-v2`[1]) to re-rank the top 50. We compared both multilingual and monolingual cross-encoders and found that the monolingual model performs better. Therefore, we translate[2] the top 50 into English before using the cross-encoder.

### 4.2 Explicit Scenario

To extract semantically relevant arguments in the explicit scenario, we applied the same method as

[1] https://www.sbert.net/docs/pretrained-models/ce-msmarco.html

[2] We use Google translate API from the following GitHub repository to translate all queries and top-50 arguments to English: https://github.com/ssut/py-googletrans.

| Factor | Algorithm | Input | Accuracy | Class num. | Prop. bounds |
|---|---|---|---|---|---|
| residence | MLP | sbert embedding | 0.93 | 2 | 0.094 - 0.906 |
| important issues | MLP | sbert embedding | 0.65 | 8 | multi-label case |
| gender | Ran. Forest | bigram POS/STOP | 0.67 | 2 | 0.377 - 0.622 |
| political spectrum | MLP | sbert embedding | 0.52 | 9 | 0.003 - 0.455 |
| civil status | MLP | sbert embedding | 0.44 | 9 | $6.8e^{05}$ - 0.367 |
| age_bin | MLP | sbert embedding | 0.43 | 4 | 0.054 - 0.330 |
| denomination | MLP | sbert embedding | 0.42 | 10 | 0.0004 - 0.409 |
| education | MLP | sbert embedding | 0.34 | 13 | 0.007- 0.281 |

Table 1: Overview of the best-performing approaches for the prediction of socio-demographic features ('Label num.' indicates the number of target labels per factor and 'Prop. bounds' shows the lower and upper bounds of label proportions).

in the baseline scenario. However, in this scenario, the retrieved arguments should not only be semantically relevant but also match the (explicitly provided) socio-demographic features of the query. While integrating these socio-demographic labels into the query and argument texts and converting the extended texts into sentence embeddings could be an option, this approach may introduce additional noise and degrade both relevance and socio-demographic matching. We instead employed exact label matching. For instance, if the query feature is `{gender: female}`, we filtered the corpus to select arguments that match this socio-demographic feature.

### 4.3 Implicit Scenario

In the implicit scenario, the explicitly provided socio-demographic features for the arguments should not be used. This means that, in order to perform the socio-demographic matching, the factors have to be extracted from the text. Our approach is to predict the socio-demographic labels for arguments and then use these predictions in the same way as in the explicit scenario. The training data was collected from the training query documents. We retrieved socio-demographic labels from the query requirement and assigned socio-demographic labels to their corresponding argument candidate lists, creating a pseudo-corpus for socio-demographic feature training.

Our preliminary analysis suggests that for categories like important issues or political spectrum, semantic information is crucial, hence sentence embeddings that capture meaning should be used as input. However, categories such as gender and age, are influenced more by lexical preferences as documented in previous research. Thus, apart from sentence embeddings, we also conducted feature engineering, focusing on German arguments (which comprise about 70% of our corpus). We extracted the following features:

- Token Length Distribution: We used the NLTK package to tokenize sentences and words for each argument, then calculated the token lengths and their distribution.
- POS & Stop Unigram Distribution: We converted all argument texts into part-of-speech (POS) tags for content words while retaining stop words in their original form.
- POS & Stop Bigram Distribution: bigrams from the POS and stop words sequences and computed their distribution.

We input these statistical features and sentence-bert embeddings into MLP, SVM, and random forest models and compare their performance in predicting the different socio-demographic factors.[3]

The best performance for each demographic category prediction is displayed in Table 1. While most categories achieved an accuracy below 0.7, the residence category showed the highest accuracy of 0.93. This performance can most likely be explained by the imbalanced data, with the majority label comprising 90.6% of the data and the minority 9.4%, leading the model to (correctly) predominantly predict the majority label. This issue of imbalance is also present in other socio-demographic factors. Poor accuracy in certain categories can also

[3]Due to time constraints of the shared task, not all combinations of model algorithms, input features, and demographic features were tested; however, MLP was used for all socio-demographic feature predictions.

| | important issues | | gender | |
|---|---|---|---|---|
| top-k | ndcg | precision | ndcg | precision |
| 4 | 0.180 | 0.182 | 0.170 | 0.172 |
| 8 | 0.181 | 0.182 | 0.171 | 0.172 |
| 16 | 0.182 | 0.180 | 0.172 | 0.170 |
| 20 | 0.182 | 0.180 | 0.171 | 0.166 |

Table 2: Prediction results from the development set when matching socio-demographic labels for 'important issues' and 'gender' requirements respectively. We left the remaining categories uncontrolled as the baseline method.

potentially be attributed to the large number of target labels, such as in education (13 labels) and denomination (10 labels). Furthermore, labels within a demographic category are not mutually exclusive, for example, 'Rechts und Konservativ-Liberal' (right and conservative-liberal), 'Rechts und Konservativ' (right and conservative), and 'Rechts und Liberal' (right and liberal) are treated as separate labels, complicating correct assignment despite statistical indicators from the texts.

The categories 'important issues' and 'gender' predicted better than others. Moreover, the accuracy for 'important issues' is underestimated by the standard accuracy score since it is a multilabel classification (one argument may correspond to more than one important issue) with each class having a binary label as its target. Separate accuracy computations for each class revealed better results, as shown in Table 3 in the Appendix.

Our approach intends to use our predictions to filter corpus arguments and then select semantically relevant arguments from the filtered corpus. For query requirements where demographic features other than 'important issues' and 'gender', we do not apply filtering and focus only on semantic matching. Results on the development set revealed that filtering based on 'important issues' was more effective than gender filtering. Indeed, gender filtering performed worse than no filtering at all when tested with the development set. Consequently, we decided to only apply demographic filtering for 'important issues' for the final submission.

## 5 Results and Discussion

Figure 1 presents the comparison between our system and Sentence-BERT. The first row shows the NDCG scores of three test sets in the baseline scenario. Our system performs similarly well in Test 1 but shows a significant advantage over Sentence-BERT in Tests 2 and 3. In test 3, it also reveals that the high accuracy can largely be attributed to the filtering procedure and the narrowing down of the corpus size. For explicit scenarios (the second row in Figure 1), our system significantly outperforms Sentence-BERT by using exact demographic label matches from queries to arguments in the corpus.

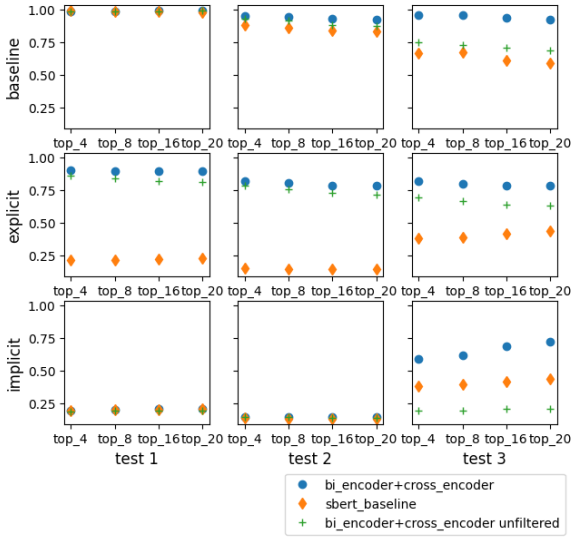


Figure 1: Relevance results (NDCG scores) for retrieved argument candidates from the top 4 to top 20 across three scenarios and three test sets.

For implicit scenarios (the third row in Figure 1), our system shows no difference from Sentence-BERT, except in Test 3, which is significantly better with the filtering procedure but much worse without filtering. The implicit scenario shows that the benefits of utilizing the 'important issues' classifier to filter first are not evident. The predicted important issues' might also be decoded by Sentence-BERT, and overall low precision may result from the mismatch of other demographic factors. Despite performing less well in the implicit scenario than in the other two, our results still achieved first place among all participated teams in Tests 2 and 3, and second in Test 1.

## 6 Conclusion

This study demonstrates the advantages of combining bi-encoder and cross-encoder models over solely using the bi-encoder (Sentence-BERT). We also found that perspective argument retrieval or inferring socio-demographic features from short arguments remains challenging, accompanied by the disadvantage of the number and quality of labels.

Moreover, longer argument texts may be necessary to decode the socio-demographic features of argument providers in the future.

## References


Ross Clement and David Sharp. 2003. Ngram and bayesian classification of documents for topic and authorship. *Literary and linguistic computing*, 18(4):423–447.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Neele Falk, Andreas Waldis, and Iryna Gurevych. 2024. Overview of perpectivearg2024: The first shared task on perspective argument retrieval. In *Proceedings of the 11th Workshop on Argument Mining*, Bangkok. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Jae-Seok Jang and Hyuk-Yoon Kwon. 2023. Question-answering pair matching based on question classification and ensemble sentence embedding. *Comput. Syst. Sci. Eng.*, 46(3):3471–3489.

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

Niklas Muennighoff. 2022. Sgpt: Gpt sentence embeddings for semantic search. *arXiv preprint arXiv:2202.08904*.

Claudia Peersman, Walter Daelemans, and Leona Van Vaerenbergh. 2011. Predicting age and gender in online social networks. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, pages 37–44.

Juan Ramos et al. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 29–48. Citeseer.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2019. Classification and clustering of arguments with contextualized word embeddings. *arXiv preprint arXiv:1906.09821*.

Christos Sardianos, Ioannis Manousos Katakis, Georgios Petasis, and Vangelis Karkaletsis. 2015. Argument extraction from news. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 56–66.

Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2020. Augmented sbert: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. *arXiv preprint arXiv:2010.08240*.

Jannis Vamvas and Rico Sennrich. 2020. X-stance: A multilingual multi-target dataset for stance detection. In *5th SwissText & 16th KONVENS Joint Conference 2020*, page 9. CEUR-WS. org.

Cathy Zhang and Pengyu Zhang. 2010. Predicting gender from blog posts.

Leixin Zhang and Çağrı Çöltekin. 2024. Tübingen-CL at SemEval-2024 task 1: Ensemble learning for semantic relatedness estimation. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1019–1025, Mexico City, Mexico. Association for Computational Linguistics.


## A Appendix

| Issue | Accuracy |
|---|---|
| Restriktive Finanzpolitik | 0.86 |
| Liberale Wirtschaftspolitik | 0.79 |
| Ausgebauter Sozialstaat | 0.76 |
| Law & Order | 0.75 |
| Restriktive Migrationspolitik | 0.73 |
| Liberale Gesellschaft | 0.73 |
| Ausgebauter Umweltschutz | 0.72 |
| Offene Aussenpolitik | 0.66 |

Table 3: Individual accuracy per class in 'important issues' (a multi-label classification problem).