GESIS-DSM at PerpectiveArg2024: A Matter of Style? Socio-Cultural Differences in Argumentation

Maximilian Martin Maurer[1], **Julia Romberg**[1], **Myrthe Reuver**[3♥], **Negash Desalegn Weldekiros**[1,4], and **Gabriella Lapesa**[1,2]
[1]GESIS - Leibniz Institute for the Social Sciences, [2]Heinrich-Heine University Düsseldorf
[3]Vrije Universiteit Amsterdam, [4]Technical University of Munich
[1]first.last@gesis.org, [3]myrthe.reuver@vu.nl, [4]negash.weldekiros@tum.de

Abstract

This paper describes the contribution of team GESIS-DSM to the Perspective Argument Retrieval Task, a task on retrieving socio-culturally relevant and diverse arguments for different user queries. Our experiments and analyses aim to explore the nature of the socio-cultural specialization in argument retrieval: (how) do the arguments written by different socio-cultural groups differ? We investigate the impact of content and style for the task of identifying arguments relevant to a query and a certain demographic attribute. In its different configurations, our system employs sentence embedding representations, arguments generated with Large Language Model, as well as stylistic features.

Our final method places third overall in the shared task, and, in comparison, does particularly well in the most difficult evaluation scenario, where the socio-cultural background of the argument author is implicit (i.e. has to be inferred from the text). This result indicates that socio-cultural differences in argument production may indeed be a matter of style.

1 Introduction

Argument retrieval is a well-established task in computational argumentation (Wachsmuth et al., 2017; Stab et al., 2018): given a query or question, e.g. on "should we have free trade agreements?", the task aims to retrieve **topically relevant** arguments. Topically relevant arguments can be heterogeneous, even for arguments with the same stance (for vs. against) - i.e. someone can focus on quality of the products as essential to free trade, while another may focus on international collaboration. These different perspectives in arguments can be relevant for the personalization and diversification of argument retrieval, and in online debate portals.

The *Perspective Argument Retrieval Task* (Falk et al., 2024) argues that the socio-cultural background (e.g. gender, religion) can be taken into account when selecting relevant arguments. Socio-cultural information can determine the dynamics of argumentation at different levels. For example, socio-demographics can be used to approximate the stance of an argument about a specific topic (e.g., women, pro abortion); additionally, the specific arguments used to support a stance can be a correlate of a specific socio-cultural attribute (religious being against abortion because *it is a sin*).

Differences between arguments of different socio-cultural groups have mostly been researched as differences in **argument content**, meaning the semantic differences in the arguments: which phrases, aspects, and points are mentioned. For instance, Spliethöver and Wachsmuth (2020) analyze how social groups differ in the social bias of their arguments, i.e., male users using terms that indicate a gender bias. However, previous research indicated that **argument style** may be different between socio-cultural groups as well. With argument style, we mean *how* something is said, e.g. features of form and not content like complexity of words, length of words and sentences, grammatical tenses, or pronoun usage. Such stylistic features have mostly been studied from the perspective of argument persuasiveness (El Baff et al., 2020) but also have been used to analyze socio-demographic differences in deliberation processes, i.e., women referring more to others than men in their arguments (Klinger and Russmann, 2015).

While the shared task does not address the issue of persuasiveness of the arguments directly, the intuition that "an argument I would write is one that is likely to resonate stronger with me", builds a potential bridge between socio-demographic retrieval and persuasion research (e.g., consider El Baff et al.

<♥> Contributions mostly completed during a visit funded by a GESIS Visiting Junior Researcher grant at GESIS – Leibniz Institute for the Social Sciences.

Proceedings of the 11th Workshop on Argument Mining, pages 169–181

(2018) on empowering vs. challenging arguments). Additionally, diversity (providing arguments with different perspectives on the same query) can be employed to maximize the reception of a certain argument. Also, such a diversity of perspectives in recommendations can be beneficial, for instance for citizens in democracies (Reuver et al., 2021).

In approaching the task of argument retrieval for specific socio-cultural profiles, we can also resort to findings from other tasks, such as author profiling (Koppel et al., 2002). This task aims to predict author characteristics from user-generated text, with these characteristics often having a socio-demographic nature, (i.e., gender or age). Successful approaches use semantic content as well as style to profile authors of user-generated texts (Rangel et al., 2021; Bevendorff et al., 2023).

Our approach The perspectivist argument retrieval task raises the question of how socio-cultural groups differ in their arguments for a given query. These differences can be semantic, i.e., groups may differ in *what* they say in their arguments, or stylistic, i.e., groups may differ in *how* they formulate them. We first explore this distinction in the shared task data in Section 3. We then describe our approaches to ranking arguments according to socio-cultural specific relevance in Section 4: One is based on semantic content similarity using 'prototypical' arguments generated with a Large Language Model (LLM). The other uses stylistic features. Our final method places third overall in the shared task, and, in comparison, does particularly well in the most difficult scenario, the one in which the socio-cultural background of the argument author is implicit (i.e. has to be inferred from the text). It implements a three-step pipeline, using semantic information in a ranking step and stylistic information to classify whether arguments are relevant for a given socio-cultural group. Our results indicate that the stylistic differences in the arguments of different socio-cultural groups are more relevant to the task of retrieving relevant arguments than semantic differences in our setup. We publicly release our code for the experiments and analyses.[1]

2 Task: Data and Evaluation Scenarios

The question at the core of the task is: Can we find the arguments that members of a given socio-cultural target group would write for this query?

[1] github.com/mmmaurer/perspective_argument_retrieval

Data The task data is a multilingual dataset in three different cycles of each +/- 30,000 arguments and +/- 300 related queries (in German, Italian, and French). The provided socio-cultural information covers gender, age, place of residence, civil status, denomination, education, political spectrum, and political issues that are of importance to the authors of queries. Additionally, the stance and political topic of the argument are provided.

Details on the size and train/dev/test splits of the three cycles is provided in Table 2 in Appendix B. In the first two cycles, politicians express their stances regarding different political issues in the context of the 2019 and 2023 Swiss elections. In contrast, the third cycle consists of voters' perspectives. For this, samples of the arguments given by politicians for the 2023 election were annotated by amateur annotators. The resulting third cycle data consists of the arguments that intuitively match their perspectives. Socio-cultural profiles were collected for both politicians and voters.

Evaluation Scenarios The systems are evaluated on three scenarios: 1. Argument retrieval without consideration of socio-cultural differences (baseline). 2. Argument retrieval for a specified socio-cultural attribute. Information about any other attribute could be used to diversify the set of retrieved arguments (explicit scenario). 3. Argument retrieval for a specified socio-cultural attribute. In contrast to explicit, information about other attributes is hidden (implicit scenario). System evaluation is based on both relevance and diversity of the selected arguments (in terms of socio-cultural attributes) to promote the diversity of opinions.

3 Data Analysis: Content Or Style?

In the development of our pipeline, data analysis played a crucial role. In the following, we summarize core findings for *content* and *style*. These analyses were conducted on the cycle 1 corpus.

Semantic content differences Firstly, we assess whether there are arguments that multiple socio-cultural groups share and find that ~ 11% of the arguments appear with the same argument text for at least two different socio-cultural profiles. The same argument may, for instance, be produced by a non-religious man and a roman-catholic woman. While this is expected in a natural setting, as groups may share views and thus arguments, this raises the question of differentiation between socio-cultural

groups, in particular for the implicit scenario.

Secondly, we cluster Sentence-BERT representations of arguments using the k-means algorithm and evaluate the resulting clusters against the ground-truth socio-cultural groupings, as well as stance and topic. Per socio-cultural attribute, we run one clustering with k equaling the number of groups in the attribute (e.g. for residence $k = 2$, as there are the two groups *city* and *countryside*). We find that all of the socio-cultural groups have an adjusted Rand score (Hubert and Arabie, 1985) of ≤ 0.1 with the respective clustering, indicating virtually no overlap of the clusters with the groups. Only the topic shows a relatively higher adjusted Rand score (for an overview of the results of our clustering experiments, see Appendix E). These analyses indicate that there is little semantic distinction between arguments of different socio-cultural groups, at least in the present semantic representation space.

Stylistic differences To examine stylistic differences in the arguments, we carried out exploratory linear regression analysis. We exclusively focused on the German-language share[2], which comprises about 22k of the in total 32k arguments. The socio-cultural attributes served as independent variables and stylistic features of the arguments as dependent variables. We tested a number of stylistic features, which can be divided into surface and syntactic features and are explained in detail in Table 5 in Appendix F. The seven surface features include measures of word and sentence length, long and complex words, the variety of vocabulary used, and two readability indices. The syntactic features cover part-of-speech (POS) tags, named entities, present tense tokens, imperative tense tokens, and first person writing. Each POS tag forms an extra feature, giving us a total of 21 syntactic features.

We ran one linear regression per stylistic feature to estimate the relationship between socio-cultural information and the particular stylistic feature. In particular, we looked into interactions between variables to take a step into the direction of socio-cultural profiles rather than single attributes. Due to the space limitations, we cannot discuss all features in the paper. However, the full set of regression outputs can be found in the project's GitHub repository. All details are outlined in Appendix D.

Our assumption is that if socio-cultural groups differ in style with respect to specific stylistic features, our regression models will be able to account for a significant amount of that variance (higher adjusted R^2). Models that explain the surface features show notably higher adjusted R^2 values than those for the syntactic ones (cf. Table 3 in Appendix D). We find the highest explanatory power for long words, words per sentence, and the Gunning Fog Index with adjusted R^2 of 0.1557, 0.1422, and 0.1256, respectively. This suggests that socio-cultural characteristics can explain the writing style of the arguments to at least some extent. A closer look into the individual effects and interactions within the socio-cultural attributes reveals several significant effects. For instance, liberals use significantly fewer long words than conservatives, which is even more pronounced in connection with civil status divorced (compared to conservative and single). Liberals in the center or right of the political spectrum exhibit a higher Gunning Fog Index than left-wing conservatives, hinting at the number of years of formal education a person needs to understand a text on the first reading.

While some of these effects certainly also depend on further factors like the context of an argument (i.e., stance and the topic of discussion), the findings add to our underlying hypothesis of different stylistic fingerprints.

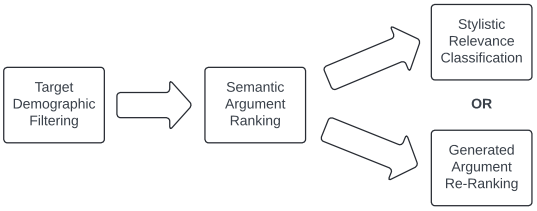

Figure 1: Illustration of our three-step pipeline.

4 System Description

As discussed in the introduction, we want to assess the impact of *content* and *style* on the perspective argument retrieval performance. We thus divide the problem into two steps covering these aspects, with an additional filtering step.[3]

Our resulting pipeline, depicted in Figure 1, consists of three steps: 1) **Target Demographic Filtering**: If arguments of a specific socio-cultural group are queried and socio-cultural information for the arguments is available, only consider the arguments from the respective socio-cultural tar-

[2]Details on the language detection are in Appendix A.

[3]All hyperparameters and implementation details of our analyses and models are given in Appendix A.

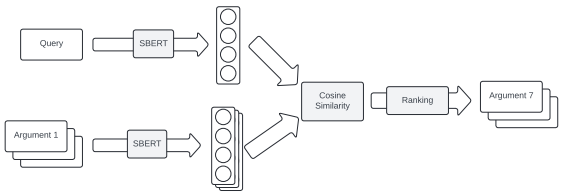

Figure 2: Semantic argument ranking step.

get group for the subsequent steps. 2) **Semantic Argument Ranking**. Find the top k arguments that are semantically closest to the query. 3) **Final Candidates Retrieval**. Option (a) **Stylistic Relevance Classification**: Select only those from the pre-selected k arguments with a stylistic *fingerprint* indicating a relevant argument given the target socio-cultural group. Option (b) **Generated Argument Re-ranking**: Per query and socio-cultural attribute, generate an argument and find the top k arguments that are semantically closest to the generated argument.

For the non-perspectivist **baseline** scenario, only step 2), semantic argument ranking, is executed. For the **explicit perspectivism** scenario, all three steps are executed. Finally, for the **implicit perspectivism** scenario, we execute solely steps 2) and 3) as no socio-cultural information is accessible.

In what follows, we detail the operationalization of these three steps.

4.1 Target Demographic Filtering

To reduce the search space in cases where only the arguments of a specific socio-cultural group are queried and the socio-cultural profiles of the authors of the arguments are known, i.e. in the *explicit perspectivist scenario*, we filter out the arguments that do not match the queried attribute.

4.2 Semantic Argument Ranking

A necessary condition for a given argument to be *relevant* for a query, both in the perspectivist and the baseline cases, is that they are semantically related, or, in other words, that the argument supports a stance towards the question stated in the query. The relevant candidate arguments for a given query should thus be selected such that their semantic similarity is as high as possible.

To operationalize this, as illustrated in Figure 2, we rely on retrieving sentence embeddings using Sentence-BERT (Reimers and Gurevych, 2019) both for the query and the arguments in a corpus and calculate the cosine similarity between the query's representation and each of the arguments' representations. Finally, the arguments are sorted according to the cosine similarity and only the top k arguments are considered.

As a backbone Sentence-BERT model, we use `paraphrase-multilingual-mpnet-base-v2`[4], a multilingual model trained on paraphrases in 50+ languages, among them the three Swiss official languages present in the dataset.

4.3 Final Candidates Retrieval

4.3.1 Stylistic Relevance Classification

A classification step to differentiate between semantically generally relevant arguments (i.e. arguments relevant to a query, regardless of socio-cultural information) and relevant arguments for a specific socio-cultural group is implemented next.

We implement a semantic selection step before the classification step and assume that the set of relevant arguments of a specific socio-cultural group given a query is a real subset of the set of relevant arguments given a query. Based on this insight, we construct positive and negative examples from the training subsets of the provided datasets: For each unique query text q, we collect the set of the overall relevant candidates $C_{q,\text{all}}$. A candidate is considered a positive example and assigned the label *relevant* for the respective socio-cultural group t if it is in the set of the relevant arguments given t and q, $C_{q,t}$. A candidate is considered a negative example and assigned the label *not relevant* if it is in $C_{q,\text{all}} \setminus C_{q,t}$. To end up with a more balanced training set, we only collect $|C_{q,t}|$ negative examples if $|C_{q,\text{all}}| \geq |C_{q,t}|$. Per example (i.e. per argument), we encode a one-hot representation of the queried socio-cultural attribute and concatenate it with surface-level stylistic features of the respective argument as input features. A full overview of the stylistic features can be found in Table 5 in Appendix F. As the majority of arguments are in German, the feature extraction assumes the language to be German. While this is sub-optimal (style may differ across languages), this serves as a first assessment of whether stylistic differences can help in this task.

For the resulting classification step of our pipeline, as visualized in Figure 3, we train a random forest classifier on the training set portion (80%) of our dataset constructed from a union of

[4] https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2

	Model	Election 2019			Election 2023			2023, User Study			Avg.		
		Baseline	Explicit	Implicit	Baseline	Explicit	Implicit	Baseline	Explicit	Implicit	Baseline	Explicit	Implicit
Relevance	SBERT	**0.986**	0.222	**0.202**	**0.855**	0.148	0.136	**0.637**	0.406	0.409	**0.826**	0.252	0.249
	STY	**0.986**	**0.835**	**0.202**	**0.855**	**0.722**	**0.139**	**0.637**	**0.616**	**0.471**	**0.826**	**0.724**	**0.271**
	GEN	**0.986**	0.645	0.185	**0.855**	0.597	0.127	**0.637**	0.493	0.348	**0.826**	0.578	0.220
Diversity	SBERT	**0.916**	0.208	**0.189**	**0.793**	0.142	0.131	**0.593**	0.400	0.397	**0.767**	0.250	0.239
	STY	**0.916**	**0.807**	**0.189**	**0.793**	**0.701**	**0.132**	**0.593**	**0.629**	**0.454**	**0.767**	**0.654**	**0.258**
	GEN	**0.916**	0.618	0.173	**0.793**	0.579	0.121	**0.593**	0.493	0.331	**0.767**	0.563	0.208

Table 1: Results for the Sentence-BERT baseline (SBERT), and our pipeline with the final step being stylistic relevance classification (STY) and a re-ranking step using generated arguments (GEN). We present mean results across k per test set (election 2019 and 2023, and the 2023 user study), scenario (Baseline, Explicit and Implicit perspectivism), and evaluation (relevance, measured by $nDCG$, and diversity, measured by αDCG). The best result per test set, scenario, and evaluation track is printed in bold.

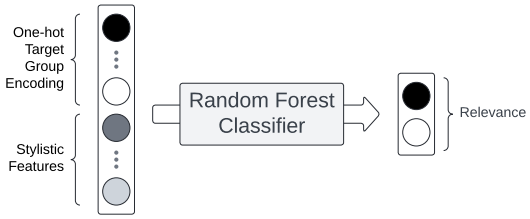

Figure 3: Stylistic relevance classification step.

the training subsets of the three datasets. The classifier achieves an $F_1 = 0.60$ for both the positive and negative classes on the held-out test set portion (20%) of our dataset. We provide a feature importance overview in Figure 6 in Appendix G.

4.3.2 Generated Argument Re-Ranking

Following the hypothesis[5] that arguments of different socio-cultural groups are differentiable by their semantic content, we implement a re-ranking step using LLM-generated arguments. We generate arguments for specific groups and queries using `occiglot-7b-eu5-instruct`[6], a `Mistral-7B-v0.1` model with continued pre-training on the five biggest languages in Europe: English, Spanish, French, German, and Italian. We generate one argument per query, which is then used to re-rank the candidates from the second step in our pipeline. Specifically, per query, we gather the Sentence-BERT representation of the generated argument and re-rank the candidates according to their cosine similarity with the generated argument. Appendix C provides more details about the generated arguments (prompts, statistics, examples, and qualitative analysis).

[5] Our initial hypothesis was that arguments differ in their semantic content across socio-cultural groups. Our downstream analysis of the semantic representation space in combination with the results did however proved our hypothesis wrong.

[6] https://huggingface.co/occiglot/occiglot-7b-eu5-instruct

5 Discussion of Results

The results of our different systems over the three test sets are displayed in Table 1. It can be seen that using stylistic relevance classification as the final step in our pipeline yields results well over the Sentence-BERT baseline across explicit scenario test sets, and for the final implicit scenario test set (Table 1: 2023, User Study).

Moreover, this approach outperforms the use of generated argument re-ranking across all perspectivist test sets. Overall, our findings show that content plays a role in pre-selecting arguments to fit the respective query, as evidenced by the comparably high baseline scenario results. For the same queries, however, socio-cultural groups appear to be less different in the content of their arguments than in their style.

6 Conclusion and Future Work

We present our approach to the Perspective Argument Retrieval Shared Task 2024. Our proposed method implements a three-step pipeline, leveraging semantic information in a ranking step and stylistic information to classify whether arguments are relevant for a given socio-cultural group. The performance of this approach, in particular for the implicit scenario, showcases the potential of including stylistic information for the task of perspectivist argument retrieval. This raises several questions for future research.

Especially with regard to the third test set, in which the perspectives of politicians and voters were flipped, we argue that investigating the reasons for differences in production and perception of arguments of different socio-cultural groups, e.g. through semantic or stylistic differences, is of interest. Consequently, how to combine this information in retrieval scenarios should be investigated.

7 Ethical Considerations and Limitations

Shared tasks have previously focused specifically on author profiling, e.g. profiling spreaders of hate speech in English and Spanish (Rangel et al., 2021) or profiling crypto influencers (Bevendorff et al., 2023), where one system contribution used LLMs and bi-encoding (Giglou et al., 2023). We acknowledge the task of authorship profiling and our approach, using stylistic features, has some established ethical harms to individuals and society at large. These harms are mostly relating to privacy and giving agency to powerful actors to track or harm individuals. However, we also found work that is meant to reduce these specific harms.

7.1 Established harms and limitations

Author profiling and related tasks on detecting user characteristics based on written content have some long-established ethical issues. Among these are concerns about privacy and revealing user identity when users write about sensitive topics (Brennan et al., 2012), and also identifying characteristics that authors may want to keep private, such as their age, gender, or religion. The perspectivist argument retrieval task is a use-case which we consider to benefit users and society: providing diverse perspectives on issues and relevant arguments, which is useful for instance for online deliberation platforms where a diversity of perspectives and interactions between different groups are important. However, this task can also be used to censor, track, or harm specific groups and individual users who write the arguments.

It is also important to be aware of legal frameworks, such as the European Unions general data protection regulation (GDPR), on datasets aimed at detecting author profiles. Rangel and Rosso (2019) have described how, for 2019 PAN shared task dataset on author profiling, all legal limitations have been followed. They also state that their interpretation of GDPR Article 22 means profiling is illegal, though with an exception for non-commercial purposes and scientific research.

7.2 Approaches to protect users from harm

The ethical and legal issues with author profiling have triggered several approaches aimed at **preventing** authorship profiling for harmful contexts. One such set of tasks is known as *adversarial stylometry* (Brennan et al., 2012) (not to be confused with adversarial learning). This set of tasks is specifically aimed at preventing user profiling based on style. For instance, in the subtask of **authorship obfuscation** the idea is to re-write the texts to such an extent that stylometric features cannot distinguish different authors or author groups anymore while leaving semantic coherence of the text intact. Successful and robust approaches across multiple models and datasets, such as by Emmery et al. (2021), use an approach of lexical substitution: changing content words strongly related to certain labels.

Other works have also looked into ethical versions of profiling tasks. For instance, Allein et al. (2023) have looked into fake news detection without author profiling: with the assumption that similar users may share similar fake news articles, they use a latent representation of a group of authors and a fake news article, without ever providing the model with direct user profile information.

Acknowledgements

Myrthe Reuver's contributions were funded by a Visiting Junior Researcher grant from GESIS – Leibniz Institute for the Social Sciences, and also the *Rethinking News Algorithms* project (grant nr 406.D1.19.073) by the Netherlands Organization of Scientific Research (NWO).

Thanks to Neele Falk, whose scripts for feature extraction[7] provided a starting point for the stylistic features.

References

Hirotogu Akaike. 1998. *Information Theory and an Extension of the Maximum Likelihood Principle*, pages 199–213. Springer New York, New York, NY.

Liesbeth Allein, Marie-Francine Moens, and Domenico Perrotta. 2023. Preventing profiling for ethical fake news detection. *Information Processing & Management*, 60(2):103206.

Ryan Bakker and Sara Hobolt. 2013. Measuring Party Positions. In *Political Choice Matters: Explaining the Strength of Class and Religious Cleavages in Cross-National Perspective*. Oxford University Press.

Janek Bevendorff, Ian Borrego-Obrador, Mara Chinea-Ríos, Marc Franco-Salvador, Maik Fröbe, Annina Heini, Krzysztof Kredens, Maximilian Mayerl, Piotr Pęzik, Martin Potthast, et al. 2023. Overview of pan 2023: Authorship verification, multi-author writing style analysis, profiling cryptocurrency influencers,

[7] https://github.com/Blubberli/featureExtraction

and trigger detection: Condensed lab overview. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 459–481. Springer.

Michael Brennan, Sadia Afroz, and Rachel Greenstadt. 2012. Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity. *ACM Transactions on Information and System Security (TISSEC)*, 15(3):1–22.

Roxanne El Baff, Henning Wachsmuth, Khalid Al-Khatib, and Benno Stein. 2018. Challenge or empower: Revisiting argumentation quality in a news editorial corpus. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 454–464, Brussels, Belgium. Association for Computational Linguistics.

Roxanne El Baff, Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2020. Analyzing the Persuasive Effect of Style in News Editorial Argumentation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3154–3160, Online. Association for Computational Linguistics.

Chris Emmery, Ákos Kádár, and Grzegorz Chrupała. 2021. Adversarial stylometry in the wild: Transferable lexical substitution attacks on author profiling. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2388–2402.

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.

Hamed Babaei Giglou, Mostafa Rahgouy, Jennifer D'Souza, Milad Molazadeh, Hadi Bayrami Asl Tekanlou Oskuee, and Cheryl D Seals. 2023. Leveraging large language models with multiple loss learners for few-shot author profiling. *Working Notes of CLEF*.

L. Hubert and P. Arabie. 1985. Comparing partitions. *Journal of Classification*, 2:193–218.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Ulrike Klinger and Uta Russmann. 2015. The sociodemographics of political public deliberation: Measuring deliberative quality in different user groups. *Communications*, 40(4):471–484.

Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. 2002. Automatically categorizing written texts by author gender. *Literary and linguistic computing*, 17(4):401–412.

M. J. Laver and Ian Budge. 1992. *Measuring Policy Distances and Modelling Coalition Formation*, pages 15–40. Palgrave Macmillan UK, London.

Stuart Lloyd. 1982. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137.

Ines Montani, Matthew Honnibal, Adriane Boyd, Sofie Van Landeghem, and Henning Peters. 2023. explosion/spaCy: v3.7.2: Fixes for APIs and requirements.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Francisco Rangel, Gretel Liz de la Peña-Sarracén, María Alberta Chulvi-Ferriols, Elisabetta Fersini, and Paolo Rosso. 2021. Profiling hate speech spreaders on twitter task at pan 2021. In *Proceedings of the Working Notes of CLEF 2021, Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st to 24th, 2021*, pages 1772–1789. CEUR.

Francisco Rangel and Paolo Rosso. 2019. On the implications of the general data protection regulation on the organisation of evaluation tasks. *Language and Law/Linguagem e Direito*, 5(2):95–117.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Myrthe Reuver, Nicolas Mattis, Marijn Sax, Suzan Verberne, Nava Tintarev, Natali Helberger, Judith Moeller, Sanne Vrijenhoek, Antske Fokkens, and Wouter van Atteveldt. 2021. Are we human, or are we users? the role of natural language processing in human-centric news recommenders that nudge users to diverse content. In *1st workshop on NLP for positive impact*, pages 47–59. Association for Computational Linguistics.

Maximilian Spliethöver and Henning Wachsmuth. 2020. Argument from old man's view: Assessing social bias in argumentation. In *Proceedings of the 7th Workshop on Argument Mining*, pages 76–87, Online. Association for Computational Linguistics.

Christian Stab, Johannes Daxenberger, Chris Stahlhut, Tristan Miller, Benjamin Schiller, Christopher Tauchmann, Steffen Eger, and Iryna Gurevych. 2018. ArgumenText: Searching for arguments in heterogeneous sources. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*,

pages 21–25, New Orleans, Louisiana. Association for Computational Linguistics.

Henning Wachsmuth, Martin Potthast, Khalid Al-Khatib, Yamen Ajjour, Jana Puschmann, Jiani Qu, Jonas Dorsch, Viorel Morari, Janek Bevendorff, and Benno Stein. 2017. Building an argument search engine for the web. In *Proceedings of the 4th Workshop on Argument Mining*, pages 49–59, Copenhagen, Denmark. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Appendix

A Hyperparameters and Implementation Details

Our models were implemented and experiments conducted using Python 3.11 unless stated otherwise.

A.1 Data Analysis

Stylistic features Stylistic features were obtained using the *readability* (https://pypi.org/project/readability/) and *SpaCy* python packages (https://pypi.org/project/spacy/; Montani et al. 2023) .

Linear Regression To select German-language arguments, we used the *langdetect* Python package (https://pypi.org/project/langdetect/). The linear regression was implemented in R 4.4.1. We used the packages *stats*, *MASS*, and *car* for the step-wise building of the regression models and inspection of variance inflation factors.

Content clustering Clustering of our Sentence-BERT embeddings was done with the *scikit-learn* (Pedregosa et al., 2011) Python package implementation of the K-means clustering algorithm (Lloyd, 1982). We clustered for each socio-cultural variable (e.g. gender, denomination) and then also combined attributes in pairs of two to test for cluster coherence. For calculating cluster coherence, we use the Adjusted Rand score as also implemented in the *scikit-learn* Python package. Detailed clustering results can be found in Appendix E.

A.2 Base Model

Our sentence embeddings model was *paraphrase-multilingual-mpnet-base-v2* (Reimers and Gurevych, 2019), implemented through the huggingface *transformers* (Wolf et al., 2020) and *sentence_transformers* (Reimers and Gurevych, 2019) Python packages. For ranking the top documents to a query, we selected k=200 using cosine similarity.

A.3 Argument Generation Model

Our generated arguments were obtained with the *occiglot-7b-eu5-instruct* model (https://huggingface.co/occiglot), a fine-tuned Mistral-7B (Jiang et al., 2023) model, called with Huggingface's *transformers* (Wolf et al., 2020) package. The model was run on an NVIDIA A100 GPU. Prompt template details, and a short analysis of these generated arguments, are provided in Appendix C.

We memory-optimize our prompting by loading an int8-quantized version of the model. For quantization, we use the *quanto* library (https://github.com/huggingface/optimum-quanto).

A.4 Stylistic Relevance Classification

Random Forest classifier The random forest classifier for detecting socio-cultural background based on the stylistic features was implemented with the scikit-learn (Pedregosa et al., 2011) package, using its default implementation and parameters: 100 trees, splitting on gini impurity, and no max depth.

B Details of the Dataset

Cycle	Corpus	Queries					
		Baseline			Perspective		
		train	dev	test	train	dev	test
Election 2019	32,387	104	134	44	5,577	1,611	2,358
Election 2023	39,093	104	134	39	5,577	1,611	1,782
2023, User Study	28,684	104	134	26	4,737	1,371	729

Table 2: Dataset sizes for the retrieval argument corpus and for the queries, divided into train/dev/test set for the baseline and the perspective scenario. The task ran for three evaluation cycles with different evaluation data.

C Prompt Formulation & Generated Arguments

We generated arguments based on queries in the corpus with the prompt in Figure 5 and the *occiglot-*

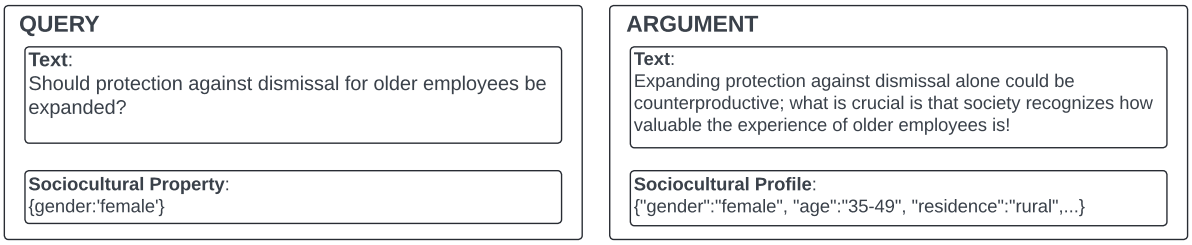

Figure 4: Example for a pair of a query and a relevant argument. Original in German, automatically translated using Google translate.

Given the question {query}, use your knowledge of the Swiss political landscape to provide a pro argument a person whose {attribute} is {group} would produce.

Figure 5: Prompt formulation.

7b-eu5-instruct model (implementation details in Appendix A). Below we provide a short description and analysis of the arguments generated with this method.

C.1 Statistics on Generated Arguments

Across the three test sets, the model produced an argument over 90% of the time. No argument was generated for 6.2%/1.2%/1.6% of prompts, respectively). While across the three test sets, 95% of the generated arguments are truncated, i.e. they end mid-sentence, the generated arguments are on average longer than the arguments in the corpus (mean raw text length of 190 characters for the corpus vs. 267/429/297 characters generated per test set, respectively).

C.2 Qualitative Analysis of Generated Arguments

Based on a qualitative inspection, we have gathered the following observations regarding the generated arguments. In Section C.3, we provide some examples of the generated arguments (German, translated into English with DeepL) to illustrate our analysis.

First of all, we notice an (unsurprising) tendency to repeat the demographics from the prompts, and additionally to generate intersectional types. Consider for example query 20191712, where the author of the (generated) argument identifies himself as young, man, and a student. Query 20191712 illustrates also another property, shared with query 201911545: the tendency of the generated arguments to be accompanied by fully fledged, unprompted, personas, e.g. descriptions of argument 'authors' (e.g. professions and other sociodemographics).

The arguments also vary greatly in complexity and overall quality. Some arguments, i.e., 201911569 (test 2), exhibit a very low degree of justification and/or resort to very abstract principles ("The federal government should invest more in education and research because it would advance us socially"), or essentially just rephrase the stance (201911545). Others, instead, display a very articulated structure with multiple supporting premises (e.g., four, explicitly listed in 201911652 (test 2)).

In addition to the degree of justification, arguments also vary greatly in terms of the type of evidence they use to support their stance. The speaker in 201911630 (test 2), for example, resorts to reporting personal experiences. The speaker in 201911777 (test 2), instead, mentions his profession "As an entrepreneur for many years" to establish himself as an authoritative source on the topic at issue.

C.3 Selected generated arguments

Query: "Should the federal government spend more or less on development cooperation?"

201911777-test 2 (Prompt: attribute: political spectrum; value: center and conservative/liberal), "As an entrepreneur for many years, I have seen how important it is to promote international relations and engage in global cooperation. Especially in times when job creation is a major challenge, it is crucial to place a strong focus on development cooperation. This is because we can help to improve economic conditions in certain regions and thus give more people access to a livelihood."

201911781-test 2 (Prompt: attribute: political spectrum; value: right and liberal) "It is not the task of the federal government to invest financial

resources in the area of development cooperation. The responsibility for development cooperation lies with the countries and economies affected by poverty and underdevelopment. The federal government has no responsibility to become active in these areas, and by increasing funding it could interfere with the sovereignty of these countries. Therefore, the federal government should invest less and not more money in development cooperation."

20191712-test 3 (Prompt: attribute: civil status; value: single) "Hello, my name is John and I'm 22 years old. I'm from Switzerland, where I'm currently studying in Zurich, and as a young man and a student at that, I firmly believe and think it's important that the federal government spends more on development cooperation, especially as we can't deny that"

Query: Should the federal government spend more or less on public transport?

201911630-test 2 (Prompt: attribute: civil status; value: married), "As a Swiss citizen who has been working for many years, I believe that the federal government needs to spend more on public transport. I am married myself and I have seen how difficult it can be to get to a meeting within a few minutes by bus. I've also seen how people can't use the train because there are too many connections that don't go to where they live. So we should spend more money and improve public transport."

201911651-test 2 (Prompt: attribute: political spectrum; value: center and conservative/liberal) "The federal government should spend more on public transport. This would improve the population's mobility options and at the same time create a more environmentally friendly way of travelling."

201911652-test 2 (Prompt: attribute: political spectrum; value: center and liberal), "I am in favour of the federal government investing more in public transport because the advantages outweigh the disadvantages. Firstly, clean air and improved quality of life are essential for citizens, and secondly, investment in public transport improves the efficiency and profitability of the transport system. Thirdly, public transport can help to unite and motivate the population towards common goals. Fourthly, investment in public transport contributes to economic development"

Query: Should the federal government spend more or less on education and research?

201911569-test 2 (Prompt: attribute: political spectrum; value: left and liberal) "Better more than less on education and research. In today's globalised society, the expansion of knowledge will not only advance us technologically, but also socially. That is why the promotion of education and research is essential."

201911571-test 2 (Prompt: attribute: political spectrum; value: center and conservative/liberal), "We should invest more money in education and research. This is the only way we can ensure that our children and grandchildren have a brilliant future. They will be the ones who have to live and work in an increasingly complex world and they need to be ready to meet these challenges, and we should also think about how we can encourage people in our communities to invest in these areas. This can be done through incentives such as tax breaks and other incentives."

201911545-test2 (Prompt: attribute: civil status; value: single), "More education and research are crucial for the progress of our society. The federal government should therefore provide more financial support in this area. We need qualified specialists and innovations to secure our prosperity.Julia, 25, ZurichJulia (25) is a young woman from Zurich with an affinity for technology and an interest in science. As a high school graduate, she dreams of doing something "meaningful" one day."

D Regression analysis

In the linear regression analysis, we include the socio-cultural attributes gender, age, residence, civil status, denomination, and political spectrum. The additionally given information about important political issues is not taken into account for now, given the large number of different profiles in this attribute.

To counter sparsity of the results, in a preprocessing step all groups in the attributes that appeared rarely (< 50) were removed. This includes "Aufgelöste Partnerschaft" and "Getrennt" in civil status, as well as "Andere Kirchen/Religionsgemeinschaften", "Christlich-orthodox", "Jüdische Gemeinschaften", and "Islamische Gemeinschaften" in denomination. Furthermore, missing value groups ("Nicht bekannt") in civil status, denomination, and education were removed. For better interpretability of levels of education, we summarized "Höhere Fachschule", "Handelsschule", and "Höhere Berufsausbildung" as "Höhere Berufsausbildung", and "Berufs-

DV	Adjusted R^2	DV	Adjusted R^2
Characters per word	0.0613	NOUN	0.0285
Words per sentence	0.1422	NUM	0.0155
Type-token ratio	0.0596	PART	0.0077
Long words	0.1557	PRON	0.0355
Complex words	0.1068	PROPN	0.0150
Flesch Reading Ease	0.0880	PUNCT	0.0412
Gunning Fog Index	0.1256	SCONJ	0.0129
ADJ	0.0165	SYM	-
ADP	0.0041	VERB	0.0232
ADV	0.0187	X	0.0022
AUX	0.0157	Named entities	0.0110
CCONJ	0.0104	Present tense	0.0123
DET	0.0121	Imperative	0.0022
INTJ	-	First person	0.0272

Table 3: Adjusted R^2 scores of the linear regression models. INTJ and SYM did not occur in the arguments and because of this, no model was built in these cases.

matura" and "Diplommittelschule" as "Berufsmatura/Diplommittelschule". Likewise, for the sake of interpretability, the given socio-cultural attribute political spectrum was divided into quasi-RILE (Laver and Budge, 1992) (an ideological scale measuring general left-to-right position; left, center, and right) and quasi-GALTAN (Bakker and Hobolt, 2013) scores (an ideological scale measuring the policy position on social issues; conservative, conservative-liberal, and liberal). The resulting dataset contains 11289 arguments.

We ran one linear regression per stylistic feature (see Table 5 for more information on the stylistic features) to estimate the relationship between socio-cultural information and the particular stylistic feature. The models were built step-wise ("forward") using the Akaike information criterion (Akaike, 1998). In particular, we looked into interactions between variables to take a step into the direction of socio-cultural profiles rather than single attributes. The formula used was *DV ~ (gender + age + residence + civil_status + denomination + education + rile + galtan)^2* where *DV* is a placeholder for any one of the dependent variables (implementation details can be found in Appendix A). Due to the large number of variables in the resulting models, it is not possible to present them in their entirety in the paper. We selected highlighted results in the paper and make the R code available in the corresponding GitHub repository.

Table 3 illustrates the explained variance of the best model selected by StepAIC for each of the 21 stylistic features. Overall, we find small adjusted R^2 scores, signaling that the socio-cultural variables we selected as predictors can explain a limited amount of the variance in our stylistic features. The fit of the models is however still highly significant and in our discussion we focus on significant effects. Looking more into detail into the models (the full set of regression outputs can be found on GitHub), we see significant effects across the different socio-cultural attributes and groups. While such effects may also be triggered by the large scale of the dataset, our findings inspire us to incorporate stylistic features into the retrieval models discussed in Section 4.

E Clustering Results

Attribute	Adjusted Rand Score
Gender	0.0009
Age	-0.0011
Denomination	0.0004
Residence	0.0007
Political Spectrum	-0.0011
Stance	0.0005
Topic	0.1947

Table 4: K-means clustering result per socio-cultural attribute, plus stance and topic.

F Stylistic Features

Table 5 describes the different stylistic features used in the linear regression analysis and the random forest classifier.

	Feature	Description
Surface Features	Characters per word	Average number of characters per word, calculated by dividing the total number of characters in a text by the total number of words. Functions as a measure of text complexity as longer words can be harder to process.
	Words per sentence	Average number of words per sentence, calculated by dividing the total number of words by the total number of sentences in a text. Functions as a measure of text complexity as longer sentences can be harder to process.
	Type-token ratio (TTR)	Indication of the diversity of vocabulary usage in a text, calculated by dividing the total number of unique words by the total number of words.
	Long words	Number of words that consist of 7 or more characters. Functions as a measure of text complexity as longer words can be harder to process.
	Complex words	Number of words that consist of 3 or more syllables. Functions as a measure of text complexity as longer words can be harder to process.
	Flesch Reading Ease	Assesses the approximate reading grade level of a text, based on average sentence length and word complexity. A higher score indicates easier readability, while lower scores indicate more difficult readability. $\text{Flesch Reading Ease} = 206.835 - 84.6 \cdot \frac{\#\text{ syllables}}{\#\text{ words}} - 1.015 \cdot \frac{\#\text{ words}}{\#\text{ sentences}}$
	Gunning Fog Index	Estimates the years of formal education required to understand a particular text on first reading. $\text{Gunning Fog Index} = 0.4\left(\frac{\#\text{ words}}{\#\text{ sentences}} + 100 \cdot \frac{\#\text{ complex words}}{\#\text{ words}}\right)$
Syntactic Features	Part-of-speech tags	Proportion of tokens tagged as a specific part-of-speech category in the text. We make use of the universal part-of-speech tagging schema and calculate a distinct score for ADJ (adjectives), ADP (adpositions), ADV (adverbs), AUX (auxiliaries), CCONJ (coordinating conjunctions), DET (determines), INTJ (interjections), NOUN (nouns), NUM (numerals), PART (particles), PRON (pronouns), PROPN (proper nouns), PUNCT (punctuations), SCONJ (subordinating conjunctions), SYM (symbols), VERB (verbs), and X (words that do not fit into the other part-of-speech categories).
	Named entities	Proportion of named entities in a text, calculated by dividing the number of named entity tokens by the total number of tokens. Functions as a measure of writing style.
	Present tense	Number of present tense verbs in a text, normalized by text length. Functions as a measure of writing style.
	Imperative	Number of imperative verb forms in a text, normalized by text length. Functions as a measure of writing style.
	First person	Number of first-person verb forms in a text, normalized by text length. Functions as a measure of writing style.

Table 5: Stylistic features and their descriptions.

G Random Forest Feature Importance

Figure 6 provides an overview of the importance of different features (stylistic, stance) in the random forest classifier.

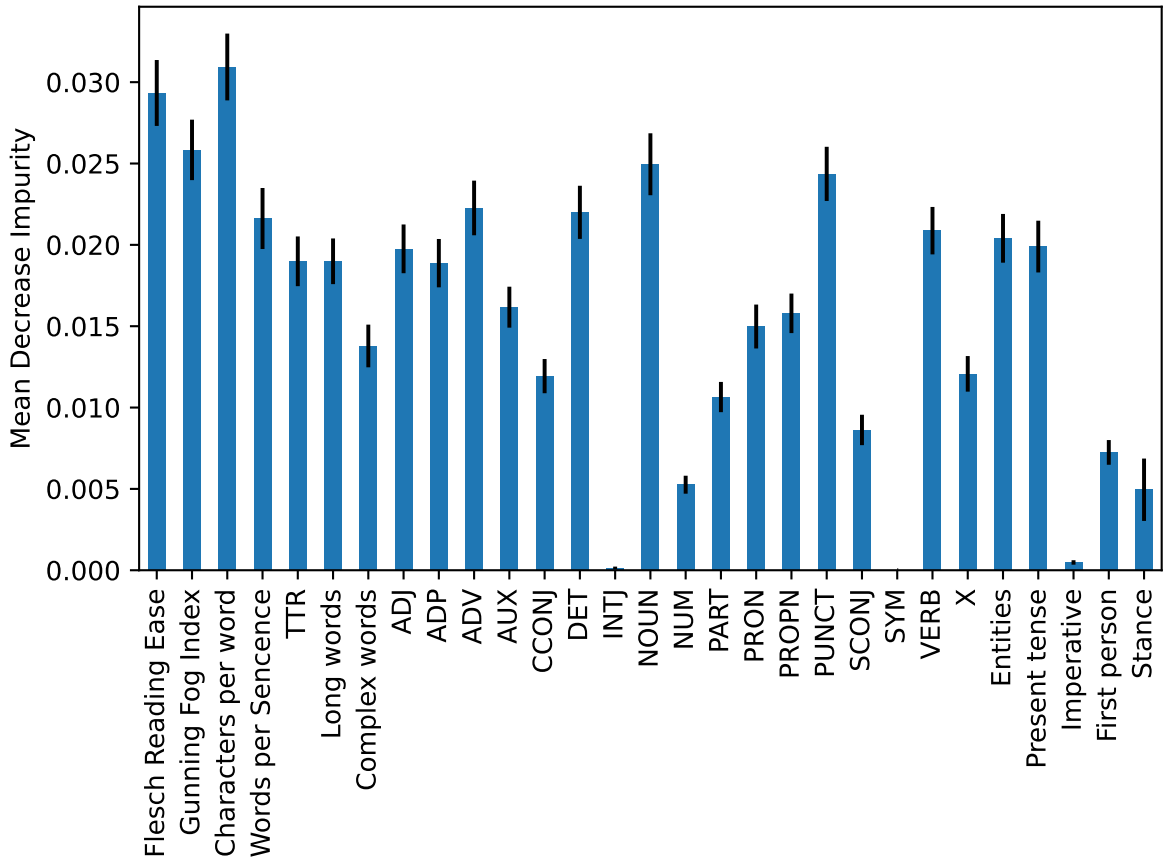

Figure 6: Random forest feature importance measured by the mean decrease impurity.