XFACT Team0331 at PerspectiveArg2024: Sampling from Bounded Clusters for Diverse Relevant Argument Retrieval

Wan Ju Kang, Jiyoung Han, Jaemin Jung, and **James Thorne**
KAIST
Republic of Korea

Abstract

This paper reports on the argument mining system submitted to the ArgMining workshop 2024 for The Perspective Argument Retrieval Shared Task (Falk et al., 2024). We combine the strengths of a smaller Sentence BERT model and a Large Language Model: the former is fine-tuned for a contrastive embedding objective and a classification objective whereas the latter is invoked to augment the query and populate the latent space with diverse relevant arguments. We conduct an ablation study on these components to find that each contributes substantially to the diversity and relevance criteria for the top-k retrieval of arguments from the given corpus.

1 Introduction

Argument retrieval remains a challenging problem in the natural language processing domain, when considered jointly with perspectives and diversity. The problem is defined as the collection of claims carrying a stance towards a query. A query may be of various topical widths ranging from an entire issue or a single statement. Arguments are a widely utilized discursive tools, and performant systems of argument recognition will prove useful in further advancing bias analyses, slant measurement (Devatine et al., 2022), content recommendation, and text generation tasks (El Baff et al., 2019). Perspectives can prove a useful clue in the argument retrieval problem but at the same time bring about an additional challenge. Especially with shorter texts, such as the ones handled in this shared task, the demographic profile of the argument author may be of use in trying to extract opinions grounded in various populations.

In this paper, we present the work conducted by our team, "XFACT team0331", for the ArgMining 2024 shared task of perspective argument retrieval. Our findings indicate feasibility of the system design across the three scenarios, helping us better understand the complexities of taking perspectives into consideration when retrieving arguments.

This shared task involves a unique challenge in terms of perspectives and socio-cultural variables. Not only are the systems required to pursue relevance to a given query, but they are also assessed in terms of the demographic profile of the argument authors.

2 Related Work

Recent years have seen rapid progress of argument retrieval along several lines of research.

Teufel et al. (1999) studied argument extraction in the scientific text domain. Later works extend the argument retrieval work to arbitrary domains and eventually the entire world wide web, such as in Rahwan et al. (2007) and similar systems discussed in the seminal text (Manning, 2008). Wachsmuth et al. (2017) designed an argument search engine involving an indexing process, which takes candidate documents over the web and indexes the assessed arguments therein into a corpus, and a retrieval process, which, upon arrival of a query, ranks and presents relevant indices of arguments. Stab et al. (2018) put together an offline component and an online component, which account for the indexing and retrieval tasks, respectively.

In the closely related problem of stance detection, Hardalov et al. (2021) proposed methods for recognizing stance across texts from multiple domains, with their design of label embeddings in the latent space that adapt to the arbitrary topic at hand. Arakelyan et al. (2023) presented a similar approach leveraging a topic-guided sampler for alleviating inherent imbalance in the data. Then, a pre-trained language model is fine-tuned against a contrastive learning objective for recognizing the in-favor and the against statements.

In news writing, Baly et al. (2018), Baly et al. (2020a), and Baly et al. (2020b) profiled media

Proceedings of the 11th Workshop on Argument Mining, pages 182–188
August 15, 2024 ©2024 Association for Computational Linguistics

sources by measuring their bias in terms of factuality reporting and any conveyed political ideology. Ko et al. (2023) solved a five-scale political stance prediction problem by incorporating texts from various sources beyond news articles, such as Reddit posts, and employs a multi-granularity hierarchy on the texts to capture any subtleties carrying stance information. In a related work by Liu et al. (2022), a triplet loss was imposed on an anchor article with two same-story versions – one leftist and the other rightist – to train a language model to identify and distinguish ideology-informed representations between articles.

Argumentative language modeling techniques have also been studied. For instance, Jo et al. (2021) proposed counterargument generation methods assisting language models with knowledge graphs such that the natural language inference process can determine the entailment/relevance of a claim more effectively. Holtermann et al. (2022) studied a similar problem but with an additional pursuit for fairness in argument generation. For every instance deemed biased, a counter-stereotypical statement is synthesized and used in training.

3 Task

The ArgMining 2024 Shared Task for Perspective Argument Retrieval consists of the following three retrieval scenarios:

- Scenario 1 is dubbed the "baseline" scenario and is a retrieval of top-k arguments from a corpus given a query alone and no additional information on the demographic profile on either of the query or the argument side.
- Scenario 2 is the explicit perspectives scenario where a demographic property is provided in addition to the query. This property may be used explicitly to filter or process arguments from the corpus.
- Scenario 3 is the implicit perspectives scenario where a demographic property is still available on top of the query, but it may not be used explicitly on the corpus-processing step. Only latent encoded information may be used to retrieve relevant arguments.

The dataset is originally from Vamvas and Sennrich (2020), where comments in the French, German, and Italian language are organized across various political issues (queries) for the 2019 Swiss federal elections.

4 Method

We propose a novel approach to retrieve relevant and diverse arguments. Major components of our design are as follows:

- an embedding model fine-tuned for two training objectives
- a large language model instructed to generate its own arguments given the query
- a topic sampler to filter the vast majority of the corpus

The overview of the proposed system is presented in Fig. 1. An embedding model produces latent vectors of arguments and a given query. These are then spread out in the latent space, to be clustered according to the nearest generated argument. More details follow in Section 4.5.

4.1 Embedding Model

We choose the PARAPHRASE-MULTILINGUAL-MPNET-BASE-V2 sentence transformer (Reimers and Gurevych, 2019), (Reimers and Gurevych, 2020) as our embedding model, given its state-of-the-art performance and capability to handle multiple languages. We further train this embedding model on a weighted combination of two losses, with the exact weighting as a hyperparameter: a contrastive loss and a classification loss. (See Appendix A)

The training of the embedding model incorporates two desirable directions of the model enhancement: we want the model (i) to produce sufficiently different latent representations for "on-set"(relevant) and "off-set"(irrelevant) arguments and (ii) to capture any underlying connections between the query and its on-set arguments sufficiently such that, when presented with a (query, irrelevant) pair, it is successfully discarded as an outlier. For these objectives, we detail the training process below.

For objective (i), we aim to enhance the embedding model's ability to widen the gap between relevant arguments and irrelevant arguments with respect to a query. For this, the multiple negatives ranking loss (Henderson et al., 2017) was chosen, and the given corpus dataset was rearranged in triplets (q, a_p^i, a_n^j), where the query q is followed by a randomly selected relevant candidate a_p and then by a randomly selected argument not in the

RELEVANT CANDIDATES list for that query, for all i in the RELEVANT CANDIDATES list.

For objective (ii), a linear layer is appended to the embedding model to from a binary classifier between FAVOR and AGAINST, labels both available in the given corpus. The corpus is rearranged in pairs (q, a_p^i), the query and its relevant candidate for all candidates i in the query's RELEVANT CANDIDATES list. The classifier outputs a real value from 0 to 1, trained on binary cross entropy loss.

For Scenario 2, the query is concatenated with its demographic property, and each corpus argument is concatenated with its demographic profile before passing through the embedding model. For Scenario 3, only the query is augmented with the socio-cultural variables information.

4.2 Argument Generator Model

We employ several open-source large language models to leverage argument generation. PHI-3-MINI-4K-INSTRUCT and GEMMA-1.1-2B-IT have each been invoked to produce 20 relevant key arguments with respect to a query in their respective instruction prompt formats: 10 favoring it and 10 against it. The purpose of these LLM-generated arguments is twofold. One is that, since the corpus contains comments from individuals who might have a rather local view on the topic at hand, the LLM, as a generic knowledge entity can provide more diverse and holistic takes on the issue. Where necessary, we make the distinction between the natural and synthesized arguments as "corpus arguments" and "LLM-generated arguments" henceforth. These two sets are produced to form the augmented corpus. The other purpose of the LLM-generated arguments is to filter out corpus arguments that are too far from the LLM-generated arguments in the latent space. The mild assumption underneath is that, if a corpus argument is relevant enough, it must be close to at least one of the LLM-generated arguments. The exact cut-off distance criterion is described in Section 4.5

4.3 Topic Filter Model

At execution time, as a first measure, we invoke a KeyBERT (Grootendorst, 2020) instance between the query and the augmented corpus to discard a large portion of the augmented corpus as irrelevant. This is a simple filter based on latent encodings from BERT (Devlin et al., 2019), which takes a document and find sub-phrases that most closely resembles a given topic by cosine similarity. The filter proceeds to retain only the arguments whose keyword set contains any of the keywords in the query's keyword set. That is, all arguments whose $k_q \cap k_a = \emptyset$ are screened away.

4.4 Clusterer

The vector representations produced by the embedding model undergo a simple clustering process equivalent to running a K-Means for one iteration. The LLM-generated arguments serve as the initial centroids, and the corpus arguments are each assigned a group it belongs to by nearest centroid. In other words, if an argument can find a cluster it can belong in, it is considered relevant. Each cluster's member argument count is calculated and used for diversity sampling as explained in Section 4.5.

4.5 Overview

In this section, we describe how the components tie in together. In Section 4.2, corpus arguments' proximity to LLM-generated arguments was chosen as a criterion for determining the relevance. We construct this criterion as a function of the classifier confidence, as measured in the classifier's final layer value. That is, if the learned classifier from Section 4.1 can barely determine whether an argument is in favor or against some given query, then that argument should be allowed to exist in a generously larger ball from the LLM-generated arguments in the latent space. The converse also applies; a confident classifier should indicate the argument's high proximity to at least one of the LLM-generated arguments. In short, the cut-off criterion is governed by the learned classifier's decision. The actual cut-off procedure then takes an indicator function over the respective ball of r^{cutoff} around each LLM-generated argument embedding a_i, that, when evaluated as all off-ball, considers the corpus argument a_c irrelevant:

$$\prod_i \mathbb{1}(\text{dist}(a_c, a_i) > r^{\text{cutoff}}) = \begin{cases} 0, \text{ if } a_c \text{ relevant} \\ 1, \text{ otherwise} \end{cases} \tag{1}$$

The relevance cut-off criterion above doubles as a diversity sampling criterion, and we capitalize on this extension by introducing a per-a_i cut-off radius instead of a uniform radius for all the balls. Each LLM-generated argument embedding is assigned its own cut-off radius r_i^{cutoff} that is inversely proportional to its member argument count. In practice, the base cut-off radius is first calculated with the

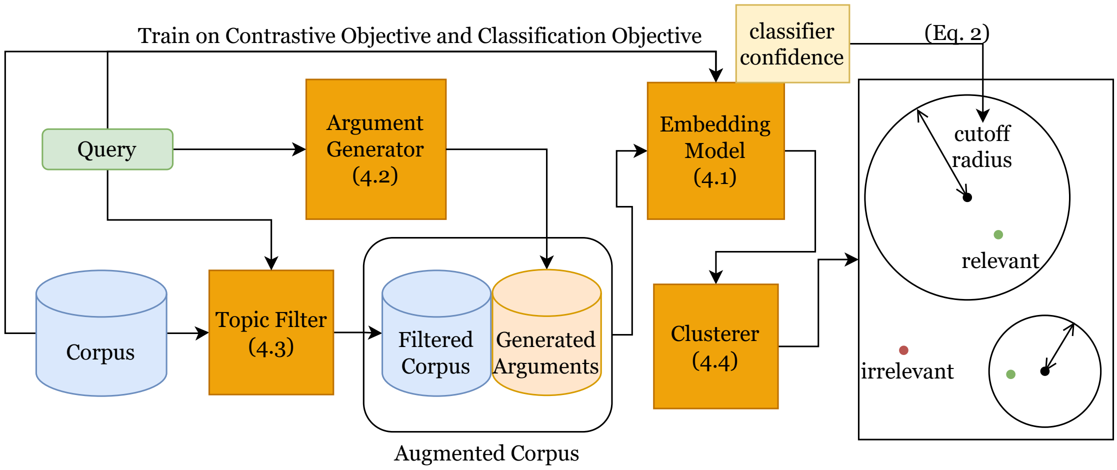

Figure 1: An overview of the proposed stance detector; sections detailing the components are in parentheses

classifier confidence and then is further adjusted by cluster member count. We re-write Equation 1 with the generalized cut-off radii. Let A_i denote the set of corpus arguments belonging to the cluster centered by the LLM-generated argument embedding a_i:

$$\prod_i \mathbb{1}(\text{dist}(a_c, a_i) > r_i^{\text{cutoff}}) = \begin{cases} 0, \text{ if } a_c \text{ relevant} \\ 1, \text{otherwise,} \end{cases} \tag{2}$$

where $r_i^{\text{cutoff}} = r^{\text{cutoff}} + r_i^{\text{adjustment}}$
and $r_i^{\text{adjustment}} \propto \frac{1}{|A_i|}$

Having a cut-off radius may be advantageous over a ranking method: one that ranks the arguments by the distance from the centroid. First, ranking takes $O(n \log n)$ time whereas cut-off executes in linear time. Second, the proposed system works regardless of whether the k value is known. That is, it is flexible enough to accommodate an arbitrary downstream ranker of k unknown *a priori*.

5 Results

Our results are presented in Tables 1 and 2. All reportings are on the dev sets, averaged across the three scenarios, due to limited print space.

5.1 Ablation

5.1.1 Effects of Relevance Cut-off

We report the results of applying only the relevance cut-off radius, as in the uniform criterion in Equation 1.

Table 3 shows that the absence of the diversity sampler compromises the diversity scores of the

k	ndcg@k	precision@k
4	0.694	0.692
8	0.679	0.671
16	0.670	0.660
20	0.677	0.673

Table 1: Relevance scores averaged across the three scenarios

k	alpha_ndcg@k	kl_divergence@k
4	0.625	0.151
8	0.618	0.134
16	0.626	0.100
20	0.638	0.091

Table 2: Diversity scores averaged across the three scenarios

k	alpha_ndcg@k	kl_divergence@k
4	0.557	0.158
8	0.565	0.140
16	0.579	0.104
20	0.580	0.094

Table 3: System follows Eq. 1. (Diversity component ablated). Diversity scores averaged across the three scenarios

k	ndcg@k	precision@k
4	0.627	0.617
8	0.625	0.619
16	0.624	0.620
20	0.618	0.613

Table 4: System follows Eq. 2 with $r_i^{\text{cutoff}} = \infty$ (Relevance component removed). Relevance scores averaged across the three scenarios

system. While KL divergence was measured (ommitted for spacing) to be similar as in Table 2, alpha_ndcg@k values have dropped substantially and seemingly relies on k value alone for diversity gains, rather than on the constituent arguments of the k retrieved passages. Since demographic profile information is latently encoded in the query and arguments, the decrease in KL divergence scores indicate that the role of the diversity sampler is indeed to take into account various demographic profiles.

5.1.2 Effects of Diversity Sampling

We report the results of applying only the diversity sampler. This is done by running K-Means on the augmented corpus for a number of iterations and then sampling from the resulting clusters inversely proportionately to each cluster size.

With no relevance cut-off and only clustering and then sampling for proportionate representation, Table 4 shows that the ndcg@k and precision@k values have both fallen by 6-7%p. This supports that the relevance threshold aids in discarding certain arguments, however diverse they may be in constituting their respective cluster.

6 Conclusion

We present an argument retrieval system combining the strengths of a fine-tuned embedding model and open-source large language models. The embedding model is trained to tell apart relevant arguments from the irrelevant arguments through triplet loss signals leveraging hard negatives. At the same time, the embedding model is also exposed to in-distribution classification examples, such that it detects out-of-distribution (irrelevant) arguments and relays the anomalous finding to the relevance cut-off criterion. LLMs provide diverse anchors to which a given corpus argument may be checked against for membership, and then the member densities are utilized for diverse representation of arguments relevant to a given query.

References

Erik Arakelyan, Arnav Arora, and Isabelle Augenstein. 2023. Topic-guided sampling for data-efficient multi-domain stance detection. *arXiv preprint arXiv:2306.00765*.

Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. 2020a. We can detect your bias: Predicting the political ideology of news articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4982–4991.

Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018. Predicting factuality of reporting and bias of news media sources. *arXiv preprint arXiv:1810.01765*.

Ramy Baly, Georgi Karadzhov, Jisun An, Haewoon Kwak, Yoan Dinkov, Ahmed Ali, James Glass, and Preslav Nakov. 2020b. What was written vs. who read it: News media profiling using text analysis and social media context. *arXiv preprint arXiv:2005.04518*.

Nicolas Devatine, Philippe Muller, and Chloé Braud. 2022. Predicting political orientation in news with latent discourse structure to improve bias understanding. In *3rd Workshop on Computational Approaches to Discourse (CODI 2022)*, pages 77–85. ACL: Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Roxanne El Baff, Henning Wachsmuth, Khalid Al Khatib, Manfred Stede, and Benno Stein. 2019. Computational argumentation synthesis as a language modeling task. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 54–64, Tokyo, Japan. Association for Computational Linguistics.

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.

Maarten Grootendorst. 2020. Keybert: Minimal keyword extraction with bert.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2021. Cross-domain label-adaptive stance detection. *arXiv preprint arXiv:2104.07467*.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. *arXiv preprint arXiv:1705.00652*.

Carolin Holtermann, Anne Lauscher, and Simone Paolo Ponzetto. 2022. Fair and argumentative language modeling for computational argumentation. *arXiv preprint arXiv:2204.04026*.

Yohan Jo, Haneul Yoo, JinYeong Bak, Alice Oh, Chris Reed, and Eduard Hovy. 2021. Knowledge-enhanced evidence retrieval for counterargument generation. *arXiv preprint arXiv:2109.09057*.

Yunyong Ko, Seongeun Ryu, Soeun Han, Youngseung Jeon, Jaehoon Kim, Sohyun Park, Kyungsik Han, Hanghang Tong, and Sang-Wook Kim. 2023. Khan: knowledge-aware hierarchical attention networks for accurate political stance prediction. In *Proceedings of the ACM Web Conference 2023*, pages 1572–1583.

Yujian Liu, Xinliang Frederick Zhang, David Wegsman, Nick Beauchamp, and Lu Wang. 2022. Politics: Pretraining with same-story article comparison for ideology prediction and stance detection. *arXiv preprint arXiv:2205.00619*.

Christopher D Manning. 2008. *Introduction to information retrieval*. Syngress Publishing,.

Iyad Rahwan, Fouad Zablith, and Chris Reed. 2007. Laying the foundations for a world wide argument web. *Artificial intelligence*, 171(10-15):897–921.

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*. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. 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.

Simone Teufel et al. 1999. *Argumentative zoning: Information extraction from scientific text*. Ph.D. thesis, Citeseer.

Jannis Vamvas and Rico Sennrich. 2020. X-stance: A multilingual multi-target dataset for stance detection. *arXiv preprint arXiv:2003.08385*.

Hyperparameter	**Value**
Embedding Model, epochs	10
Embedding Model, ratio (contrastive-to-classification multi-objective relative weighting)	5
Embedding Model, batch size	64
Embedding Model, learning rate	2e-05
Embedding Model, warmup steps	10000
Embedding Model, weight decay	0.01
Clusterer, K (full system)	1
Clusterer, K (ablated (Section 5.1.2))	100

Table 5: Hyperparameters of the system

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.

A Appendix

A.1 Table of hyperparameters

Table 5 shows our system components and its hyperparameters, where relevant.

A.2 Instruction prompt and response examples

Below is a sample from invoking a PHI-3-MINI-4K-INSTRUCT.

Prompt Sample:
Can you provide 10 arguments that favor the following statement and 10 arguments that are against it?: Faut-il renforcer la protection contre le licenciement des travailleuses et travailleurs plus âgé·e·s?

Response Sample:
Arguments in favor of strengthening protection against the dismissal of older workers:

1. Age discrimination: Strengthening protection against dismissal helps prevent age discrimination, ensuring that older workers are not unfairly targeted for layoffs. 2. Experience and knowledge: Older workers often have valuable experience and knowledge that can benefit the company, and protecting them from dismissal helps retain this expertise. 3. Economic stability: Older workers are more likely to have dependents and financial obligations, so protecting them from dismissal can help maintain economic stability for families. 4. Reduced turnover costs: Retaining older workers can reduce

turnover costs, as it is generally more expensive to hire and train new employees. 5. Diversity and inclusion: Protecting older workers from dismissal promotes diversity and inclusion in the workplace, which can lead to better decision-making and innovation. 6. Social justice: Strengthening protection against dismissal of older workers can be seen as a step towards social justice, ensuring that all workers are treated fairly regardless of age. 7. Improved morale: Protecting older workers from dismissal can improve morale and job satisfaction, as employees feel valued and secure in their positions. 8. Reduced unemployment: By protecting older workers from dismissal, the number of unemployed individuals may decrease, as they are less likely to be laid off. 9. Long-term planning: Companies that protect older workers from dismissal can benefit from long-term planning, as they can rely on the experience and knowledge of these employees. 10. Legal compliance: Strengthening protection against dismissal of older workers can help companies comply with existing labor laws and regulations.

Arguments against strengthening protection against the dismissal of older workers:

1. Merit-based decisions: Some argue that dismissals should be based on merit and performance, rather than age, and that protecting older workers from dismissal may hinder this process. 2. Job opportunities for younger workers: Critics argue that protecting older workers from dismissal may limit job opportunities for younger workers, potentially leading to higher unemployment rates among the youth. 3. Innovation and adaptability: Some believe that companies need to be able to adapt quickly to changing market conditions, and protecting older workers from dismissal may hinder this process. 4. Financial burden: Strengthening protection against dismissal of older workers may place a financial burden on companies, as they may be required to provide additional benefits or compensation. 5. Workforce flexibility: Critics argue that protecting older workers from dismissal may limit a company's ability to restructure its workforce and adapt to changing business needs. 6. Ageism concerns: Some argue that strengthening protection against dismissal of older workers may inadvertently reinforce ageism, as it may be perceived as favoring older workers over younger ones. 7. Skill gaps: Critics argue that protecting older workers from dismissal may lead to skill gaps in the workforce, as companies may be less willing to invest in training and development for older employees. 8. Competitive disadvantage: Companies that protect older workers from dismissal may be at a competitive disadvantage compared to those that can more easily restructure their workforce. 9. Employee motivation: Some argue that protecting older workers from dismissal may reduce employee motivation, as younger workers may feel that their opportunities for advancement are limited. 10. Economic efficiency: Critics argue that strengthening protection against dismissal of older workers may reduce economic efficiency, as companies may be less able to respond to market changes and adjust their workforce accordingly.