Detecting Discrimination in Job Vacancies: A Critical Reflection on the Potential of AI Language Models

STEVEN VETHMAN, Sciences Po Law School, France and TNO, The Netherlands
AJAYA ADHIKARI, TNO, The Netherlands
ELIZA D.K. HOBO, TNO, The Netherlands
MAAIKE H.T. DE BOER, TNO, The Netherlands
JOOST A.G.M. VAN GENABEEK, TNO, The Netherlands
COR J. VEENMAN, Leiden University, LIACS, The Netherlands and TNO, The Netherlands

Explicit discrimination in job vacancies by using terms that refer to the candidate's background is illegal. Yet, it is still present in numerous vacancies, as was recently observed in the Netherlands. Labour market authorities have organized efforts for the detection of explicit discrimination, which are based on the detection of terms such as "young" or "male". However, many non-discriminatory phrases also contain these terms, such as "we are a young company" or "working with male patients". This results in a labour-intensive task to identify discriminatory job vacancies and act on them. AI language models are seen as promising innovations that may improve efficiency. Yet, their use by governmental bodies raises concerns and requires caution.

In this paper, we critically examine the potential of AI and language models to support labour market authorities in detecting explicit discrimination. We do this through an investigation of the potential efficiency gain whilst centring user needs. For this, we first create a labelled data set concerning gender discrimination and investigate a variety of models in their ability to detect known and unforeseen discriminating terms in context. Results show that these methods can support detecting explicit gender discrimination by bringing substantial gains to precision and make sensible suggestions for new terms to detect in vacancies. We complement this with a critical reflection based on interviews with ten experts. They state that considerations on responsibly using AI and language models go beyond efficiency, emphasizing the importance of the underlying goal of discrimination detection. Is this goal reached within a reasonable investment and with acceptable side-effects? In conclusion, this applied use case demonstrated that AI and language models could meaningfully bring efficiency to labour market authorities' efforts to detect explicit discrimination in job vacancies. However, we advocate that, even for technologies used for common good, critical reflections beyond efficiency are needed to decide between AI and non-AI alternatives.

Keywords: Discrimination, AI, Algorithm, Language Model, NLP, Bias, Critical Reflection, Thematic Analysis, Responsible AI

Reference Format:
Steven Vethman, Ajaya Adhikari, Eliza D.K. Hobo, Maaike H.T. De Boer, Joost A.G.M. van Genabeek, and Cor J. Veenman. 2025. Detecting Discrimination in Job Vacancies: A Critical Reflection on the Potential of AI Language Models. In *Proceedings of Fourth European Workshop on Algorithmic Fairness (EWAF'25).* Proceedings of Machine Learning Research, 20 pages.

Authors' Contact Information: Steven Vethman, Sciences Po Law School, Paris, France and TNO, The Hague, The Netherlands, steven.vethman@sciencespo.fr; Ajaya Adhikari, TNO, The Hague, The Netherlands; Eliza D.K. Hobo, TNO, The Hague, The Netherlands, eliza.hobo@tno.nl; Maaike H.T. De Boer, TNO, The Hague, The Netherlands; Joost A.G.M. van Genabeek, TNO, The Hague, The Netherlands; Cor J. Veenman, Leiden University, LIACS, Leiden, The Netherlands and TNO, The Hague, The Netherlands.

EWAF'25, June 30–July 02, 2025, Eindhoven, NL

1 Introduction

A job vacancy is written to attract the most suitable job candidate. Favouring a vacancy towards specific job seekers is only allowed if it is directed towards objective criteria, such as required skills. You cannot ask for someone who *has the balls* to do it or another *lady* to complete your team. Specification of socio-demographic properties is generally disapproved by labour market authorities and forbidden in discrimination law [41, 55, 56].

Labour market authorities such as the Netherlands Labour Authority (NLA) [42] and the Netherlands Employment Insurance Agency (UWV) [53] facilitate, supervise and orientate research on the recruitment process on the Dutch job market. Their interests include *explicit discrimination* as defined in the Dutch Equal Treatment Act [56], which is enforceable by law. They have organized efforts against explicit discrimination by automatically detecting certain words or short phrases (search terms) from a list of potentially discriminatory terms, called a *forbidden list* (FL). Vacancies that contain terms of the FL are flagged for manual inspection to assess whether they are used in a discriminatory context. One cannot ask for a "male" prison guard, but one can specify the job concerns "male" prisoners. Many job vacancies containing non-discriminatory uses of search terms are flagged, which distracts from finding the job vacancies that do discriminate. Next to that, the FL approach is restrictive as it does not enable finding explicit discrimination beyond the included terms. These limitations make efforts to continuously monitor and prevent discrimination in job vacancies impractical and allows explicit discrimination to slip through. To illustrate the magnitude of this issue, just for age discrimination, more than 70.000 Dutch job vacancies were estimated to explicitly discriminate in 2019 [24].

Labour market authorities, as many governmental bodies, are therefore incentivized to use digital innovation to make their processes more efficient [43, 45]. Since the rise of AI and language models, a governmental vision on generative AI was published. It aims for a value-driven AI ecosystem where public bodies do not fall behind [40]. However, government use of AI and algorithms is met with caution and concern as they can be (and have been) in conflict with human rights [3, 4, 15, 46].

Therefore, we set out to critically investigate the potential for AI and language models to support labour market authorities with the detection of explicit discrimination in job vacancies. We centre the needs and context of labour market authorities through a mixed method approach with two related actions: a *quantitative* investigation to gauge the potential meaningful efficiency gain of AI and language models in detecting explicit discrimination and a *qualitative* examination to critically reflect on the necessary considerations beyond efficiency to decide between AI and non-AI alternatives.

Concretely, our mixed-methods approach consists of the following three contributions: First, we *collect and annotate a data set* in consultation with domain experts from UWV, the NLA and the Netherlands Institute for Human Rights (NIHR) [22]. Second, we *demonstrate the efficiency gain* that language models and machine learning can achieve relative to the baseline of the FL approach. We assess the performance of the models on known forms of explicit discrimination and explore whether they can detect new and unforeseen potentially discriminatory terms. Third, we *critically reflect* on the potential efficiency gain by reporting the key considerations from interdisciplinary experts with relevant experience. We do this through qualitative research consisting of ten interviews and a reflexive thematic analysis based on the methodology of Braun and Clarke [11].

In the next section, we elaborate on related work to position our contribution. Section 3 describes our methodology for the quantitative research concerning the creation of the data set and the experimental design choices for detecting

known and unknown forms of explicit discrimination. In Section 4, we demonstrate the results: the data set and the estimated gain of AI and language models to detect the discriminatory context of known and unknown forms of explicit discrimination. Section 5 elaborates on our qualitative research approach, the interviews and thematic analysis. Section 6 reports on the distilled considerations from the critical reflection. We finalize the paper with a discussion and conclusion.

2 Related Work

In this section, we position our contribution in research on discrimination in job vacancies, research on bias in language models, AI tooling for inclusive writing, and critical studies on the use of AI and language models.

Scientific studies into discriminatory formulations of job vacancies are scarce. This may be related to the general trend of discrimination in hiring having moved from being explicit to implicit [57]. Still, explicit discrimination in job vacancies has not vanished. Job advertisements have been shown to request a specific gender in China [33, 34], and to request specific age limits or preferences on physical appearance in Indonesia [7, 44]. Efforts to detect these discriminatory practices ranges from manual content or field analysis to text mining with natural language processing (NLP). Also in the Netherlands, NLP techniques were used to detect explicit age discrimination in online vacancies published in 2017, 2018 and 2019 [24, 25, 31]. They compiled a list of age-related formulations that are discriminatory, similar to the search terms and the previously referred "forbidden list" (FL). Using the method of regular expressions they estimated that the lower bound on the prevalence of age discrimination was estimated around 70.000 vacancies, which was shown to be just the tip of the iceberg. Our contribution extends their work on Dutch job vacancies through investigating AI and language models alternatives beyond the FL approach, exploring their potential to find unforeseen or unknown discrimination as well as our additional effort to critically reflect on this opportunity through a qualitative study.

Furthermore, our contribution is distinct from many NLP bias studies and inclusive writing tooling as we focus on explicit discrimination that is actionable for labour market authorities. AI research of the last ten years has expanded on the detection of biased language that is strongly associated with demographics (age, gender, ethnicity) [10, 14, 27], although their reliability has been contested [9, 19]. Likewise, AI tooling for inclusive writing, such as Textio [52] and Develop Diverse [20] highlight and suggest improvements for biased wording in job vacancies that may make the text more appealing to wider range of applicants. In alignment with the call for actionability of [18] , we focus on what information labour market authorities need to take action: that is the explicit discrimination, enforceable by law [56]. This is in contrast with implicit biased wording, that is generally not illegal, but still influences the diversity of applicants through an implied sense of belonging [28].

Finally, our work builds on insights from critical studies on AI language models (e.g. [8, 9]) and critically applied use cases with AI and language models (e.g. [26, 51]). We predominantly incorporate their overarching message to not immediately assume that AI and language models will bring the desired improvement. Specifically, we align our research to the seminal work of Bender et al. [8] by applying their following recommendations to our use case: (1) spend time on and clearly motivate assembling data sets suited for the tasks at hand rather than increasing data or model size, (2) rather than aiming for the highest scores, focus on efficiency gains that are meaningful for the task at hand, and (3) consider risks and limitations of language models relative to current design or alternatives.

3 Method for Data Creation and Discrimination Detection in Context

This section delineates our methodology for the quantitative investigation on the potential efficiency gain of AI and language models. To start, we motivate our selection of gender discrimination and describe the creation process of the annotated data set. Thereafter, the design choices concerning language models and classification methods are described. This is followed by the experimental designs for known and unknown explicit discrimination detection in context. Finally, we specify the performance metrics used for evaluation.

3.1 Creation of the Annotated Data Set

For the design of the annotated data set for explicit discrimination detection, we did a preliminary analysis based on interviews with experts from the NLA, UWV and NIHR and an exploratory data analysis on Dutch vacancies.

Based on interviews, four types of discrimination were selected for further exploration: age, gender, ethnicity and nationality. These types were primarily chosen by their expected frequency in job vacancies, such that the potential of AI and language models is evaluated for where the investment may be worth it in terms of impact.

Given that the frequency of age discrimination based on search terms was already investigated in the Netherlands [31], we focused the data exploration on the grounds of gender, ethnicity and nationality. With the help of NLA, UWV and NIHR, we gathered potentially discriminating words as search terms (certain words or short phrases within regular expressions) for these three discrimination grounds based on their experience or court cases. Search terms concerned explicit discrimination based on the definition of direct and indirect discrimination in the Dutch Equal Treatment Act [56], which due to the legal status is enforceable for authorities. The domain experts shared that cases can directly refer to demographic membership, e.g. “young” (direct discrimination) or that a seemingly “neutral” requirement disproportionately targets or excludes a certain demographic, e.g. “starter" or “student" excluding older workers (indirect discrimination). To be clear, this means that more implicit forms of discrimination are outside of our scope.

Then, to explore the results of using a FL for gender, ethnicity and nationality, we applied the search terms to 2.4 million Dutch vacancies posted on public websites in 2018 (which were available to the authorities). A small random sample of 100 sentences of job vacancies for each search term was inspected by the authors to get an indication of the precision of the FL. Precision here refers to the proportion of sentences with a search term that was likely to be discriminatory. We observed that for the grounds of ethnicity and nationality, precision was already high. To illustrate, the use of the terms “without an accent” (100%) and “mother tongue” (72%) was most often unambiguously discriminating to those that had acquired proficient speech but were not born speaking the language required. For gender discrimination, we observed lower precision for terms such as “female” (22%), “male”(19%), “woman/women”(8%), “man/men”(4%). We noticed that a discriminatory context often related to whether the word directly applies to the candidate (“looking for talented saleswomen”) or other aspects of the job opening (“providing expert advice to women interested in our product”). Gender discrimination had six search terms that occurred more than 500 times in 2018. In the end, gender discrimination was selected as the scope of this research. It had sufficient frequency, variety in possible search terms, room for improvement in precision and was preferred over age discrimination as that was recently investigated [24].

In consultation with NLA, UWV and NIHR, the scope of the data collection and annotation was further specified. Annotation of gender discrimination was performed on vacancy texts that contained known potentially

discriminating terms from the FL, the baseline method. The experts annotated with the aim to find vacancies with *Highly Suspected Discrimination* (HSD) in contrast to a label that may imply *certain* discrimination, because only courts can provide a definite judgement. This aim on "highly suspected" was put here to limit the role of AI and language models to support users and leave users the responsibility on interpreting the job vacancy and to decide the appropriate action. Moreover, we decided to do the annotation at the sentence level. This was chosen based on the trade-off between having sufficient context available (text around the search term) and the speed of annotation (longer texts take longer to read and annotate).

Two experts from NLA and three experts from UWV were selected (two women, three men) to annotate 5947 sentences of Dutch job vacancies that contained the explicit search terms. The possible labels were "yes, I highly suspect the sentence to be discriminatory", "no, I do not highly suspect the sentence to be discriminatory", or "?: I do not know or would take longer than 30 seconds to decide". The annotators received a general instruction together with a human rights expert from NIHR for any questions beforehand. Afterwards, all five annotators provided their annotation independently. To determine the inter-annotator agreement between the five annotators, an identical subset of 600 random sentences (from the 5947) was given to all five annotators. The chosen measure for inter-annotator agreement is the Fleiss' Kappa [23], which measures the level of agreement between two or more annotators, when the annotation is measured on a categorical scale. The remaining 5347 sentences were evenly distributed among the five annotators.

3.2 Methods to detect explicit discrimination in context

Among the sentences detected by the FL approach, there are too many that contain terms but in the end are non-discriminatory. From a machine learning perspective, the non-discriminatory sentences are false positives (FP) and the discriminatory sentences are true positives (TP). The precision of the FL approach, defined as $TP/(FP+TP)$, is low. The sentences from vacancies that do not contain search terms are the negatives, consisting of true negatives (TN) that are non-discriminatory and false negatives that are discriminatory unrelated to the FL.

Below, we describe experiments that follow two aims of AI and language models: 1) to find known explicit discrimination, or to increase the precision of the FL method, 2) to find unknown explicit discrimination, or to detect discriminatory context of terms they have seen before. The second aim looks for potential to reduce the false negatives, thereby increasing the recall: $TP/(FN+TP)$.

We gauge the potential of AI and language models for discrimination detection by evaluating a variety of modelling techniques for language (feature extraction methods) and machine learning methods. Given the setting of labour market authorities, it is essential that technologies are compliant with regulations and that authorities have sufficient control and accountability on their use. Therefore, we focus here on the relatively small language models (SLM): bag-of-words [36], Word2Vec [39] and BERT [21] (descriptions below). Due to their size and power requirements, these SLMs can more easily be hosted and run on-premise relative to large language models (LLMs) such as ChatGPT. LLMs are less of a viable option, as they are more complicated to develop in compliance with existing and upcoming legislation [49]. This caution is reiterated in recent reports by the Dutch Data Protection Authority (Autoriteit Persoonsgegevens) and the government's vision on generative AI [6, 40].

The first SLM uses Bag of Words (BoW), a technique that models language via frequency of word(s) combinations [32]. Concretely, for BoW, we use 1-gram and 2-gram to limit the explosion of amount of features, which we implemented with Python package Scikit-learn 0.24.2. The second uses Word2Vec (W2V) [39], a technique that

encodes text fragments in a semantic space. These semantic vector embeddings of words are produced by using a neural network trained on a large document set. For W2V, we modeled a complete sentence by averaging the vectors of the words in the sentence [30] using Python package spaCy 3.1.1. For BoW and W2V, an additional machine learning model is needed to decide whether discrimination is highly suspected. Three state-of-the-art types of machine learning models were used from the Scikit-learn 0.24.2 library [48]: Logistic Regression (LR), XGBoost (XG) and Random Forest (RF) [29].

The third SLM model concerns BERT [21], which stands for Bidirectional Encoder Representation from Transformers. Transformers are a specific type of neural network architecture that can exploit sequentially correlated data, such as written language [54]. Most relevant is that BERT creates contextual embeddings, which means that the same word gets different vectors depending on the context in which it occurs. For BERT, the pretrained Dutch version BERTje [17] was used. A BERT model can apply feature extraction as well as training for a classification task with an integrated logistic like softmax function, called *fine-tuning*. This results in seven methods.

3.2.1 Detection of Known *Explicit Discrimination*

For the detection of known gender discrimination, we compare the performance of methods whilst using a grid search on hyperparameters. We split the data into a train (70%) and a test set, while stratifying on the label (HSD) as well as the search terms. This ensures that the proportion of discriminatory sentences in the train and test set are similar for each known form of gender discrimination. In the train set, we apply a 4-fold cross-validation to optimize a set of hyperparameters over a search space (see Table 5 in Appendix A). For BERTje, we used the recommended hyperparameter search space by the authors of BERT [21]. The best models per the chosen performance metrics (see section 3.3) are selected based on the cross-validation and performance is evaluated on the test set.

We do an additional experiment to analyse the dependency of each model on the amount of available data. For this, we created a learning curve that evaluates performance as a function of the data set size. This is done by evaluating the models with a 10-fold cross validation. For each iteration, the training data set increases by a twenty point logarithmic scale sampled from the nine training folds and tested on the corresponding test fold. The proportion of the labels (HSD, non-HSD) are stratified in the training and test sets for reliable comparison.

3.2.2 Detection of Unknown *Explicit Discrimination*

For the detection of unknown forms of gender discrimination, we set up two experiments. The first experiment focuses on detecting explicit discrimination given that unknown (to the model) explicit terms are used. We do this in a leave-one-search-term-out fashion, i.e. we split the data into n groups according to the n search terms. Then, each group functions in turn as a test set, while the models are trained on the sentences of the remaining $n-1$ groups. For example, the models are trained on sentences with search terms “male”, “woman/women”, etc., but excluding those with the term “female” and are consequently tested on detecting the discriminatory context in sentences with the term “female”. Accordingly, we test the ability of the models to generalise the detection of the difference between a discriminatory and a non-discriminatory context to unforeseen explicit terms.

The second experiment aims to detect explicit discrimination given that no known explicit terms are used. In this set-up, we applied the models on a set of 300,000 job vacancy sentences that do not contain any of the search terms. Subsequently, for each additional model we collect the 100 that have been assigned the largest predicted probabilities to contain explicit discrimination. Finally, for this exploration, the authors labelled the sentences as to

Translated Search Terms	Original Dutch Terms	Frequency	HSD
boy(s)	jongen(s)	997	10.9%
man/men	man(nen)	997	25.6%
male	mannelijk(e)	505	47.1%
lady/ladies	dame(s)	985	18.1%
woman/women	vrouw(en)	993	20.0%
female	vrouwelijk(e)	1059	60.7%
other	overige	191	26.3%
informal	informeel	220	17.7%
Total		5947	28.8%

Table 1. Description of the collected search terms used to create the annotated data set in English and Dutch. The number of sentences per search term are shown as well as the proportion Highly Suspected to be Discriminating (HSD). The category *other* includes: “you are our guy, with balls, enthusiastic boy(s), enthusiastic girl(s), young god(dess)" which translates to “ben jij onze man, met ballen, enthousiaste jongen(s), enthousiaste meid(en), jonge god(in)". The category *informal* includes “dude/guy, gall”, i.e. in Dutch “gozer/kerel/vent, griet"

whether they were highly suspected to be discriminatory. As a result, we explore the models’ abilities to find new potentially discriminatory terms that are not on the current FL.

3.3 Evaluation Metrics

The aim is to gauge the potential of AI and language models to improve on the baseline FL approach. Here, the biggest potential gain is through higher precision, whilst not missing vacancies that may contain explicit discrimination (also known as recall). Therefore, we evaluated our models using the area under the precision-recall curve or Average Precision (AP). This threshold independent metric summarizes all possible combinations of precision and recall values [16]. In addition, we use the Area Under the receiver operating characteristic Curve (ROC): AUC. The ROC shows the trade-off between the false positive rate and the true positive rate. The AUC is another common and robust threshold independent measure to compare machine learning models [37]. For model comparison, AP and AUC were used to select optimal hyperparameters and to evaluate performance on the test set. For hyperparameter optimization, we applied a standard gridsearch procedure.

4 Data Set and Experimental Results

This section presents the results for the two contributions related to the quantitative research to assess the potential efficiency gain of AI and language models: the annotated data and the measured performance of the methods to detect known and unknown forms of gender discrimination.

4.1 Annotated Data

A descriptive overview of the search terms for gender discrimination both in English and Dutch are shown in Table 1. Note that multiple less frequent formulations and search terms are grouped in search terms *other* and *informal*, as is described in the caption. *Boy(s)* has the lowest HSD percentage (10.9%), i.e. baseline precision, while *female*

Vectorizer	Model	AP	AUC
BoW	LR	74,0%	88,5%
	XG	76,8%	88,7%
	RF	72,4%	87,1%
W2V	LR	64,9%	82,4%
	XG	61,8%	79,7%
	RF	56,6%	78,5%
BERT	BERT	**83,3**%	**92,6**%

Table 2. Cross-validation Performance of all models

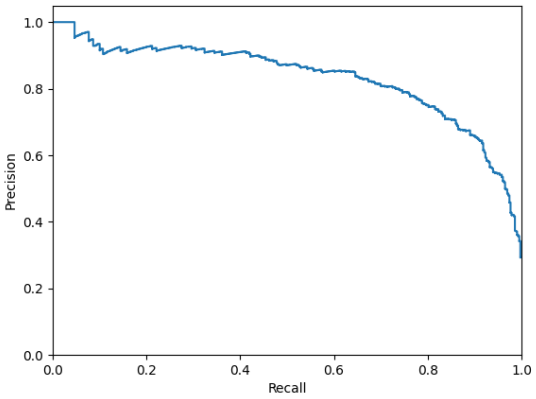
and *male* have the highest: 60.7% and 47.1%. In total, 5947 sentences are annotated of which 28.8% were labelled HSD, which is therefore the precision of the baseline FL method.

Moreover, the Fleiss' Kappa showed that in general there was a moderate agreement between the annotators: Kappa =.557 (95% CI, .565 to .611 and p-value < 0.0005). The agreement was lower for the label concerning "?", Kappa = 0.038, p-value = 0.003, while for the labels of interest: "yes" and "no" the Kappa values were 0.602 and 0.657 with p-value < 0.0005 signalling good agreement. According to guidelines of Altman [2] moderate and good agreement here signals that it is unlikely the annotators agreed by chance. This level of agreement is in line with the subjectivity of interpreting texts to decide on a label as sensitive as discrimination. Given the aim of signalling vacancies with highly *suspected* discrimination, we deem one annotator's expert opinion valuable enough to decide whether a vacancy needs a follow-up check and pool all sentences into one data set. For any sentence in the subset of 600 sentences where the provided labels of the annotators differ, the label is appointed by majority vote. From the whole data set, 51 sentences received the label "?" and were excluded from the data for subsequent experiments. For each search term, this accounted for at most 2% of the sentences. The data set is available on request for research purposes that align with the labour market authorities.

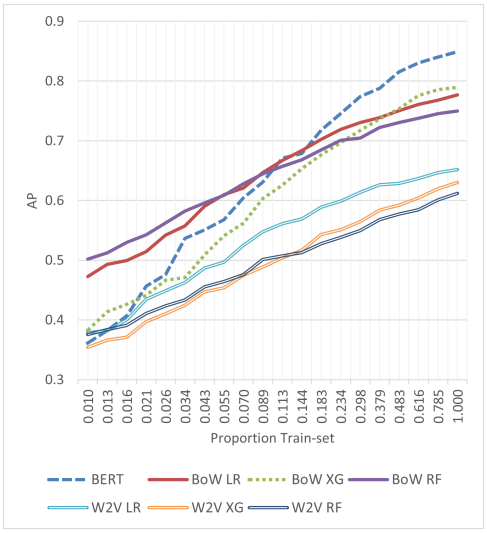
4.2 Detection of *Known* Explicit Discrimination in Context

Table 2 shows the cross-validation performance of the different models. The models based on BoW show high scores above 72% AP. This indicates that just the occurrence of context words in a sentence already has value in predicting whether a gender related term is used in a discriminatory way. W2V models perform worse than BoW with AP values below 65%. BERT stands out with AP = 83.3%, which is 6.5% higher than the best BoW model. The order of the models according to the AUC performance is the same as the AP performance. Due to this similarity, the remaining experiments were conducted with the trained models that were optimized for AP.

To put the performance of the models in perspective, recall that 28.8% of the sentences are HSD such that the FL approach (on this data set) would have that level of precision. Figure 1a shows the possible precision and recall combinations that the BERT model could provide. Assuming that the authority wants high recall of at least 80% or 90%, the suggestions from the model would have precision of 75% or 66%, respectively. In Figure 1b, the learning curve indicates the performance of the methods for varying size of training data set. Most interestingly, we see that BoW XG & BERT benefit most from increasingtraining data. They start with similar performance as the W2V



(a) Precision Recall-curve of the BERT model



(b) Learning-curve of the seven models

Fig. 1. Detection of Known Explicit Discrimination. In (a) the precision recall-curve of the BERT model. In (b) the learning curve of the seven model: the x-axis shows the proportion of the train-set used to train the models and the y-axis the AP value on the test-set. The proportions are in a logarithmic scale from 0 to 1.

models, but around 18% for BERT and 50% for BoW XG they start outperforming BoW LR & RF. Expectedly, the gap between BoW XG & BERT and the other models and also the baseline method will further increase with more data.

4.3 Detection of *Unknown* Explicit Discrimination

Table 3 shows that BERT outperforms the other models in detecting discriminatory context in each scenario where one of the known search terms is treated as an unforeseen potentially discriminating term (leave-one-search-term-out). The AP scores for the adjectives “male” (0.90) and “female” (0.93) stand out. BERT shows great potential in surpassing the FL baseline (with HSD: 47.1% for “male” and 60.7% for “female”) in finding these and similar search terms with high precision, even when not trained with sentences containing those terms.

Furthermore, for our exploration of the methods’ capabilities to detect new forms of discrimination, BERT showed clearly the most potential. BERT’s 100 first suggestions for discriminatory sentences contained 37 sentences that the authors would annotate as HSD. The method BoW XG came second, with 7 sentences out of 100 that were given the label of HSD. All new forms found either concerned occupations that ended with the syllable “man" such as “vakman" (craftsman) or “voorman" (foreman), which contains the same meaning in Dutch as English. Or, they concerned a sentence where explicitly only the female denomination of an occupation was requested, “medewerkster" or “stagiare", the female-only formulation for “employee” and “intern". Noteworthy was also that classifiers (mainly BERT and BoW XG) had picked up on societal gender bias in occupation participation. They

Search Term	BoW XG	W2V RF	BERT
male	0.73	0.67	0.90
man/men	0.59	0.52	0.64
informal	0.56	0.31	0.61
female	0.82	0.78	0.93
lady/ladies	0.47	0.35	0.67
woman/women	0.41	0.39	0.53
boy(s)	0.35	0.31	0.53
other	0.44	0.52	0.67

Table 3. AP scores for detection of unforeseen potentially discriminating search terms. See Table 1 for definitions of *other* and *informal*. Only BoW XG and W2V RF are shown as the configurations of the BoW and W2V models performed similarly.

flagged sentences mentioning occupations that are either dominated by men or by women ("machine operator, plumber, forklift driver, nurse, kindergarten teacher"). More details can be found in Table 6 in Appendix A.

5 Critical Reflection Approach

Incorporating AI and language models within a task for labour market authorities may bring efficiency gains. However, governmental AI use requires caution as their adverse impact has also been demonstrated. Hence, we perform a qualitative study by interviewing experts to create a critical reflection on what needs to be considered beyond efficiency. In this section, we describe how we collect and analyse the input from experts.

5.1 Interviews: Aim, Protocol and Expert Selection Criteria

First, we formulate the key interview question that decides how we structure the interviews and select the experts. To make sure that gathered insights for the critical reflection on potential efficiency gain are actionable, we want to provide a scenario where experts need to compare, make a choice, such that the practical considerations arise. For this reason, we ask the experts to compare SLMs we have studied to the previous FL approach as well as to LLMs. We expect that LLMs, given their hype and potential cannot be ignored when critically reflecting on language models, and therefore are considered as an alternative in this part of the study. Thus, we set to answer the following question in the interviews: *What considerations are relevant for deciding between using the FL, SLMs and LLMs as decision support for discrimination detection by labour market authorities?*

Given the specificity of our research, we are highly selective in inviting experts to share their deliberations. Interviewees are selected based on their expertise or experience with three core components of our use-case: (I) AI language models, (II) discrimination in the labour market domain and (III) use or experimentation of AI by a governmental organisation, preferably a labour market authority. This selection makes finding suitable interviewees challenging, yet set a high bar for the quality of each contribution. We set out to continue reaching out to potential candidates until interviews reached a level of theoretical data saturation. Theoretical data saturation is a concept in qualitative data analysis which is defined as the phase or end-point in which the researcher collects data until no new data appear, theories are well developed and concepts clearly described [1]. We argue that our question requires relatively few interviews to reach data saturation: it is niche and specific, requires experienced experts and

needs depth and time to ask for nuances and reasoning . After a first round of inviting eight experts and conducting six interviews, we had not reached data saturation. So we invited six more experts and, after ten interviews in total, we reached data saturation in February 2025. Signs for saturation were that the new interviews mainly reinforced what had been said before, and that we had sufficient examples to support key findings.

The interview protocol was designed as follows. Each interview starts with introductions and informed consent on the anonymous contribution, recording and transcription of the interview for this research. Then, the setting of the task within labour market authorities is discussed. The approach with FL and SLM is introduced through a slide show with examples and the main results on the potential efficiency gain. Subsequently, a scenario of the embedding is sketched, based on prior interactions and discussions with labour market authorities. In the scenario, the inspector of job vacancies, who currently applies the FL approach to detect explicit discrimination, is envisioned as the user of the AI tooling for decision support. The inspector remains in control and needs to read the suggested job vacancy personally in order to formulate the reasoning for whether or not to reach out to the creator of the vacancy. Reaching out to an organisation is done through an open, inviting approach rather than with blame and a fine. Consequently, with a clear view on the use case, we spend 30 to 45 minutes on the interview question mentioned above.

5.2 Thematic Analysis

We extract the key considerations from the interviews through a thematic analysis (TA), which is a qualitative research method used to identify, analyse and report on the patterns in written text or speech [11, 12]. Applied to our use case, this means that we *code* (label through interpretation) the *data* (transcripts of the interviews), and subsequently decipher *themes* (key considerations) through an analysis on the code patterns across interviews.

There are multiple variants of TA, where coding reliability, codebook and inductive reflexive TA are common [13]. We opt for the latter that focuses on letting the data speak for themselves without categorisation from previous frameworks. We let experts with relevant experience determine priorities and framing of their contribution. Inductive reflexive TA acknowledges the value of subjective interpretation where each coder brings their experiences to their interpretation rather than aiming for homogeneity. Given the presence of two of the authors during the interviews and heterogeneity in the backgrounds of us authors, this could not be avoided regardless. Whilst acknowledging the inherent flexibility and subjectivity, a structured rigorous approach is proposed to distill key insights, which is defined in a six phase process [11]. We demonstrate each step in application to our use case below.

The TA is performed with support from qualitative data analysis software ATLAS.ti (v9.12.1) [5], which functions as an interface and dashboard for a smoother coding and analysing experience. All analysis is still manually performed through the suggested six phase process [11]. We started with familiarising ourselves with the data by reading the transcripts of the interviews and iteratively decided on a coding strategy (1). We generated initial codes by noting keywords and paraphrasing the core of the interviewee's contributions (2). Through collective discussion, we grouped these codes into a first set of twenty seven themes (3). We re-analysed the initial themes to arrive at eight overarching themes and formulated sub-themes with keywords (4). We discussed these eight themes with the co-authors and determined relevant quotes per theme (5). We reported on the seven considerations of the interdisciplinary experts through writing them in this article (6).

6 Critical Reflection Results

The interviewed experts explicitly shared their considerations for choosing between the approaches for discrimination detection by labour market authorities. Through the six phases of the thematic analysis we structured 160 codes from ten interviews into seven themes of critical considerations. These themes show that considerations need to be made beyond efficiency before bringing AI and language models into practice. We note specific insights for FL, SLM or LLM in each theme whenever experts differentiated their views based on these approaches.

Set-up, Deployment and Environmental Costs

Most experts explicitly mentioned costs for set-up and implementation of AI and language models as a main consideration. For SLMs, the costs were front-loaded, which was attributed to the need for "staff with data-science and domain expertise for development as well as the manual data annotation process". For LLMs, multiple experts argued that the set-up costs may be smaller as "LLMs such as ChatGPT are more intuitive and do not require coding to use". However, other experts added that investment in terms of robust IT infrastructure and expertise are required to fine-tune such models to the specific use case and to put the necessary guardrails into place for responsible use. Regarding deployment, experts shared that costs for LLMs were expected to be higher. They emphasized the higher costs for usage or hosting as a result of the higher computational power as well as the legal costs to comply with the still new and upcoming regulation on LLMs. Five experts also mentioned the costs to the environment of which the burden is often not included in the price and predominantly experienced by the Global South.

Biased Models and Harmful Effects

All experts raised concerns about biased language models and harmful effects of using language models. Societal stereotypes are embedded in the way we speak and write and thus also end up in language models. Moreover, especially for LLMs, production and training data are skewed to the United States, with DeepSeek (China) and Mistral (France) as exceptions. Experts noted that discrimination detection in Dutch vacancies requires "cultural, geographical nuances" and "knowledge of the Dutch labour market", that may not be captured by these models.

In terms of harmful impact of bias from AI, experts mentioned the risk for "a tunnel vision on certain sectors, jobs or even specific companies", or: a hyper-focus "on certain types of discrimination, while other remain overlooked". In one interview, it was gathered that biased societal connotations of wording may be useful as input for the language models in order to recognize discriminatory context in job vacancies. Bias concerns were also present for the FL approach. One interviewee emphasized that inspectors, as any human, are prone to systemic societal biases. Moreover, they may have personal biases towards a sector or company. Overarching to these considerations, multiple experts noted that the perceived harm (from bias) in our task was small due to the inspector making the final decision and the informative (instead of punitive) goal of the discrimination detection. The experience of being approached with an intention to offer support (aimed at preventing the exclusion of people at this initial stage of recruitment) was expected to differ from being approached with the stricter intention of issuing fines.

Transparent and Accountable Practices

Transparency and accountability were considerations mentioned in all conversations. For SLM and LLM approaches, transparency about the use of technology in itself was considered crucial, where The Algorithm Register [47] was given as an example. Next to that, the lack of transparency in the content of LLM training data was highlighted in

many interviews, as well as the fact that so much of their data are captured without consent. Moreover, in terms of accountability, one expert noted that all language models remain stochastic models, whilst SLMs may score better due to their reproducibility. This was attributed to the fact that LLMs are continuously updated and sometimes taken offline, see GEITje-7B not too long ago. GEITje-7B is a Dutch Large Language model based on Mistral 7B, which (without clear documentation to verify training data) was deemed unlikely to be solely trained on exclusively copyright-free material [50]. Therefore, with SLMs, it is easier to have reproducible output, and running them on-premise allows issues raised much later to be consulted upon. Interviewees commended that, no matter the used technology, it was humans deciding and providing reasoning to these decisions when organisations are approached based on their job vacancy. Two interviewees highlighted that accountability also includes active adaptation for any issues raised. Here, FL and SLM were seen as preferred as one can add a regular expression or alter the training data concerning the error. Adapting a prompt for an LLM may be easy, but cannot guarantee a correct (continual) interpretation of the prompt and thus cannot guarantee to prevent similar issues.

Deterioration of Critical Thinking

All experts independently expressed that deterioration of critical thinking should be an active consideration when AI and language models may be applied, and especially by governmental bodies including labour market authorities. Interviewees recounted their own experiences of AI adoption. The technology was either immediately over-relied upon or people were very critical, with any error debunking the total utility of the system. Applied to our use case, most interviewees expected a risk of deterioration of the user's critical interpretation of discriminatory context after a few experiences where the AI correctly prioritised job vacancies that are highly suspected to be discriminatory. That is, they emphasized "a risk for automation bias may kick in and a state of autopilot may be reached".

This risk was especially attributed to LLMs, due to the ease of using them, the hype and the fact you receive answers in human language. Consideration and action are needed to prevent this from happening or to choose not to use it at all. The interviewees mentioned increasing AI literacy as well as adaptation of the user interface of the tooling to stimulate critical reflection. Examples include inviting users to challenge suggestions of the AI and selecting sentences for review that were not flagged by the AI.

Ease of Adoption

Eight out of the ten interviewees mentioned that the ease of adoption of the proposed AI and language models is decisive for whether they will improve the current way of working. Experts shared their experiences where enthusiasm, a straightforward interface and clear communication on motivation were prerequisites for innovation adoption. For LLMs, an interviewee highlighted that "guardrails to restrict use to intended goals are required. Otherwise, the risk of a slippery slope to also use it for other work activities may lead to accidents, where sensitive and/or personal information may be breached." Another aspect of adoption mentioned was that alternatives of SLM and LLM may bring about the datafication of the inspector's job. With the promised increase in performance of AI, an expert stated that: "higher targets and pressure may cause also the productivity of inspectors to become more scrutinized. Too much monitoring and focus on quantity over quality may reduce any enjoyment of the job."

Resilience to Big and New Tech

Resilience to dependencies on big tech and future trends were also frequently mentioned as considerations. Five of the ten interviewees raised that using LLMs in public services of any governmental body creates an undesirable dependency on big tech. Experts elaborated that this so-called *vendor-lock-in* was especially attributed to the infrastructure that companies such as OpenAI provide. Moreover, two experts on LLMs stated that competitors of Big Tech are disadvantaged due to the capital-intensive technology, computational resources and manual content moderation. SLMs were therefore regarded as more opportunistic from a resilience angle due to the open-source language models and classification methods. However, two experts expected that the hype and increasing functionalities of LLMs, will lead to a decreased attention for SLMs. This would challenge the long term sustainability of the investment into SLMs. One interviewee even stated: “It [big tech dependency in large language models] is inevitable, it is a lost race, we need to focus on how we cope with it."

The Underlying Goal

Overarching to all the six above considerations, eight of the ten experts emphasised that choices based on these considerations can only be made in the context of a clearly defined end-goal. Although we shared the motivations from the consulted authorities UWV, NIHR and NLA and the potential embedding of the AI and language models in the use case, experts still raised critical questions to dig deeper on this motivation and embedding. This was best summarized by the following question: *What is underlying the goal of explicit discrimination detection?*

If the goal is to address discrimination in recruitment and work, the amount of resources and efforts to focus on job vacancies may be disproportional. Interviewees noted that there is more evidence on discrimination based on foreign-sounding names in resumes or experienced discrimination in the workplace. One interviewee said: “Always critically reflect whether an initiative directly helps people with distance to the labour market”. On the other hand, if the goal is to prevent any discriminatory vacancies on publicly hosted “werk.nl” of the UWV, the use case may serve a more direct purpose to fulfil their duty of hosting job vacancies that do not explicitly discriminate.

In other words, specifying the underlying goal helps to decide whether investing in the challenge to responsibly use AI and language models is proportional to what you gain. Experts agreed that, based on the information available, SLMs can bring substantial efficiency gain and should be prioritised over LLMs: “Why use something large and complex, when something simple works?” and “You are more independent, it is cheaper, better for the environment and you don’t always need the newest technology to make impact". Marginal benefits of LLMs were considered improbable to be worth the additional challenges or investment.

7 Discussion

This section discusses overarching insights, recommendations and limitations of our research.

To start, the challenging complexity of discrimination was noted during the process of choosing the classification label. If the legal nature of discrimination is so complex, what role does AI have to influence decision making of inspectors? We tackled this complexity by choosing to label vacancies for “highly suspected to be discriminatory” (HSD). This limits the influence of AI and leaves inspectors to decide which organisations to approach and formulate the reasoning to do so. This aligns with the call of [18] for actionability, as it focuses NLP efforts on what they can bring to support actions against discrimination.

Theme	Key Insight
Set-up, Deployment and Environmental Costs	While SLMs have higher initial costs, LLMs incur greater ongoing costs in usage, compliance, and environmental impact.
Biased Models and Harmful Effects	Bias risks involves nuances and tunnel vision, whereas harm depends more on context and embedding than model size.
Transparent and Accountable Practices	Reproducibility and adaptability to errors are crucial to maintain transparency and accountability over time.
Deterioration of Critical Thinking	AI should not erode users' critical thinking, where the user interface and AI literacy are potential factors.
Ease of Adoption	User enthusiasm, protections against misuse as well as over-monitoring of users are key to adoption.
Resilience to Big and New Tech	Investments should consider sustainability and future technological shifts, including the risk for vendor lock-in.
The Underlying Goal	The model's purpose must be justified and simpler solutions may already fulfill the task with fewer risks.

Table 4. Overview of Themes and Key Insights

Similarly, the complexity of the subjective nature of discrimination was noted during the creation of the data set and was reflected in the moderate to good inter-annotator agreement. Since annotators occasionally diverged in their judgements on HSD, AI cannot be expected to achieve perfect precision for all perspectives. This places an upper bound on the level of precision an inspector can experience when using such an AI-tooling. We addressed this issue by reframing it around user needs: is the goal to achieve a substantial improvement over current FL approaches using smaller, simpler language models, or to aim for near-flawless precision with complex, larger models? The interviewed experts were clear in their view: if it can be done with simpler models, why bring in the risks of bigger and more complex ones. This perspective aligns with the call to question the need for ever-larger models [8].

Furthermore, the findings of the unknown discrimination detection experiments suggested that the relatively more complex and larger language model, BERT, has most potential to generalise discriminatory contexts and find new forms of explicit discrimination. This may hint that LLMs could provide even better results. However, the critical reflection showed the importance of the embedding and the goal of such models. A future research direction may examine a responsible use of LLMs to periodically find new forms of explicit discrimination to add to a FL. As such, labour market authorities can reduce the influence of LLMs to the current way of working and choose to use alternatives with more human control, FL or SLM, to give suggestions to inspectors.

In terms of limitations, we acknowledge that the estimated efficiency gain achieved in this work, is a lower bound. The estimates of AP and AUC may be improved with different approaches. For example, the low performance of W2V can be improved through other aggregation methods such as weighting the word vectors with their tf-idf values [38] or doc2vec [35]. Performance may also improve with a model-per-searchterm approach. An error analysis could help identify which synthetic examples can be added to help models classify cases where the discriminatory context is nuanced. Furthermore, our findings suggest that SLMs can already provide a substantial

efficiency gain for labour market authorities and that therefore LLMs are not worth the investment. Further research that also compares the performance of LLMs could make this finding more robust.

Next to that, we recognize that we missed out on relevant voices due to stakeholder involvement being limited to labour market authorities and the strict interview criteria. Future research is advised to include civil society organisations representing those impacted by labour market discrimination. Due to the lack of these voices, we have not sufficiently validated that the only known gender discrimination in job vacancies is related to binary genders. Involving LGBTQIA+ rights groups could have ensured that their experienced discrimination were considered within our investigation of gender discrimination. See [51], for a practical approach to centre civil society in AI development.

8 Conclusions

In this paper, we have investigated an opportunity for labour market authorities to use AI and language models for improving their supervision on explicit discrimination in job vacancies. We took on this research by centring the needs and expertise of labour market authorities by co-creating an annotated data set. Based on this data set, we found that even smaller language models (especially BERT and bag-of-words with XGBoost) could be an upgrade relative to the baseline approach of using a forbidden list of potentially discriminating terms. These smaller language models increase the precision and allow for the exploration of new discriminatory formulations.

Critically reflecting with experts on the use case, positioned this opportunity within considerations beyond efficiency gain. The experts highlighted for example the importance of accountability and transparency of labour market authorities. They also emphasised that use of any innovation should hone, rather than deteriorate, a skill such as critical thinking. Even when AI and language models bring meaningful efficiency gains for the common good, we need to remain critical. In the case of detecting discrimination, we must centre the question of whether developing and using this technology reflects the actual needs of those facing discrimination in the labour market.

Positionality Statement We are a research team combining expertise in the labour market, language, statistics, and AI. For most of this project, we were based in the Netherlands and share academic backgrounds. Although some of us have experienced discrimination in recruitment, we acknowledge our privileged position in society that has shaped both our insights and our blind spots. By collaborating with institutions and consulting diverse experts, we aimed to broaden our perspectives, skills, and knowledge. We remain open for critical feedback.

Acknowledgments

This research was funded through three projects: Skills Matching, Vaardig met Vaardigheden; part of the National Growth Fund funded by the Dutch Ministry of Education, Culture and Science, and DIVERSIFAIR. The DIVERSIFAIR project is co-funded by the European Union through Erasmus+ grant (101107969). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or European Educative and Culture Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. We sincerely appreciate our close collaborators within Skills Matching: the Netherlands Labour Authority, the Netherlands Employment Insurance Agency (UWV) and the Netherlands Institute for Human Rights. We are especially grateful to the experts who contributed to the annotation process and the experts who participated in the interviews. Finally, we thank the anonymous reviewers for their valuable feedback.

References

[1] Khaldoun M. Aldiabat and Carole L. Le Navenec. 2018. Data Saturation: The Mysterious Step in Grounded Theory Methodology. *The Qualitative Report* 23, 1 (2018), 245–261. https://doi.org/10.46743/2160-3715/2018.2994

[2] Douglas G Altman. 1990. *Practical statistics for medical research*. CRC press.

[3] Amnesty International. 2020. Xenophobic Machines: Discrimination through Unregulated Use of Algorithms in the Dutch Welfare System. https://www.amnesty.org/en/documents/eur35/4686/2021/en/ Accessed: 2025-01-10.

[4] Amnesty International. 2023. Unchecked Power: ID Checks and Collection of Data from Peaceful Protesters in the Netherlands. https://www.amnesty.org/en/documents/eur35/6650/2023/en/ Accessed: 2025-03-14.

[5] ATLAS.ti. 2025. *ATLAS.ti: Qualitative Data Analysis Software*. https://atlasti.com Web Version v9.12.1-2025-03-06.

[6] Autoriteit Persoonsgegevens | Department for the Coordination of Algorithmic Oversight (DCA). 2025. AI & Algorithmic Risks Report Netherlands: Winter 2024/2025. https://autoriteitpersoonsgegevens.nl/en/documents/ai-algorithmic-risks-report-netherlands-arr-february-2025 Accessed: 2025-03-07.

[7] Siti Awaliyah. 2016. The effectiveness of anti-discrimination laws for job seekers in Indonesia. *Journal of Law, Policy & Globalization* 51 (2016), 41.

[8] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. ACM, 610–623. https://doi.org/10.1145/3442188.3445922

[9] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 5454–5476. https://doi.org/10.18653/v1/2020.acl-main.485

[10] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Eds.), Vol. 29. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf

[11] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[12] Virginia Braun and Victoria Clarke. 2022. *Thematic analysis: A practical guide*. Sage Publications, Los Angeles, USA.

[13] Virginia Braun and Victoria Clarke. 2024. Thematic analysis. In *Encyclopedia of quality of life and well-being research*. Springer, 7187–7193.

[14] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186.

[15] Caitlin Chandler, Andrew Couts, Crofton Black, Ariadne Papagapitos, and Daniel Howden. 2025. Computer Says No Fly. https://www.lighthousereports.com/investigation/computer-says-no-fly/ Accessed: 2025-03-14.

[16] Jesse Davis and Mark Goadrich. 2006. The Relationship between Precision-Recall and ROC Curves. In *Proceedings of the 23rd International Conference on Machine Learning* (Pittsburgh, Pennsylvania, USA) *(ICML '06)*. Association for Computing Machinery, New York, NY, USA, 233–240. https://doi.org/10.1145/1143844.1143874

[17] Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. Bertje: A dutch bert model. *arXiv preprint arXiv:1912.09582* (2019).

[18] Pieter Delobelle, Giuseppe Attanasio, Debora Nozza, Su Lin Blodgett, and Zeerak Talat. 2024. Metrics for What, Metrics for Whom: Assessing Actionability of Bias Evaluation Metrics in NLP. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 21669–21691. https://doi.org/10.18653/v1/2024.emnlp-main.1207

[19] Pieter Delobelle, Ewoenam Kwaku Tokpo, Toon Calders, and Bettina Berendt. 2021. Measuring Fairness with Biased Rulers: A Survey on Quantifying Biases in Pretrained Language Models. *CoRR* abs/2112.07447 (2021). arXiv:2112.07447 https://arxiv.org/abs/2112.07447

[20] Develop Diverse. 2025. Develop Diverse: Inclusive Writing for Diverse Workplaces. https://developdiverse.com Accessed: 2025-03-12.

[21] 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). Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[22] ENNHRI. 2025. Netherlands - ENNHRI Member Information. https://ennhri.org/our-members/netherlands/ Accessed: 2025-03-12.

[23] Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76, 5 (1971), 378.

[24] Antske Fokkens and Camiel Beukeboom. 2020. Leeftijdsdiscriminatie in vacatureteksten: Een herhaalde geautomatiseerde inhoudsanalyse naar verboden leeftijd-gerelateerd taalgebruik in vacatureteksten uit 2017 en 2019. [Rapport in opdracht van het College voor de Rechten van de Mens.]. https://publicaties.mensenrechten.nl/file/61ac0049-4137-32d8-549c-efcec2c68fdb.pdf Accessed at 19-05-2025.

[25] Antske S. Fokkens, Camiel J. Beukeboom, and Isa Maks. 2018. Leeftijdsdiscriminatie in vacatureteksten: Een geautomatiseerde inhoudsanalyse naar verboden leeftijd-gerelateerd taalgebruik in vacatureteksten: Rapport in opdracht van het College voor de Rechten van de Mens. https://publicaties.mensenrechten.nl/file/bd556622-9cad-4416-acd9-58258a520ab4.pdf Accessed at 19-05-2025.

[26] Vinitha Gadiraju, Shaun Kane, Sunipa Dev, Alex Taylor, Ding Wang, Remi Denton, and Robin Brewer. 2023. "I wouldn't say offensive but...": Disability-Centered Perspectives on Large Language Models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 205–216. https://doi.org/10.1145/3593013.3593989

[27] Ismael Garrido-Muñoz, Arturo Montejo-Ráez, Fernando Martínez-Santiago, and L Alfonso Ureña-López. 2021. A Survey on Bias in Deep NLP. *Applied Sciences* 11, 7 (2021), 3184.

[28] Danielle Gaucher, Justin Friesen, and Aaron C Kay. 2011. Evidence that gendered wording in job advertisements exists and sustains gender inequality. *Journal of personality and social psychology* 101, 1 (2011), 109.

[29] Aurélien Géron. 2019. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems* (2nd ed.). O'Reilly Media.

[30] Sam Henry, Clint Cuffy, and Bridget T McInnes. 2018. Vector representations of multi-word terms for semantic relatedness. *Journal of biomedical informatics* 77 (2018), 111–119.

[31] ISZW. 2020. Leeftijdsdiscriminatie in vacatureteksten (Age discrimination in job vacancy texts). https://www.rijksoverheid.nl/documenten/rapporten/2020/04/23/leeftijdsdiscriminatie-in-vacatureteksten.

[32] Thorsten Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. In *European Conference on Machine Learning*. 137–142.

[33] Peter Kuhn and Kailing Shen. 2013. Gender discrimination in job ads: Evidence from china. *The Quarterly Journal of Economics* 128, 1 (2013), 287–336.

[34] Peter Kuhn, Kailing Shen, and Shuo Zhang. 2020. Gender-targeted job ads in the recruitment process: Facts from a Chinese job board. *Journal of Development Economics* 147 (2020), 102531. https://doi.org/10.1016/j.jdeveco.2020.102531

[35] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*. PMLR, 1188–1196.

[36] Qian Li, Hao Peng, Jianxin Li, Congying Xia, Renyu Yang, Lichao Sun, Philip S Yu, and Lifang He. 2020. A survey on text classification: From shallow to deep learning. *arXiv preprint arXiv:2008.00364* (2020).

[37] Charles X Ling, Jin Huang, Harry Zhang, et al. 2003. AUC: a statistically consistent and more discriminating measure than accuracy. In *Ijcai*, Vol. 3. 519–524.

[38] Cai-zhi Liu, Yan-xiu Sheng, Zhi-qiang Wei, and Yong-Quan Yang. 2018. Research of text classification based on improved TF-IDF algorithm. In *2018 IEEE International Conference of Intelligent Robotic and Control Engineering (IRCE)*. IEEE, 218–222.

[39] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Adv. in neural information processing systems*. 3111–3119.

[40] Ministry of the Interior and Kingdom Relations. 2024. Government-wide Vision on Generative AI of the Netherlands. https://www.government.nl/documents/parliamentary-documents/2024/01/17/government-wide-vision-on-generative-ai-of-the-netherlands Accessed: 2025-03-07.

[41] Nederlandse Arbeidsinspectie. 2021. Discriminatie via online vacatureplatforms. https://www.nlarbeidsinspectie.nl/actueel/nieuws/2021/01/22/discriminatie-via-online-vacatureplatforms Accessed: 2025-03-12.

[42] Netherlands Labour Authority. 2025. Netherlands Labour Authority - Home Page. https://www.nllabourauthority.nl The Netherlands Labour Authority works for fair, healthy, and safe working conditions and socio-economic security for everyone. Accessed: 2025-03-12.

[43] Netherlands Labour Authority - Ministry of Social Affairs and Employment. 2025. Onze werkwijze. https://www.nlarbeidsinspectie.nl/werken-bij/over-ons/onze-werkwijze Accessed: 2025-03-12.

[44] Panggih Kusuma Ningrum, Tatdow Pansombut, and Attachai Ueranantasun. 2020. Text mining of online job advertisements to identify direct discrimination during job hunting process: A case study in Indonesia. *Plos one* 15, 6 (2020), e0233746. https://doi.org/10.1371/journal.pone.0233746

[45] NL Digital Government. 2025. Value-Driven Digitalisation Work Agenda. https://www.nldigitalgovernment.nl/overview/digitalisation-policy/value-driven-digitalisation-work-agenda/ Accessed: 2025-03-12.

[46] Netherlands Court of Audit. 2024. Focus on AI in Central Government. https://english.rekenkamer.nl/publications/reports/2024/10/16/focus-on-ai-in-central-government Accessed: 2025-03-14.

[47] Algoritme Register Ministry of the Interior and Kingdom Relations (BZK). 2025. The Algorithm Register of the Dutch Government. https://algoritmes.overheid.nl/en Accessed: 2025-03-12.

[48] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[49] Rijksoverheid. 2023. Kamerbrief over voorlopig standpunt voor rijksorganisaties bij het gebruik van generatieve AI. https://www.rijksoverheid.nl/documenten/kamerstukken/2023/12/11/kamerbrief-over-voorlopig-standpunt-voor-rijksorganisaties-bij-het-gebruik-van-generatieve-ai Accessed: 2025-03-12.

[50] Stichting BREIN. 2025. BREIN Takes Down a Large Language Model for the Second Time in as Many Weeks. https://stichtingbrein.nl/brein-takes-down-a-large-language-model-for-the-second-time-in-as-many-weeks/ Accessed: 2025-03-12.

[51] Harini Suresh, Rajiv Movva, Amelia Lee Doğan, Rahul Bhargava, Isadora Cruxén, Ángeles Martinez Cuba, Giulia Taurino, Wonyoung So, and Catherine D'Ignazio. 2022. Towards Intersectional Feminist and Participatory ML: A Case Study in Supporting Feminicide Counterdata Collection. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1010–1022. https://doi.org/10.1145/3531146.3533132

[52] Textio. 2025. Textio: Augmented Writing for Better Hiring and Communication. https://textio.com Accessed: 2025-03-12.

[53] UWV. 2025. Employee Insurance Agency (UWV). https://www.uwv.nl/en/about-uwv Accessed: 2025-05-19.

[54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[55] Werk.nl. 2025. Discriminatie en ongelijke behandeling in vacatures. https://www.werk.nl/werkgevers/over/spelregels/discriminatie/index.aspx Accessed: 2025-03-12.

[56] Overheid.nl Wettenbank. 2020. Algemene Wet Gelijke Behandeling (Dutch Equal Treatment Act). https://wetten.overheid.nl/BWBR0006502/2020-01-01. Accessed: 2022-01-13.

[57] Zara Whysall. 2018. Cognitive biases in recruitment, selection, and promotion: The risk of subconscious discrimination. *Hidden inequalities in the workplace* (2018), 215–243.

A Additional Figures and Tables

Model	HP Search Space	Best BoW HP	Best W2V HP
Logistic Regression	C: (0.01, 0.05, 0.1 , 0.5, 1, 5, 10, 50, 100) Solver: (lbfgs, liblinear) Penelty: l2 Class weight: balanced	C: 0.5 Solver: liblinear Penelty: l2 Class weight: balanced	C: 0.1 Solver: lbfgs Penelty: l2 Class weight: balanced
XGBoost	Min child weight: (2, 5, 10) Learning rate: (0.3, 0.2, 0.1, 0.05, 0.01, 0.005) Max depth: (10, 50, 100) Scale positive weight: 2.5	Min child weight: 2 Learning rate: 0.3 Max depth: 10 Scale positive weight: 2.5	Min child weight: 2 Learning rate: 0.2 Max depth: 50 Scale positive weight: 2.5
Random Forest	Max depth: (10, 50, 100) Number of estimators: (200, 600, 1000, 1400, 2000) Minimum samples split: (2, 5, 10, 50) Class weight: balanced	Max depth: 100 Number of estimators: 200 Minimum samples split: 5 Class weight: balanced	Max depth: 50 Number of estimators: 1400 Minimum samples split: 2 Class weight: balanced
BERT	Learning rate: (2e-5, 3e-5, 5e-5) Epoch amount: (2, 3, 4) Batch size: (16, 32)	**Best HP:** Learning rate: 3e-5 Epoch amount: 3 Batch size: 16	

Table 5. The hyperparameter (HP) search space for LR, XB, RF and BERT models together with their optimal values according to the AP metric. For LR, XB and RF the hyperparameters are determined for both BoW and W2V vectorizers. The default values were chosen for all other hyperparameters that are not present in this table.

Method	HSD	M	F	Examples
BoW LR	6	5	1	carpenter* (M)
BoW RF	2	1	1	craftsman (M)
BoW XG	7	2	5	saleswoman (F)
W2V LR	0	0	0	-
W2V RF	3	3	0	foreman (M)
W2V XG	1	0	1	intern* (F)
BERT	37	25	9	employee* (F)

Table 6. Capability of finding unknown forms of discrimination. The count of sentences which are highly suspected to be discriminatory (HSD) are further categorized by either being an occupation ending with “man” (M) or that only the female format of the occupation was requested (F). An asterisk (*) denotes that the term’s discriminatory connotation is lost in translation.