Universiteit Leiden
The Netherlands

Opleiding Informatica

Detection of indirect gender bias

in job vacancy texts

Guido de Vries

Supervisors:
Cor Veenman & Ajaya Adhikari

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

19/08/2022

Abstract

Discrimination comes in many forms and is present in many places in modern society, even though it is agreed upon that discrimination is damaging to society. One such place is job vacancies and job interviews, where candidates may have a reduced chance simply based on their ethnicity, sexuality, gender or other discriminatory factors. This research aims to explore one of these aspects, namely gender discrimination in job vacancies. The research will try to help identify key aspects in job vacancies that could be indicative of gender discrimination. Previous research has already focused on if gendered keywords like ’female’ or ’man’ are being used in a discriminatory context using machine learning. ’Female’ could for example indicate the gender of the job applicant, but also a different person who is part of the job. In this paper, the focus is instead placed on indirect gender discrimination in direct bias context. Job vacancies which are identified as highly suspected discriminatory based on gender are combined into a data set with other job vacancies which are chosen based on certain keywords or phrases. The goal is to correctly identify which job vacancies are highly suspected discriminatory with regards to gender when a certain sentence which is used to mark the job vacancy as discriminatory, has been left out. This means job vacancies will be labeled based on the left-over text. The paper will demonstrate that although indirect context does contain some identifiable hints that are indicative of gender bias, the current approach is not sufficient for practical use in industries. The research is conducted in collaboration with TNO, the Netherlands Organisation for Applied Scientific Research, as part of larger research that explores discrimination in job vacancies.

Contents

1 Introduction **1**

2 Previous work **2**

3 Methods **2**

3.1 Models 2

3.2 Evaluation metrics 3

3.3 The data-set 4

3.4 Reproduction & further training 4

3.5 Model construction 5

3.6 Model training 5

4 Results **6**

4.1 Logistic Regression coefficients in different contexts 7

4.2 Feature importance: Random Forest 8

4.3 Gendered nouns in combination 12

5 Conclusion **12**

6 Discussion **14**

6.1 The word-2-vec approach 14

6.2 Hyper-parameter optimization 14

6.3 Data reproduction 14

6.4 Pairs of opposite genders 14

6.5 Data set completeness 15

6.6 Duplicate data 15

1 Introduction

Gender discrimination is apparent in many places of modern society. Even though dozens of countries have laws in place to reduce or eliminate discrimination [2] it still occurs frequently. One such place where gender discrimination is still apparent is within companies. In 2018, one of the largest video-game companies Riot Games faced a lawsuit for mistreating its female workers. The company was ordered to pay out 100 million dollars to its female employees as a result [8]. According to video-game website Kotaku, many female employees at Riot Games, which only made up 21% of its employees [5], described the company as a place where women are treated unfairly [10]. Research suggests [7] that having more gender diversity reduces workplace harassment. On top of the negative culture a non-diverse workplace manifests, similar legal consequences may also hinder the growth of a business.

As such, it becomes important to hire more diverse candidates for jobs in order to reduce potential problems. One way of achieving this is by writing inclusive job descriptions and vacancies. However, these may not always be written by professionals or they may be written by people who unconsciously discriminate. If these texts can be identified as discriminatory, proper action can be taken to reduce discrimination. In the US for example, about 4% of men and about 7% of women have experienced being turned down for a job based on their gender in 2017 [17]. Job vacancies may also contain language which is unappealing towards a certain gender. This will in turn reduce the diversity of the potential hiring pool.

On top of that, recent laws are increasing the pressure on companies to hire more diverse candidates. In 2021, the Dutch government introduced a new quotum for listed companies that supervisory boards must be made up for at least one-third by men and for one one-third by women [1]. Non-discriminatory job listings can help negate the need for companies to seek out men or women specifically to satisfy different laws and they may encourage a healthier work relationship that boosts productivity.

The aim of this research is to analyze job vacancies in the Netherlands and to correctly classify them as either non-discriminatory or highly suspected discriminatory. By using computer models that are able to precisely classify job vacancies, new insights may be found into what writing styles and terminology may contribute to gender discriminatory job vacancies. With these insights, better vacancies can be written to improve diversity in hirings. This research will be focusing on indirect context specifically. This means that job vacancies which have been hand-classified as highly suspected discriminatory with regards to gender will have the sentences that were indicative of the discrimination filtered out. The remaining vacancy is then used for classification.

The main question this research will aim to answer is if we can train an algorithm to detect gender discrimination in job vacancies when the sentences used to hand-mark the vacancies as suspected gender discriminatory or not, are filtered out of the vacancy during training. This will be done by first replicating results from similar research. Then, using modern models such as a bag-of-words model and a word-2-vec model, we will show that these algorithms can in fact be used to train models. However, we will also show that for practical uses, more elaborate approaches may be needed to improve performance. This is because current approaches yield results which may are likely insufficient to automate practical tasks. Different approaches for potentially improving performance will be elaborated on in the discussion. There we will also show what factors may be hindering the learning process of the algorithms.

2 Previous work

The paper ”Context-Aware Discrimination Detection in Job Vacancies using Computational Language Models” [20] has previously already explored the detection of gender bias in Dutch job vacancies. In that paper, different vectorizers and models were combined to classify job vacancies. The research looked at a direct context, meaning individual sentences which were indicative of being potentially discriminatory based on certain keywords were analyzed. They were labeled as either ’most likely discriminatory’ or ’non-discriminatory’. The paper shows that machine learning can be used effectively in detecting explicit discrimination in job postings. However, the research did not thoroughly research implicit discrimination which may be hidden job vacancies. This paper will focus on that specifically, by analyzing indirect contexts.
In the paper ”Text mining of online job advertisements to identify direct discrimination during job hunting process: A case study in Indonesia” [16] the occurrence of discrimination - of different types - in job vacancies, including gender discrimination, is also explored. Their approach to classification is through applying word pattern templates, which replaces words with tags that indicate the type of word. Classification then occurs by applying predetermined queries to the word patterns. Their approach, too, uses a dictionary of keywords for detecting discrimination. As a result, subtle or inconspicuous phrases that indicate discrimination will remain unexplored.

3 Methods

This research will build on top of previous findings from the paper ”Context-Aware Discrimination Detection in Job Vacancies using Computational Language Models” [20] by analyzing the indirect context of Dutch job vacancies; The residue of each vacancy after filtering out direct context will be used by models for classification. The *Models* section goes over the different models and vectorizers used in the research. *Evaluation metrics* provides an overview of the metrics used to measure the performance of the models. The *Model Construction* section shows different pre-processing steps and settings that were used to help the models. And the *Model training* section mentions the used cross-validation set-up. The software and packages used in this project includes:

- Python version 3.8.6
- Scikit-learn [18] version 1.0.2
- Spacy [14] version 3.2.4 with pipeline *nl_core_news_lg* (3.2.0)

3.1 Models

To classify job vacancies, a combination of different vectorizers and models are used. The vectorizers are used to transform an input text into a machine-readable list of data of a fixed length. How these vectorizers transform data into an input matrix depends on the type of vectorizer. The algorithms are then used to process this data and output a label, either ’0’ or ’1’. The two used vectorizers are a ’bag-of-words’ (BoW) model and a ’word-2-vec’ (W2V) model. The bag-of-words model counts the number of occurrences of each word in a list of job vacancies and compiles them into a matrix to feed into a model. So each number in the matrix represents a word and its value is the number of times the word occurs inside a job vacancy. In the word-2-vec model, each word is represented as a vector with hundreds of dimensions. Words with a similar meaning have vectors that point in a similar direction. The distance between two word vectors indicates how closely the two words are related, with a smaller magnitude being a semantically similar. In this research, the vectors of each word from an input text are averaged and the resulting vector is used as an input for the models. Word-2-vec models are pre-trained, meaning the vectors of each word are defined beforehand by analysis of real-life texts.

The three used models to train the input data on are the Logistic Regression (LR) model, the Random Forest (RF) model and the Gradient Boosting (GB) model.

Logistic Regression works by modelling an S-shaped curve, where input values in the equation are combined with weights, also called coefficients [9]. The output of the model is a probability, where values above 0.5 are associated with the label '1' and other values are associated with the label '0'. Coefficients that are assigned to terms can be either positive or negative and their signs indicate the direction of its influence. Their magnitudes indicate how important the feature is in the classification task.

Random Forest works by combining many different decision trees into a 'forest' of classifiers. Each tree outputs their own prediction for a given classification task. The Random Forest model then outputs the majority vote class among the trees as its final decision [21]. To get a better understanding of the Random Forest decisions, their feature weights can be exposed, which uses a Gini importance. The Gini importance is defined as "the total decrease in node impurity averaged over all trees of the ensemble" [15].

Gradient Boosting works through gradient descent. It is an iterative process which tries to minimize prediction errors by slightly changing the previous iteration of the model [13]. The smaller the prediction error, the smaller the next changes have to be, as the model is expected to be closer to its optimum.

3.2 Evaluation metrics

The performance of a model is evaluated by using the 'average precision' (AP) metric. The average precision is a metric which combines the precision and recall metrics [23]. Precision is the ratio of entries that are labeled '1' in the data set that are also correctly labeled '1' by a model. Recall is the ratio of entries correctly labeled '1' by a model out of all entries that are labeled '1' in the data set. If a model optimizes the precision, its recall will go down as the model will be more likely to produce false negatives. Likewise, by improving the recall of a model, its precision will go down as the model will output more false positives [12]. The average precision therefore acts as a balanced middle-ground by optimizing the area under the precision-recall curve (AUC), which is a curve displaying how precision and recall compare at different decision thresholds.

$$AP = \sum_{i=1}^{n} P_i \cdot (R_i - R_{i-1})$$

$$P_i = \frac{TP_i}{TP_i + FP_i}$$

$$R_i = \frac{TP_i}{TP_i + FN_i}$$

In the average precision formula, AP is the average precision. TP are true positives, FP are false positives and FN are false negatives. P_i and R_i are the precision and recall respectively at certain thresholds [22].

3.3 The data-set

The data set used in this research is a data set labeled by domain experts from the Netherlands Labour Authority (NLA), the Dutch Employment Insurance Agency (UWV) and the Netherlands Institute of Human Rights NIHR. The NLA monitors whether or not companies follow the law and provide a safe and healthy work atmosphere [19]. The UWV provides an efficient execution of employement insurance and also provides job market services [3]. And the NIHR is involved with promoting and protection of human rights in many different ways [4].

In the data set, potentially discriminatory sentences based on gender are labeled by these experts. A label of '0' indicates no discrimination. A label of '1' indicates highly suspected discrimination and a label of '2' indicates that it took longer than 15 seconds to identify if the sentence was highly suspected discriminatory or not. A full explanation of how the data came to be and how it was labeled can be found in the paper "Context-Aware Discrimination Detection in Job Vacancies using Computational Language Models" [20]. The full data set contains 5948 entries and roughly 29% of entries are labeled with a '1'. The number of entries labeled with a '2' is negligibly small. All other entries have the label '0'. A job posting may have multiple sentences that are labeled and thus those job vacancies will appear multiple times in the data set, once for each labeled sentence.

For this research, the data-set is duplicated into three different data sets by splitting the text of each job vacancy into different parts. One set considers only the sentences that are labeled, which is considered the 'direct context'. Another set considers the full vacancy including sentences that are labeled. The last set contains job vacancies with a labeled sentence filtered out, thus containing only the residue of the job vacancy, referred to as the 'indirect context'. The focus is on training a model on the indirect context. To make the explanation less confusing, figure 5 in the appendix visualizes this process of classifying indirect context and how it will be used to answer the research question.

3.4 Reproduction & further training

As a first step, the findings in the paper "Context-Aware Discrimination Detection in Job Vacancies using Computational Language Models" [20] are reproduced using the direct context. This is to help ensure that our methods are executed correctly. Then, to best judge the performance of the models when dealing with gender bias in indirect context, the hyper-parameters of each model are optimized. This is to further improve the average precision of each model. The parameters used in the hyper-parameter optimization are found in table 3, alongside the best parameters for the bag-of-words model and the word-2-vec model. These results are compared against the default parameters used by Scikit-learn.

It is also worth exploring the worst average precision a model can realistically output, which is done by constructing a worst-case model; a model will guess for each input the label '0' or '1' at random. This model is used to put the average precision of the models in indirect context into perspective. A model that performs only marginally better than the baseline model is considered a bad model.

Additionally, to get a better understanding of how the models work, the coefficients from the Logistic Regression model and the feature importances from the Random Forest model are analyzed. We will briefly explore why certain weights - be it positive or negative - are assigned to certain terms. This will be done by exploring how often certain terms is present in indirect contexts.

3.5 Model construction

Data preprocessing

For the bag-of-words vectorizer, a list of Dutch stop words is used [11]. This is to ensure that a model does not accidentally assign high weights to words that do not contribute to the meaning of a text. A couple words are excluded from the stop word list, namely *geen* (no / none) - as it strongly influences the intent of the text - and *ervaring* (experience) - as that word is applicable in the context of job vacancies.
For the word-2-vec model, Spacy's *nl_core_news_lg* pack already has its own default list of stop-words and as such no additional stop-word filtering is used in that vectorization.
In the case of the bag-of-words model, any kind of interpunction has been removed. This is to ensure that words with a trailing interpunction character are not identified as different words. The pre-processing is done using the default regular expression exposed by the Scikit-learn API.
In the case of the word-2-vec model, interpunction has also been removed, namely the characters ’?’, ’!’, ’;’, ’,’, ’-’, ’—’, ”*” and ’.’. For both models, capitalization has also been removed. Each upper-case letter has been replaced with its corresponding lower-case letter.

Settings

For the bag-of-words model, both 1-grams and 2-grams are considered, meaning both individual words and pairs of words are considered as features for the models. The maximum number of features is set to 5000 to improve run-time duration and to prevent over-fitting. Whether or not a feature is included in this list of 5000 features depends on how often they appear in total in the text. The least common features are excluded whereas the top 5000 most common features are included. During cross-validation on the full vacancies and indirect contexts, Scikit-learn's GroupKFold feature is used to put vacancies with the same identifier in the same group. This is because vacancies may contain multiple sentences that are labeled. This would otherwise result in some of the data appearing in both the training and test sets.

3.6 Model training

For the training of the models, GroupKFold cross-validation using a k-fold of 7 is used, where six folds are used for training and one fold is used for testing. These fold are equal in size. Reported numbers are always the average of the individual k-folds unless specified otherwise. For the hyper-parameter optimization, the data-set is split into 30% test data and 70% training data. The training data is further split into a 5-fold GroupKFold cross-validation to find the optimal hyper-parameters. The best hyper-parameter combination is then applied to the 30% test data to evaluate how well they perform on the indirect context.

4 Results

The results of the difference in performance of the paper ”Context-Aware Discrimination Detection in Job Vacancies using Computational Language Models” [20] and the reproduced results in this research can be seen in table 1. The results for the BERT model are not reproduced in this research. The reproduced results all fall within one standard deviation of the original results. A standard deviation is also displayed to show the potential range of deviations.

Original				Reproduced	
Vectorizer	Model	AP	AUC	AP mean	St. dev.
BoW	LR	74,0%	88,5%	74,4%	6,0%
	GB	76,8%	88,7%	72,9%	7,1%
	RF	72,4%	87,1%	75,8%	6,6%
W2V	LR	64,9%	82,4%	62,1%	4,4%
	GB	61,8%	79,7%	60,3%	4,7%
	RF	56,6%	78,5%	59,6%	5,7%
BERT	BERT	83,3%	92,6%	-	-

Table 1: The original results and the reproduced results. No hyper-parameter optimization has been used yet. AUC refers to the ’area under the ROC curve’, which has not been reproduced in this research.

The average precision was computed for the full vacancy and the indirect context using the same vectorizers and models with the same settings. The results are displayed in table 2. It is apparent that direct context yields a higher average precision for all models than the full vacancy models. And the full vacancy models also yield a higher average precision than the indirect context models. In order to improve the average precision of the indirect context, hyper-parameter optimization is used. The specific parameters alongside their chosen range of values to consider can be found in table 3. The best parameters per vectorizer and model combination are also shown.
After optimizing the hyper-parameters, the bag-of-words and word-2-vec models are re-ran using optimal parameters to identify to what extent the optimized parameters are better than the default parameters defined by Scikit-learn. The results are displayed in table 4. After optimization, the

Vectorizer	Model	AP		
		Direct bias	Full vacancy	Indirect context
BoW	LR	74,4%	59,5%	47,1%
	GB	72,9%	68,4%	52,8%
	RF	75,8%	64,9%	55,8%
W2V	LR	62,1%	46,0%	43,9%
	GB	60,3%	48,5%	43,2%
	RF	59,6%	49,9%	45,0%

Table 2: The average precision across all three different contexts; the direct context which only observes the highly suspected discriminatory sentence, the full vacancy, and the rest of the vacancy with the sentence of the direct context filtered out. Default parameters were used in these tests.

Model	Hyperparameter Search Space	Best BoW Hyperparameters	Best W2V Hyperparameters
Logistic Regression	C: [0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 5, 10, 50, 100] solver: ['lbfgs', 'liblinear'] penalty: ['l2'] class_weight: ['balanced']	C: 0.01 solver: 'lbfgs' penalty: 'l2' class_weight: 'balanced'	C: 0.001 solver: 'liblinear' penalty: 'l2' class_weight: 'balanced'
Gradient Boosting	learning_rate: [0.3, 0.2, 0.1, 0.05, 0.01, 0.005] max_depth: [2, 5, 10, 20, 50]	learning_rate: 0.2 max_depth: 2	learning_rate: 0.05 max_depth: 10
Random Forest	max_depth: [10, 50, 100] min_samples_split: [2, 10, 50] n_estimators: [50, 200, 1000, 2000] class_weight: ['balanced']	max_depth: 100 min_samples_split: 10 n_estimators: 2000 class_weight: 'balanced'	max_depth: 50 min_samples_split: 10 n_estimators: 1000 class_weight: 'balanced'

Table 3: Hyperparameter optimization for the three vectorizers and the two models. The left column displays the search space and the two right columns display per parameter the best values that were found.

	Default BoW AP	Optimized BoW AP	Increase
LR	47,1% ($\bar{\sigma} = 4,2\%$)	53,3% ($\bar{\sigma} = 4,4\%$)	+ 6,2%
GB	52,8% ($\bar{\sigma} = 2,0\%$)	52,7% ($\bar{\sigma} = 2,9\%$)	- 0,1%
RF	55,8% ($\bar{\sigma} = 3,7\%$)	57,4% ($\bar{\sigma} = 3,2\%$)	+ 1,6%
	Default W2V AP	Optimized W2V AP	Increase
LR	43,9% ($\bar{\sigma} = 3,0\%$)	44,3% ($\bar{\sigma} = 3,2\%$)	+ 0,4%
GB	43,2% ($\bar{\sigma} = 3,7\%$)	42,6% ($\bar{\sigma} = 3,9\%$)	- 0,6%
RF	45,0% ($\bar{\sigma} = 3,0\%$)	46,3% ($\bar{\sigma} = 2,8\%$)	+ 1,3%

Table 4: Percentages may differ by a few decimals depending on the data splits during training

bag-of-words plus Random Forest model perform the best, with an average precision of 57,4% in indirect context.

Learning curve graphs in the appendix in figure 1, 2 and 3 show how each model trains with increasing training set sizes. The averages of all six graphs are combined into figure 4. The bag-of-words plus Random Forest combination performs the best after learning from all the data, however at lower fractions of data, the bag-of-words plus Logistic Regression combination performs better. To identify how well of an average precision the models generated, a baseline model was created, labeling classes at random. Random labeling yields a lower bounds average precision of 0.2868.

4.1 Logistic Regression coefficients in different contexts

Table 5 displays the top 20 1-grams and 2-grams with highest weights, either positive or negative, for the direct context only when using the Logistic Regression model and the bag-of-words vectorizer. Table 6 displays the same, but for the full vacancy with the bag-of-words and Logistic Regression combination instead. Table 7 also displays the weights, but for the indirect context only. So it considers the full vacancy, but with the highly suspected discriminatory sentences filtered out. The parameters used when training the Logistic Regression model are the best hyper-parameters found

Direct bias term	Weight	Ratio discriminatory	Ratio non-discriminatory
vrouwelijke	0.684	0.2776	0.0792
vrouwelijk	0.578	0.1114	0.0189
voorkeur	0.517	0.1411	0.0291
gastvrouw	0.441	0.0478	0.0061
vrouwelijk maatje	0.393	0.0659	0.0031
jongens	-0.362	0.0431	0.1158
vrijwilliger	0.344	0.1440	0.0376
mannelijke	0.311	0.1224	0.0567
jongen	-0.292	0.0344	0.0917
mannelijke vrijwilliger	0.271	0.0385	0.0012
mevrouw	-0.271	0.0251	0.0586
timmermannen	0.262	0.0531	0.0132
man	-0.242	0.0548	0.1205
handige	0.199	0.0222	0.0035
bestaat	-0.199	0.0047	0.0314
vrijwilligster	0.198	0.0332	0.0040
maatje	0.196	0.1009	0.0378
enthousiaste	0.186	0.0659	0.0388
liefst	0.185	0.0309	0.0038
opzoek	0.184	0.0262	0.0118

Table 5: Top 20 1-grams and 2-grams in direct context only with the highest weights, either positive or negative.

during the grid search. The weights in the direct context are the largest on average, whereas the weights in the indirect context are the smallest on average. Many terms with a high weight like *vrouwlijke* (female), *vrouwelijk* (feminine) and *voorkeur* (preference) do not appear in the top 20 of the indirect context. There are a few terms however, such as *jongens* (boys) and *gastvrouw* (hostess), that do appear in both. The only 2-grams within the three top 20's are *vrouwelijk maatje* (female buddy) and *mannelijke vrijwilliger* (male volunteer).

4.2 Feature importance: Random Forest

Table 8 shows the Random Forest feature importances alongside the ratio of non-discriminatory vacancies and highly suspected discriminatory vacancies containing these words at least once. The Random Forest model has been trained using the optimized hyper-parameters that are documented in table 3. The table with importances shares terminology found in the Logistic Regression top-20 highest weights, for example *fysiotherapeut* (fysiotherapist) and *jongens* (boys). There are also new terms however, such as *vca* and *ontwikkelen* (develop).

Full vacancy term	Weight	Ratio discriminatory	Ratio non-discriminatory
vrouwelijke	0.519	0.2991	0.0969
vrouwelijk	0.455	0.1236	0.0258
jongens	-0.294	0.0519	0.1283
gastvrouw	0.284	0.0711	0.0302
mannelijke	0.279	0.1359	0.0662
vrouwelijk maatje	0.269	0.0711	0.0069
voorkeur	0.239	0.2630	0.2046
jongen	-0.233	0.0449	0.1000
mannelijke vrijwilliger	0.232	0.0414	0.0052
vrijwilligster	0.210	0.0455	0.0111
fysiotherapeut	0.191	0.0402	0.0033
vrijwilliger	0.175	0.2210	0.1191
professionele	-0.165	0.0443	0.0945
handige	0.164	0.0245	0.0066
samenstelling	0.163	0.0461	0.0116
bestaat	-0.160	0.0560	0.1030
oppas	-0.160	0.0169	0.0747
figuranten	-0.158	0.0012	0.0130
algemeen	0.156	0.0630	0.0328
schoonmaak	0.150	0.0280	0.0113

Table 6: Top 20 1-grams and 2-grams in the full vacancy with the highest weights, either positive or negative.

Indirect bias term	Weight	Ratio discriminatory	Ratio non-discriminatory
jongens	-0.223	0.0140	0.0522
fysiotherapeut	0.215	0.0397	0.0031
jongen	-0.179	0.0192	0.0437
algemeen	0.169	0.0531	0.0243
emailadr	0.147	0.1382	0.1040
gastvrouw	0.142	0.0461	0.0286
bestuur	0.135	0.0187	0.0180
dame	-0.135	0.0501	0.0444
instelling	0.124	0.0583	0.0584
individuele	0.123	0.0280	0.0158
developer	-0.122	0.0017	0.0194
21	0.120	0.0688	0.0425
bel	0.117	0.0542	0.0291
40 uur	0.114	0.1061	0.0867
cursussen	-0.113	0.0192	0.0338
basis	-0.111	0.0851	0.1082
aandacht	-0.110	0.0630	0.0759
hotel	0.109	0.0210	0.0069
erop	0.108	0.0606	0.0255
12	-0.105	0.0006	0.0154

Table 7: Top 20 1-grams and 2-grams in indirect context only with the highest weights, either positive or negative.

Term	Importance	Total occurrences	Ratio discriminatory*	Ratio non-discriminatory*
'fysiotherapeut'	0.00480	167	0.0397	0.0031
'timmerman'	0.00409	37	0.0548	0.0125
'vca'	0.00330	315	0.0810	0.0281
'ontwikkelen'	0.00317	816	0.0618	0.1293
'maatje'	0.00287	1049	0.1394	0.0685
'trainer'	0.00270	427	0.0047	0.0272
'klasse'	0.00250	117	0.0000	0.0177
'kennis'	0.00226	1627	0.1487	0.2131
'lokatie'	0.00222	156	0.0070	0.0340
'jongens'	0.00220	399	0.0140	0.0522
'communicatie'	0.00213	463	0.0309	0.0659
'processen'	0.00211	166	0.0041	0.0291
'team'	0.00210	2827	0.2292	0.3003
'hbo'	0.00209	1271	0.1172	0.1817
'jongen'	0.00207	374	0.0192	0.0437
'vrouw'	0.00203	677	0.1114	0.0666
'professionele'	0.00202	458	0.0402	0.0827
'richting'	0.00201	289	0.0169	0.0529
'oppas'	0.00196	1344	0.0169	0.0742
'type'	0.00196	336	0.0268	0.0624

Table 8: Resulting importances in the Random Forest model. The Discriminatory ratio in this context refers to the ratio of vacancies labeled as highly indicative of discrimination in which the given word occurs at least once. The non-discriminatory ratio refers to the ratio of non-discriminatory vacancies in which the given word occurs at least once in the indirect context.

Mentioned phrases	Ratio discriminatory	Total occurrences
Male keywords	0.247	977
Female keywords	0.346	1137
Both genders	0.249	370
Either gender	0.295	2234
'Boys and girls'	0.206	63

Table 9: The ratio of job vacancies labeled as discriminatory in the data set when their indirect context contains specific gendered words that indicate a man or woman.

4.3 Gendered nouns in combination

Table 9 shows for indirect contexts containing specific gender terminology how often they are labeled as discriminatory. The *Male keywords* row observes indirect contexts including specific keywords that specify males, namely *man(nen), jongen(s)* and *mannelijk(e)*. The *Female keywords* row observes indirect contexts that includes specific female keywords, namely *vrouwe(en), dame(s), meisje(s)* and *vrouwelijk(e)*. The row *Both genders* observes indirect contexts including terminology from both lists and the row *Either gender* observes indirect contexts including terminology from either lists. The *'Boys and girls'* row observes combinations of *jongen(s) en/of meid(en), man(nen) en/of vrouw(en)*.

Indirect contexts that only mention one gender have relatively large ratios of discrimination. Indirect contexts mentioning multiple genders have a relatively low ratio of discrimination.

5 Conclusion

Data reproduction

The slight difference in performance between the reproduced results and the original results can be explained by a slight difference in methodology. In this research, stop-words were filtered out beforehand, which is not the case in the previous research. Despite that, results are very similar which is a strong indication that the methods in this research and the previous research can be compared. Especially given the standard deviation of the reproduced results, the average AP falls within one standard deviation.

Main findings

Initial results for the average precision in full context and indirect context yield worse results than the direct context, with the indirect context yielding the worst average precision. The largest drop in average precision when comparing the direct context to indirect context, is for the Logistic Regression model with the bag-of-words vectorizer. This combination experiences a 27,3% drop in average precision. The smallest drop in average precision between the direct and indirect context is for the Random Forest model combined with the Word-2-Vec vectorizer, sitting at 14,6%. These are also the largest and smallest relative drops in performance respectively.

The large drops in performance between direct and indirect context suggest that many key factors used to classify gender discrimination are lost when filtering out the direct context. Especially when considering the high baseline performance of 0,2868 returned by the baseline model, a worst-case drop in performance of 27,3% is translated to a relative 38,3% drop in performance when the baseline model is treated as the AP range's zero point.

Additionally, when optimizing each models' hyper-parameters, the average precision only improves by about 1,47% on average. This indicates that with the current data set and methods there is little room to improve the performance in indirect context.

That said, the baseline model still only has half the average precision of the indirect context models. This means that even though many key factors from the direct context are lost, there are still some subtle hints that the models were able to pick up on and correctly classify vacancies on. However, a best-case average precision of 57,4% after hyper-parameter optimization does not sound as a useful result for practical use. Additional or different approaches will be necessary to improve this performance.

Learning curve trends

The learning curves show that starting at a sixteenth of the data, the different vectorizer and model combinations all improve their performance at a consistent rate whenever the amount of data during training is doubled. There is also no clear indication of the models plateauing at higher intervals. This may indicate that the average precision in indirect context may significantly improve if the data set were larger.

Gendered nouns weights

Results show that many nouns that reference a gendered person (man, woman, boy, girl) have negative weights assigned to them by the Logistic Regression model, which is indicative of 'no discrimination'. Intuitively you might expect such words to have positive weights however. The reason the weights are negative though becomes clear when observing the context in which they are used. There are 63 vacancies which include the phrase 'boy(s) and/or girl(s)' or '(man/men) and/or (woman/women)'. The ratio of discriminatory vacancies among those 63 is 0.206, which is lower than ratio discriminatory vacancies in the whole data-set on the other hand, which is 0.295.

When broadening the search of gendered people to vacancies referencing both at least one male and one female person, there are 370 vacancies. This totals roughly 6,2% of vacancies mentioning both men and women in the indirect context. Within those 370 vacancies, the ratio discriminatory vacancies is 0.249, which is again lower than the 0.295 ratio within the whole set of indirect contexts.

Another important factor that influences the weight of gendered nouns, is that those nouns may reference something related to the job itself rather than the candidate. For example, a babysitter might have to babysit a group of boys. In this context, the gendered noun could be part of a non-discriminatory job vacancy without being associated with the gender of the potential candidate.

Using this data, it makes sense that gendered terms tend to have negative weights assigned to them. Despite them being associated with a specific gender, the context in which they are used is not related to the job's candidate. The models cannot identify that these gendered words are referencing things unrelated to the candidate. As such, it makes sense that the models assign negative weights to these gender terms.

6 Discussion

6.1 The word-2-vec approach

The word-2-vec vectorizer in this research could be considered a poor choice as a vectorizer. In larger texts with more words, each word is given a smaller weight and the average vector of the whole text becomes less susceptible to outliers. Consequently, the average vector of two random large texts are more likely to be similar than the average vector of two smaller texts. In direct context, only one sentence is considered and therefore the word-2-vec vectorizer may be a suitable choice. However, in indirect context additional considerations may be useful to take into account. For example, the use of inverse document frequency could be used to assign higher weights to uncommon words, allowing for more distinction in vectors between larger texts.

6.2 Hyper-parameter optimization

After hyper-parameter optimization, performance of the models went up on average. That said, gradient boosting performance did decrease slightly. This can be explained by the high standard deviation of the tests. A higher standard deviation can cause more accidental outliers in the resulting average precision. With a standard deviation of 2,9% and 3,9%, an accidental reduction in performance by 0,1% and 0,6% respectively is feasible.

6.3 Data reproduction

The Scikit-learn version used in this research is 1.0.2. The version used in previous research is 0.24.2. Although the Scikit-learn documentation for the Logistic Regression, Random Forest and Gradient Boosting models do not note any changes between those versions, it may be possible that subtle differences exist between the two. As such, the reproduced results may differ slightly. It is also unclear whether or not the previous research uses the same Spacy language package when training the models.

6.4 Pairs of opposite genders

Results show that gendered terms are given negative weights. Results also show that indirect contexts mentioning multiple genders or a pair of genders are less likely to be discriminatory. This pattern of paired genders has not been registered by the models however. One reason why a model would be unable to pick up on this connection is the specific n-grams used in the bag-of-words vectorizer. In this research, 1-grams and 2-grams were used. This means that combinations such as 'boy or girl' and 'man or woman' would not be considered by the vectorizer, as those are 3-grams. However, these 3-grams may be very important in differentiating between the discriminatory use of these gendered nouns and non-discriminatory use. Future research should therefore also consider looking into 3-grams.

Another reason why pairs of words that indicate opposite genders are not considered by the model, is the maximum vocabulary size, which was set to 5000. For every occurrence of a 2-gram, the two 1-grams making up the 2-gram would be present at least as often in the data set as that 2-gram.

This means that for every 2-gram that made it into the bag-of-words vocabulary, two 1-grams would also make it into the vocabulary, reducing the relative frequency of 2-grams.
Considering these two factors, it might be worth exploring a bag-of-words model that only considers 3-grams. Potentially in combination with more thorough pre-processing and word filtering, computation time could be kept manageable and performance in this aspect could improve.

6.5 Data set completeness

The data set used for this research did not contain full job vacancies; after a certain number of characters the paragraphs were cut off. As a result, the last sentence in a job vacancy would often not be complete, which might influence the essence of a text.
Some job vacancies also looked very similar to each other. This happened when a company or entity would post the same job vacancy multiple times, each with slight modifications. This could lead to an over-representation of certain terminology. For example, there were 29 job vacancies containing the word *fysiotherapeut* (physiotherapist) which all referenced the same phone number. And there were only 92 job vacancies total containing the word *fysiotherapeut*. So around 31.5% of occurrences came from the same entity.
On top of that, all job vacancies in the data set were posted by Dutch entities. The Dutch job market however, may differ from other countries with similar cultures. Therefore, data from this research may not represent the job market of other countries.

6.6 Duplicate data

The data set contained 5948 total job vacancies. However, job vacancies could contain multiple sentences that are flagged as potentially discriminatory. Therefore, the total number of unique vacancies in the data set was only 5350, which is roughly 9% lower. The total number of unique job vacancies containing multiple flagged sentences is 108. Additionally, companies would list dozens of similar job postings. This impacted the quality of the data, as some job postings would look much alike. This reduced the diversity of words used within the data set as a whole.

References

[1] Betere man-vrouw verhouding geregeld voor top bedrijfsleven. https://www.rijksoverheid.nl/actueel/nieuws/2021/09/28/betere-man-vrouw-verhouding-geregeld-voor-top-bedrijfsleven. Accessed: June 21st 2022.

[2] Is sexual harassment explicitly prohibited in the workplace? https://www.worldpolicycenter.org/sites/default/files/media-server/maps/1079.png.

[3] Wat is uwv?: Uwv: Over uwv. https://www.uwv.nl/overuwv/wat-is-uwv/index.aspx, Dec 2014.

[4] About national human rights institutions. https://ennhri.org/about-nhris/, May 2020.

[5] Diversity and inclusion progress report. https://www.riotgames.com/en/work-with-us/diversity-and-inclusion/diversity-and-inclusion-progress-report, Apr 2020.

[6] Openmoji. https://openmoji.org/, 2022. Icons were provided by OpenMoji. They are licensed under the Creative Commons Attribution 4.0 International License. A copy of the license is available at http://creativecommons.org/licenses/by/4.0/.

[7] Shiu-Yik Au, Andreanne Tremblay, and Leyuan You. Does board gender diversity reduce workplace sexual harassment? https://deliverypdf.ssrn.com/delivery.php?ID=6620051221180011100801171171030170300240720850410370200220671160140640960310920170070550630020171140340380000760070191190771130240150170230511160290030010801070681230840390871041230950850740931040710001201140961110911161201141120850200240901041141241248amp;EXT=pdf&amp;INDEX=TRUE, Sep 2020.

[8] Kellen Browning. Riot games to pay $100 million in gender discrimination case. *The New York Times*, Dec 2021.

[9] Jason Brownlee. Logistic regression for machine learning. https://machinelearningmastery.com/logistic-regression-for-machine-learning/, Aug 2020.

[10] Cecilia D'Anastasio. Inside the culture of sexism at riot games. https://kotaku.com/inside-the-culture-of-sexism-at-riot-games-1828165483, Aug 2018.

[11] Alex de Wekker. Stopwoorden. https://help.carerix.com/nl/articles/2074382-stopwoorden. Accessed: June 22nd 2022.

[12] Ahmed Fawzy Gad. Mean average precision (map) explained. https://blog.paperspace.com/mean-average-precision/, Apr 2021.

[13] Gaurav. An introduction to gradient boosting decision trees. https://www.machinelearningplus.com/machine-learning/an-introduction-to-gradient-boosting-decision-trees/, Mar 2022.

[14] Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.

[15] Ceshine Lee. Feature importance measures for tree models - part i. https://medium.com/the-artificial-impostor/feature-importance-measures-for-tree-models-part-i-47f187c1a2c3, Sep 2020.

[16] Panggih Kusuma Ningrum, Tatdow Pansombut, and Attachai Ueranantasun. Text mining of online job advertisements to identify direct discrimination during job hunting process: A case study in indonesia. https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0233746, Jun 2020.

[17] Kim Parker and Cary Funk. Gender discrimination comes in many forms for today's working women. https://www.pewresearch.org/fact-tank/2017/12/14/gender-discrimination-comes-in-many-forms-for-todays-working-women/, Aug 2020.

[18] 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. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[19] Ministerie van Sociale Zaken en Werkgelegenheid. Home. https://www.nlarbeidsinspectie.nl/, Jul 2022.

[20] S. Vethman, A. Adhikari, M. H. T. de Boer, J. A. G. M. van Genabeek, and C. J. Veenman. Context-aware discrimination detection in job vacancies using computational language models, 2022.

[21] Tony Yiu. Understanding random forest. https://towardsdatascience.com/understanding-random-forest-58381e0602d2, Sep 2021.

[22] Shivy Yohanandan. Map (mean average precision) might confuse you! https://towardsdatascience.com/map-mean-average-precision-might-confuse-you-5956f1bfa9e2, Jun 2020.

[23] Ethan Zhang and Yi Zhang. *Average Precision*, pages 192–193. Springer US, Boston, MA, 2009.

Appendix

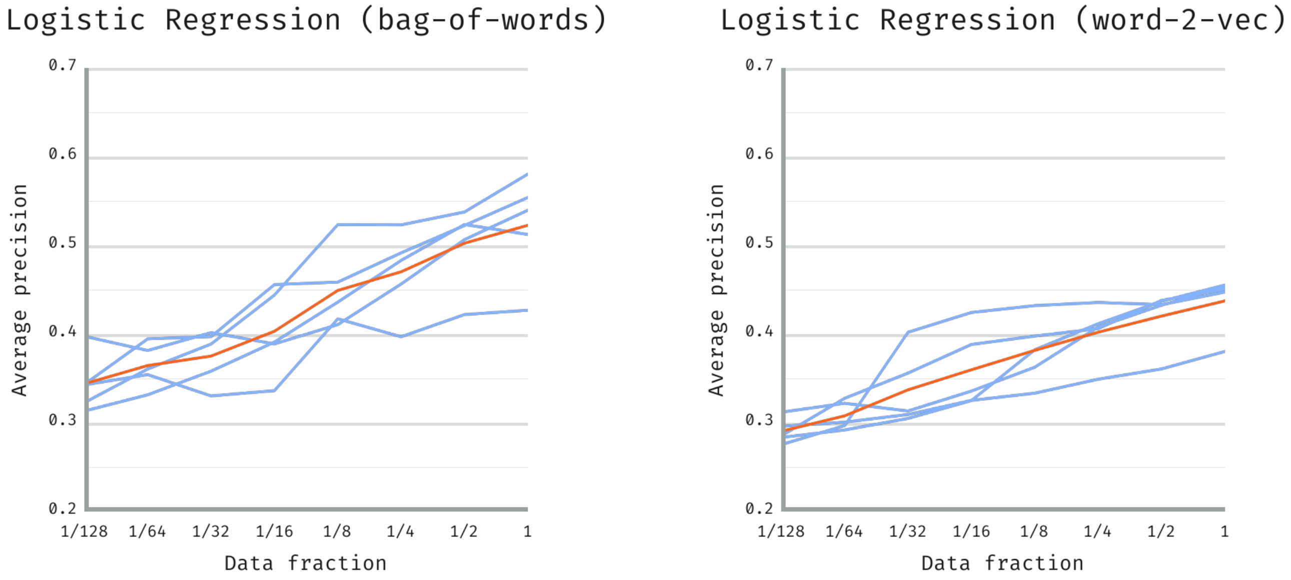


Figure 1: The learning curves of the Logistic Regression models. The orange line is the average learning curve of the blue lines representing individual folds. In this case a 5-fold is used.

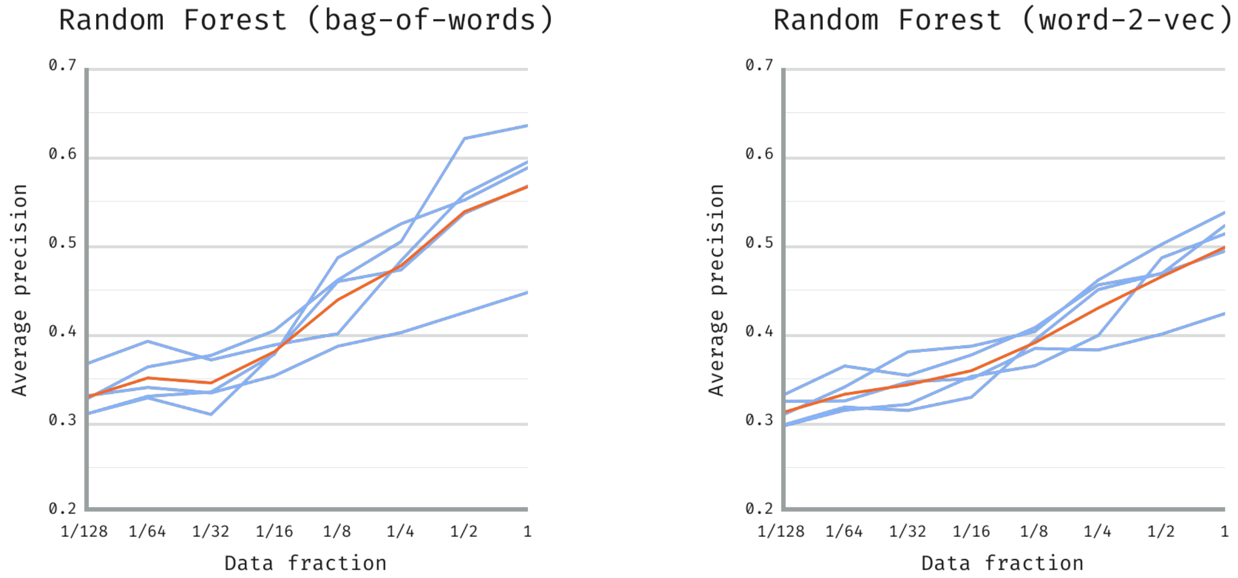


Figure 2: The learning curves of the Random Forest models. The orange line is the average learning curve of the blue lines representing individual folds. In this case a 5-fold is used.

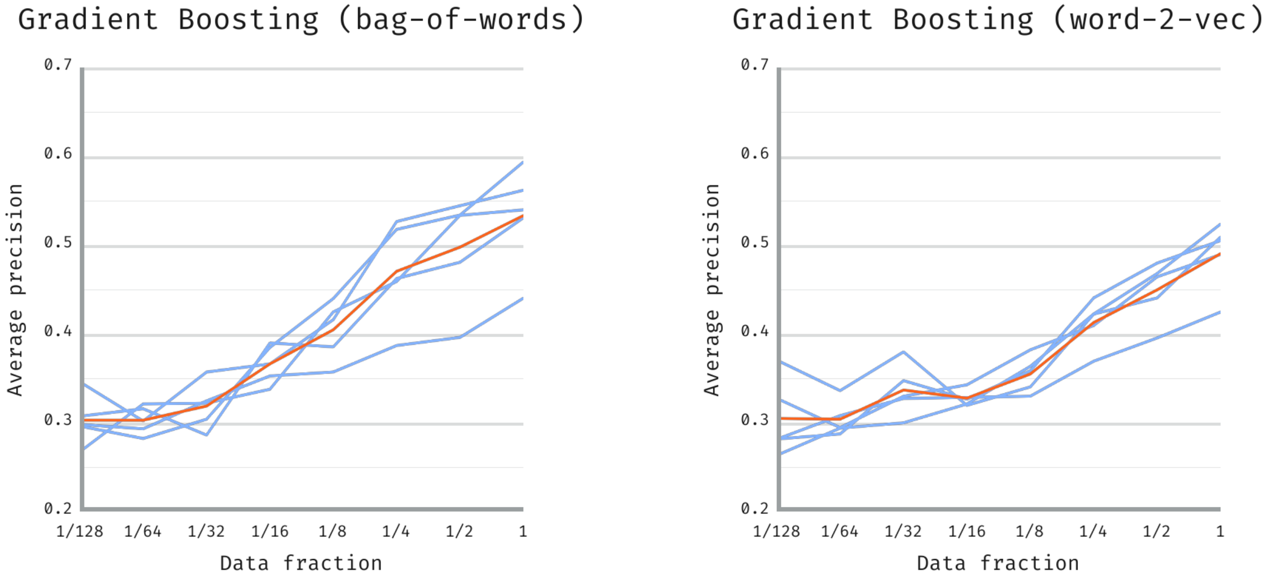


Figure 3: The learning curves of the Gradient Boosting models. The orange line is the average learning curve of the blue lines representing individual folds. In this case a 5-fold is used.

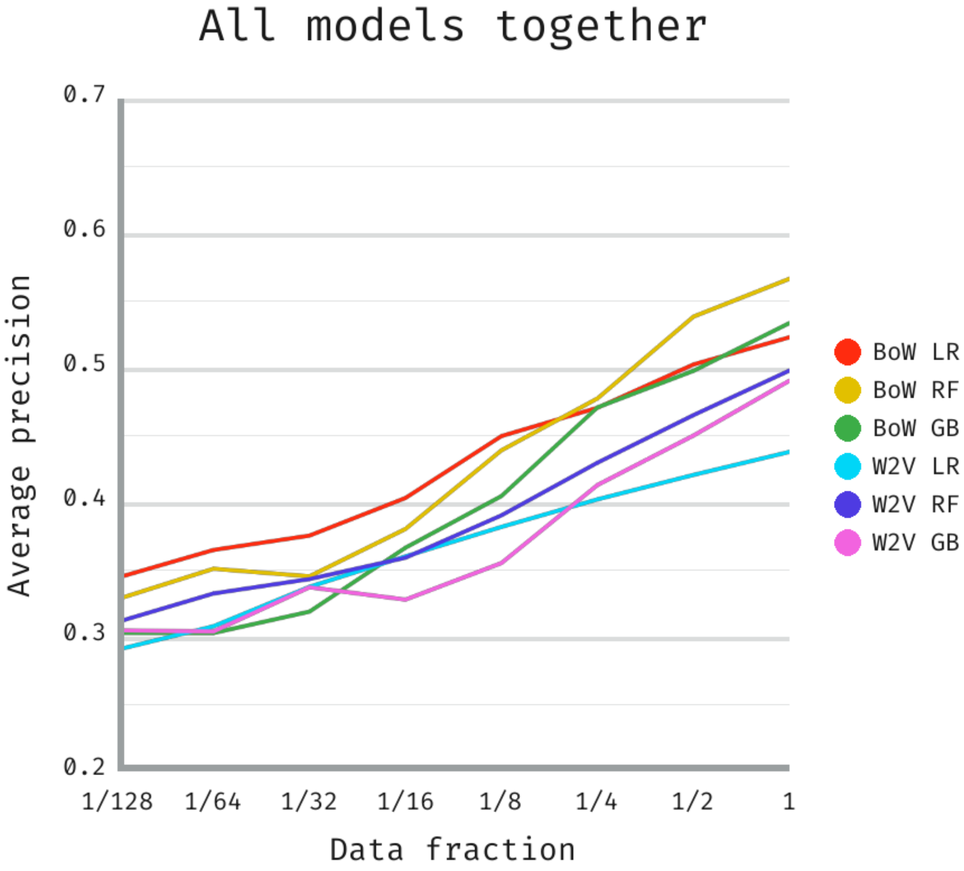
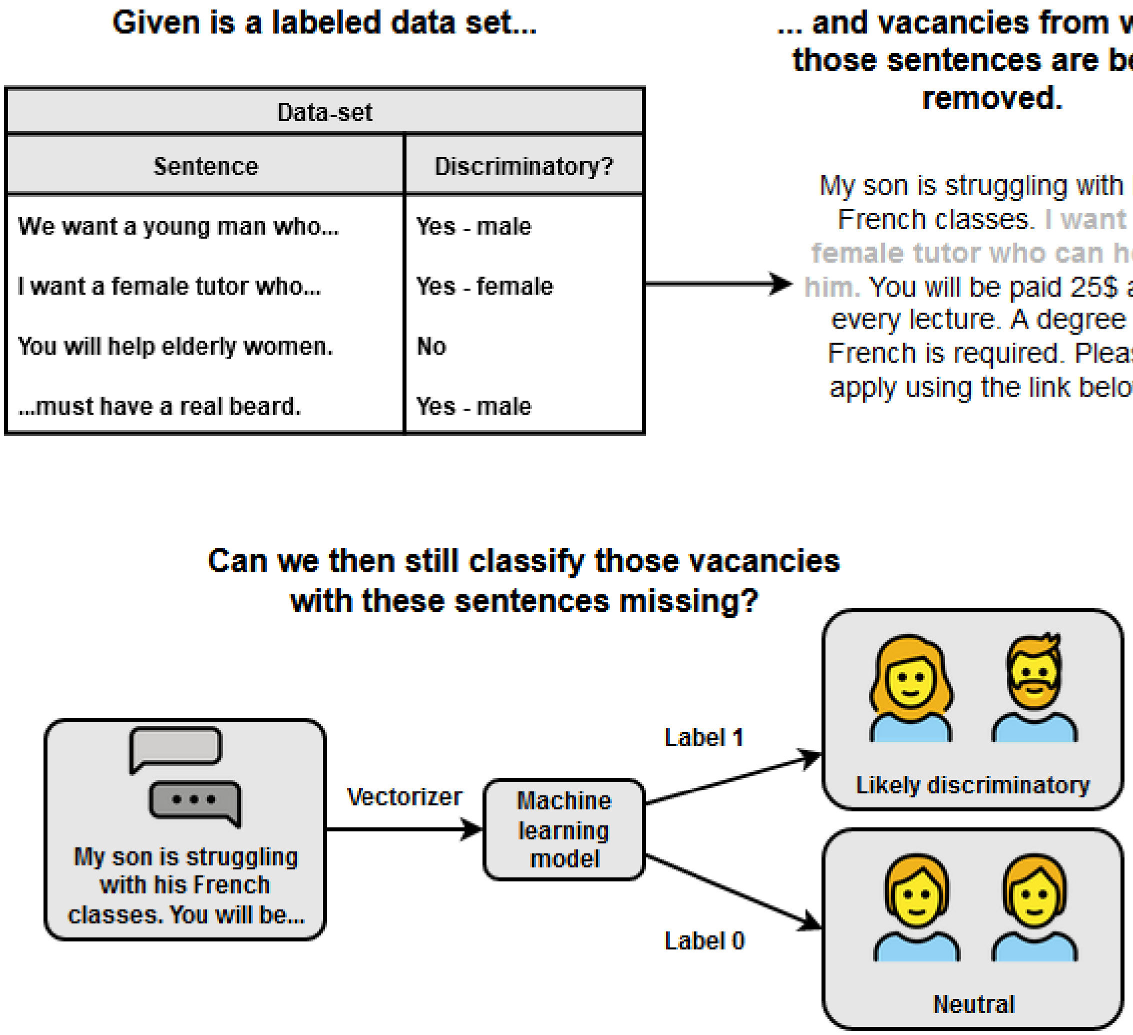


Figure 4: All learning curves combined.



Given is a labeled data set...
... and vacancies from which those sentences are being removed.

Data-set	
Sentence	Discriminatory?
We want a young man who...	Yes - male
I want a female tutor who...	Yes - female
You will help elderly women.	No
...must have a real beard.	Yes - male

My son is struggling with his French classes. I want a female tutor who can help him. You will be paid 25$ after every lecture. A degree in French is required. Please apply using the link below.

Can we then still classify those vacancies with these sentences missing?
My son is struggling with his French classes. You will be...
Vectorizer
Machine learning model
Label 1
Label 0
Likely discriminatory
Neutral

Figure 5: A schematic illustration of the research question and how the classifiers and vectorizers will be used to classify indirect context. The icons were provided by OpenMoji [6].