Humanized Computing and Communication
Vol. 1, No. 1 (2020) 1-22

DOI: 10.35708/HCC1868-148362

Analyzing Sentiments of German Job References

Finn Folkerts, Vanessa Schreck, Shirin Riazy, and Katharina Simbeck

Hochschule für Technik und Wirtschaft Berlin (HTW), Berlin, Germany
{folkerts,schreckv,riazys,simbeck}@htw-berlin.de

Received (11/02/2019)
Revised (02/19/2020)
Accepted (02/28/2020)

Abstract. Filling a vacancy takes a lot of (costly) time. Automated preprocessing of applications using artificial intelligence technology can help to save time, e.g., by analyzing applications using machine learning algorithms. We investigate whether such systems are potentially biased in terms of gender, origin, and nobility. Using a corpus of common German reference letter sentences, we investigate two research questions. First, we test sentiment analysis systems offered by Amazon, Google, IBM and Microsoft. All tested services rate the sentiment of the same template sentences very inconsistently and biased at least with regard to gender. Second, we examine the impact of (im-)balanced training data sets on classifiers, which are trained to estimate the sentiment of sentences from our corpus. This experiment shows that imbalanced data, on the one hand, lead to biased results, but on the other hand, under certain conditions, can lead to fair results.

Keywords: Machine Learning; Discrimination; Fairness; Sentiment Analysis.

1 Introduction

In order to improve cost efficiency in recruiting, companies automate their processes by using either external software solutions or self-developed tools. Besides quality of and cost per recruitment, staffing time is an important key performance indicator [35]. For this reason, Natural Language Processing (NLP) systems are sometimes used to pre-select candidates [4, 31]. The use of such artificial intelligence (AI) models is expected to increase productivity [1], but often lacks transparency [22, 29]. The systems are used as black boxes. This – at times purposeful – non-transparency (see e.g. [20]) may obfuscate ethical issues such as discrimination [1, 2, 31]. Automation and digitization could contribute to making HR processes less prone to prejudices or stereotypes. AI systems are

commonly expected to be objective and therefore help to avoid potential discrimination. However, when systems learn from biased data they will reproduce those biases [8]. Sentiment analysis systems have been shown to assess sentences inconsistently, depending on gender and race of the subject [17].

In this research we first analyzed the fairness of commercial sentiment analysis systems. We have extended the work of [17] to the human resources (HR) domain by testing the sentiment analysis systems provided by Google Natural Language API[1], Amazon Web Service Comprehend[2], IBM Watson Natural Language Understanding[3] and Microsoft Azure Cognitive Service[4] with sentences from German job reference letters. For this purpose, we compiled a test corpus with typical German job reference sentences. We explored whether the sentiment scores for almost identical sentences differentiate between male and female subjects, German and Turkish surnames as well as German surnames with and without nobiliary particle.

This experiment shows that all tested sentiment analysis services evaluate almost identical sentences unpredictably differently, depending on the subject used in a sentence.

Further, we have replicated sentiment analysis systems in order to examine the impact of (im-)balanced training data sets on Machine Learning (ML) models in computational linguistics. Seven different classifiers were trained on the basis of our corpus with sentences from German job reference letters. As each sentence in our corpus is labeled with a grade, the models were able to learn how to classify the sentences' sentiments. In order to test the effects of imbalanced training data sets, we have designed three scenarios with skewed data in contrast to the balanced training set. We have derived two realistic scenarios from studies [24,34] on job references referring the distribution of grades between men and women. In addition, we created an exaggerated scenario in which the training data consists of a significantly larger amount of positive sentences with male subjects than with female subjects.

Findings from this experiment suggest that imbalanced training data can lead to biased results depending on the chosen ML models, but also that a balanced training data set produces fair results if a suitable model is used.

2 Related Work

This research paper is an expanded version of [13], where we showed that four commercial sentiment analysis services rate sentences very inconsistently in terms of gender, origin, and nobility (see Section 4).

Several other studies that are concerned with bias of NLP systems regarding gender or origin have shown before how algorithms reproduce stereotypes. Comparing 200 sentiment analysis systems [17] found evidence for strong bias. In [8], the authors trained a system that learned word associations from the

[1] https://cloud.google.com/natural-language/
[2] https://aws.amazon.com/comprehend
[3] https://www.ibm.com/watson/services/natural-language-understanding/
[4] https://azure.microsoft.com/en-en/services/cognitive-services/

Common Crawl[5] corpus via *word embedding* techniques, resulting in replication of stereotypes like women getting more associated with family and arts, while men get more associated with career and sciences. Bolukbasi et al. [5] attempted to develop a methodology to reduce bias in applications using word embedding. The reproduction of bias has also been detected in HR settings [16], where data mining techniques are used as aids for personnel selection, promotions, etc. Even though it is often expected for algorithms to work independently of human bias, the opposite has been observed [10]. Occasionally, an algorithm might even amplify a bias that exists within the training set [25]. With a growing metrization of modern work reality [23], data protection and other ethical concerns have been raised [31]. Particularly, unreflective use of new technologies in the HR area is advised against [21].

However, algorithmic bias is not limited to NLP but also affects computer vision systems. In [36] it turned out that a certain training set, which contained 33% more pictures with women associated with the activity cooking than men, leads to amplified reproduction of gender based stereotypes when models are being trained on this data set. It turned out that one particular training data set containing 33% more pictures of women associated with the activity cooking than men, led to an amplification of gender based stereotypes when models trained with this data set are being used [36]. The *Gender Shades* study [7] identified a high misclassification rate for face recognition algorithms with respect to the skin tone of a person. Compared to light-skinned individuals, dark-skinned individuals were up to 34.4% more likely to be misclassified.

Non-algorithmic HR decisions, which are often used as a basis for algorithmic HR applications, have been shown to be biased. Several studies have shown that applicants with uncommon or foreign sounding names are discriminated against by potential employers [3, 9, 15].

3 The German Job Reference Corpus (GJRC)

We compiled a test corpus with typical German job reference letter sentences from German books on how to write job reference letters [12, 14, 18, 19, 30]. We combined those template sentences with subjects of varying gender, origin, and nobility.

We focused on comparing sentences with different surnames, expecting to receive lower sentiment scores for non-German names. We chose to compare German surnames with Turkish surnames as citizens from Turkish origin represent the largest migration group in Germany [32].

Furthermore, we expected higher sentiment scores for names that indicate nobility ("von", "zu"). However, a nobiliary particle does not grant any legal privileges in Germany since 1919 [33]. Consequently, we examined the differences for ten German surnames with nobiliary particle as well as ten Turkish surnames and ten common German surnames.

To find suitable surnames for our experiment, we looked up the lists of members of the German state parliaments. We then mapped the names to their

[5] https://commoncrawl.org/

origins and randomly picked ten German surnames, ten German surnames with nobiliary particle and ten Turkish surnames. In Table 1 we listed all surnames that we used to compile the GJRC.

Table 1. Surnames used to compile the corpus.

German	**German with nobiliary particle**	**Turkish**
Becker	von Eyb	Aras
Dürr	von Halem	Erikli
Gruber	vom Bruch	Bozoğlu
Haußmann	von Berg	Çağlar
Klein	von Breitenbuch	Taş
Pfeiffer	von Brunn	Dogan
Sänze	von Danwitz	Demirel
Pohl	von Angern	Özdemir
Stettner	von Kalben	Yilmaz
Zimmermann	von Pein	Yüksel

Following a literature review [12, 14, 18, 19, 30], we collected 843 different sentences that are commonly used in German job references. In order to generate multiple versions of the same sentence, we modified each one so that it can be used as a template: all words that are gender-specific or require gender-specific declension were substituted with a suitable placeholder. Four examples of such template sentences are shown in Table 2.

Note that by law, German reference letters must be phrased favorably to employees [6], even if they did not perform well. Therefore, a generally positive sentiment can be expected.

To compile the German Job Reference Corpus, we combined each template sentence with each of the 30 different surnames and both gender specific titles. This yields 60 distinct sentences originating from the same template. Additionally, we altered each template sentence by replacing the title and surname with the corresponding male or female pronoun, thus adding another two sentences per template to the corpus.

Eventually, the corpus consists of 52,266 sentences in total, out of which 1,686 sentences are formed with a pronoun instead of a name. We have made the German Job Reference Corpus available on *GitHub*[6].

4 Experiment A: Sentiment Analysis

We conducted an experiment using the cloud services for sentiment analysis from Amazon, Google, IBM and Microsoft because they hold the biggest market

[6] https://github.com/iug-htw/GJRC

Table 2. Sentence templates used to compile the corpus (excerpt).

ID	Template	Grade
743	Die Qualität von <title_dat_acc> <name_s> Arbeit lag stets deutlich über dem Standard <poss_gen_m_n> Teams.[1]	Good
	The quality of <title_dat_acc> <name_s> work was always considerably above <poss_gen_m_n> team's standard.	
291	Bei wichtigen Aufgaben war <title> <name> zuverlässig und pflichtbewusst.[1]	Sufficient
	In important tasks, <title> <name> was reliable and dutiful.	
814	Wir bedauern <title_dat_acc> <name_s> Ausscheiden, bedanken uns für <poss_nom_w_pl_acc_w_pl> konstruktive Mitarbeit und wünschen <pers_pron_dat> für <poss_nom_w_pl_acc_w_pl> berufliche und private Zukunft weiterhin viel Erfolg und alles Gute.[1]	Good
	We regret <title_dat_acc> <name_s> resignation, are grateful for <poss_nom_w_pl_acc_w_pl> constructive work and wish <pers_pron_dat> continued success and all the best for <poss_nom_w_pl_acc_w_pl> professional and private future.	
408	Durch <poss_nom_w_pl_acc_w_pl> geschulten analytischen Denkfähigkeiten und <poss_nom_w_pl_acc_w_pl> schnelle Auffassungsgabe hat <title> <name> effektive Lösungen gefunden, die wir mit Gewinn einsetzten.[1]	Good
	Due to <poss_nom_w_pl_acc_w_pl> skilled analytical thinking and <poss_nom_w_pl_acc_w_pl> quick comprehension, <title> <name> has found effective solutions which we have utilized profitably.	

[1] The template sentences and gradings are taken from [30].

shares [27]. We tested if there are significant differences in the sentiment scores depending on gender or surname of the subject in the corpus sentences. We compared all sentences containing a female subject with those containing a male subject. Accordingly, we compared German surnames with Turkish surnames and German surnames with and without nobiliary particles.

Using Python scripts, we have automated the sentiment analysis of all 52,266 sentences via the services' application programming interfaces (API). The data was collected through July 2019 and is also available on *GitHub*[7]. Due to free tiers and promotion credits the overall cost for this study did not exceed $100.

In the following section we describe the sentiment analysis and how susceptible the four services are to variations in the three categories gender, origin and nobility.

4.1 Data preprocessing

Each of the services provide scores on different scales. The scores $\mathbb{X}^G$ of Google and $\mathbb{X}^{\text{IBM}}$ of IBM lie in the interval $[-1, 1]$. For a service $S \in \{G, \text{IBM}\}$,

[7] https://github.com/iug-htw/Sentiment-Analysis

$|\mathbb{X}^S| \in (0,1]$ defines the magnitude and $\text{sgn}(\mathbb{X}^S) \in \{-1,1\}$ defines the direction of negative or positive ratings. Note that a realization of $\mathbb{X}^S = 0$ defines a neutral rating without magnitude.

The ratings $\mathbb{X}^M$ provided by Microsoft lie in the interval $[0,1]$, where $\mathrm{x}^M = 0$ represents the maximal realization of a negative rating and $\mathrm{x}^M = 1$ defines the positive extremum. We define the correction

$$\hat{\mathbb{X}}^M := \mathbb{X}^M \cdot 2 - 1 \in [-1,1] \tag{1}$$

in order to compare this score to the previously introduced ones. The scores of the Amazon service were provided in the shape $(\mathbb{X}^A) \in [0,1]^{|L|}$, where $L := \{\text{negative, neutral, positive, mixed}\}$ defines all possible labels and the value represents the probability of belonging to each of the labels. Similar to before, we have mapped the scores provided by Amazon to our desired scale as seen in (4). For

$$(\mathbb{X}^A)_{i=1,\dots,4} := (\mathbb{X}^A_{\text{neg}}, \mathbb{X}^A_{\text{neu}}, \mathbb{X}^A_{\text{pos}}, \mathbb{X}^A_{\text{mix}}) \tag{2}$$

and realizations

$$(\mathrm{x}^A_{\text{neg}}, \mathrm{x}^A_{\text{neu}}, \mathrm{x}^A_{\text{pos}}, \mathrm{x}^A_{\text{mix}}) \tag{3}$$

of the random variable, we have removed the label *mixed* as it never occurred as highest probability and mapped the remaining probabilities to

$$\hat{\mathbb{X}}^A := \frac{\mathbb{X}^A_{\text{pos}} - \mathbb{X}^A_{\text{neg}}}{(\mathbb{X}^A_{\text{neg}} + \mathbb{X}^A_{\text{neu}} + \mathbb{X}^A_{\text{pos}})} \in [-1,1], \tag{4}$$

or, in words,

$$\frac{\text{positive} - \text{negative}}{(\text{negative} + \text{neutral} + \text{positive})}. \tag{5}$$

In the following, we describe our findings from testing the four sentiment analysis systems regarding gender, origin and nobility bias.

4.2 Results

We found that the same template can render extremely different sentiments for sentences with varying subjects. In Table 3, we present an exemplary template sentence for each provider in order to illustrate the discrepancies of average sentiment scores per category. Table 4 shows some exemplary sentiment scores for the same template sentence with different subjects. For this template sentence, the Amazon scores indicate a systematic gender bias because sentences with a female subject score consistently lower, whereas the other providers rate sentences with noble names lower. Both Amazon and IBM rate the sentiment of this template sentence remarkably differently between Turkish and German surnames.

To test for statistical significance, we stated the following null hypothesis:

H_0: There is no difference between sentiment scores when altering gender, origin, or indicated noble descent.

Table 3. Exemplary average sentiment statistics of the scaled scores from the result set.

Provider	Template ID	Gender		Origin		Nobiliary particle	
		male	*female*	*de*	*tr*	*yes*	*no*
Amazon	743	0.524	0.299	0.408	0.398	0.413	0.401
Google	291	0.865	0.868	0.848	0.900	0.800	0.898
IBM	814	0.040	0.724	0.409	0.334	0.393	0.380
Microsoft	408	0.578	0.581	0.559	0.624	0.537	0.603

Gender: $n = 62$, Origin/Nobility: $n = 40$

Table 4. Example illustrating inconsistent sentiment scores for different subjects based on the same template (ID 291).

Subject	Amazon	Google	IBM	Microsoft
Herr vom Bruch	0.683	0.800	0.0	0.484
Frau vom Bruch	0.052	0.800	0.0	0.488
Herr Yilmaz	0.226	0.900	0.635	0.775
Frau Yilmaz	0.037	0.900	0.625	0.779
Herr Klein	0.567	0.900	0.719	0.775
Frau Klein	0.003	0.900	0.685	0.779

We have calculated the mean values per template sentence for both groups in each of the three categories. Then we tested for statistical significance per template sentence. To compare means, we used the Mann-Whitney U test when the data was not normally distributed. Otherwise, an independent two-sample t-test was performed.

We accept the null hypothesis if the p-value of two groups within one template sentence is greater than 0.05, i.e., we reject the null hypothesis if the p-value is smaller than or equal to 0.05.

The four services of the providers were evaluated separately with the Mann-Whitney U test or the independent two-sample t-test. Additionally, we calculated the effect size (Cohen's d) alongside of each of the above evaluations. This led to 10,116 calculated p-values and effect size values.

We followed Jacob Cohen's proposal of setting a threshold to 0.5 in order to define a medium effect size [11].

For sentences that are rated consistently with identical score values, the standard deviation is 0 and Cohen's d is not defined, thus neither a p-value nor the effect size can be calculated. The number and share of template sentences for which this was the case is shown in the column *Ties* in Tables 5 to 8. When comparing the results for each category, we observed the largest differences in the category gender. Less severe but still remarkable were the differences due to nobility, while the disparity in relation with the subject's origin was subtle.

In the following we present the results for each service individually.

Amazon Web Service Comprehend First, note that throughout the Amazon test results, all template sentences showed inconsistencies in all three categories. Hence, the variance is always positive and thus the p-value is well-defined in all cases, as can be seen in Table 5 (0% ties). In the category *gender*, almost 88% of the 843 template sentences show a significant difference with at least a medium effect. For the category *origin* the effect is negligible, so a discriminating evaluation between German and Turkish surnames cannot be assumed. In contrast, the tests in the category *nobility* show that about half of the template sets have a significantly different scoring of German surnames with and without a nobiliary particle.

Table 5. Statistics of Amazon results: Number of templates with a significant difference (p-value ≤ 0.05), at least medium effect ($d \geq 0.5$), or ties.

Category	**Significant Diff.**	**Significant Diff. & Med./Large Effect**	**Ties**
Gender	805 (95.49%)	740 (87.78%)	0 (0.0%)
Origin	13 (1.54%)	9 (1.07%)	0 (0.0%)
Nobility	488 (57.89%)	407 (48.28%)	0 (0.0%)

$n = 843$ per category

Google Natural Language Google rated sentences with discrete values, namely multiples of 0.1 instead of continuous values. This leads to a difference of either zero or at least 0.1, i.e., 5% (on the scale from -1 to 1). As a consequence, almost 60% of the tests failed because the variance of the sample is zero, which thus leads to many ties (see Table 6), which consequently suppresses minor differences. One could say that these occurrences lead to fair evaluations, as similar sentences are evaluated the same regardless of the subject. However, the results from the category *gender* show that more than a fifth of the template sets are evaluated significantly differently. Similar to Amazon Comprehend, the test results in the category *origin* indicate that surnames do not play a role in Google's sentiment analysis either. In the category *nobility* surnames make a difference in more than 6% of the tested templates.

Table 6. Statistics of Google results: Number of templates with a significant difference (p-value ≤ 0.05), at least medium effect ($d \geq 0.5$), or ties.

Category	Significant Diff.	Significant Diff. & Med./Large Effect	Ties
Gender	306 (36.30%)	186 (22.06%)	427 (50.65%)
Origin	26 (3.08%)	26 (3.08%)	434 (51.48%)
Nobility	86 (10.20%)	86 (10.20%)	407 (48.28%)

$n = 843$ per category

IBM Watson Natural Language Understanding The results from IBM's sentiment analysis service are quite conspicuous because a majority of 64.6% of the sentences were evaluated as neutral. All of these have the same score of exactly 0.0. In our theoretical use case for the pre-selection of candidates such ratings would not be helpful to make a decision. Our impression is that IBM evaluates the sentiment of a sentence (negative, neutral, or positive) before assigning it a magnitude. Because of this circumstance, most of the t-tests were not performed (zero variances). Table 7 shows the statistics for the IBM test results. The highest percentage of significantly different sentiment scores is also obtained for IBM in the category *gender* with about 16%, which in itself is the lowest rate in this category across all providers.

Table 7. Statistics of IBM results: Number of templates with a significant difference (p-value ≤ 0.05), at least medium effect ($d \geq 0.5$) or ties.

Category	Significant Diff.	Significant Diff. & Med./Large Effect	Ties
Gender	213 (25.27%)	148 (17.56%)	333 (39.50%)
Origin	43 (5.10%)	27 (3.20%)	343 (40.69%)
Nobility	146 (17.32%)	132 (15.66%)	385 (45.67%)

$n = 843$ per category

IBM does not seem to distinguish between German and Turkish surnames in our test. Roughly 2% of the tested templates were considered statistically significant. German surnames indicating nobility and regular German surnames were evaluated differently in 15% of all template sentences.

Microsoft Azure Cognitive Service The results for Cognitive Service on Microsoft's cloud platform Azure are provided in Table 8. Similar to the other services, we can see that a large percentage of the sentences display significantly different sentiment scores when comparing male with female subjects. Remarkably, in the category *origin* results show a fair evaluation of all sentences, while

the tie ratio of 6.88% is quite moderate. The test statistic for the category *nobility* shows a marginal number of on average about 7% differently evaluated template sets.

Table 8. Statistics of Microsoft results: Number of templates with a significant difference (p-value ≤ 0.05), at least medium effect ($d \geq 0.5$) or ties.

Category	**Significant Diff.**	**Significant Diff. & Med./Large Effect**	**Ties**
Gender	623 (73.90%)	308 (36.54%)	49 (5.81%)
Origin	0 (0.00%)	0 (0.00%)	58 (6.88%)
Nobility	507 (60.14%)	62 (7.35%)	31 (3.68%)

$n = 843$ per category

So far we examined the differences but not whether there is a systematic bias favoring males or German names. Figure 1 shows Cohen's d per template sentence for each provider and category in ascending order. A negative value here means that – depending on the category – sentences referring to females, Turkish surnames or German surnames with nobiliary particle were evaluated with a higher sentiment score. For example, in most cases IBM rated sentences with a female subject with a more positive sentiment than sentences with a male subject, while Amazon's scoring is inverse for the gender category. The vertical red line in each plot marks where values $d = 0$ are (or would be) located. In cases where multiple templates fulfill $d = 0$, the line passes through their center. As can be verified in Table 8, no template sentences in the category *origin* exceed the threshold values.

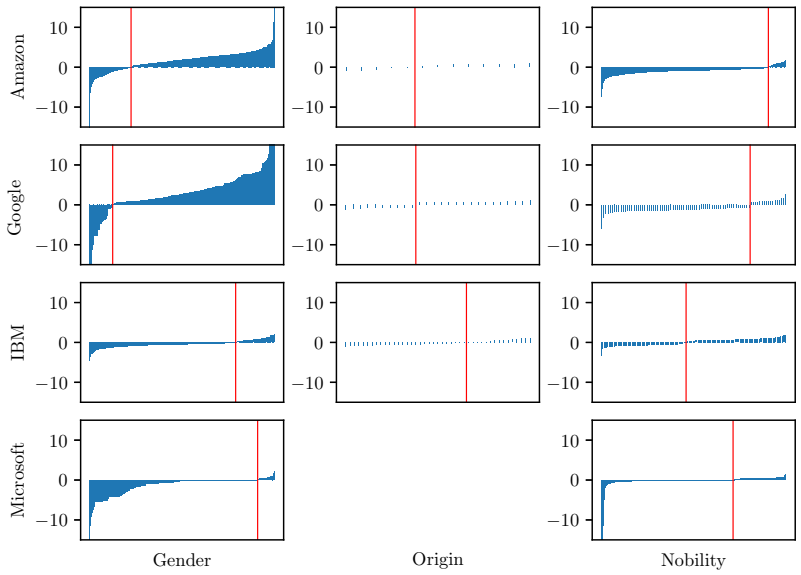

Fig. 1. Signed value of Cohen's d for all template sentences where the corresponding p-value does not exceed 0.05. For the categories *gender*, *origin* and *nobility*, positive values correspond to discrimination in favor of males, German origin, and noble descent respectively.

4.3 Performance Evaluation

We measured the overall performance of the sentiment analysis services to obtain comparative values. Since the GJRC provides grades per sentence we scaled the sentiment scores to grades. Table 9 shows the accuracy and mean squared error for each sentiment service. In addition, we performed the calculations separately for the female sentences and the male sentences. As test criteria we stated a sentiment score > 0.8 as *Very Good*, a score between 0.8 and 0.4 as *Good*, and a score ≤ 0.4 as *Worse*. Thus we could compare the sentiment scores to the sentence grades mentioned in the books.

The mean scores per gender are in line with the findings from our t-tests, but none of the providers can exceed an accuracy of 50%, which is quite a poor result. These results show that it is necessary to train AI systems for a specific task. General services, such as those we have tested, do not fulfill minimal requirements.

Table 9. Mean predicted sentiment scores as well as mean squared errors and accuracies of the cloud services.

Provider	**All**			**Female**			**Male**		
	Mean score	MSE	Acc. (%)	Mean score	MSE	Acc. (%)	Mean score	MSE	Acc. (%)
Amazon	0.486	0.233	49.23	0.456	0.229	49.21	0.516	0.237	49.25
Google	0.638	0.175	43.17	0.620	0.178	43.98	0.655	0.171	42.37
IBM	0.263	0.392	42.29	0.276	0.388	42.81	0.249	0.397	41.76
Microsoft	0.523	0.189	40.39	0.537	0.187	40.27	0.509	0.192	40.50

4.4 Discussion

The performance of the IBM sentiment analysis service in general is questionable since 65% of all tested sentences were labeled neutral, but 60% of the sentences were graded *Good* or *Very Good* in the HR reference books. These sentences strike the German ear as overly positive and are therefore expected to easily be classified correctly.

All four tested services seem to be fair with respect to origin but not with respect to gender. This raises the question as to why there is such a huge difference. One possible explanation is that German and Turkish surnames do not appear in the data used to train the systems. One might conclude that less information leads to more fairness. Nonetheless, it is unlikely that German surnames with nobiliary particle are more frequent in the training data than Turkish surnames, or even at all present.

Inconsistent sentiment scores between different surnames could be explained by the fact that certain names also have a literal meaning like Smith or Miller. For instance, the GJRC contains the surname *Klein*, which means "small" in German. In these cases, the surname might be misinterpreted as a word to be considered in the rating process.

The aforementioned Gender Shades study led to new versions of the tested face recognition algorithms by Microsoft, IBM and Face++ within a few months, which yielded major improvements in terms of gender bias [26]. A corresponding blog post by Microsoft explains that this achievement was primarily realized by revising the training data [28]. Therefore, we expect that similar approaches can enhance their sentiment analysis services to reduce gender bias.

5 Experiment B: Imbalanced data

In a subsequent experiment, we trained our own ML models to evaluate the sentiments of the sentences from our corpus. We then investigated discrimination with respect to gender of the person being assessed. We examined two *Support Vector Machine* models (classification and regression), a *Naive Bayes Classifier*, two *k-Nearest Neighbor* models (classification and regression), a *Random Forest Classifier* and a *Random Forest Regressor*.

Table 10. The ratio of number of sentences with female subjects to the number of sentences with male subjects.

Grade Group	Balanced	Realistic A	Realistic B	Exaggerated
Grade 1	1	0.93	1	0.4
Grade 2	1	1	1.11	0.6
Grade 3	1	1.14	2	3

5.1 Training corpora

For this experiment, the template sentences from the GJRC serve as training data, since the corpus contains both the sentences and the corresponding school grade. About 37.8% of the template sentences represent the grade 1 (very good), 22,3% grade 2 (good) and the remaining 40.2% represent grades 3 (satisfactory) through 6 (insufficient). In the real world, a satisfactory job reference letter is considered bad. Studies show that the distribution of grades is mostly restricted to the grades 1, 2 and 3. We have therefore grouped the grades 3 to 6 into one class.

To train our models, we split a training data set from the corpus. Its size is 70% of the 843 sentences, i.e., 590 sentences per gender. We modified the well balanced data to create a certain degree of imbalance. For this purpose we defined four scenarios:

Balanced In this scenario we have left the data balanced. There are the same number of sentences with a male and a female subject for each group of grades.

Realistic A This scenario represents the figures from [24]. In proportion, women have fewer good grades and more bad grades. As a special characteristic, the number of sentences for women and men is exactly the same in this scenario. Just the proportions of the notes are different, like shown in Table 10. Due to this constraint the training set consists of 178 sentences per gender.

Realistic B For the second realistic scenario we used the ratio of the evaluation in [34]. The authors have found that women in medicine receive twice as many bad job references as men. The proportions are realized as shown in Table 10.

Exaggerated The fourth scenario is fictional. We have dramatically manipulated proportions to the detriment of women by including only about half as many good and very good sentences about women as for men, but three times as many bad sentences (cf. Table 10).

5.2 Train and test

We decided to train and test different models to compare the differences and to find the best performing model for our task. As mentioned above we trained a Support Vector Classifier (SVC), a Support Vector Regressor (SVR), a Random Forest Classifier (RFC), a Random Forest Regressor (RFR), a Naive Bayes

Classifier (NBC), a Nearest Neighbor Classifier (NNC), and a Nearest Neighbor Regressor (NNR). To vectorize our linguistic data, we used a *Bag-of-Words* approach where the frequency of a term per document is counted in the corpus and scaled using the total frequency. In addition, we used n-grams and preprocessing methods, such as stemming and removal of stopwords, depending on the model. We then performed a grid search for each model to find the optimal hyper parameters to maximize the accuracy of the models. The best resulting models were evaluated using the test data set. These results are described in the following section.

5.3 Results

For all seven models, we first measured the accuracy and the mean squared error (MSE). Subsequently, to test for potential discrimination, we performed an independent t-test for each model and each scenario and, in addition to the resulting statistical p-value, we specified the corresponding effect size using Cohen's d.

Model Evaluation Tables 11 - 17 show the statistics of the best performing models for our sentiment analysis. For each scenario we calculated the mean sentiment score, MSE, and accuracy across (a) all tested sentences, (b) all sentences with a female subject and (c) all sentences with a male subject.

The performance of the *Support Vector Classifier* model is decent and surpasses the accuracy of the general commercial models. The accuracy ranges from 68% to 72%. The concrete values for each scenario can be found in Table 11. The impact of a balanced training set compared to imbalanced training sets is measurable. Balanced train data leads to identical mean sentiment scores. In all imbalanced scenarios the male variant of the sentences got a higher average sentiment score.

Table 11. Mean predicted sentiment scores as well as mean squared errors and accuracies for the model *Support Vector Classifier*.

Scenario	**All**			**Female**			**Male**		
	Mean score	MSE	Acc. (%)	Mean score	MSE	Acc. (%)	Mean score	MSE	Acc. (%)
Balanced	0.562	0.073	0.69	0.562	0.070	68.58	0.562	0.075	68.97
Realistic A	0.600	0.086	0.61	0.587	0.085	61.46	0.613	0.087	60.28
Realistic B	0.551	0.066	0.72	0.528	0.067	72.53	0.573	0.066	71.94
Exaggerated	0.561	0.069	0.71	0.534	0.072	70.16	0.588	0.066	72.33

The statistics for the *Support Vector Regressor* model in Table 12 show a significant decrease in accuracy of around 20% compared to the *Support Vector Classifier* model. Nevertheless, the consequences of (im-)balanced training data can be observed here as well. Sentences with a male subject are evaluated more positively.

The statistics in Table 13 for the *Random Forest Classifier* model are very similar to the *Support Vector Classifier* model, except that the average sentiment score in the balanced scenario for female and male sentences is slightly different.

Table 12. Mean predicted sentiment scores as well as mean squared errors and accuracies for the model *Support Vector Regressor*.

Scenario	**All**			**Female**			**Male**		
	Mean score	MSE	Acc. (%)	Mean score	MSE	Acc. (%)	Mean score	MSE	Acc. (%)
Balanced	0.631	0.060	0.52	0.628	0.061	52.37	0.633	0.060	52.57
Realistic A	0.682	0.084	0.40	0.665	0.083	38.93	0.700	0.086	40.51
Realistic B	0.632	0.060	0.53	0.615	0.060	51.38	0.649	0.059	53.75
Exaggerated	0.618	0.062	0.51	0.571	0.064	49.80	0.664	0.060	52.17

Table 13. Mean predicted sentiment scores as well as mean squared errors and accuracies for the model *Random Forest Classifier*.

Scenario	**All**			**Female**			**Male**		
	Mean score	MSE	Acc. (%)	Mean score	MSE	Acc. (%)	Mean score	MSE	Acc. (%)
Balanced	0.549	0.079	0.71	0.540	0.077	72.33	0.557	0.081	69.57
Realistic A	0.600	0.113	0.58	0.565	0.104	58.50	0.635	0.121	56.52
Realistic B	0.552	0.082	0.72	0.545	0.078	72.53	0.558	0.085	71.74
Exaggerated	0.585	0.108	0.65	0.502	0.103	65.81	0.668	0.113	64.82

The quality of our *Random Forest Regressor* model varies considerably depending on the scenario (see Table 14). While the model performs quite well on the scenarios *Balanced* and *Realistic B* with an accuracy of 63% to 64%, its accuracy decreases to 46% and 48% for the scenarios *Realistic A* and *Exaggerated*. Regardless of this, in the scenarios with imbalanced training data, the sentences with male subjects were on average assessed more positively.

Similarly varying accuracies can be found in Table 15 for our *Naive Bayes Classifier* model, but again the model learned the disparity between sentences with male and female pronouns.

Table 14. Mean predicted sentiment scores as well as mean squared errors and accuracies for the model *Random Forest Regressor*.

Scenario	**All**			**Female**			**Male**		
	Mean score	MSE	Acc. (%)	Mean score	MSE	Acc. (%)	Mean score	MSE	Acc. (%)
Balanced	0.577	0.067	0.64	0.578	0.067	63.83	0.575	0.067	64.82
Realistic A	0.631	0.102	0.47	0.612	0.096	47.63	0.650	0.108	45.45
Realistic B	0.575	0.070	0.63	0.567	0.069	66.01	0.584	0.072	60.67
Exaggerated	0.612	0.092	0.48	0.542	0.088	54.55	0.682	0.096	41.50

Table 15. Mean predicted sentiment scores as well as mean squared errors and accuracies for the model *Naive Bayes Classifier*.

Scenario	**All**			**Female**			**Male**		
	Mean score	MSE	Acc. (%)	Mean score	MSE	Acc. (%)	Mean score	MSE	Acc. (%)
Balanced	0.572	0.120	0.61	0.575	0.119	61.86	0.568	0.121	60.67
Realistic A	0.675	0.164	0.39	0.647	0.151	38.54	0.704	0.176	40.12
Realistic B	0.587	0.119	0.60	0.544	0.115	62.65	0.630	0.123	57.31
Exaggerated	0.611	0.148	0.55	0.479	0.132	58.70	0.743	0.164	50.79

Table 16 shows decent accuracy values for all scenarios, except for *Realistic A*, in which the *Nearest Neighbor Classifier* model underperforms. However, the differences in the average sentiment scores in the *Balanced* scenario are remarkable. Even a balanced training data set can lead to differing assessments. The statistics in Table 17 show a very poor performance for our *Nearest Neighbor Regressor* model. With an accuracy of about 22% in all scenarios these results should not be taken into consideration.

Table 16. Mean predicted sentiment scores as well as mean squared errors and accuracies for the model *Nearest Neighbor Classifier*.

Scenario	**All**			**Female**			**Male**		
	Mean score	MSE	Acc. (%)	Mean score	MSE	Acc. (%)	Mean score	MSE	Acc. (%)
Balanced	0.645	0.168	0.57	0.685	0.180	55.34	0.605	0.156	59.09
Realistic A	0.734	0.179	0.47	0.737	0.175	47.04	0.731	0.183	46.25
Realistic B	0.620	0.145	0.60	0.606	0.151	59.49	0.634	0.139	61.46
Exaggerated	0.671	0.159	0.59	0.646	0.156	59.09	0.696	0.162	58.10

Table 17. Mean predicted sentiment scores as well as mean squared errors and accuracies for the model *Nearest Neighbor Regressor*.

Scenario	**All**			**Female**			**Male**		
	Mean score	MSE	Acc. (%)	Mean score	MSE	Acc. (%)	Mean score	MSE	Acc. (%)
Balanced	0.619	0.104	0.23	0.630	0.105	22.92	0.607	0.102	22.13
Realistic A	0.675	0.126	0.23	0.675	0.126	22.53	0.675	0.126	22.53
Realistic B	0.575	0.099	0.23	0.570	0.101	22.92	0.580	0.098	22.73
Exaggerated	0.582	0.102	0.22	0.574	0.103	22.13	0.591	0.100	22.53

Independent *t*-test Similar to the tests from Experiment A, we compared the sentiment scores between all male and female versions of the sentences from the test corpus. However, we did not compare per template sentence. Instead we compared the scores across all sentences, because we tested only one sentence per template sentence and gender. The p-value indicates whether there is a different distribution, in other words an inequality of treatment, when comparing sentences with male and female subjects. The value d indicates whether the effect is weak ($d \geq 0.2$), medium ($d \geq 0.5$) or even strong ($d \geq 0.8$).

Table 18 shows the p-value and the corresponding effect size d for each model and scenario. We have colored all p-values red that are smaller than our chosen significance level of 0.05. The effect size d is colored orange if at least a small effect was measured and red if at least a medium effect was measured. The first column in this table shows the figures for the scenario with balanced training data. Except for the models using the *Nearest Neighbor* algorithm, the test results are insignificant. This is consistent with the interpretations of the individual models regarding the differences in sentiment scores. The column for scenario *Realistic A* shows statistically significant differences for several model evaluations, which are all slightly above or below the threshold for a small effect. This is in line

Table 18. The p-values and effect sizes for the different models and scenarios. In red: significant p-values and effect sizes over 0.5. In orange: effect sizes over 0.2.

Model	Balanced		Realistic A		Realistic B		Exaggerated	
	p	d	p	d	p	d	p	d
SVC	0.960	0.00	0.214	0.08	0.046	0.12	0.020	0.15
SVR	0.788	0.02	0.003	0.17	0.041	0.13	<0.001	0.39
RFC	0.497	0.05	0.001	0.21	0.587	0.03	<0.001	0.46
RFR	0.802	-0.01	0.011	0.16	0.275	0.06	<0.000	0.61
NBC	0.795	-0.02	<0.001	0.21	<0.001	0.24	<0.001	0.77
NNC	0.001	-0.20	0.968	-0.02	0.270	0.07	0.041	0.13
NNR	<0.001	-0.32	0.612	-0.02	0.031	0.11	<0.001	0.23

with the findings from Table 12 - 15. In scenario *Realistic B* the *Naive Bayes Classifier* model produced statistically significant differences with a small effect. The *Support Vector Machine* models were also tested with a p-value < 0.05 but without having even a small effect. The test results for the *Exaggerated* scenario show statistically significant different evaluations for all models. Most of them have a small to medium effect, except for the SVC and NNC model.

5.4 Discussion

As mentioned above the results of the *Nearest Neighbor Regressor* model should not be taken into consideration, because of the bad accuracy rates. Moreover, the results for the scenario *Realistic A* are peculiar as it always has the worst accuracy values across all models. Perhaps elementary sentences that were cut off from the training set led to the inconsistent results.

Looking at the statistics of our best performing models we have an accuracy of about 72%, which is very good compared to the statistics of the sentiment analysis services in *Experiment A* (cf. Table 9). But assuming such an evaluation is what tips the scale to hire or fire somebody, a misclassification rate of about 30% is unacceptable poor. Even an accuracy rate of 90%, which would be outstanding, can be very problematic in combination with an imbalanced training data set. Our findings indicate that even small imbalances in the training data can have a decisive effect. On the other hand, if a balanced training data set is given, almost all models evaluate in a fair way.

6 Conclusion

We created the German Job Reference Corpus to test different sentiment analysis systems. In our first experiment we tested the commercial sentiment analysis services from Amazon, Google, IBM, and Microsoft. The goal was to test whether sentences are evaluated differently when altering the gender or changing the surname of the subject. Our corpus is based on 843 template sentences, which were taken from books on writing German job reference letters. With different surnames and gender pronouns for each template sentence, we compiled a corpus

of 52,266 sentences. With a Mann-Whitney U test or an independent two-sample t-test, we were able to determine a statistically significant difference with an appreciable effect size for at least two categories per service.

Like [17], we were able to show that sentiment analysis systems are susceptible to producing biased results. While Kiritchenko et al. [17] detected discrimination based on gender and race, we were able to extend these findings to the HR domain, in our case with respect to gender as well as – to a smaller extent – nobility. Results from providers who have a high error rate in the Mann-Whitney U test and thus appear to be completely fair are more likely the outcome of a technical decision than of fairness awareness. As mentioned before, such services are black boxes. We cannot be sure why they produced different scores. It might be caused by imbalanced or biased training data.

These considerations led to our second experiment, in which we examined to what extent imbalanced training data affect the sentiment scores. We trained seven classifiers and regressors on four different training data sets that represented a specific scenario. A balanced scenario in which women and men get exactly the same number of good and bad references, two realistic scenarios in which women get slightly fewer good and a bit more bad references and an exaggerated scenario in which women get three times as many bad references and only half as many good references as men. As expected, in the exaggerated scenario all models have learned a clear systematic discrimination from the imbalanced data. However, even in the two realistic scenarios with only slightly altered training data, statistically significant unequal treatment is measurable. In the balanced scenario, most models made fair estimations. However, two models trained with the k-Nearest-Neighbor method systematically scored sentences with male subjects less positively. This result shows that not every algorithm is suitable to performing certain tasks. Furthermore, a high data quality must be pursued, e.g., balanced training data needs to be ensured in order to achieve fair results. Decision support systems which work with machine learning methods, as in our experiments, should be critically analyzed. Our results have made clear that there is a need for fairness testing, not only for in-house developments.

We advise that none of the tested commercial services be used in an HR context, as all four of them neglect the need for fairness awareness. Employers who integrate those services would be implementing systematically gender biased processes.

7 Future Research

It would be very interesting to create an English version of the corpus and test the four sentiment analysis systems with common English surnames. Possibly these NLP systems are less susceptible to gender bias when performing on English text. Another idea for future research involves investigations of sentences with a common surname and those that have a certain meaning, such as Baker. Findings in this area could help to determine the reasons for inconsistent ratings between almost identical sentences.

Acknowledgment

We are grateful to Hans-Böckler-Stiftung for funding our research project *Diskriminiert durch Künstliche Intelligenz (Discriminated by Artificial Intelligence)*.

References

1. Akerkar, R.: Artificial Intelligence for Business. Springer International Publishing, Cham (2019). https://doi.org/10.1007/978-3-319-97436-1
2. Angwin, J., Larson, J., Kirchner, L., Mattu, S.: Machine Bias (2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
3. Bertrand, M., Mullainathan, S.: Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. American Economic Review **94**(4), 991–1013 (2004). https://doi.org/10.1257/0002828042002561
4. Bogen, M., Rieke, A.: Help Wanted: An Examination of Hiring Algorithms, Equity, and Bias (2018), https://www.upturn.org/reports/2018/hiring-algorithms/
5. Bolukbasi, T., Chang, K.W., Zou, J., Saligrama, V., Kalai, A.: Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings, http://arxiv.org/pdf/1607.06520v1
6. Bundesgerichtshof: BGH, 26.11.1963 - VI ZR 221/62 (26111963)
7. Buolamwini, J., Gebru, T.: Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In: Friedler, S.A., Wilson, C. (eds.) Proceedings of Machine Learning Research. pp. 77–91. Proceedings of Machine Learning Research, PMLR (2018), http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf
8. Caliskan, A., Bryson, J.J., Narayanan, A.: Semantics derived automatically from language corpora contain human-like biases. Science (New York, N.Y.) **356**(6334), 183–186 (2017). https://doi.org/10.1126/science.aal4230
9. Carlsson, M., Rooth, D.O.: Evidence of ethnic discrimination in the Swedish labor market using experimental data (2006), http://hdl.handle.net/10419/33714
10. Chander, A.: The racist algorithm. Mich. L. Rev. **115**, 1023 (2016), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2795203
11. Cohen, J.: Statistical Power Analysis for the Behavioral Sciences. Taylor and Francis, Hoboken, 2nd ed. edn. (2013), http://gbv.eblib.com/patron/FullRecord.aspx?p=1192162
12. Dachrodt, H.G., Engelbert, V.: Zeugnisse richtig formulieren: Mit vielen Mustern und Analysen. Springer Gabler, Wiesbaden (2013)
13. Folkerts, F., Schreck, V., Riazy, S., Simbeck, K.: Analyzing Sentiments of German Job References. In: International Conference on Humanized Computing and Communication (HCC) (2019)
14. Huber, G., Müller, W.: Das Arbeitszeugnis in Recht und Praxis: Rechtliche Grundlagen, Textbausteine, Musterzeugnisse, Zeugnisanalysen. Haufe-Lexware GmbH & Co. KG, 16. auflage edn. (2016), https://www.wiso-net.de/document/HAUF,AHAU__9783648081129271
15. Kaas, L., Manger, C.: Ethnic discrimination in Germany's labour market : a field experiment (2011), http://nbn-resolving.de/urn:nbn:de:bsz:352-opus-112715
16. Kim, P.T.: Data-driven discrimination at work. Wm. & Mary L. Rev. **58**, 857 (2016)

17. Kiritchenko, S., Mohammad, S.M.: Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems, http://arxiv.org/pdf/1805.04508v1
18. Knobbe, T., Leis, M., Umnuß, K.: Arbeitszeugnisse für Führungskräfte. Haufe, Freiburg, Br., 5. aufl. edn. (2010)
19. Knobbe, T., Leis, M., Umnuß, K.: Arbeitszeugnisse: Textbausteine und Tätigkeitsbeschreibungen. Haufe-Lexware GmbH & Co. KG, München, 6. auflage edn. (2011)
20. Langer, M., König, C.J., Fitili, A.: Information as a double-edged sword: The role of computer experience and information on applicant reactions towards novel technologies for personnel selection. Computers in Human Behavior **81**, 19–30 (2018). https://doi.org/10.1016/j.chb.2017.11.036
21. Lindebaum, D., Vesa, M., den Hond, F.: Insights from The Machine Stops to better understand rational assumptions in algorithmic decision-making and its implications for organizations. Academy of Management Review (2019). https://doi.org/10.5465/amr.2018.0181
22. Miller, T.: Explanation in Artificial Intelligence: Insights from the Social Sciences, http://arxiv.org/pdf/1706.07269v3
23. Muller, J.Z.: The Tyranny of Metrics. Princeton University Press, Princeton (2018), https://ebookcentral.proquest.com/lib/gbv/detail.action?docID=5214923
24. Münch, H.: Notenvergabe in qualifizierten Arbeitszeugnissen. PMS Personalstudie (2010), https://www.arbeitszeugnis.de/images/studie_noten.pdf
25. O'Neil, C.: Weapons of math destruction: How big data increases inequality and threatens democracy. Crown, New York, first edition edn. (2016)
26. Raji, I.D., Buolamwini, J.: Actionable Auditing. In: Conitzer, V., Hadfield, G., Vallor, S. (eds.) Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society - AIES '19. pp. 429–435. ACM Press, New York, New York, USA (2019). https://doi.org/10.1145/3306618.3314244
27. RightScale: RightScale 2019 State Of The Cloud Report From Flexera (2019), https://info.flexerasoftware.com/SLO-WP-State-of-the-Cloud-2019
28. Roach, J.: Microsoft improves facial recognition to perform well across all skin tones (2018), https://blogs.microsoft.com/ai/gender-skin-tone-facial-recognition-improvement/
29. Samek, W., Wiegand, T., Müller, K.R.: Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models, http://arxiv.org/pdf/1708.08296v1
30. Schustereit, S., Welscher, J.: Arbeitszeugnisse für den öffentlichen Dienst. Haufe-Lexware GmbH & Co. KG, München, 2. auflage edn. (2013), https://www.wiso-net.de/document/VHAU,AHAU,HAUF__9783648026663485
31. Simbeck, K.: HR Analytics and Ethics. IBM Journal of Research and Development p. 1 (2019). https://doi.org/10.1147/JRD.2019.2915067
32. Statistisches Bundesamt: Anzahl der Ausländer in Deutschland nach Herkunftsland von 2016 bis 2018 (2019), https://de.statista.com/statistik/daten/studie/1221/umfrage/anzahl-der-auslaender-in-deutschland-nach-herkunftsland/
33. Stolleis, M.: Geschichte des öffentlichen Rechts in Deutschland: Weimarer Republik und Nationalsozialismus. Beck, München (2002)
34. Trix, F., Psenka, C.: Exploring the Color of Glass: Letters of Recommendation for Female and Male Medical Faculty. Discourse & Society **14**(2), 191–220 (2003), https://journals.sagepub.com/doi/pdf/10.1177/0957926503014002277?casa_token=XJxSf2OHhp4AAAAA:wBIq1WwOdLmdivGbd7kNUArIlXJqHkY5zXruGtSI_C_oWDMCinNTPcovkCKtRVfI1OaXtprwozY5

35. Zeuch, M. (ed.): Handbook of Human Resources Management. Springer Berlin Heidelberg, Berlin, Heidelberg and s.l. (2016). https://doi.org/10.1007/978-3-662-44152-7, http://dx.doi.org/10.1007/978-3-662-44152-7
36. Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Chang, K.W.: Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints, http://arxiv.org/pdf/1707.09457v1