Analyzing Sentiments of German Job References

(Invited Paper)

Finn Folkerts*, Vanessa Schreck†, Shirin Riazy‡ and Katharina Simbeck§
Hochschule für Technik und Wirtschaft Berlin
Treskowallee 8, 10318 Berlin, Germany
*Email: folkerts@htw-berlin.de
†Email: schreckv@htw-berlin.de
‡Email: riazys@htw-berlin.de
§Email: simbeck@htw-berlin.de

***Abstract*—Filling a vacancy takes a lot of (costly) time. Automated pre-processing of applications using artificial intelligence technology can help to save time, e.g., by analyzing applications using machine learning algorithms. We investigated whether such systems are potentially biased in terms of gender, origin and nobility. For this purpose we created a corpus of common German reference letter sentences on which we performed sentiment analysis using the cloud services by Amazon, Google, IBM and Microsoft. We established that all tested services rate the sentiment of the same template sentences very inconsistently and biased at least with regard to gender.**

I. Introduction

In order to improve the efficiency, companies automate their recruiting process by using either external software solutions or self-developed tools. Besides quality of and cost per recruitment, staffing time is an important performance indicator [1]. For this reason, Natural Language Processing (NLP) systems are sometimes used to pre-select candidates [2], [3]. The use of such artificial intelligence (AI) models is expected to to increase productivity [4], but often lacks transparency [5], [6]. The systems are used as black boxes, this non-transparency may obfuscate ethical issues such as discrimination [2], [4], [7]. 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 [9].

In this research we analysed the fairness of commercial sentiment analysis systems. We have extended the work of [9] 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 letter 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.

Our experiments show that all tested sentiment analysis services evaluate almost identical sentences unpredictably different, depending on the subject used in a sentence.

II. Related Work

Several studies that are concerned with bias of NLP systems regarding gender or origin have shown how algorithms reproduce stereotypes. Comparing 200 sentiment analysis systems [9] found evidence for strong bias. In [8], the authors trained a system that learned word associations from the *Common Crawl*[5] corpus via *word embedding* techniques, resulting in replicating stereotypes like women getting more associated with family and arts, while men get more associated with career and sciences. Bolukbasi et al. [10] attempted to develop a methodology to reduce bias in applications using word embedding. However, algorithmic bias is not limited to NLP but also affects computer vision systems. It turned out that one particular training data set contained 33% more pictures with women associated with the activity cooking than men, which amplifies the reproduction of gender based stereotypes when models trained with this data set are being used [11].

The *Gender Shades* study [12] 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.

However, traditional HR processes are biased, too. Several studies have shown that applicants with uncommon or foreign sounding names are discriminated against by potential employers [13]–[15].

III. 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 [16]–[20]. We combined those template sentences with subjects of varying gender, origin and nobility.

[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/
[5]https://commoncrawl.org/

TABLE I
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

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 minority in Germany [21].

Further, we expected higher sentiment scores for names that indicate nobility ("von", "zu"). However, a nobiliary particle does not grant any privileges by law in Germany since 1919 [22]. 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. Then we mapped the names to their origins and randomly picked ten German surnames, ten German surnames with nobiliary particle and ten Turkish surnames. In Table I we listed all surnames that we used to compile the GJRC.

Following a literature review, 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 II.

Note that by law, German reference letters must be phrased favorably to the employee [23], even if the employee 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].

IV. EXPERIMENT

We conducted an experiment using the cloud services for sentiment analysis from Amazon, Google, IBM and Microsoft, because they own the biggest market shares [24]. 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.

A. Data pre-processing

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 $S \in \{G, \text{IBM}\}$, $|\mathbb{X}^S| \in (0, 1]$ 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 $\text{x}^M = 0$ represents the maximal realization of a negative rating and $\text{x}^M = 1$ 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

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

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

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

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

TABLE II
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 utilized profitably.	

[1] The template sentences and gradings are taken from [18].

TABLE III
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

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.

B. Results

We found that the same template can render extremely different sentiments for sentences with varying subjects. In Table IV, we present an exemplary template sentence for each provider in order to illustrate the discrepancies of average sentiment scores per category. Table III 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 the 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 different 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.

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 [25].

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 NaN in Tables V to VIII. 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 solely.

a) 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 V (0% NaN). 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 IV
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 Web Service Comprehend	743	0.524	0.299	0.408	0.398	0.413	0.401
Google Natural Language	291	0.865	0.868	0.848	0.900	0.800	0.898
IBM Watson Natural Language Understanding	814	0.040	0.724	0.409	0.334	0.393	0.380
Microsoft Azure Cognitive Service	408	0.578	0.581	0.559	0.624	0.537	0.603

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

TABLE V
STATISTICS OF AMAZON RESULTS:
NUMBER OF TEMPLATES WITH A SIGNIFICANT DIFFERENCE
(p-VALUE ≤ 0.05), AT LEAST MEDIUM EFFECT ($d \geq 0.5$) OR ERRORS

Category	Significant Diff.	Significant Diff. & Med./Large Effect	NaN
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

TABLE VI
STATISTICS OF GOOGLE RESULTS:
NUMBER OF TEMPLATES WITH A SIGNIFICANT DIFFERENCE
(p-VALUE ≤ 0.05), AT LEAST MEDIUM EFFECT ($d \geq 0.5$) OR ERRORS

Category	Significant Diff.	Significant Diff. & Med./Large Effect	NaN
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

TABLE VII
STATISTICS OF IBM RESULTS:
NUMBER OF TEMPLATES WITH A SIGNIFICANT DIFFERENCE
(p-VALUE ≤ 0.05), AT LEAST MEDIUM EFFECT ($d \geq 0.5$) OR ERRORS

Category	Significant Diff.	Significant Diff. & Med./Large Effect	NaN
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

TABLE VIII
STATISTICS OF MICROSOFT RESULTS:
NUMBER OF TEMPLATES WITH A SIGNIFICANT DIFFERENCE
(p-VALUE ≤ 0.05), AT LEAST MEDIUM EFFECT ($d \geq 0.5$) OR ERRORS

Category	Significant Diff.	Significant Diff. & Med./Large Effect	NaN
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

b) 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 a division error (see Table VI) which consequently suppresses minor differences. One could say that this is completely fair, 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 different. 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.

c) IBM Watson Natural Language Understanding: The results from IBM's sentiment analysis service are extremely 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 are not defined, again due to a division by zero error. Table VII 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.

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.

d) Microsoft Azure Cognitive Service: The results for Cognitive Service on Microsoft's cloud platform Azure are provided in Table VIII. 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 error rate 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.

So far we examined the differences but not whether there is

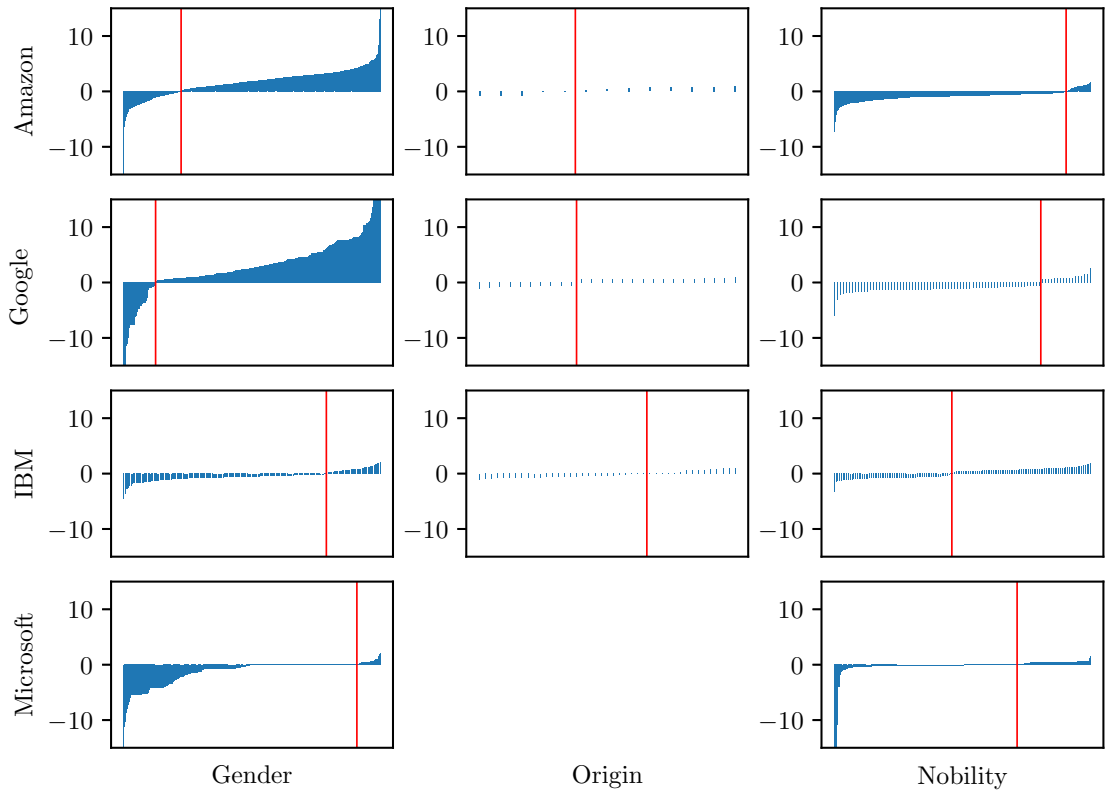

Fig. 1. Absolute value of Cohen's d for all template sentences where the corresponding p-value does not exceed 0.05. For the categories *gender*, *origin* and *nobility*, positive values correspond to discrimination in favor of males, German origin and noble descent respectively.

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 VIII, no template sentences in the category *origin* exceed the threshold values.

V. 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 iteral 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 [27]. Therefore, we expect that similar approaches can enhance their sentiment analysis services to reduce gender bias.

VI. Conclusion

We created the German Job Reference Corpus to test four commercial sentiment analysis systems from Amazon, Google, IBM and Microsoft. The corpus contains 52,266 sentences compiled from 843 template sentences taken from books on writing German job reference letters. Our goal was to test whether sentences are evaluated differently when altering the gender or changing the surname of the subject. 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 [9], we were able to show that sentiment analysis systems are susceptible to producing biased results. While Kiritchenko et al. [9] 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 unbalanced or biased training data.

We advise that none of the tested 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.

VII. 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.

Acknowledgement

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

References

[1] M. Zeuch, Ed., *Handbook of Human Resources Management*. Berlin, Heidelberg and s.l.: Springer Berlin Heidelberg, 2016.

[2] K. Simbeck, "HR Analytics and Ethics," *IBM Journal of Research and Development*, p. 1, 2019.

[3] M. Bogen and A. Rieke. (2018) Help Wanted: An Examination of Hiring Algorithms, Equity, and Bias. [Online]. Available: https://www.upturn.org/reports/2018/hiring-algorithms/

[4] R. Akerkar, *Artificial Intelligence for Business*. Cham: Springer International Publishing, 2019.

[5] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models," 2017.

[6] T. Miller, "Explanation in Artificial Intelligence: Insights from the Social Sciences," 2017.

[7] J. Angwin, J. Larson, L. Kirchner, and S. Mattu. (2016) Machine Bias. [Online]. Available: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[8] A. Caliskan, J. J. Bryson, and A. Narayanan, "Semantics derived automatically from language corpora contain human-like biases," *Science (New York, N.Y.)*, vol. 356, no. 6334, pp. 183–186, 2017.

[9] S. Kiritchenko and S. M. Mohammad, "Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems," 2018.

[10] T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama, and A. Kalai, "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings," 2016.

[11] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang, "Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints," 29.07.2017.

[12] J. Buolamwini and T. Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," in *Proceedings of Machine Learning Research*, ser. Proceedings of Machine Learning Research, S. A. Friedler and C. Wilson, Eds. PMLR, 2018, pp. 77–91. [Online]. Available: http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

[13] M. Bertrand and S. Mullainathan, "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination," *American Economic Review*, vol. 94, no. 4, pp. 991–1013, 2004.

[14] L. Kaas and C. Manger, "Ethnic discrimination in Germany's labour market : a field experiment," 2011.

[15] M. Carlsson and D.-O. Rooth, "Evidence of ethnic discrimination in the Swedish labor market using experimental data," Bonn, 2006.

[16] H.-G. Dachrodt and V. Engelbert, *Zeugnisse richtig formulieren: Mit vielen Mustern und Analysen*. Wiesbaden: Springer Gabler, 2013.

[17] G. Huber and W. Müller, *Das Arbeitszeugnis in Recht und Praxis: Rechtliche Grundlagen, Textbausteine, Musterzeugnisse, Zeugnisanalysen*, 16th ed. Haufe-Lexware GmbH & Co. KG, 2016.

[18] S. Schustereit and J. Welscher, *Arbeitszeugnisse für den öffentlichen Dienst*, 2nd ed. München: Haufe-Lexware GmbH & Co. KG, 2013.

[19] T. Knobbe, M. Leis, and K. Umnuß, *Arbeitszeugnisse für Führungskräfte*, 5th ed. Freiburg, Br.: Haufe, 2010.

[20] ——, *Arbeitszeugnisse: Textbausteine und Tätigkeitsbeschreibungen*, 6th ed. München: Haufe-Lexware GmbH & Co. KG, 2011.

[21] Statistisches Bundesamt. (2019) Anzahl der Ausländer in Deutschland nach Herkunftsland von 2016 bis 2018. [Online]. Available: https://de.statista.com/statistik/daten/studie/1221/umfrage/anzahl-der-auslaender-in-deutschland-nach-herkunftsland/

[22] M. Stolleis, *Geschichte des öffentlichen Rechts in Deutschland: Weimarer Republik und Nationalsozialismus*. München: Beck, 2002.

[23] Bundesgerichtshof, "BGH, 26.11.1963 - VI ZR 221/62," 1963.

[24] RightScale. (2019) RightScale 2019 State Of The Cloud Report From Flexera. [Online]. Available: https://info.flexerasoftware.com/SLO-WP-State-of-the-Cloud-2019

[25] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Hoboken: Taylor and Francis, 2013.

[26] I. D. Raji and J. Buolamwini, "Actionable Auditing," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society - AIES '19*, V. Conitzer, G. Hadfield, and S. Vallor, Eds. ACM Press, 2019, pp. 429–435.

[27] J. Roach. (2018) Microsoft improves facial recognition to perform well across all skin tones. [Online]. Available: https://blogs.microsoft.com/ai/gender-skin-tone-facial-recognition-improvement/