10-9-2023

IMPACT OF DATA COLLECTION ON ML MODELS: ANALYZING DIFFERENCES OF BIASES BETWEEN LOW- VS. HIGH-SKILLED ANNOTATORS

Johannes Schneider
University of Liechtenstein, Germany, johannes.schneider@uni.li

Daniel Eisenhardt
Ruhr-Universität Bochum, Germany, daniel.eisenhardt@rub.de

Christian Utama
Freie Universität Berlin, Germany, christian.utama@fu-berlin.de

Christian Meske
Ruhr-Universität Bochum, Germany, christian.meske@rub.de

IMPACT OF DATA COLLECTION ON ML MODELS: ANALYZING DIFFERENCES OF BIASES BETWEEN LOW- VS. HIGH-SKILLED ANNOTATORS

Research Paper

Johannes Schneider[1], Daniel Eisenhardt[2], Christian Utama[3], and Christian Meske[2]

[1] University of Liechtenstein, Institute of Information Systems, Vaduz, Liechtenstein
{johannes.schneider}@uni.li
[2] Ruhr-University Bochum, Faculties of Mechanical Engineering and Computer Science, Bochum, Germany
{daniel.eisenhardt, christian.meske}@rub.de
[3] Freie Universität Berlin, Department of Information Systems, Berlin, Germany
{christian.utama}@fu-berlin.de

Abstract. Labeled data is crucial for the success of machine learning-based artificial intelligence. However, companies often face a choice between collecting few annotations from high- or low-skilled annotators, possibly exhibiting different biases. This study investigates differences in biases between datasets labeled by said annotator groups and their impact on machine learning models. Therefore, we created high- and low-skilled annotated datasets measured the contained biases through entropy and trained different machine learning models to examine bias inheritance effects. Our findings on text sentiment annotations show both groups exhibit a considerable amount of bias in their annotations, although there is a significant difference regarding the error types commonly encountered. Models trained on biased annotations produce significantly different predictions, indicating bias propagation and tend to make more extreme errors than humans. As partial mitigation, we propose and show the efficiency of a hybrid approach where data is labeled by low-skilled and high-skilled workers.

Keywords: annotators, machine learning models, bias, labeling.

1 Introduction

The phrase “data is the new oil” was coined already in 2006 by Clive Humby. It has ever since been confirmed by the area of big data and AI, which learns decision making capability from data (Rodrigues & Pereira, 2018; Schneider et al., 2022). Especially, labeled data has proven useful in the context of supervised learning enabling a variety of applications (Hötter et al., 2022; ; Saravanan & Sujatha, 2018). Data labeling is the process of assigning labels to raw data. It is essential for supervised machine learning as it provides the necessary information on the ground truth of a data instance for training algorithms to recognize patterns and make accurate predictions.

18th International Conference on Wirtschaftsinformatik,
September 2023, Paderborn, Germany

While unlabeled data is often abundant, data labeling can be costly, since humans typically do it. Humans bare potential biases, alone poor quality data can have a negative impact on the annotator's performance (Cabrera et al., 2014).

Furthermore, human characteristics such as biased decision-making can adversely affect data quality (Barbosa & Chen, 2019; Ding et al., 2022; Hube et al., 2018; Kafkalias et al., 2022). Since personal characteristics and experiences can influence how annotators label data, the skill level of annotators can also influence labeling behavior. For example, highly qualified annotators may have a stronger opinion and thus tend to label data in a more extreme way. In contrast, less qualified annotators may tend to label data more cautiously and thus show a tendency towards the middle. In turn, machine learning models trained on such data can exhibit poor performance, e.g. increasing noise in the data can result in decreasing model accuracies (Caiafa et al., 2020; Barbosa & Chen, 2019; Geva et al., 2019; Zhang et al., 2014; Zhu & Wu, 2004). While there is rich literature on crowdsourcing and biases of annotators, the impact of biases originating from low-skilled and high-skilled annotators on model performance has received only limited attention (e.g., Snow et al., 2008; Wassem, 2016). So far, results have been inconclusive showing that non-expert annotations can outperform expert annotations (Waseem, 2016), can yield acceptable data quality (Irshad et al., 2014; O'Neil et al., 2017; Warby et al., 2014; Zhang et al., 2020), show only little differences to expert annotations (See et al., 2023) or get outperformed by experts (Snow et al., 2008). This calls for additional investigation to understand how biases impact overall performance. There is also a lack of works that analyze how data biases are reflected in model predictions in multi-class settings, where different types of errors might have different consequences. As an example, a website classifier misclassifying an adult site as a children's entertainment site is much more harmful than misclassifying it as a gambling site. Similarly, failing a student that should have passed has different implications than letting a student pass that should have failed (Schneider et al., 2022). With the steadily increasing influence of AI on more and more critical areas, the relevance of the training data, which determines the origin of the tendencies and the performance of such systems, increases. With the previously mentioned increasing need for labeled data, the effects of potential biases of annotators must be investigated in order to avoid the development of potentially harmful systems as early as possible. In this work, we are particularly interested in understanding:

RQ1: How do biases of high-skilled annotators differ during labeling from biases of low-skilled annotators in text sentiment classification tasks?

RQ2: To what extent do biases in annotated training data lead to biases in predictions of ML models and how can these biases be mitigated?

That is, we aim to understand if machine learning models amplify biases in data, dampen them or shift one bias to another. All the evidence so far points to the existence of bias in annotations provided by different groups and its negative effect on model performance but to our best knowledge, very few of the existing studies have looked

into its subsequent propagation into model predictions, i.e., whether biases in annotations are reflected in predictions. Therefore, we set out to find evidence in support of this statement by comparing crowdsourced annotations provided by low- and high-skilled annotators. As another (minor) difference to most of the existing studies we segregate annotators based on their apparent skill level on the task at hand, whereas prior compared experts and non-experts or people from different demographic groups. Finally, we also discuss the mitigation of biases.This question is of high practical relevance, since companies can impact data biases through their choice of annotators, i.e., crowd workers. They can employ high or low skilled workers in typical crowdsourcing settings, e.g., by demanding a certain level of education of crowd workers. To answer our question, we employ an empirical approach. This seems well-suited in the context of machine learning, since machine learning models are notoriously difficult to understand (Meske et al., 2022) making a deductive approach challenging. That is, despite significant efforts in the research community to explain machine learning models, many challenges remain (Meske et al., 2022).

We explore the presence of biases in crowdsourced annotations and investigate whether these biases are the same for low- and high-skilled annotators. We also provide empirical evidence to understand to what extent biases and errors in annotations are transferred to ML models. Finally, we show that biases can be reduced by using a mix of high- and low-skilled workers. Thus, our work has tangible implications for practitioners facing the choice of what skill level of workers they should demand. However, in this paper, we provide findings of our work in progress for one common task in machine learning, i.e., annotation of textual data with sentiments, using a state-of-the-art model in NLP, e.g., a transformer-based deep learning model.

The paper is structured as follows. We start with the background and related work on the topics of crowdsourcing in machine learning research, comparing annotations from different sources and training models with biased data. This is followed by a description of the methods we used to derive our analysis, a summary of our results and finally the conclusions drawn and an outlook for future research.

2 Background and Related Work

2.1 Crowdsourcing in Machine Learning Research

In the context of machine learning, crowdsourcing has a multitude of application areas, namely data generation, model evaluation and debugging, hybrid intelligence systems, and behavioral experiments (Vaughan, 2018). For data generation purposes, crowdsourcing is seen as a quick and efficient way to annotate large amounts of data, which could in turn be used to train supervised machine learning models. Although crowdsourcing alleviates the problem of data quantity, there are questions regarding the quality of the data. Annotations or labels collected might be noisy due to spammers providing arbitrary labels to maximize their financial gains (Eickhoff et al., 2012), annotators' malicious behavior (Wang et al., 2013), and cognitive biases (Eickhoff, 2018).

In this paper, we adopt the definition of biases as annotators' cognitive biases, i.e., systematic deviation patterns in thinking from the rational causing systematic errors (Haselton et al., 2015). In addition to potential malicious data biases, Liu et al. (2021) found that demographic groups of annotators show similarities in biases within their group and differences across different groups. Meaning different unconscious factors of annotators can influence their labeling behaviour and therefore be the origin of biases that transfer into the data. In recent years, various ways to get around this issue have been studied, including label aggregation techniques (Jagabathula et al., 2014; Zhang et al., 2016), designing tasks and incentive schemes to induce high quality answers (He et al., 2013; Radanovic et al., 2016), and training models with noisy labels (Cordeiro & Carneiro, 2020; Han et al., 2018; Lee et al., 2022; Nomura et al., 2021; Rodrigues and Pereira, 2018). Most works on label aggregation techniques have been based on the general expectation-maximization (EM) framework proposed by Dawid & Skene (1979), although simple majority voting has also been shown to perform robustly in some cases (Saab et al., 2019; Van Atteveldt et al., 2021).

2.2 Comparing Crowdsourced (Non-Expert) and Expert Annotations

Evaluating the quality of crowdsourced annotations is seen as standard practice before using them in downstream applications. To do so, it is common to measure the inter-annotator agreement using some statistical measure such as Krippendorff's alpha (Hayes & Krippendorff, 2007). Snow et al. (2008) evaluated the quality of non-expert annotations and compared them to expert annotations for five natural language tasks. They found that across all tasks, between 4 to 10 non-expert annotations are required to match an expert annotator's performance and for a particular task, a supervised machine learning model trained with non-expert annotations can outperform one trained with expert annotations, indicating the presence of a strong bias in individual labelers. Waseem (2016) compared expert and non-expert annotations collected for a hate speech labeling task and showed that non-expert annotators are more likely to label items as hate speech compared to experts. Comparing the performances of downstream classification models trained on these annotations, it was found that models trained on expert annotations showed superior performance. Long et al. (2021) showed that non-expert annotations from crowdsourcing services could have negative effects on a supervised machine learning model's performance. The study of Shakeri Hossein Abad et al. (2022) supports these findings by showing a fall of at least 8% across all performance measures when using non-expert annotations. Barbera et al. (2020) looked at the task of content truthfulness assessment and explored how the adopted response scale and annotators' own bias affect their responses. Their results showed that annotators tend to be biased towards their beliefs, i.e., items aligned to their political beliefs were marked as truthful more often. These studies hint at the difference in annotation behavior between experts and non-experts. They also suggest that machine learning models trained on either expert or non-expert annotations might vary performance-wise.

2.3 Training with Biased Data

Binns et al. (2017) conducted empirical experiments in the context of content and showed that the classifiers trained on men-annotated labels are less sensitive to women-annotated labels and vice-versa. Similarly, Al Kuwatly et al. (2020) investigated annotator bias by training hateful content classifiers on annotations provided by groups segregated by demographic features such as gender, age, education and first language. The results show that except for gender, all other demographic features considered had a significant effect on the predictions, e.g., classifiers trained on annotations provided by native English speakers demonstrate significantly better performance in classifying personal attack comments compared to those trained on non-native speakers-provided annotations. Geva et al. (2019) explored annotator bias in three natural language tasks by training models either with or without annotator identifiers as features and found that model performance is superior in the former case, indicating that models could use the additional information to better replicate annotators' behavior. Moreover, models trained from a pool of annotations provided by certain annotators were also found to perform poorly when tested on data provided by another subset of annotators.

Existing studies have therefore shown the effects of demographic characteristics and domain knowledge by annotators on the accuracy of the resulting models. However, our primary area of interest lies in investigating the transmission of bias, stemming from the labeling procedure of the training data to the corresponding machine learning models.

3 Methodology

Ultimately, we aim to empirically analyze how biases in data impact model predictions using multiple datasets and multiple machine learning models. In this paper, we use one dataset and one machine learning model to gather findings by tackling the problem of sentiment analysis in the realm of natural language processing (NLP) using a state-of-the-art deep learning model.

For our analysis, we utilized a dataset of crowdsourced annotations for weather tweet sentiment analysis collected through Crowdflower. The tweets, which have been presented to the annotators have a ground truth label collected from domain experts outside the crowdsourcing setup, meaning a total of 102 experts were involved in the ground truth labeling process. The raw dataset contains 1,000 tweets, and each tweet was annotated by 20 annotators to be classified into one of the following classes based on their sentiment: (1) Negative, (2) Neutral, (3) Positive, (4) I can't tell and (5) Tweet not related to weather condition. Previous research has shown that having many sentiment classes is a challenging task for machine learning models (Bouazizi and Ohtsuki, 2019), which is why we decided to remove the last two classes from the dataset in order to achieve a respectable performance from our subsequent classifier. After removing the last two classes, our final dataset contains 763 tweets and 13,484 annotations. The class distributions for the tweets and annotations are provided in Table 1. It can be seen that both the tweets and annotations are fairly evenly distributed between the three classes.

Table 1. Distribution of tweets and annotations

Class	No. of tweets	No. of annotations
Negative	271	4,668
Neutral	261	4,486
Positive	231	4,330

We classified the 102 annotators into low- and high-skilled annotators based on their annotation performance as measured by the average weighted F1 score, which indicates the skill, and the number of tweets they annotated, i.e., for high-skilled workers we also demanded a minimum amount of experience. Figures 1 and 2 illustrate the distributions of the annotators based on these two variables. For our purposes, we intended to have at least 30 annotators in both groups in order to enable reliable statistical tests for our subsequent analyses and ultimately set a minimum value of 0.8 for the average weighted F1 score and 60 for the number of annotated tweets for an annotator to be classified as highly skilled. This led to 31 high-skilled annotators and 71 low-skilled annotators.

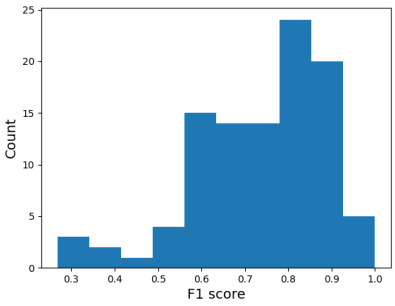

Figure 1. Annotators' F1 score distribution

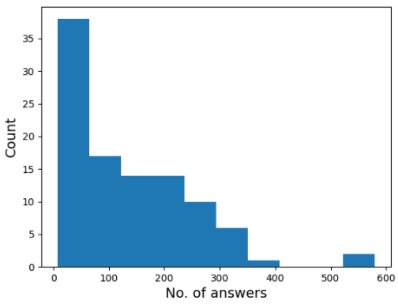

Figure 2. Annotators' number of answers distribution

To measure overall biases in the annotations, we utilized the concept of normalized entropy in information theory (Shannon, 1948), the formula for which is given below.

$$H = -\sum_{i=1}^{n} P(x_i) \log_n P(x_i),$$

Where H is the normalized entropy with values in [0,1], i denotes the outcome class and ranges from 1 to 6, n is the number of possible outcome classes and P(x_i) is the probability of observing the outcome of class i. Meaning P(x_i) equals the relative frequency with which the annotator at hand contributed to the different types of existing errors. A value of one indicates no bias and a value of 0 large bias. If the annotators were not biased, according to our previous definition, we would expect that their errors in confusing classes are uniformly distributed, i.e., there is no tendency to make certain

types of errors more commonly than others. In our context, this is best illustrated with the help of the confusion matrix shown in Table 2. There are 6 types of errors that can occur. If the error probabilities are equal then the number of errors made for each error type is proportional to the number of observations belonging to the true class. Consider an unbiased annotator who errs in 10% of their annotations. If they were to annotate 300 tweets made up of 160 negative, 100 neutral and 40 positive tweets, they would make 16, 10, and 4 errors for the three classes, respectively. Since there are two error types per true class, the errors would also have to be equally distributed between the two types, i.e., 8 Type 1 and Type 2 errors each, and so on. Dividing the numbers of errors made by the number of observations belonging to its true class, e.g., 8/160 (Type 1 and 2), 5/100 (Type 3 and 4) and 2/40 (Type 5 and 6), we get equal values for all error types and normalizing the values such that they sum up to 1 leads to the error probabilities. Calculating the entropy from the error probabilities in this case would lead to a value of 1. If the annotator in question were somewhat biased, e.g., more prone to making Type 1 errors than Type 2, the number of Type 1 errors might be higher than that of Type 2 – 12 vs. 4. In this case, the entropy would be equal to 0.976. Hence, the lower the entropy of an annotator is, the more biased their annotations are.

Table 2. Example confusion matrix for the dataset

True Class	**Predicted Class**		
	Negative	Neutral	Positive
Negative	-	Type 1	Type 2
Neutral	Type 3	-	Type 4
Positive	Type 5	Type 6	-

For each high-skilled annotator, we calculated their error probabilities and entropy on the tweets they annotated. As our results are dependent on the dirstibution of ground truth labels among the different annotator groups, we evenly dispensed the true classes across them. To provide a fair comparison, i.e., to ensure the overall number of correct labels is very similar for low- and high-skilled workers, we created ensembles of non-skilled annotators and aggregated their annotations through majority voting. The number of non-skilled annotators ensembles was chosen to equal the number of high-skilled annotators. The number of annotators in each ensemble is set to be the fewest possible as such so that their aggregated annotation achieves a comparable F1 score to that of the corresponding high-skilled annotator (difference < 0.03). The error probabilities and entropies of the ensembles of low-skilled annotators were then calculated in the same way. To find out whether the annotations of the two groups exhibit bias, we conducted a one sample t-test with the population mean set to the entropy of the raw annotations (0.9885). This procedure was chosen to find out whether there is no significant difference between the mean entropy of the high-skilled annotators' (ensembles of low-skilled annotators') annotations and the raw annotations' entropy. Finally, we ran Kolmogorov-Smirnov tests on the following variables to further investigate two scenarios:

- Entropy of the high-skilled annotators' vs. ensembles of low-skilled annotators' annotations should indicate if1: there is no bias-wise difference between the high-skilled annotators and the ensembles of low-skilled annotators.
- Error probabilities of the high-skilled annotators vs. ensembles of low-skilled annotators should show if2: high-skilled annotators and ensembles of low-skilled annotators are prone to the same types of errors, i.e., there is no significant difference in probabilities across all error types.

To investigate the propagation of annotation bias into model predictions, we fine-tuned smallBERT (Devlin et al., 2019) for 20 epochs with the learning rate set to 1e-4 to classify the tweets using three different sets of annotations as training data: (1) aggregated annotations from all high-skilled annotators, (2) aggregated annotations from all low-skilled annotators and (3) half of the tweets annotated by high-skilled annotators and the other half by low-skilled annotators. For each of the 30 annotators sets (of each skill level), we trained a separate model. We used 5-fold cross-validation and recorded the cross-validated accuracies and error probabilities on the predictions. Lastly, we ran pair-wise Kolmogorov-Smirnov tests to compare the error probabilities across all error types for the three annotation sets, with the goal to investigate if: there is no significant difference between the three models' predictions in terms of error probabilities.

4 Results and Discussion

The t-test results for the detection of biases are presented in Table 3. For both low- and high-skilled annotators, biases are detected at a 1% significance level, meaning that both groups tend to make certain types of errors more often as opposed to having equal probabilities for all types of errors. Comparing the two groups, no statistically significant difference is found, suggesting that low- and high-skilled annotators exhibit about the same level of bias. It is intuitive to think that ensembles of low-skilled annotators show less bias due to the aggregation of annotations, i.e., different biases "cancel". However, biases within the group of low-skilled workers exist. Low-skilled workers picked the "Neutral" and "Positive" labels slightly more often than their high-skilled counterparts, while the opposite applies to the "Negative" label. The overall distribution of replies of low and high-skilled workers is given in Table 4.

Table 3. One sample t-test results for bias detection

Group	Mean entropy +/- std	t-statistic	p-value
High-skilled annotators	0.714 +/- 0.179	-8.42	2.13e-9
Ensembles of low-skilled annotators	0.719 +/- 0.157	-9.36	2.07e-10

Table 4. Distributions of annotations from low- and high-skilled workers

Class	No. of annotations	
	High-skilled annotators	Low-skilled annotators
Negative	1,896 (36.8%)	2,590 (31.1%)
Neutral	1,707 (33.1%)	2,961 (35.5%)
Positive	1,550 (30.1%)	2,780 (33.4%)

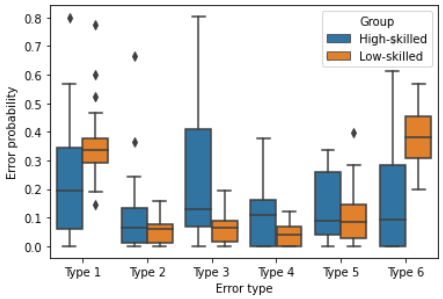

Figure 3. Distributions of error probabilities for low- and high-skilled annotators

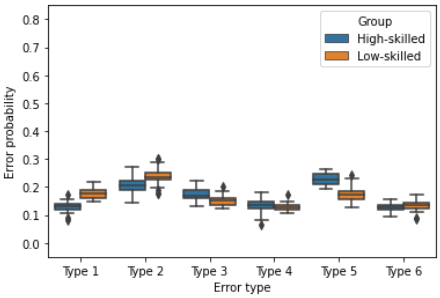

Figure 4. Distributions of error probabilities for classifiers trained on the low- and high-skilled annotation sets

Figure 3 shows the distributions of error probabilities for annotations provided by high-skilled and low-skilled annotators. The results show that in general, high-skilled annotators demonstrate large variances across all error types, i.e., a certain annotator A might be highly susceptible to Type 1 error and not to other error types. For another annotator B high susceptibility to Type 3 error might be observed. In contrast, ensembles of low-skilled annotators seem to be relatively consistent with regard to the errors they make, in that Type 1 and 6 errors are the most common by far and only small variances are observed. The latter is a consequence of aggregating outcomes of multiple workers.

While it might be tempting to attribute the irregularities of errors as a regression to the mean, i.e., towards “Neutral”, it is not the case because having “Neutral” as the aggregated annotation for a certain tweet is only possible if the majority of the annotators actually answered ”Neutral” and not when “Positive” and ”Negative” annotations are observed in equal measure. Instead, it might be attributed to the tendency of low-skilled annotators to choose the “safe” answer in order to minimize large errors in an estimate. That is, in our case the error is maximal if “Positive” and “Negative” are confused. Choosing “Neutral” avoids such large errors. This tendency is also apparent from the distributions of annotations from the two groups shown in Table 4, where low-skilled workers have a slight preference for neutral labels.

Table 5. KS-test results for the annotators' error probabilities

Error type	Statistic	p-value
Type 1	0.484	0.001*
Type 2	0.258	0.256
Type 3	0.452	0.003*
Type 4	0.452	0.003*
Type 5	0.194	0.615
Type 6	0.677	4.62e-7*

The Kolmogorov-Smirnov test results for comparing the error probabilities between low and high-skilled workers are shown in Table 5. The probabilities of Type 1, 3, 4 and 6 errors between high-skilled and low-skilled annotators are significantly different (at 1% significance level). As is evident in Figure 3, high-skilled annotators tend to make Type 3 and 4 errors more often than low-skilled counterparts, while the opposite is true for Type 1 and 6 errors. Based on this evidence, it could be argued that high-skilled annotators are more confident in their judgment than low-skilled ones, leading the former to choose the "Negative" and "Positive" labels more often than the latter, although erroneously in some cases. This is in line with the findings of Hube et al. (2019), which state that annotators with strong opinions tend to provide non-neutral options more often, leading to biased annotations. These behavioral tendencies might have their origin in the way experts and novices approach topics and problem solving tasks, as Haerem & Rau (2007) stated that experts predominantly rely on deep structures of the problem and novices on the surface structure. We can also rule out the possibility of this finding being an artifact of the data, as Table 4 shows that in general, high-skilled annotators annotate tweets as "Positive" less often compared to low-skilled annotators.

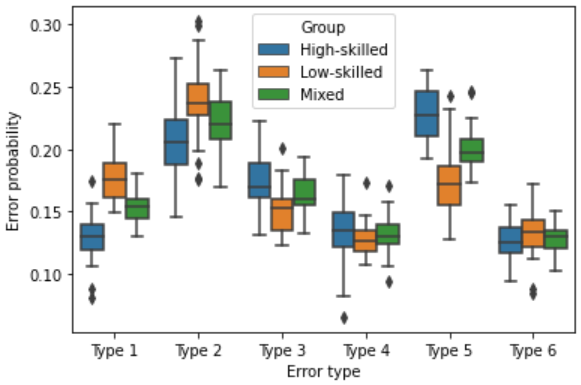

Figure 5. Distributions of error probabilities for classifiers trained on the low-, high-skilled and mixed sets of annotations

Table 6. KS-test results for the model predictions' error probabilities

Error type	Predicted Class		Low vs. Mixed		High vs. Mixed	
	Statistic	**p-value**	**Statistic**	**p-value**	**Statistic**	**p-value**
Type 1	0.9	5.78e-13*	0.533	2.93e-4*	0.633	5.79e-6*
Type 2	0.567	8.74e-5*	0.433	0.007*	0.3	0.135
Type 3	0.567	8.74e-5*	0.367	0.035	0.333	0.071
Type 4	0.367	0.035	0.233	0.393	0.333	0.071
Type 5	0.867	8.25e-12*	0.667	1.28e-6*	0.567	8.74e-5*
Type 6	0.233	0.393	0.267	0.239	0.167	0.808

Results of the empirical experiment on classifiers are summarized in Figure 4, Figure 5 and Table 6. First of all, it can be seen that the trends of the error probabilities (Figure 4) are consistent with the ones shown in Figure 3 – classifiers trained on high-skilled annotators' annotations produce Type 3 and 4 errors in their predictions more often than those trained on annotators provided by low-skilled annotators, while the latter group of classifiers is more susceptible to Type 1 and 6 errors. Although the Kolmogorov-Smirnov test results show that only the differences in Type 1 and Type 3 errors are detected at 1% significance level, the overall trend remains consistent when comparing the means of the error probabilities. This is while both classifiers exhibit almost the same level of accuracy: 0.71 +/- 0.01 (low-skilled) vs. 0.72 +/- 0.01 (high-skilled). Based on these results, it can be concluded that biases in annotations are reflected in model predictions to a considerable extent, indicating that caution should be exercised when gathering crowdsourced annotations for training models. It is also interesting to see that the model makes extreme errors, i.e., Types 2 and 5 confusing "Positive" and "Negative" much more often than humans. We attribute this to the fact that the classifier is trained on few data and on a specific usecase in contrast to humans, i.e., humans have a more-fine granular understanding of sentiment (and language) originating from similar tasks and other domains (a broughter knowledgebase) reducing the likelihood of large errors. From Figure 5, it can be seen that the mixed annotation strategy has an averaging effect on the error probabilities – across all error types, the error probabilities of these classifiers are in between the values exhibited by the classifiers trained on the other two annotation sets. In applications where avoiding any extreme error probability is desirable, the results indicate that implementing a mixed annotation strategy might help machine learning models to produce more balanced, hence less biased, predictions.

5 Conclusion and Future Work

We have shown that both low- and high-skilled annotators exhibit certain biases in their annotations. Although there is no evidence of significant difference regarding the extent of bias between the two groups, it was found that they are susceptible to different types of errors. Assuming the same number of errors, high-skilled annotators tend to avoid the "Neutral" label and provide polarized annotations instead, whereas the opposite is

true for low-skilled annotators. By training classifiers on biased annotations, we found evidence that biases in the annotations also manifest in model predictions, calling for caution when training models on crowdsourced annotations and different costs are attached to the various types of error. In particular, classifiers tended to show more extreme errors than found in human data, i.e., confusing sentiments of opposite sentiment. As partial mitigation, practitioners might employ a mix of low- and high-skilled workers to minimize biases in models as an ex-post optimization approach.

Our results provide a first glance at the presented issue, as they are based on a single dataset and classifier limiting the generalizability of our findings. We plan to conduct similar experiments using further datasets from different tasks and classifiers. In addition, we want to examine different annotator groups e.g. selecting them prior to their annotation task based on their experience and expertise in a task specific field. Furthermore, we want to look into mitigation strategies and measure their effects to further validate our mentioned mixed annotation strategy with possibly other datasets. We hope to contribute to the information systems research community by providing a concise but thorough overview of the propagation of biases into ML model predictions.

References

Al Kuwatly, H., Wich, M. & Groh, G. (2020), Identifying and Measuring Annotator Bias Based on Annotators' Demographic Characteristics, *in* 'Proceedings of the fourth workshop on online abuse and harms', pp. 184-190.

Barbera, D. L., Roitero, K., Demartini, G., Mizzaro, S. & Spina, D. (2020), Crowdsourcing Truthfulness: The Impact of Judgment Scale and Assessor Bias, *in* 'European Conference on Information Retrieval', Springer, pp. 207-214.

Barbosa, N. M., & Chen, M. (2019). Rehumanized crowdsourcing: A labeling framework addressing bias and ethics in machine learning. *in* 'Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems', pp. 1-12.

Binns, R., Veale, M., Van Kleek, M. & Shadbolt, N. (2017), 'Like Trainer, Like Bot? Inheritance of Bias in Algorithmic Content Moderation', *Springer International Publishing*, pp. 405-415.

Bouazizi, M. & Ohtsuki, T. (2019). 'Multi-Class Sentiment Analysis on Twitter: Classification Performance and Challenges', *Big Data Mining and Analytics* **2**(3), pp. 181-194.

Cabrera, G. F., Miller, C. J., & Schneider, J. (2014). Systematic labeling bias: De-biasing where everyone is wrong. 22nd International Conference on Pattern Recognition, pp. 4417-4422.

Caiafa, C. F., Solé-Casals, J., Marti-Puig, P., Zhe, S., & Tanaka, T. (2020). Decomposition methods for machine learning with small, incomplete or noisy datasets. Applied Sciences, **10**(23), pp. 8481-8501.

Cordeiro, F. R., & Carneiro, G. (2020). A survey on deep learning with noisy labels: How to train your model when you cannot trust on the annotations?. *in* '2020 33rd SIBGRAPI conference on graphics, patterns and images, pp. 9-16.

Dawid, A. P. & Skene, A. M. (1979), 'Maximum Likelihood Estimation of Observer Error-Rates Using the Em Algorithm', *Journal of the Royal Statistical Society: Series C (Applied Statistics),* **28**(1), pp. 20-28.

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2019), Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding, *in* 'Proceedings of NAACL-HLT', pp. 4171-4186.

Ding, Y., You, J., Machulla, T. K., Jacobs, J., Sen, P., & Höllerer, T. (2022). Impact of Annotator Demographics on Sentiment Dataset Labeling. *in* 'Proceedings of the ACM on Human-Computer Interaction', pp. 1-22.

Eickhoff, C. (2018), Cognitive Biases in Crowdsourcing, *in* 'Proceedings of the eleventh ACM international conference on web search and data mining', pp. 162-170.

Eickhoff, C., Harris, C. G., de Vries, A. P. & Srinivasan, P. (2012), Quality through Flow and Immersion: Gamifying Crowdsourced Relevance Assessments, *in* 'Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval', pp. 871-880.

Geva, M., Goldberg, Y. & Berant, J. (2019), Are We Modeling the Task or the Annotator? An Investigation of Annotator Bias in Natural Language Understanding Datasets, *in* 'Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing', pp. 1161-1166.

Haerem, T., & Rau, D. (2007). The influence of degree of expertise and objective task complexity on perceived task complexity and performance. Journal of Applied Psychology, **92**(5), pp. 1320-1331.

Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., & Sugiyama, M. (2018). Co-teaching: Robust training of deep neural networks with extremely noisy labels. Advances in neural information processing systems, pp. 8527-8537.

Haselton, M. G., Nettle, D. & Murray, D. R. (2015), 'The Evolution of Cognitive Bias', *The handbook of evolutionary psychology*, pp. 1-20.

Hayes, A. F. & Krippendorff, K. (2007), 'Answering the Call for a Standard Reliability Measure for Coding Data', *Communication methods and measures* **1**(1), pp. 77-89.

He, J., van Ossenbruggen, J., & de Vries, A. P. (2013). Do you need experts in the crowd? A case study in image annotation for marine biology. *in* 'Proceedings of the 10th Conference on Open Research Areas in Information Retrieval', pp. 57-60.

Hötter, J., Wenck, H., Feuerpfeil, M., & Witzke, S. (2022). Kern: A Labeling Environment for Large-Scale, High-Quality Training Data, *in* 'Proceedings of the Natural Language Processing and Information Systems: 27th International Conference on Applications of Natural Language to Information Systems', pp. 502-507

Hube, C., Fetahu, B. & Gadiraju, U. (2018), Limitbias! Measuring Worker Biases in the Crowdsourced Collection of Subjective Judgments, *in* 'Proceedings of the 1st Workship on Subjectivity, Ambiguity and Disagreement in Crowdsourcing'.

Hube, C., Fetahu, B. & Gadiraju, U. (2019), Understanding and Mitigating Worker Biases in the Crowdsourced Collection of Subjective Judgments, *in* 'Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems', pp. 1-12.

Irshad, H., Montaser-Kouhsari, L., Waltz, G., Bucur, O., Nowak, J. A., Dong, F., ... & Beck, A. H. (2014). Crowdsourcing image annotation for nucleus detection and segmentation in computational pathology: evaluating experts, automated methods, and the crowd. *in* 'Pacific symposium on biocomputing Co-chairs', pp. 294-305.

Jagabathula, S., Subramanian, L., & Venkataraman, A. (2014). Reputation-based worker filtering in crowdsourcing. Advances in Neural Information Processing Systems, pp. 2492-2500

Kafkalias, A., Herodotou, S., Theodosiou, Z., & Lanitis, A. (2022). Bias in Face Image Classification Machine Learning Models: The Impact of Annotator's Gender and Race, *in* 'Proceedings of the Artificial Intelligence Applications and Innovations: 18th IFIP WG 12.5 International Conference', pp. 89-100.

Lee, D. M., Kim, Y., & Seo, C. G. (2022). Context-based Virtual Adversarial Training for Text Classification with Noisy Labels.

Liu, H., J. Thekinen, S. Mollaoglu, Da Tang, J. Yang, Y. Cheng, H. Liu & J. Tang (2021), 'Toward Annotator Group Bias in Crowdsourcing'.

Long, H. L., A. O'Neil & S. Kübler (2021), On the Interaction between Annotation Quality and Classifier Performance in Abusive Language Detection, *in* 'Proceedings of Recent Advances in Natural Language Processing', pp. 868–875.

Meske, C., Bunde, E., Schneider, J. & Gersch, M. (2022), 'Explainable Artificial Intelligence: Objectives, Stakeholders, and Future Research Opportunities', *Information Systems Management* **39**(1), pp. 53-63.

Nomura, Y., & Kurita, T. (2021). Robust Training of Deep Neural Networks with Noisy Labels by Graph Label Propagation. *in* 'Frontiers of Computer Vision: 27th International Workshop', pp. 281-293.

O'Neil, A. Q., Murchison, J. T., van Beek, E. J., & Goatman, K. A. (2017). Crowdsourcing labels for pathological patterns in CT lung scans: can non-experts contribute expert-quality ground truth?. *in* 'Intravascular Imaging and Computer Assisted Stenting, and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis: 6th Joint International Workshops, CVII-STENT 2017 and Second International Workshop', pp. 96-105.

Radanovic, G., Faltings, B. & Jurca, R. (2016), 'Incentives for Effort in Crowdsourcing Using the Peer Truth Serum', *ACM Transactions on Intelligent Systems and Technology (TIST)* **7**(4), pp. 1-28.

Rodrigues, F. & Pereira, F. (2018), Deep Learning from Crowds, *in* 'Proceedings of the AAAI Conference on Artificial Intelligence'.

Saab, F., Elhajj, I. H., Kayssi, A. & Chehab, A. (2019), 'Modelling Cognitive Bias in Crowdsourcing Systems', *Cognitive Systems Research* **58**, pp. 1-18.

Saravanan, R. & Sujatha, P. (2018), A State of Art Techniques on Machine Learning Algorithms: A Perspective of Supervised Learning Approaches in Data Classification, *in* '2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS): IEEE', pp. 945-949.

Schneider, J., Richner, R. & Riser, M. (2022), 'Towards Trustworthy AutoGrading of Short, Multi-lingual, Multi-type Answers*', Int Journal of Artificial Intelligence in Education*.

See, L., Comber, A., Salk, C., Fritz, S., Van Der Velde, M., Perger, C., Schill, C., McCallum, I., Kraxner, F. & Obersteiner, M. (2013). Comparing the quality of crowdsourced data contributed by expert and non-experts, **8**(7).

Shakeri Hossein Abad, Z., G. P. Butler, W. Thompson & J. Lee (2022), 'Crowdsourcing for Machine Learning in Public Health Surveillance: Lessons Learned From Amazon Mechanical Turk', *Journal of medical Internet research* **24**(1).

Shannon, C. E. (1948), 'A Mathematical Theory of Communication', *The Bell system technical journal* **27**(3), pp. 379-423.

Snow, R., O'connor, B., Jurafsky, D. & Ng, A. Y. (2008), Cheap and Fast–but Is It Good? Evaluating Non-Expert Annotations for Natural Language Tasks, *in* 'Proceedings of the 2008 conference on empirical methods in natural language processing', pp. 254-263.

Van Atteveldt, W., van der Velden, M. A. & Boukes, M. (2021), 'The Validity of Sentiment Analysis: Comparing Manual Annotation, Crowd-Coding, Dictionary Approaches, and Machine Learning Algorithms', *Communication Methods and Measures* **15**(2), pp. 121-140.

Vaughan, J. W. (2018), 'Making Better Use of the Crowd: How Crowdsourcing Can Advance Machine Learning Research', *Journal of Machine Learning Research* **18**.

Wang, T., Wang, G., Li, X., Zheng, H. & Zhao, B. Y. (2013), Characterizing and Detecting Malicious Crowdsourcing, *in* 'Proceedings of the ACM SIGCOMM 2013 conference on SIGCOMM', pp. 537-538.

Warby, S. C., Wendt, S. L., Welinder, P., Munk, E. G., Carrillo, O., Sorensen, H. B., Jennum, P., Peppard, P., E., Perona, P. & Mignot, E. (2014). Sleep-spindle detection: crowdsourcing and evaluating performance of experts, non-experts and automated methods. Nature methods, **11**(4), pp. 385-392.

Waseem, Z. (2016), Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter, *in* 'Proceedings of the first workshop on NLP and computational social science', pp. 138-142.

Zhang, J., Wu, X., & Sheng, V. S. (2014). Imbalanced multiple noisy labeling. IEEE Transactions on Knowledge and Data Engineering, **27**(2), pp. 489-503.

Zhang, J., Wu, X. & Sheng, V. S. (2016), 'Learning from Crowdsourced Labeled Data: A Survey', *Artificial Intelligence Review* **46**(4), pp. 543-576.

Zhang, T., Yu, L., Hu, N., Lv, S., & Gu, S. (2020). Robust medical image segmentation from non-expert annotations with tri-network. *in* 'Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference', pp. 249-258.

Zhu, X., & Wu, X. (2004). Class noise vs. attribute noise: A quantitative study. The Artificial Intelligence Review, **22**(3), pp. 177-210.