Inter-annotator agreement is not the ceiling of machine learning performance: Evidence from a comprehensive set of simulations

Russell Richie[1] **Sachin Grover**[1] **Fuchiang (Rich) Tsui**[1,2]
[1]Tsui Lab, Department of Biomedical and Health Informatics
Children's Hospital of Philadelphia, Philadelphia, PA
[2]Department of Anesthesiology and Critical Care
Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA
{richier,grovers1,tsuif}@chop.edu

Abstract

It is commonly claimed that inter-annotator agreement (IAA) is the ceiling of machine learning (ML) performance, i.e., that the agreement between an ML system's predictions and an annotator can not be higher than the agreement between two annotators. Although Boguslav and Cohen (2017) showed that this claim is falsified by many real-world ML systems, the claim has persisted. As a complement to this real-world evidence, we conducted a comprehensive set of simulations, and show that an ML model can outperform IAA even if (and especially if) annotators are noisy and differ in their underlying classification functions, as long as the ML model is reasonably well-specified. Although the latter condition has long been elusive, leading ML models to underperform IAA, we anticipate that this condition will be increasingly met in the era of big data and deep learning. Our work has implications for (1) maximizing the value of machine learning, (2) adherence to ethical standards in computing, and (3) economical use of annotated resources, which is paramount in settings where annotation is especially expensive, like biomedical natural language processing.

1 Introduction

It is standard when conducting machine learning (ML) and natural language processing (NLP) work to calculate inter-annotator agreement (IAA) metrics like Cohen's Kappa (Cohen, 1960). This is done not just for annotation quality control, but also as a comparison for machine learning models' performance. In particular, it has commonly been claimed – by some of the most prominent researchers in ML and NLP (Boguslav and Cohen, 2017) – that *IAA places an upper bound or ceiling on the performance of machine learning models*. When researchers claim this and their model reaches IAA, they are implicitly suggesting (or at least it follows) that the model has performed as well as possible or has solved the task for that dataset, and/or that the dataset cannot be used to drive further development of ML models. Despite the prominence of this claim, however, Boguslav and Cohen (2017) reported that neither they nor a professional literature search service could find evidence in support of it. This is concerning for at least two reasons.

First, as Boguslav and Cohen (2017) say, "if the assumption [that IAA bounds ML] turns out not to be supported. . . we may be mis-estimating the actual performance of our [ML] systems. In particular, we may be over-estimating the quality of their performance by under-estimating how good [performance] could potentially be" (pg 298). This underestimation of the maximum possible performance may lead to the development of poorer models under the belief that they have achieved maximum capacity. Moreover, as noted by Boguslav and Cohen (2017), such misestimation may violate ethical standards concerning accurate characterization of the limitations of computer systems (e.g., ACM Code of Ethics and Professional Conduct 2.7 Anderson, 1992; see also Petersen et al., 2021 on recommendations for safe, effective use of clinical decision support systems).

Second, and relatedly, if a modeler stops using an annotated dataset to drive ML development once ML performance on that dataset reaches IAA, they may be underutilizing those annotations. This could be an enormous waste of money, since annotation is often one of the most expensive components of an ML/NLP project, especially in biomedical NLP where the time of annotators (often biomedical experts) is especially expensive. For example, Hill et al. (2015) noted that then state-of-the-art word embeddings had reached IAA on existing word relatedness benchmark datasets (e.g., WordSim-353, Finkelstein et al., 2001). Believing that IAA was the upper bound of ML, they therefore believed that such datasets could no longer be used to drive development of different word embedding models. This led them to collect new annotations of word similarity, yielding the benchmark SimLex-999. Their abstract lays out this logic:

> "Further, unlike existing gold standard evaluations, for which automatic approaches have reached…the inter-annotator agreement ceiling, state-of-the-art models perform well below this ceiling on SimLex-999. There is therefore plenty of scope for SimLex-999 to quantify future improvements to distributional semantic models, guiding the development of the next generation of representation-learning architectures." (pg. 1)

If IAA does not in fact bound ML, then the older word relatedness benchmarks *could have actually been used to "guide… the development of the next generation of representation-learning architectures"*, and there would have been less need to spend time and money annotating SimLex-999. Given that Hill et al. (2015) has been cited over 1000 times according to Google Scholar, other researchers may have absorbed and replicated their logic, which would be concerning if the claim is not really true. Indeed, a Stack Exchange post (tomas , https://stats.stackexchange.com/users/84364/tomas) roughly contemporaneous with Hill et al. (2015) suggests that this logic may be widespread.

Despite the popularity and stakes of the claim that IAA bounds ML, Boguslav and Cohen (2017) found, across 6 papers, 20 ML systems that outperform IAA, on tasks ranging from entity recognition in clinical notes (Roberts et al., 2008), to deception detection (Pérez-Rosas et al., 2015) (and see Wilbur, 1998 for earlier evidence that ML can outperform IAA in information retrieval[1]). However, claims that IAA bounds ML performance have persisted. This is seen in both biomedical and broader ML/NLP, in (1) papers that are often cited much more than Boguslav and Cohen (2017) and published in high impact outlets including JAMA Network, AMIA, Nature Human Behavior, and ACL (e.g., Grčar et al., 2017; Pilehvar et al., 2018; Amidei et al., 2018; Sarker et al., 2019; Pustu-Iren et al., 2019; Ribeiro et al., 2019; Richie et al., 2019; O'Connor et al., 2020; Hebart et al., 2020; Mayfield and Black, 2020; Basile, 2020; Bevilacqua et al., 2021; Li et al., 2021; Higashinaka et al., 2021; Goldberg et al., 2021), (2) machine learning lectures at well-known universities including University of Pittsburgh (Han, 2017), University of Edinburgh (Cohen, 2020), and City University of New York (CUNY, Gorman, 2020) and in slides by noted NLP textbook authors Jurafsky and Martin (Jurafsky and Martin, 2022), and (3) online posts and social media discussions by prominent machine learning users (e.g., Ruder, 2021).

It is not entirely clear why the claim that IAA bounds ML survived Boguslav and Cohen's counterexamples, but we suspect at least two factors are at play. First, Boguslav and Cohen (2017) was published in a fairly specialized journal (*Studies in Health Technology and Informatics*), and therefore may not have reached as many ML/NLP practitioners as it could or should have. Consistent with this, as of March 29, 2022, Boguslav and Cohen (2017) has been cited only 7 times, according to Google Scholar. Second, we suspect that the issue deserves a broad proof of concept based on simulations, in addition to the empirical examples raised by Boguslav and Cohen. Simulations would be complementary to the real-world evidence brought by Boguslav and Cohen, in at least two ways. First, simulations allow us to simplify the problem to its essence, which may be clarifying in ways that real-world studies, with all their potentially distracting idiosyncrasies, are not. Second, simulations allow us to precisely control and test different potentially relevant annotation and modeling factors, and therefore better understand when/how/why a model can or can't beat IAA. Therefore, the aim of this study is to use a comprehensive set of computational simulations to bolster the evidence that IAA is not the upper bound on ML performance.

The rest of this paper is organized as follows. We start with simplified simulations that capture the basic elements of an ML pipeline with two annotators and train a supervised model with these annotations (Experiment 1). We then relax various assumptions of this setup to simulate potentially more realistic settings, and to better understand the conditions under which an ML model can outperform IAA (Experiment 2). In both experiments, the general approach is to (a) simulate two annotators' annotations on a test set (where both annotators label all samples of this set); (b) simulate their annotations on disjoint halves of a training set; (c) train an ML model on the training set annotations; and

[1] We thank an anonymous reviewer for this suggestion.

(d) compare the agreement between the ML model and each annotator to the annotators' IAA on the test set. We conclude with a general discussion interpreting our work and its implications.

2 Experiment 1

2.1 Simulations

We consider a binary classification task conducted by two annotators, $A1$ and $A2$ who probabilistically classify the *i*-th sample, x_i (composed of a single variable), into one of two classes, $y_i \in \{0, 1\}$, according to a logistic function of x_i, as in:

$$p(y_i = 1) = \frac{1}{1 + e^{-x_i}} \tag{1}$$

In a simulation, we first sample the independent variable x from a standard normal distribution (zero mean, unit standard deviation), producing an X_{train} and an X_{test} with some number of samples each. Through Eq 1, $A1$ and $A2$ independently annotate every sample in X_{test}, which yields $y_{test,A1}$ and $y_{test,A2}$. Then $A1$ annotates the first half of X_{train}, and $A2$ annotates the second half of X_{train}, and we concatenate their annotations into a single y_{train}. We then train a logistic regression on (X_{train}, y_{train}), and generate this model's predictions on X_{test}, i.e., we generate $\hat{y}_{test}$.

We can then calculate the inter-annotator agreement $f(y_{test,A1}, y_{test,A2})$, and the model's average performance using both $A1$'s and $A2$'s annotations as ground truth, as in $average(f(\hat{y}_{test}, y_{test,A1}), f(\hat{y}_{test}, y_{test,A2}))$, where f is either F1-score or Cohen's Kappa. F1-score is defined as:

$$F_1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{2}$$

where $Precision$ is $\frac{TP}{TP+FP}$ and $Recall$ is $\frac{TP}{TP+FN}$. Cohen's Kappa is defined as:

$$\kappa \equiv \frac{p_o - p_e}{1 - p_e} \tag{3}$$

where p_o is the observed agreement among annotators, and p_e is the probability of chance agreement, which is often calculated using the base rates of each label in observed annotations.

Conventionally, Cohen's Kappa is used to measure IAA because it controls for chance annotator agreement, and F1 is used to measure model performance because it balances precision and recall and punishes (inappropriately) 'extreme' models (a setting with 1 positive sample and infinite negative samples, and a model that always assigns the positive class, will have $Recall = 1$, $Precision = 0$, and $F1 = 0$). However, Boguslav and Cohen (2017) suggest that in many linguistic annotation tasks, especially named entity recognition or others involving phrase extraction, where there are a very large number of potential spans that no annotator ever extracts, it is often the case that $p_e = 0$ in Equation 3. In this case, *Kappa is equivalent to F1* (Hripcsak and Rothschild, 2005), which arguably justifies the commonly conducted direct comparison between (1) IAA measured with Kappa and (2) model performance measured with F1. In our simulations, however, it is more straightforward to simply calculate and compare IAA and ML performance in the same metric(s), and we opt for this here.

Finally, we note here that our goal is not to critique these particular measures, their usage, or the paradigm of inter-annotator agreement more generally (for such critique, see for example Amidei et al., 2018). Rather, our focus is merely to demonstrate the falsity of the claim that IAA bounds ML, i.e., that $f(y_{test,A1}, y_{test,A2})$ >= $average(f(\hat{y}_{test}, y_{test,A1}), f(\hat{y}_{test}, y_{test,A2}))$. Although we chose Kappa and F1-score here because of their common usage in ML and NLP, we expect our results to generalize to other measures (e.g., Matthews Correlation Coefficient).

2.2 Results

In our simulations, X_{test} contains 100 samples and X_{train} contains 1000 samples, and the simulations were repeated 100 times. Figure 1a shows the results. As can be seen, on average, the model achieves F1=0.67 when comparing to the annotators on the test set, while the annotators score only F1=0.58 when comparing to each other (t=12.44, p < 10-25). Likewise, the model 'agrees' with the annotators at a Cohen's Kappa of about 0.35, while the annotators agree with each other at only 0.16 (t=13.60, p < 10-29). Clearly, inter-annotator agreement does not provide an upper bound on ML performance in this simple setting.

While IAA clearly doesn't provide an upper bound on model performance, it is also clear (see Figure 1b) that the two are positively correlated (r = 0.48, p < 10-6). The correlation arises because when an annotator happens to assign a positive class to samples whose $p(y_i = 1) > 0.5$, the annotator's predictions will be closer to both (a) the model's predictions (which always assigns the positive class to samples when $p(y_i = 1) > 0.5$, and (b) the other annotator's predictions, because the latter annotator will also usually assign the positive class when $p(y_i = 1) > 0.5$. Thus, one might fairly say that although IAA doesn't *bound* ML performance, IAA predicts ML performance. This relationship may be partly what underlies the appeal or intuitiveness of the notion that IAA bounds ML performance, an issue we return to in the general discussion. At the same time, Figure 1b also makes clear that, at the level of individual simulations, the ML model tends to outperform IAA, since most points are above the line $y = x$.

2.3 Discussion

The simulations above show that, contrary to many claims in the ML and NLP communities, IAA does not bound ML performance, *at least in this simple case*. At the same time, because IAA and ML performance are positively correlated, IAA does give some indication of the level of ML performance that can be expected, which could explain why many people believe that IAA bounds ML performance. It may even be that when two authors use the same term ('bound', 'ceiling', or 'limit', the keywords we used to find claims about the relationship between IAA and ML performance), one author may intend that IAA is a strict ceiling on ML performance, and another may intend that low IAA merely *predicts* low ML performance (although based on our reading of the literature, we tend to think most writers intend the first meaning). Failure to carefully distinguish these meanings may be contributing to confusion in the field, and we hope these results clarify the distinction.

Also note that in this simulation, the annotators $A1$ and $A2$ have identical classification functions, so this simulation can be equivalently viewed as having a single annotator classify *all* of the training samples once, and all of the test samples twice. Having a single annotator classify the test samples twice allows us to calculate not inter-annotator agreement, but *intra*-annotator agreement (also known as test-retest reliability in the psychometrics literature, Guttman, 1945). Therefore, concluding that IAA does not bound ML performance is also applicable to *intra*-annotator agreement.

3 Experiment 2

The conditions of Experiment 1 are intentionally oversimplified from real-world conditions. To better understand the range of conditions under which ML can or cannot outperform IAA, we next introduce some additional complexity in the simulations.

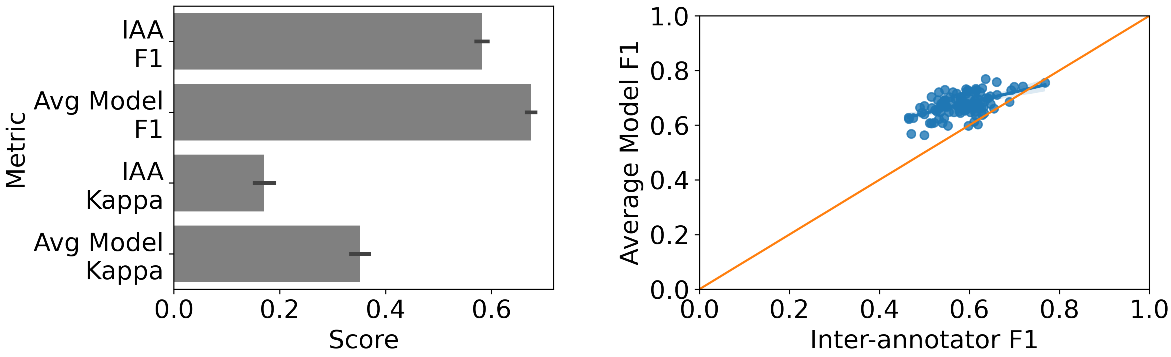
3.1 Simulations

First, it seems unrealistic that two different annotators, with different experiences and perceptual and cognitive systems, will ever understand and perform an annotation task in exactly the same way. In a sentiment analysis task, for example, annotators may have different thresholds for what is considered a 'positive' text. In other words, it seems unrealistic that two annotators will have the exact same classification function. One straightforward way to relax this assumption is to allow different annotators to have different intercepts in the linear component of the logistic function, as in:

$$p(y_i = 1) = \frac{1}{1 + e^{-(x_i + b_j)}} \tag{4}$$

where b_j is the intercept for annotator $j \in \{1, 2\}$. If, for example, $b_2 > b_1$, Annotator 2 will generally be more likely than Annotator 1 to assign a sample to the positive class. In our experiments, we will simply assume $b2 \geq 0$ and $b1 = -b2$ (other combinations of intercepts, like setting $b_1 = 0$ and varying b_2, were also tested and the general pattern of results did not change). This will of course decrease IAA. It may also seem intuitive that, when annotators systematically disagree about how to approach the task, it will be more difficult for a model to learn anything coherent, decreasing model performance, possibly to a performance worse than IAA.

Second, it also seems possible, in practice, that annotators' judgments are more deterministic (less noisy) than implied in Experiment 1, where the IAA Cohen's Kappa averaged only 0.18. This likely strikes most ML practitioners as much lower than what is seen and accepted in empirical ML applications. It therefore seems reasonable that a given annotator, facing a sample twice, would generally classify it the same way each time (i.e., that *intra*-annotator agreement is high). There are various ways to parameterize determinism in annotation, but we opt to simply exponentiate the outputs of the logistic function by a parameter, γ,



(a) Bars show means across simulations, and error bars display 95% confidence intervals.

(b) Scatterplot of inter-annotator F1 (x-axis) against Average Model-Annotator F1 (y-axis), across simulations. Each dot represents a single simulation

Figure 1: Experiment 1 results

and then divide these values by their sum so they add to 1, as proper probabilities, as in:

$$a_{1,i} = \frac{1}{1+e^{-(x_i+b_j)}} \tag{5}$$

$$p(y_i = 1) = \frac{a_{1,i}^{\gamma}}{a_{1,i}^{\gamma} + (1-a_{1,i})^{\gamma}} \tag{6}$$

Thus, instead of viewing the logistic function as producing probabilities, we can view it (Equation 5) as producing 'activations' of the possible annotations in the annotator's mind, i.e., $a_{1,i}$ and $a_{0,i}$ refer to the activations of labels 1 and 0, respectively, for the i-th sample. These activations are then converted into probabilities by Equation 6. When $\gamma = 0$, the choice is completely random (i.e., $p(y_i = 1) = 0.5$) and does not depend on x_i and b_j. When $\gamma = 1$, then annotation probabilities of Equation 6 are identical to the activations produced by Equation 4. When $\gamma > 1$ and approaches positive infinity, choice becomes more deterministic, such that with an extremely high γ, an annotator will almost always classify a sample x_i as positive if $a_{1,i} > 0.5$. One might argue that, to more accurately model commonly seen levels of IAA metrics, we need to test $\gamma > 1$, which ought to boost IAA and perhaps therefore make it harder for ML to outperform IAA.

Third, and perhaps most importantly and obviously, a machine learning model will always be misspecified in some way (Box, 1976). That is, the ML model will almost always lack some of the variables that influence an annotator's judgment, or the ML model may be purely linear while annotators are actually using some nonlinear combination of variables. Although it may seem obvious that, if the model is misspecified enough, ML performance will fall short of IAA, we also simulate this condition to show that *the model does not need to be perfectly specified to beat IAA*. To simulate misspecification, we simply augment Equation 5 with a second independent variable, x_2, as in:

$$a_{1,i} = \frac{1}{1+e^{-(x_{1,i}+m*x_{2,i}+b_j)}} \tag{7}$$

We assume that, like x_1, x_2 is sampled (independently) from the standard normal distribution (i.e., x_1 and x_2 together constitute a standard multivariate normal distribution). We also assume that there is a coefficient m on x_2 controlling the relative importance of x_2 to annotator decisions. We then assume that annotators' judgments follow from Equations 7 and 6. To misspecify an ML model, we simply withhold x_2 from X_{train} and X_{test} when fitting the model and generating $\hat{y}_{test}$, respectively. That is, annotators make decisions with both x_1 and x_2, but the model only has access to x_1. When m is large, reflecting great importance of x_2 to annotator decisions, then the ML model is greatly misspecified and this misspecification will have large negative impacts on $average(f(\hat{y}_{test}, y_{test,A1}), f(\hat{y}_{test}, y_{test,A2}))$. When $m = 0$, of course, x_2 is ignored in annotators' decisions and the ML model is not misspecified at all – the omission of x_2 from the ML model has no effect on its performance.[2]

[2] We note that increasing m can also increase IAA because it will push the output of Equation 7 toward 0 or 1, which in turn makes annotators' labels less noisy. Although it may be undesirable for m to influence both misspecification and

We emphasize that m is just one simple way to introduce misspecification in the simulations. In more complex real world tasks with more complex models (such as deep learning), misspecification can take many different forms. Exploring this further may be a useful avenue in future work.

Finally, although it is generally intentional and desirable that ML models classify samples deterministically (i.e., $\hat{y}_i = 1$ if and only if $p(y_i = 1) \geq 0.5$), we can simulate a noisy ML model to better understand the conditions under which ML can or cannot beat IAA. That is, it seems intuitive that one advantage an ML model has over human annotators, is that an ML model can make decisions with perfect consistency. To simulate a noisy ML model, we simply pass a trained ML model's predicted probabilities through Equation 6 and sample its predictions accordingly.

3.2 Results

We sample b_2 from $\{0, 0.25, 0.5\}$, γ from $\{1, 3, 6\}$, and m from $\{0, 0.25, 0.5\}$. The ranges of b_2 and γ were chosen so that reasonably high IAA could be achieved despite individual differences, while the range of m was chosen so that we had values of m that lead to ML > IAA, and values of m such that ML < IAA. We also simulate both fully deterministic and noisy model predictions. In the latter case, the ML model uses the same value of γ that simulated annotators use. We simulate all possible combinations of parameters and conditions. As in Experiment 1, our simulations involve X_{test} of 100 samples and X_{train} of 1000 samples, but now we run 400 simulations per combination of parameters. Because F1 and Cohen's Kappa show the same general pattern of results, we only use F1 to compare IAA and model performance in Experiment 2.

Figure 2 shows, for each combination of parameter values, the difference between average model F1 and annotator F1, such that bars above $y = 0$ indicate that the model outperforms IAA. Table 1 shows the same results as Figure 2, but transposes Figure 2's arrangement of parameter combinations, and just shows whether ML outperforms IAA or vice versa. As can be seen in both Figure 2 and Table 1, ML outperforms IAA across a broad range of conditions.

First, perhaps contrary to intuition, ML can outperform IAA when annotator classification functions differ, i.e., when $b_2 \neq b_1$. In fact, the larger the difference $b_2 - b_1$, the larger the margin by which the model beats IAA (e.g., compare the 1st, 2nd and 3rd blue or orange bars in any subplot of Figure 2). Rather than causing the model to learn something incoherent, $b_2 \neq b_1$ causes the model to learn a $\hat{b}$ that compromises between b_2 and b_1. For example, in the simple case $b_1 = -0.5$ and $b_2 = 0.5$ (and $\gamma = 1$), the model will tend to learn $\hat{b} = 0$. This causes the model's predictions to be, on average, closer to either annotator's predictions than the annotators' predictions are to each other.

Second, even if we increase determinism in annotator judgments (via γ in Equation 6) such that IAA reaches levels typically seen in empirical applications (e.g., Kappa = 0.6 or 0.7, see bottom row of Figure 2 subplots), ML can still beat IAA.

Third, ML can outperform IAA even under some model misspecification ($m = 0.25$ or $m = 0.5$), although misspecification reduces the margin by which ML outperforms IAA (e.g., compare top row subplots of Figure 2, or more strikingly, middle or bottom row subplots).

Fourth, although determinism in model predictions is clearly an advantage ML has over noisy human annotators (blue bars are generally higher than orange bars in Figure Figure 2), it is not necessary for ML to beat IAA. Systematic differences in annotator behavior are sufficient, as can be seen in the right most bars of the first and second subplots in the top row of Figure 2. Although these differences between ML and IAA are quite small, they are statistically significant, as indicated by the 95% confidence intervals excluding 0.

Most importantly, we note that ML beats IAA in a realistic combination of conditions, i.e., when annotators have good IAA ($\gamma = 6$, Kappa=0.61) despite (small) systematic differences in behavior ($-b_1 = b_2 = 0.25$), and the ML model is mildly misspecified ($m = 0.25$). In Figure 2, this situation is represented in the middle bars of the subplot of the third row and second column, which is surrounded by a black box.

4 General Discussion

In a comprehensive range of simulations, we showed that, contrary to popular belief (Boguslav and Cohen, 2017), *inter-annotator agreement is not the upper bound on machine learning performance*. We showed this is the case even if (and especially if) annotators are noisy and differ in their under-

IAA, it is not immediately obvious how to better parameterize misspecification, and in any case, we don't think this property affects our conclusions.

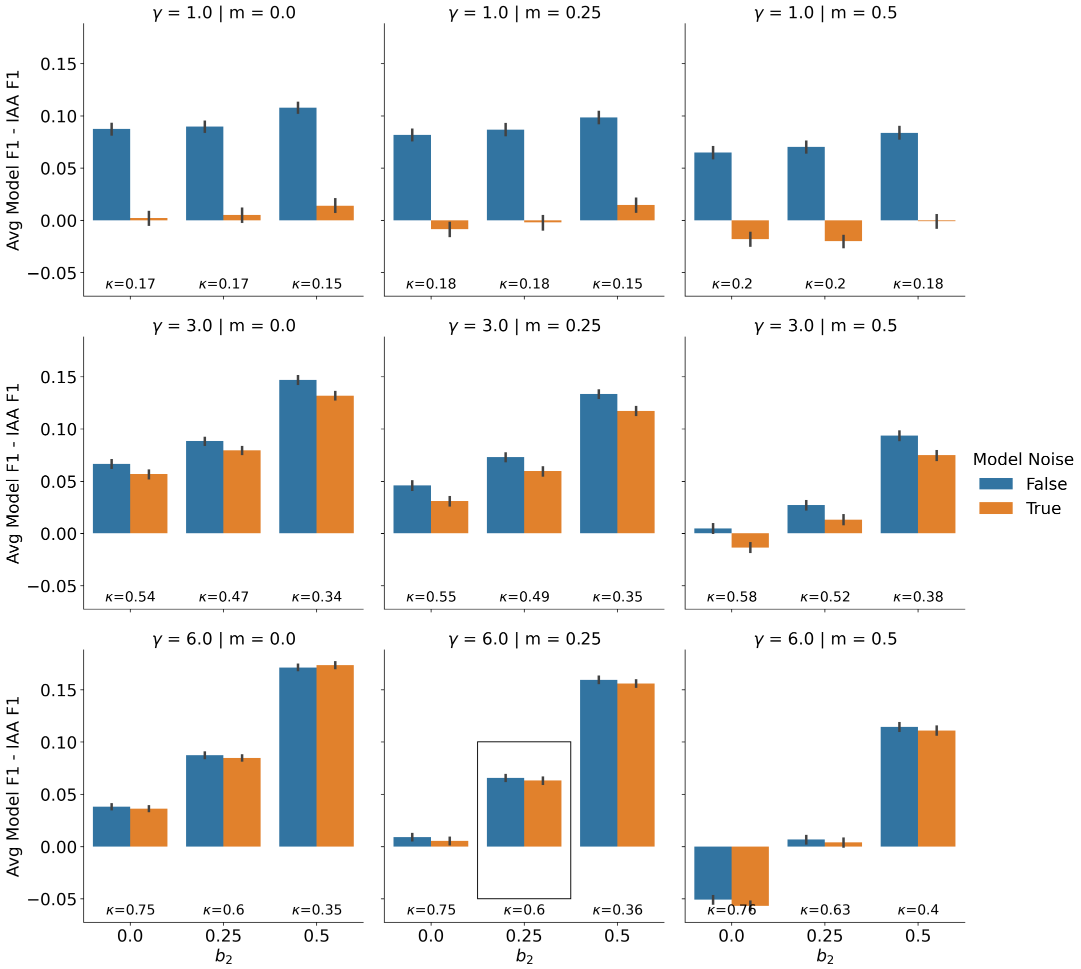


Figure 2: Experiments 2 results. Bars show mean differences between average Model F1 scores and IAA F1 score, i.e., bars above $y = 0$ indicate ML outperforming IAA. Error bars represent 95% confidence intervals. Below the bars are inter-annotator Cohen's Kappa's at each level of γ, m, and b_2. The black box in the middle column and bottom row represents a realistic condition, where ML still beats IAA.

lying classification functions, as long as the ML model is reasonably well-specified. While we think noisy annotators with (possibly small) systematic individual differences are the norm rather than the exception, well-specified models have been elusive for a long time in domains with unstructured data like (biomedical) NLP or machine learning. This was especially true in decades past, when the belief that IAA bounded ML proliferated, and this ill-specification likely led ML models to underperform IAA. However, reasonably well-specified models are likely to be increasingly attainable in today's era of big data, increased computing power, and correspondingly complex nonlinear models like deep neural networks. Although these real-world cases involve much more complex data and models than we simulated here, we believe our conclusions still apply, and we therefore expect to see more empirical cases of ML outperforming IAA (like those in Boguslav and Cohen, 2017). Likewise, although we focused on binary classification here, we expect our results to generalize straightforwardly to other settings, like multiclass classification or regression.

On the other hand, whether and how much a model will beat IAA depends, as we have shown, on the degree of model misspecification, the degree of noise in annotators' judgments, the degree of individual differences in the annotators, and possibly other factors. An ML practitioner might therefore wish to determine, given a particular annotated dataset, how well-specified the model must be in order to beat IAA by a given margin. Modeling one's annotators (e.g., Passonneau and Carpenter, 2014), and their noise levels and individual differences, may be useful here. Beyond this, it is unclear how best to perform such an analysis, and thus we leave this to future work. For the time being, then, we simply recommend that researchers not claim

			γ		
m	b_2	Model Noise	1	3	6
0	0	False	ML >IAA	ML >IAA	ML >IAA
		True	ML >IAA	ML >IAA	ML >IAA
	0.25	False	ML >IAA	ML >IAA	ML >IAA
		True	ML >IAA	ML >IAA	ML >IAA
	0.5	False	ML >IAA	ML >IAA	ML >IAA
		True	ML >IAA	ML >IAA	ML >IAA
0.25	0	False	ML >IAA	ML >IAA	ML >IAA
		True	**IAA >ML**	ML >IAA	ML >IAA
	0.25	False	ML >IAA	ML >IAA	ML >IAA
		True	**IAA >ML**	ML >IAA	ML >IAA
	0.5	False	ML >IAA	ML >IAA	ML >IAA
		True	ML >IAA	ML >IAA	ML >IAA
0.5	0	False	ML >IAA	ML >IAA	**IAA >ML**
		True	**IAA >ML**	**IAA >ML**	**IAA >ML**
	0.25	False	ML >IAA	ML >IAA	ML >IAA
		True	**IAA >ML**	ML >IAA	ML >IAA
	0.5	False	ML >IAA	ML >IAA	ML >IAA
		True	**IAA >ML**	ML >IAA	ML >IAA

Table 1: Experiment 2 results. Cells indicate whether the mean ML model F1 outperforms IAA F1, or vice versa. Bolded are the few settings in which ML does not outperform IAA.

that IAA is the ceiling of ML performance on their dataset. (Relatedly, for consideration of what, if not IAA, constitutes the upper bound on ML performance, we refer the reader to the discussion section of Boguslav and Cohen, 2017).

We realize that the simulations are so simple that our results and their implications may seem obvious. To an extent, we share this impression. At the same time, the persistence of the belief that IAA bounds ML performance, *despite any evidence or argument in support of this claim, and despite empirical evidence contrary to the claim* (Boguslav and Cohen, 2017), suggests that the results are *not* intuitive, at least for a large number of practicing ML users (the smaller number of theoretical statistics and machine learning researchers may not be surprised by the present results). We are not entirely certain why the belief that IAA bounds ML has persisted – and, to some extent, this is a psychological and sociological question outside the scope of our work – but we suspect there are at least a few culprits. First, as Boguslav and Cohen (2017) pointed out, this belief makes our models appear better than they are, and it may be the case that ML users were therefore eager to believe that IAA bounded ML. Second, as noted above, most models to-date have been (enormously) misspecified, so most models will tend to fall short of IAA. Third, as we showed in Experiment 1, IAA positively correlates with ML performance. These latter two facts combined may give the appearance of IAA "pushing down on" ML performance (see especially Mozetič et al., 2016 and their Figure 1 or Richie et al., 2019 and their Figure 3 for possible cases of this reasoning).

Regardless of the reasons that the belief IAA bounds ML persisted in the past, our results ought to help dispel this belief in the future, and thereby help researchers realize the full potential of machine learning models, adhere to ethical standards in reporting the performance of computational systems, and use expensive annotated resources more efficiently.

Acknowledgements

Thanks to Luiz Eduardo, Victor Ruiz, Mayla Boguslav, and Kevin Cohen for helpful discussion, as well as three anonymous reviewers for their feedback. This work supported is by the National Science Foundation under Grant No. 2031150 and Children's Hospital of Philadelphia.

References

Jacopo Amidei, Paul Piwek, and Alistair Willis. 2018. Rethinking the agreement in human evaluation tasks. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3318–3329.

Ronald E Anderson. 1992. ACM code of ethics and professional conduct. *Communications of the ACM*, 35(5):94–99.

Valerio Basile. 2020. It's the end of the gold standard as we know it. On the impact of pre-aggregation on the evaluation of highly subjective tasks. In *2020 AIxIA Discussion Papers Workshop, AIxIA 2020 DP*, volume 2776, pages 31–40. CEUR-WS.

Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, Roberto Navigli, et al. 2021. Recent trends in word sense disambiguation: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. International Joint Conference on Artificial Intelligence, Inc.

Mayla Boguslav and Kevin Bretonnel Cohen. 2017. Inter-annotator agreement and the upper limit on machine performance: Evidence from biomedical natural language processing. *Studies in health technology and informatics*, 245:298–302.

George EP Box. 1976. Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Shay Cohen. 2020. Methods in annotation and evaluation. https://www.inf.ed.ac.uk/teaching/courses/fnlp/lectures/10_slides-2x2.pdf.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, pages 406–414.

Simon B Goldberg, Michael Tanana, Zac E Imel, David C Atkins, Clara E Hill, and Timothy Anderson. 2021. Can a computer detect interpersonal skills? Using machine learning to scale up the Facilitative Interpersonal Skills task. *Psychotherapy Research*, 31(3):281–288.

Kyle Gorman. 2020. Inter-annotator Agreement. http://m.mr-pc.org/t/ling83800/2020sp/lecture11handout.pdf.

Miha Grčar, Darko Cherepnalkoski, Igor Mozetič, and Petra Kralj Novak. 2017. Stance and influence of Twitter users regarding the Brexit referendum. *Computational social networks*, 4(1):1–25.

Louis Guttman. 1945. A basis for analyzing test-retest reliability. *Psychometrika*, 10(4):255–282.

Na-Rae Han. 2017. Linguistic annotation. https://sites.pitt.edu/~naraehan/ling1340-2017/Lecture15.pdf.

Martin N. Hebart, Charles Y. Zheng, Francisco Pereira, and Chris I. Baker. 2020. Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature Human Behaviour*, 4:1173–1185.

Ryuichiro Higashinaka, Luis F D'Haro, Bayan Abu Shawar, Rafael E Banchs, Kotaro Funakoshi, Michimasa Inaba, Yuiko Tsunomori, Tetsuro Takahashi, and João Sedoc. 2021. Overview of the dialogue breakdown detection challenge 4. In *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction*, pages 403–417. Springer.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

George Hripcsak and Adam S Rothschild. 2005. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American medical informatics association*, 12(3):296–298.

Dan Jurafsky and James H Martin. 2022. Word sense disambiguation. https://web.stanford.edu/~jurafsky/slp3/slides/Chapter18.wsd.pdf.

Xinhang Li, Hao Liu, Fabrício Kury, Chi Yuan, Alex Butler, Yingcheng Sun, Anna Ostropolets, Hua Xu, and Chunhua Weng. 2021. A Comparison between Human and NLP-based Annotation of Clinical Trial Eligibility Criteria Text Using The OMOP Common Data Model. In *AMIA Annual Symposium Proceedings*, volume 2021, page 394. American Medical Informatics Association.

Elijah Mayfield and Alan W Black. 2020. Should you fine-tune BERT for automated essay scoring? In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 151–162.

Igor Mozetič, Miha Grčar, and Jasmina Smailović. 2016. Multilingual Twitter sentiment classification: The role of human annotators. *PloS one*, 11(5):e0155036.

Karen O'Connor, Abeed Sarker, Jeanmarie Perrone, Graciela Gonzalez Hernandez, et al. 2020. Promoting reproducible research for characterizing nonmedical use of medications through data annotation: Description of a Twitter corpus and guidelines. *Journal of medical Internet research*, 22(2):e15861.

Rebecca J Passonneau and Bob Carpenter. 2014. The benefits of a model of annotation. *Transactions of the Association for Computational Linguistics*, 2:311–326.

Verónica Pérez-Rosas, Mohamed Abouelenien, Rada Mihalcea, and Mihai Burzo. 2015. Deception detection using real-life trial data. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 59–66.

Carolyn Petersen, Jeffery Smith, Robert R Freimuth, Kenneth W Goodman, Gretchen Purcell Jackson, Joseph Kannry, Hongfang Liu, Subha Madhavan, Dean F Sittig, and Adam Wright. 2021. Recommendations for the safe, effective use of adaptive CDS in the US healthcare system: An AMIA position paper. *Journal of the American Medical Informatics Association*, 28(4):677–684.

Mohammad Taher Pilehvar, Dimitri Kartsaklis, Victor Prokhorov, and Nigel Collier. 2018. Card-660: Cambridge rare word dataset - A reliable benchmark for infrequent word representation models. *arXiv preprint arXiv:1808.09308*.

Kader Pustu-Iren, Markus Mühling, Nikolaus Korfhage, Joanna Bars, Sabrina Bernhöft, Angelika Hörth, Bernd Freisleben, and Ralph Ewerth. 2019. Investigating correlations of inter-coder agreement and machine annotation performance for historical video data. In *International Conference on Theory and Practice of Digital Libraries*, pages 107–114. Springer.

Vinicius Ribeiro, Sandra Avila, and Eduardo Valle. 2019. Handling inter-annotator agreement for automated skin lesion segmentation. *arXiv preprint arXiv:1906.02415*.

Russell Richie, Wanling Zou, Sudeep Bhatia, and Simine Vazire. 2019. Predicting high-level human judgment across diverse behavioral domains. *Collabra: Psychology*, 5(1).

Angus Roberts, Robert Gaizasukas, Mark Hepple, and Yikun Guo. 2008. Combining terminology resources and statistical methods for entity recognition: an evaluation. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.

Sebastian Ruder. 2021. Challenges and Opportunities in NLP Benchmarking. `http://ruder.io/nlp-benchmarking`.

Abeed Sarker, Graciela Gonzalez-Hernandez, Yucheng Ruan, and Jeanmarie Perrone. 2019. Machine learning and natural language processing for geolocation-centric monitoring and characterization of opioid-related social media chatter. *JAMA network open*, 2(11):e1914672–e1914672.

tomas (https://stats.stackexchange.com/users/84364/tomas). Inter-rater agreement of a gold standard dataset - a ceiling for reliable evaluation of algorithms? Cross Validated. URL:https://stats.stackexchange.com/q/165096 (version: 2020-10-28).

W John Wilbur. 1998. The knowledge in multiple human relevance judgments. *ACM Transactions on Information Systems (TOIS)*, 16(2):101–126.