HHAI 2025
D. Pedreschi et al. (Eds.)

doi:10.3233/FAIA250647

Underperformance or Pluralism: A Machine Learning Perspective on Inter-Annotator Agreement

Luana BULLA [a,b,1], Misael MONGIOVÌ [a,b] and Aldo GANGEMI [b,c]
[a] *University of Catania, Italy*
[b] *Consiglio Nazionale delle Ricerche-ISTC, Catania, Italy*
[c] *University of Bologna, Italy*
ORCiD ID: Luana Bulla https://orcid.org/0000-0003-1165-853X, Misael Mongiovì https://orcid.org/0000-0003-0528-5490, Aldo Gangemi https://orcid.org/0000-0001-5568-2684

Abstract. Standard metrics for measuring the performance of machine learning models work well when the data are objective, meaning there is only one correct output for a given input. However, for tasks where the outcome is subjective—such as opinions or morality—these metrics can only reference an average gold standard, potentially underestimating the model's performance. Inter-annotator agreement metrics, such as Fleiss' Kappa and Cohen's Kappa, can help estimate the subjectivity of a task. However, they do not provide a way to quantify what level of performance is satisfactory. Moreover, they are not always applicable in different settings, such as multi-label data, where annotators can assign multiple labels, or when the annotator set varies across samples. To address this challenge, we propose a novel inter-annotator agreement metric based on the F1-score, designed to support binary, multi-class, and multi-label data while accounting for incomplete annotations. Unlike traditional metrics, our approach enables direct comparison between human and machine annotations within a unified framework and accommodates varying levels of annotator coverage. Empirical validation on synthetic datasets demonstrates the proposed metric's robustness against annotation imbalances and incomplete sample coverage. Additionally, a case study on multi-label data shows that the proposed metric effectively highlights how modern LLMs match human performance in moral value classification.

Keywords. Inter-annotator agreement, Classification metrics, LLM evaluation, Machine learning

1. Introduction

The evolution of human-centered AI systems demands reevaluating traditional validation paradigms that treat human annotations and machine learning models as separate entities. While human annotators remain essential for creating benchmark datasets, modern workflows increasingly position AI models as active collaborators in annotation pro-

[1] Corresponding Author: Luana Bulla, luana.bulla@phd.unict.it

cesses [11]. This symbiotic relationship necessitates validation frameworks that operate on a unified plane, where human and machine performance can be directly compared through interpretable metrics. Current evaluation methodologies suffer from a fundamental dichotomy: machine learning models are assessed through detection-oriented metrics like F1-score, while human annotation quality is measured through agreement statistics such as Fleiss' Kappa. This creates an artificial separation between two intrinsically connected aspects of AI development. The F1-score optimizes for precision-recall tradeoffs but ignores task subjectivity, while Fleiss' Kappa measures consensus without quantifying concrete detection performance against a ground truth. This disparity impedes meaningful comparisons and obscures opportunities for human-AI quality synergies. A key challenge in model evaluation is the lack of a standardized framework for directly comparing machine learning performance with human annotation patterns. Traditional inter-annotator agreement (IAA) metrics like Fleiss' Kappa, while valuable for assessing annotation consistency, are not suitable for evaluating model performance. Conversely, model evaluation metrics provide no insights into how machine performance align with human annotation variability. This methodological gap is particularly problematic in subjective annotation tasks where model interpretability gaps and annotation uncertainty hinder robust validation. An example of this evaluative discrepancy emerges in the domain of moral value detection, a task that is inherently affected by subjectivity and interpretive ambiguity. In a recent model assessment conducted on a benchmark dataset in this field [1], the best-performing zero-shot large language model (i.e. GPT-4) achieves an overall F1-score of 0.55. However, the inter-annotator agreement, as measured by Fleiss' Kappa, is 0.36. Since the two values cannot be compared, whether the model's performance is satisfactory is unclear, considering that even human would not perform greatly in the same task. Other significant areas of application include sentiment and emotion analysis, as well as the investigation of social norms and human behavior through machine learning models. These domains are often hindered by a lack of comprehensive benchmark datasets, and are characterized by highly subjective annotations, which consequently result in low inter-annotator agreement.

To bridge this divide, we propose a unified validation framework that synthesizes model performance metrics and human agreement measures into a single evaluative plane. Our approach introduces a hybrid metric, F1-kappa, which harmonizes the precision-recall balance of F1-score with Fleiss' Kappa's expectation-adjusted design. By embedding agreement uncertainty into model evaluations (and vice versa), F1-kappa enables direct comparison between human and machine performance while quantifying their alignment. This metric addresses two critical limitations of current methodologies: the inability to compare model outputs with human annotations using a common scale, and the flexibility to accommodate incomplete annotation distributions and varying levels of annotator participation, thus facilitating a more accurate and comprehensive assessment of dataset heterogeneity.

We validate our approach through large-scale experiments spanning binary, multi-class, and multi-label classification tasks on synthetic and real datasets. Results demonstrate that the F1-kappa behave similarly to Fleiss' Kappa at various conditions. Furthermore F1-kappa supports any kind of settings including multi-label (more than one choice per annotator) and can focus on a specific class (e.g., positive samples), besides making human agreement comparable with model performance.

This paper outlines the following contribution:

- We propose a novel Inter-Annotator Agreement score, named F1-kappa, which generalizes F1 to more than two annotators and normalizes by the expected value in a similar way as Fleiss' Kappa. F1-kappa can be applied to any kind of categorical data and enables direct comparison with model performance.
- We propose a direct model-human comparison framework to evaluate models in classification tasks, addressing the subjectivity limitations of gold standards and enhancing the applicability of traditional machine learning metrics such as precision, recall, and F1 score.
- We perform a comprehensive assessment of the proposed metric across binary, multi-class, and multi-label scenarios, demonstrating its versatility and robustness under different conditions.
- We present a case study on a multi-label moral value dataset and show how the proposed framework is able to highlight that modern LLMs match human performance.

The paper is structured as follows. Section 2 presents related work. Section 3 briefly introduces Fleiss' Kappa, then describes our F1-kappa metric. Section 4 present results on both synthetic and real data. Eventually Section 5 concludes the work.

2. Related Work

The evaluation of machine learning models relies on well-established metrics such as precision, recall, and F1-score, which quantify performance based on the trade-off between false positives and false negatives [2]. However, these metrics are not directly applicable to human annotations, which are typically assessed using inter-annotator agreement metrics such as Cohen's Kappa [19], Fleiss's Kappa [6,5] and Krippendorff's alpha [12]. Cohen's kappa measures the agreement between two annotators while accounting for the probability of random agreement. It is widely used for binary and categorical classification tasks but it is not applicable when multiple annotators are involved. Fleiss' Kappa extends the metric to multiple annotators, computing an overall agreement score across a dataset while adjusting for chance agreement. However, it does not capture fine-grained differences in label distributions. Krippendorff's alpha generalizes agreement measurement to different data types, including ordinal, interval, and ratio scales, making it more flexible for diverse annotation scenarios. These metrics present limitations when applied to complex annotation tasks, as they do not capture the nuances of annotation subjectivity and variability across annotators [17]. Specifically, traditional agreement metrics assume a fixed ground truth and measure consistency among annotators rather than performance relative to a standard reference. This approach does not accommodate cases where ambiguity or granularity differences exist in the labeling process, leading to an incomplete representation of annotation reliability.

An alternative approach, the pairwise F1-score [10], has been proposed to address some of these limitations. Unlike Fleiss' Kappa, which measures overall agreement by considering categorical agreement rates, the pairwise F1-score evaluates consistency between pairs of annotators at the instance level. This method captures finer-grained differences in annotations by comparing label overlap rather than treating disagreements as binary mismatches. However, the pairwise F1-score does not support more than two annotators or a variable number of annotators per sample. Moreover, it does not account for

agreement occurring by chance, making it difficult to distinguish significant agreement from random chance. Our proposed metric builds upon these principles by introducing a more holistic framework that accounts for inter-annotator variability while maintaining a direct alignment with model evaluation methodologies.

Other approaches address the challenges associated with multi-class and multi-label tasks overcoming the limitations of both inter-annotator metrics and F1-score in fully capturing annotation inconsistencies. In multi-label and multi-class tasks, annotation discrepancies often arise due to the subjective interpretation of class boundaries and the varying levels of granularity employed by different annotators [13]. In Yang et al. [20], for example, segmenting structures with high inter-observer variability has highlighted the limitations of traditional agreement metrics, underscoring the need for more robust evaluation methodologies. While several approaches have been proposed to mitigate these discrepancies, such as crowdsourcing consensus models [3,4,16,7] and probabilistic annotation techniques [21,15], they remain insufficient for addressing the full complexity of human annotation behaviors. Specifically, recent studies have explored techniques such as probabilistic annotation models that assign confidence scores to labels rather than treating them as discrete categories, and crowdsourcing-based consensus methods that aggregate multiple annotator judgments to produce more reliable gold standards. However, these approaches primarily focus on optimizing annotation quality rather than directly aligning human annotations with model performance evaluation.

Some recent works attempt to integrate human annotations into the training process to overcome limitations related to insufficient training data, sparse and subjective annotations. Golazizian et al. [8] introduce a novel framework for efficient annotation collection and modeling. Their two-stage approach first uses a small group of annotators to build a multitask model. Then, they select and annotate a few samples per annotator to incorporate their unique perspective. To evaluate their framework at scale, they released the Moral Foundations Subjective Corpus, a dataset of 2,000 Reddit posts annotated by 24 annotators for moral sentiment. Their findings demonstrate that the framework surpasses previous state-of-the-art methods in capturing individual annotator perspectives while reducing the annotation budget by 75% across two datasets. Additionally, their approach fosters more equitable models by reducing performance disparities among annotators. In a complementary approach, van der Meer et al. [14] propose Annotator-Centric Active Learning (ACAL). This framework incorporates an annotator selection strategy alongside data sampling. ACAL aims to efficiently capture the full spectrum of human judgments in subjective tasks while evaluating model performance using annotator-centric metrics that prioritize minority perspectives over the majority. The authors experiment with various annotator selection strategies across seven subjective NLP tasks, using both traditional and newly developed human-centered evaluation metrics. Their results suggest that ACAL improves data efficiency and excels in capturing diverse human perspectives. However, its effectiveness depends on having a large and diverse pool of annotators for selection.

Despite these advancements, existing approaches remain limited in their ability to establish a direct comparison between human annotations and model performance. Consensus-driven methods prioritize agreement over capturing individual annotator biases, which can obscure important variations in human judgments. Additionally, while active learning strategies improve data efficiency, they do not fundamentally address the challenge of evaluating models and annotators within a unified framework. Our proposed

methodology extends beyond these limitations by introducing an evaluation metric that explicitly aligns annotation quality assessment with model performance measures, providing a more interpretable and robust comparison framework for complex annotation tasks.

3. A Unified Approach to Inter-Annotator Agreement and Evaluation

The proposed metric adapts the F1 score to measure inter-annotator agreement with more than two annotators while accounting for agreement by chance, taking inspiration from the popular Fleiss' Kappa score. In the following sections, we first describe Fleiss' Kappa, then introduce our F1-kappa score. Finally, we detail how to make the performance of a classifier comparable to F1-kappa.

3.1. Background

Perhaps the most known metric for measuring inter-annotator agreement for categorical data is Fleiss' Kappa [6]. Introduced by Joseph L. Fleiss in 1971, Fleiss' Kappa is a statistical measure designed to evaluate the degree of agreement among more than two annotators when assigning categorical ratings to a fixed number of items. It extends Cohen's Kappa, which is limited to two annotators, by accommodating multiple annotators and handling cases where different annotators evaluate different subsets of items. Compared to alternative reliability measures such as Krippendorff's Alpha or intra-class correlation coefficients (ICCs), Fleiss' Kappa is advantageous in cases where annotators do not necessarily rate all subjects and where categorical rather than continuous data are analyzed.

Fleiss' Kappa (κ) quantifies inter-annotator agreement by comparing the observed agreement among annotators to the agreement expected by chance. The measure accounts for variations in category distributions and is particularly useful in scenarios where ratings are nominal (categorical without a meaningful order). The value of Fleiss' Kappa ranges from -1 to 1, where higher values indicate stronger agreement: a κ of 1 signifies perfect agreement, 0 indicates agreement equal to chance, and negative values suggest disagreement greater than what would be expected randomly.

Let S be the set of samples, A be the set of annotators and k be the number of categories. The formula for Fleiss' Kappa is defined as:

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{1}$$

where P_o is the observed agreement, calculated as:

$$P_o = \frac{1}{|S|} \sum_{s \in S} \left(\frac{\sum_{j=1}^{k} n_{sj}(n_{sj} - 1)}{n_s(n_s - 1)} \right) \tag{2}$$

and P_e is the expected agreement by chance, given by:

$$P_e = \sum_{j=1}^{k} p_j^2 \tag{3}$$

where n_{sj} is the number of annotators who assigned category j to sample s, n_s is the total number of ratings for sample s and p_j is the proportion of all assignments to category j across all samples:

$$p_j = \frac{1}{|S|} \sum_{s \in S} \frac{n_{sj}}{n_s} \tag{4}$$

Note that in this formulation κ is undefined if there are samples without annotations or just one annotation, since eq. 2 would produce a division by zero. Therefore, these samples are removed from the dataset before computing κ.

3.2. A novel inter-annotator agreement metric based on F1-score

As described in the introduction, despite being a robust metric for measuring inter-annotator agreement, Fleiss' Kappa is not comparable to classification metrics widely used in machine learning. Therefore, it does not enable us to determine whether a given metric score is satisfactory, that is, if it is comparable to the performance of annotators. To fill this gap, we define a novel metric based on the F1-score. We remind that the F1-score is defined as the harmonic average between precision and recall, where precision is the fraction of correctly identified instances among all instances predicted as positive, and recall is the fraction of correctly identified instances among all actual positive instances. The F1-score is a versatile metric that can be applied to various classification scenarios, including binary, multi-class, and multi-label classification. The F1-score can be used to evaluate the agreement between two annotators, by considering one annotator as the gold standard and the other one as the predictor. However, there is no trivial way to accommodate more than two annotators and to manage a different set of annotators per sample. Moreover, the F1-score tends to overestimate the performances when the categories are balanced or unbalanced towards the negative class, as we show in our experimental analysis.

We start by describing our metric for binary data. The main idea is to use the F1-score for classification to evaluate the performance of each annotator and then normalize the result in a manner similar to how Fleiss' Kappa operates. In this way, we take advantage of both sides: we can support any kind of settings including multi-label (more than one choice per annotator), focus on a specific class (e.g., positive samples), and, at the same time, have a clear understanding of how much our performance stands out from chance. We consider a set of annotators A and a set of samples S. Each annotator $a \in A$ covers a subset of samples that we denote with S_a. We also denote as $f(a,s)$ the (0 or 1) value given by annotator a to sample s and with A_s the set of annotators that have annotated sample s.

To compute F1 we consider one annotator at a time and compute their classification performance considering the other annotations as the gold standard[2]. We adapt the precision and recall concepts to handle disagreement among the "gold" annotators by replacing the true values with the average "gold" annotations $Avg_{\neg a}$. We discard from S the set of samples s such that $|A_s| \leq 1$, since no agreement can be computed on them. We then update the sets S_a accordingly. Given a sample s we define:

[2] an equivalent alternative is compare each pairs of annotators, then averaging

$$Avg_{\neg a}(s) = \frac{1}{|A_s| - 1} \sum_{a' \in A_s / a} f(a', s) \tag{5}$$

F1 is computed for each annotator $a \in A$. We define it in terms of (adapted) precision (p_a) and recall (r_a) .

$$p_a = \frac{\sum_{s \in S_a} \min(Avg_{\neg a}(s), f(a,s))}{\sum_{s \in S_a} f(s,a)} \tag{6}$$

$$r_a = \frac{\sum_{s \in S_a} \min(Avg_{\neg a}(s), f(a,s))}{\sum_{s \in S_a} Avg_{\neg a}(s)} \tag{7}$$

$$\mathrm{F1}_a = 2 \cdot \frac{p_a \cdot r_a}{p_a + r_a} \tag{8}$$

We then average among all annotators:

$$\mathrm{F1} = \frac{1}{\sum_{a \in A} |S_a|} \sum_{a \in A} |S_a| \cdot \mathrm{F1}_a \tag{9}$$

For multi-class and multi-labels we consider each class separately and then compute the weighted average among all classes. Therefore, we substitute f with f_c where $f_c(a,s)$ is 1 if the annotator a assigns class $c \in \mathbf{C}$ (set of classes) to sample s, zero otherwise. In this case, F1 is computed as:

$$multiclass\_\mathrm{F1} = \frac{\sum_{c \in \mathbf{C}} w_c \cdot \mathrm{F1}^c}{\sum_{c \in \mathbf{C}} w_c} \tag{10}$$

where w_c is computed as:

$$w_c = \sum_{a \in A} \sum_{s \in S_a} Avg^c_{\neg a}(s) \tag{11}$$

with $\mathrm{F1}^c$ and $Avg^c_{\neg a}$ computed as F1 and $Avg_{\neg a}$, but replacing f with f_c.
As for Fleiss' Kappa we normalize F1 by computing its expected value, i.e., the value we would obtain by chance. To do so we define the expected precision $\hat{p}_a$ and expected recall $\hat{r}_a$ as:

$$\hat{p}_a = \frac{\sum_{s \in S_a} Avg_{\neg a}(s)}{|S_a|} \tag{12}$$

$$\hat{r}_a = \frac{\sum_{s \in S_a} f(a,s)}{|S_a|} \tag{13}$$

and compute the remaining expected metrics $\hat{\mathrm{F1}}_a$, $\hat{\mathrm{F1}}$ etc. accordingly.
Finally we define our metric F1-kappa as:

$$\mathrm{F1}\text{-}\kappa = \frac{\mathrm{F1} - \hat{\mathrm{F1}}}{1 - \hat{\mathrm{F1}}} \tag{14}$$

As for Fleiss' κ, the value of F1-κ ranges from -1 to 1, where higher values indicate stronger agreement: a F1-κ of 1 signifies perfect agreement, 0 indicates agreement equal

to chance, and negative values suggest disagreement greater than what would be expected randomly.

Now we describe how to evaluate the performance of a classifier in a way that is comparable with F1-κ. The proposed F1-κ metric can be interpreted as a measure of the average human performance in annotating a dataset and, therefore, serves as a reference for classification performance. The classifier evaluation is done by computing the same F1-κ with the only difference of substituting the annotations $f(a,s)$ of annotator a in eq. 6 and eq. 7 with the results of the classifier. We still need to exclude one annotator at a time from the gold standard to ensure that the evaluation is performed under the same conditions, thereby guaranteeing that the result can be directly compared with the inter-annotator agreement F1-κ.

4. Results and Evaluation

We conduct an experimental analysis to assess the effectiveness of the F1-κ metric in both binary and multi-label scenarios. In Section 4.1, we compare the performance of our metric with the adapted Fleiss' Kappa (cf. Sect. 3.1), aiming to assess their consistency. We conduct our evaluation on different synthetic datasets designed to simulate various binary classification scenarios. In Section 4.2, we show how F1-κ can be applied on a real-world multi-label classification scenario, specifically addressing the highly subjective domain of moral value identification.

4.1. Binary classification task

To assess the robustness of our metric, we construct synthetic datasets that emulate a binary classification task under varying annotation conditions. These datasets are designed to reflect real-world scenarios characterized by various degrees of agreement, label uncertainty, and varying rates of missing annotations. The data generation process begins with the creation of a reference truth vector, where binary labels are assigned to samples based on a predefined probability distribution (probability of positive). Annotators then provide labels with a certain probability of agreement with the reference truth. In cases of not agreement, the annotator generates a random label with the same probability distribution of the reference. Additionally, a proportion of annotations is intentionally omitted to simulate missing data. We compare F1-κ with the modified version of Fleiss' kappa (cf. Sect. 3.1), which enables handling missing values.

Figures 1 and 2 present the results of our proposed metric, F1-κ, alongside Fleiss' Kappa on a synthetic dataset of 500 samples annotated by five different annotators. Additionally, Figure 3 reports the performance of F1-score (Eq. 9), which lacks the normalization process introduced in F1-κ (Eq. 14). This comparison highlights the impact of our balancing mechanism, demonstrating the independence of the F1-κ metric from the percentage of positive cases under varying conditions. Unlike the conventional F1-score, which is influenced by the distribution of positive and negative instances, the F1-κ provides a more robust assessment of inter-annotator agreement, unaffected by class imbalance. We evaluate metrics on datasets with varying agreement probabilities (ranging from 0.1 to 1.0) and probabilities of positive labels set at 0.1, 0.5, and 0.9. Additionally, we report the confidence intervals calculated over 10 independent runs to account for variability in the results.

Figures 2 and Figure 1 show the F1-κ and Fleiss' Kappa scores across varying agreement probabilities and with different probability of positive values ($p = 0.3$, $p = 0.5$ and $p = 0.9$). Both metrics demonstrate a similar, generally linear, increase in agreement as the threshold is raised. Specifically, at a positive class probability of 0.1, both F1-κ and Fleiss' Kappa initiate near zero and incrementally approach values of approximately 0.3 and 0.8 at thresholds of 0.5 and 0.9, respectively, ultimately converging towards perfect agreement (1.0). Moreover, values are independent of the probability of positive values, showing stability (i.e. independence from the percentage of positive cases) at different degrees sparsity in the data. Notably, the importance of the balancing mechanism inherent in F1-κ becomes evident when contrasted with the F1-score without normalization by expected value (Figure 3). The F1-score exhibits a marked sensitivity to the probability distribution of classes. At low agreement probabilities (i.e., 0.1), the weighted F1-score closely mirrors the distribution of positive instances within the synthetic dataset. That is, F1-score values approximate the positive probability (e.g., $\sim$0.1 for p = 0.1, $\sim$0.5 for p = 0.5, and $\sim$0.9 for p = 0.9). This bias decreases with increasing agreement probabilities as the weighted F1-score asymptotically approaches 1.0. The behaviour depends on the fact that when the percentage of positive samples is high the chance of observing a true positive is high even in random data. This demonstrates how the balancing mechanism by expected value empowers F1-κ to provide a robust and reliable measure of inter-annotator agreement, exhibiting a behaviour comparable to Fleiss' Kappa in terms of resilience to class distribution imbalances within the dataset.

This trend is also observed in the presence of missing data. Figures 5, 4, and 6 illustrate the performance of F1-κ, Fleiss' Kappa, and the unbalanced F1-score, respectively, under varying percentages of missing data (0.1, 0.5, and 0.9) and agreement probabilities, with the probability of positive labels fixed at $p = 0.5$. The behaviour of Fleiss' Kappa and F1-κ remains highly similar, replicating the trends observed in the absence of missing data (Figures 2 and 4). However, the standard deviation of both metrics increases notably with a missing data rate of 0.9. Furthermore, at a low agreement probability of 0.1, both Fleiss' Kappa and F1-κ exhibit slightly lower values compared to the complete-data scenario (below the 0 threshold). This is due to the reduction in labeled data at high missing data rates. Comparison with the unbalanced F1-score (Figure 6) reveals results consistent with those presented in Figure 3. Specifically, the unbalanced F1-score initiates near 0.5, reflecting the balanced class distribution in the synthetic dataset, and converges towards 1.0 with increasing agreement probabilities.

4.2. Case study on moral values identification

To assess the performance of our metric on real-world multi-label data, we present a case study on moral value classification, a highly subjective task. Specifically, we examine the work of Bulla et al. [1], which assesses the performance of several large language models (LLMs) in detecting moral values within the Moral Foundation Reddit Corpus (MFRC) [18]. This dataset is grounded in Moral Foundations Theory (MFT) [9], which conceptualizes morality through five moral dyads: Care/Harm, Fairness/Cheating, Loyalty/Betrayal, Authority/Subversion, and Purity/Degradation. Each dyad represents a core dimension of human morality, offering a structured framework for analyzing moral expressions in textual data. In Bulla et al. GPT-4 achieves the highest overall performance, surpassing human annotators in all moral dimensions. However, the model ex-

Figure 1. Fleiss' κ with complete annotations

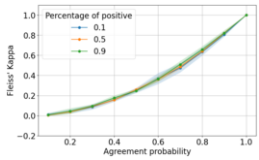


Figure 2. F1-κ with complete annotations

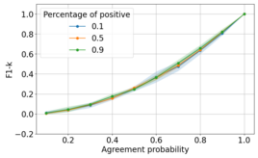


Figure 3. F1-score with complete annotations

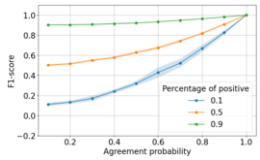


Figure 4. Fleiss' κ with missing annotations ($p = 0.5$)

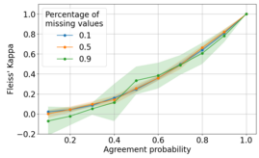


Figure 5. F1-κ with missing annotations ($p = 0.5$)

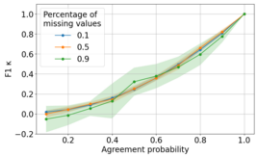
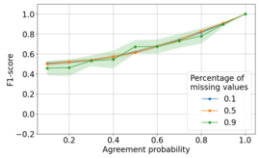


Figure 6. F1 with missing annotations ($p = 0.5$)



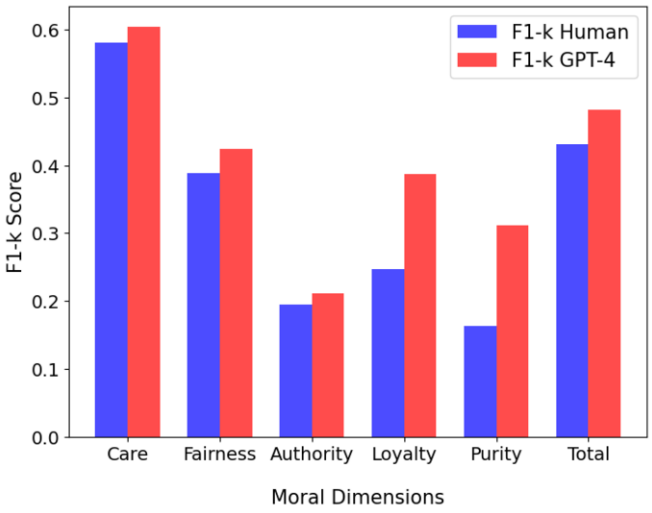
hibits challenges in distinguishing between dyads that are inherently more ambiguous in textual contexts (e.g., Authority/Subversion). To ensure a fair comparison between human and AI performance, the study implements an evaluation methodology that mitigates biases introduced by annotation aggregation, which is often susceptible to low inter-annotator agreement. Specifically, the authors adopt an F1-score-based approach, wherein model predictions are compared against aggregate annotations from the MFRC while systematically excluding one annotator at a time. This methodology enables a direct and equitable comparison between human annotators and model performance, with final results obtained by averaging across all annotators. The gold-standard annotations are determined by applying a threshold of 0.50 to individual annotation scores.

In our analysis, we compute the F1-κ metric for each moral dyad, enabling a direct comparison and assessing the consistency between GPT-4 predictions and annotator-assigned labels. We do not compare the F1-κ metric with Fleiss' Kappa, as the latter is not directly comparable to the former. Fleiss' Kappa measures inter-annotator agreement, whereas the F1-κ metric incorporates both precision and recall, which introduces a distinct evaluative dimension that cannot be directly aligned with Fleiss' Kappa in this context. This approach offers insights into our metric's adequacy in a real-world multi-label scenario, providing a setting for evaluating its effectiveness. To achieve this, we apply the F1-κ metric to estimate the inter-annotator agreement for each individual dyad as described in Section 3.2. We then compute the analogous metric by substituting the annotations of the reference annotator with the model's predictions, as described in the end of Section 3.2.

Figure 7 presents the F1-κ scores for all moral dyads on the MFRC dataset, comparing the predictions of GPT-4 with those of human annotators. For this comparison, we select the highest-performer model from Bulla et al., specifically the GPT-4 outcomes in a double-prompt scenario. In this scenario, GPT-4 is first prompted to determine whether the input conveys moral content and, if so, is prompted again to identify which moral dyads are present in the text through a two-step inference process. As depicted in figure (Fig. 7), our results are consistent with those reported in Bulla et al., where GPT-4 outperforms human annotators in all moral dyads, achieving the highest performance in Care and Fairness dyads. On average across all moral dyads, GPT-4 outperforms human annotators with an improvement of 5 percentage points in terms of F1-κ score (i.e. 0.48

vs 0.43). Specifically, GPT-4 achieves an F1-κ of 0.61 and 0.42 for Care and Fairness, compared to 0.58 and 0.39 for human annotations. However, for the more opaque dyads (i.e. Authority, Loyalty, and Purity), GPT-4's F1-κ scores are 0.21, 0.39, and 0.31, respectively, in contrast to 0.19, 0.25, and 0.16 for human annotators. The F1-κ values in Figure 7 are slightly lower than the F1 scores reported in Bulla et al., primarily due to the normalization process applied in our F1-κ metric (cf. Sect.3.2), which adjusts the results according to the expected F1-score, calculated based on the label distribution across the entire annotation corpus. This normalization enables adapting the scale such that a random level of agreement is lowered to 0. In contrast, the standard F1 score does not account for agreement by chance. Notably, the comparison with human annotations is crucial for accurately evaluating model performance in this inherently subjective task. By performing a direct comparison, we establish a scale that contextualizes the performance of both models and human annotators in relation to the subjectivity of the task. This comparative framework allows for a more nuanced interpretation of the results, positioning model and human performance with respect to the chance threshold and providing deeper insights into the model's alignment with human moral judgment.

Figure 7. The proposed F1-κ metric enables directly comparing inter-annotator agreement with the model's performance. In moral value detection, the model outperforms human annotators. This result is consistent with earlier findings.



5. Conclusion

We introduce F1-kappa, a novel inter-annotator agreement metric designed to bridge the gap between traditional agreement measures, such as Fleiss' Kappa, and standard machine learning evaluation metrics. By incorporating elements of both precision-recall trade-offs and chance-adjusted agreement calculations, our approach enables a unified framework for assessing both human annotations and model performance, particularly in tasks characterized by subjectivity. Empirical validation on both synthetic and real-

world datasets demonstrates the robustness of F1-kappa in handling annotation imbalances and incomplete coverage. This study is part of a broader research framework on evaluation metrics that incorporates human annotation variability into machine learning performance assessments. Potential directions include extending the methodology to account for annotators with differing levels of expertise, adapting the metric for hierarchical or structured annotation tasks, and incorporating it into active learning pipelines to optimize data collection. By aligning model evaluation with human-centered annotation processes, our approach contributes to a more comprehensive and interpretable assessment of AI systems in subjective classification tasks. A further direction goes to associating this work to the "harnessing disagreement" research [**?**], which is working on multilevel annotation models that separate task ambiguity from annotator variance, training models to predict the distribution of human opinions rather than point estimates, and using disagreement patterns to identify edge cases and improve robustness.

Acknowledgement

We acknowledge financial support from the Next Generation EU Program with the Future Artificial Intelligence Research (FAIR) project, code PE00000013, CUP 53C22003630006, and by the Italian Research Center on High Performance Computing, Big Data and Quantum Computing (ICSC).

References

[1] Bulla, L., De Giorgis, S., Mongiovì, M., Gangemi, A.: Large language models meet moral values: A comprehensive assessment of moral abilities. Computers in Human Behavior Reports p. 100609 (2025)
[2] Derczynski, L.: Complementarity, f-score, and nlp evaluation. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). pp. 261–266 (2016)
[3] Dumitrache, A., Inel, O., Aroyo, L., Timmermans, B., Welty, C.: Crowdtruth 2.0: Quality metrics for crowdsourcing with disagreement. arXiv preprint arXiv:1808.06080 (2018)
[4] Dumitrache, A., Inel, O., Timmermans, B., Ortiz, C., Sips, R.J., Aroyo, L., Welty, C.: Empirical methodology for crowdsourcing ground truth. Semantic Web **12**(3), 403–421 (2021)
[5] Falotico, R., Quatto, P.: Fleiss' kappa statistic without paradoxes. Quality & Quantity **49**, 463–470 (2015)
[6] Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychological bulletin **76**(5), 378 (1971)
[7] Genta, L., et al.: Consensus-based crowdsourcing: Techniques and applications (2015)
[8] Golazizian, P., Omrani, A., Ziabari, A.S., Dehghani, M.: Cost-efficient subjective task annotation and modeling through few-shot annotator adaptation. arXiv preprint arXiv:2402.14101 (2024)
[9] Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S.P., Ditto, P.H.: Moral foundations theory: The pragmatic validity of moral pluralism. In: Advances in experimental social psychology, vol. 47, pp. 55–130. Elsevier (2013)
[10] Hripcsak, G., Rothschild, A.S.: Agreement, the f-measure, and reliability in information retrieval. Journal of the American medical informatics association **12**(3), 296–298 (2005)
[11] Hsieh, C.Y., Li, C.L., Yeh, C.k., Nakhost, H., Fujii, Y., Ratner, A., Krishna, R., Lee, C.Y., Pfister, T.: Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In: Findings of the Association for Computational Linguistics: ACL 2023. pp. 8003–8017 (2023)
[12] Krippendorff, K.: Computing krippendorff's alpha-reliability (2011)
[13] Marchal, M., Scholman, M., Yung, F., Demberg, V.: Establishing annotation quality in multi-label annotations. In: Proceedings of the 29th international conference on computational linguistics. pp. 3659–3668 (2022)

[14] van der Meer, M., Falk, N., Murukannaiah, P.K., Liscio, E.: Annotator-centric active learning for subjective nlp tasks. arXiv preprint arXiv:2404.15720 (2024)
[15] Paun, S., Artstein, R., Poesio, M.: Statistical methods for annotation analysis. Springer Nature (2022)
[16] Pyatkin, V., Yung, F., Scholman, M.C., Tsarfaty, R., Dagan, I., Demberg, V.: Design choices for crowdsourcing implicit discourse relations: revealing the biases introduced by task design. Transactions of the Association for Computational Linguistics **11**, 1014–1032 (2023)
[17] Rainio, O., Teuho, J., Klén, R.: Evaluation metrics and statistical tests for machine learning. Scientific Reports **14**(1), 6086 (2024)
[18] Trager, J., Ziabari, A.S., Davani, A.M., Golazizian, P., Karimi-Malekabadi, F., Omrani, A., Li, Z., Kennedy, B., Reimer, N.K., Reyes, M., et al.: The moral foundations reddit corpus. arXiv preprint arXiv:2208.05545 (2022)
[19] Vieira, S.M., Kaymak, U., Sousa, J.M.: Cohen's kappa coefficient as a performance measure for feature selection. In: International conference on fuzzy systems. pp. 1–8. IEEE (2010)
[20] Yang, F., Zamzmi, G., Angara, S., Rajaraman, S., Aquilina, A., Xue, Z., Jaeger, S., Papagiannakis, E., Antani, S.K.: Assessing inter-annotator agreement for medical image segmentation. IEEE Access **11**, 21300–21312 (2023)
[21] Zhou, Y., He, W., Hou, W., Zhu, Y.: Pianno: a probabilistic framework automating semantic annotation for spatial transcriptomics. Nature Communications **15**(1), 2848 (2024)