Soft metrics for evaluation with disagreements: an assessment

Giulia Rizzi[1,2], Elisa Leonardelli[3], Massimo Poesio[4,5], Alexandra Uma, Maja Pavlovic[4], Silviu Paun, Paolo Rosso[2,6], Elisabetta Fersini[1]

[1]University of Milano-Bicocca, [2]Universitat Politècnica de València, [3]Fondazione Bruno Kessler
[4]Queen Mary University of London, [5]University of Utrecht
[6]ValgrAI - Valencian Graduate School and Research Network of Artificial Intelligence
g.rizzi10@campus.unimib.it, eleonardelli@fbk.eu, {m.poesio, m.pavlovic}@qmul.ac.uk,
{alexandra.uma2, spaun3691}@gmail.com, prosso@dsic.upv.es, elisabetta.fersini@unimib.it

Abstract

The move towards preserving judgement disagreements in NLP requires the identification of adequate evaluation metrics. We identify a set of key properties that such metrics should have, and assess the extent to which natural candidates for soft evaluation such as Cross Entropy satisfy such properties. We employ a theoretical framework, supported by a visual approach, by practical examples, and by the analysis of a real case scenario. Our results indicate that Cross Entropy can result in fairly paradoxical results in some cases, whereas other measures Manhattan distance and Euclidean distance exhibit a more intuitive behavior, at least for the case of binary classification.

1. Introduction

As the realization grows that disagreement between subjects in many natural language tasks may be the result of genuine differences in interpretation rather than of unclear guidelines or poor quality annotators (Poesio and Artstein, 2005; Passonneau et al., 2012; Plank et al., 2014; Aroyo and Welty, 2015; Akhtar et al., 2019; Basile et al., 2021; Uma et al., 2021b,a; Davani et al., 2022; Sap et al., 2022; Leonardelli et al., 2023), many researchers have started investigating methods for learning and evaluating models from datasets in which such differences in interpretation are preserved, particularly for subjective tasks (Basile et al., 2021; Uma et al., 2021b,a; Leonardelli et al., 2023). However, our understanding of this form of evaluation is still only at the beginning.

In this paper, we argue that soft evaluation metrics – metrics to evaluate the ability of NLP models to predict not just the preferred interpretation of an item, but also its probability and the probability of alternative interpretations according to human judgements, that Uma et al. called soft label (Uma et al., 2021b) – should satisfy a number of properties, that we define within a theoretical framework.

We then analyze four candidate metrics with respect of this set of formal properties. The metrics analysed include Cross Entropy, possibly the most widely used among such metrics, and which was also the main soft evaluation metric in the two recent Learning With Disagreements (LeWiDi) SemEval shared tasks (Uma et al., 2021a; Leonardelli et al., 2023). The other considered candidates are Manhattan Distance, Euclidean Distance and the Jensen-Shannon Divergence. For the binary label case, we also provide empirical examples and graphical visualizations of the metrics' behavior. Moreover we analyze how the metrics behave in a real case scenario, namely the LeWiDi shared task. Finally we discuss the case of multi-class labels.

One key result is that the widely used Cross Entropy metric has several counterintuitive properties, which other metrics considered do not suffer from, at least for the binary classification case. The situation is more complex for multi-label classification.

2. Soft Evaluation Metrics

The fundamental characteristic required of a soft evaluation metric is the ability to compare two probability distributions: the target distribution obtained from annotator judgments, and the distribution predicted by a model. In this Section, we introduce four metrics that have been used or could be used for such soft evaluation (Uma et al., 2021b; Basile et al., 2021; Uma et al., 2021a; Leonardelli et al., 2023).

Cross Entropy Cross Entropy is a common measure used in information theory and machine learning to quantify the difference between two probability distributions.
Given two distributions p, and q, their Cross Entropy is defined as:

$$\mathbb{H}(p, q) = \mathbb{E}_p\,[log q] = -\sum_k p(k)\log(q(k)) \quad (1)$$

Where $\mathbb{E}_p$ is the expected value operator with respect to the distribution p.

In the binary classification case, Cross Entropy simplifies to:

$$\mathbb{H}(p, q) = -[p\log(q) + (1-p)\log(1-q)] \quad (2)$$

NLPerspectives 2024@LREC-COLING 2024, pages 84–94

Manhattan Distance The Manhattan distance, also known as $\mathbb{L}_1$ distance measures the absolute differences between corresponding elements of two distributions. Given two distributions p and q, the Manhattan distance is defined as:

$$\mathbb{L}_1(p, q) = \sum_k |p(k) - q(k)| \quad (3)$$

Euclidean Distance The Euclidean distance, also known as $\mathbb{L}_2$ distance measures the the straight-line distance between two points in Euclidean space. Given two distributions p and q, the Euclidean distance is defined as:

$$\mathbb{L}_2(p, q) = \sqrt{\sum_k (p(k) - q(k))^2} \quad (4)$$

Jensen-Shannon Divergence (JSD) The Jensen-Shannon Divergence is a symmetrized and smoothed version of the Kullback-Leibler Divergence (KL Divergence). Given two distributions p and q, the Jensen-Shannon Divergence is defined as:

$$\mathbb{JSD}(p, q) = \frac{1}{2}\left(D_{KL}(p \parallel m) + D_{KL}(q \parallel m)\right) \quad (5)$$

Where D_{KL} is the Kullback-Leibler Divergence and $m = \frac{1}{2}(p + q)$. That corresponds to:

$$\mathbb{JSD}(p, q) = \frac{1}{2}\left(\sum_k p(k) \log\left(\frac{p(k)}{m(k)}\right) + \sum_k q(k) \log\left(\frac{q(k)}{m(k)}\right)\right) \quad (6)$$

where $m(k) = \frac{1}{2}(p(k) + q(k))$.

Although the Wasserstein distance is commonly used to quantify the difference between two probability distributions, it was not included in our analysis: it is crucial to highlight that, in the specific case of two binary distributions that are not rearranged, the Wasserstein distance reduces to the Manhattan distance.

3. Desirable properties

In this Section, we identify a set of properties that soft evaluation metrics should satisfy. We will use $q(k)$ to indicate the probability of an item k having the positive label according to the model, and $p(k)$ to indicate the real probability of k having the positive label according to the gold (soft) standard. Finally, we use $\mathbb{M}$ to indicate the general measure to quantify the difference between two probability distributions.

Property 1 [Symmetry] *Given two probability distributions $q(k)$ and $p(k)$ representing the probability of an item k being classified with the positive label and the corresponding real value associated with k in the golden standard,*

$$\mathbb{M}(p(k),q(k)) = \mathbb{M}(q(k),p(k))$$

Property 2 [Boundedness] *Given two probability distributions $q(k)$ and $p(k)$ representing the probability of an item k being classified with the positive label and the corresponding real value associated with k in the golden standard, there exist constants a and b such that, for every item k,*

$$a \leq \mathbb{M}(p(k),q(k)) \leq b$$

Property 3 [Triangle Inequality] *Given three probability distributions $q(k)$, $r(k)$, and $p(k)$ representing the probability of an item k being classified with the positive label by two different models ($q(k)$ and $r(k)$) and the corresponding real value associated with k in the golden standard ($p(k)$),*

$$\mathbb{M}(p(k),q(k)) + \mathbb{M}(q(k),r(k)) \geq \mathbb{M}(p(k),r(k))$$

Property 4 [Transitivity] *Given three probability distributions $q(k)$, $r(k)$, and $p(k)$ representing the probability of an item k being classified with the positive label by two different models ($q(k)$ and $r(k)$) and the corresponding real value associated with k in the golden standard ($p(k)$),*

$$\mathbb{M}(p(k),q(k)) < \mathbb{M}(p(k),r(k)) \Rightarrow \mathbb{M}(q(k),r(k)) < \mathbb{M}(p(k),r(k))$$

Property 5 [Sum invariant] *Given two probability distributions $q(k)$, and $p(k)$ representing the probability of an item k being classified with the positive label by a model ($q(k)$) and the corresponding real value associated with k in the golden standard ($p(k)$). A divergence $\mathbb{M}$ is sum invariant if whenever c is independent from p, q*

$$\mathbb{M}(c + p(k), c + q(k)) <= \mathbb{M}(p(k), q(k))$$

This property is strictly related to the following three subproperties:

Property 5.a [Minimum penalization at perfect match] *Given three probability distributions $q(k)$, $r(k)$, and $p(k)$ representing the probability of an item k being classified with the positive label by two different models ($q(k)$ and $r(k)$) and the corresponding real value associated with k in the golden standard ($p(k)$), if $p(k) = q(k)$ and $r(k) \neq p(k)$, then*

$$\mathbb{M}(p(k),q(k)) < \mathbb{M}(p(k),r(k))$$

Property 5.b [Fair penalization] *Given three probability distributions q(k), r(k) and p(k) representing the probability of an item k being classified with the positive label by two different models (q(k) and r(k)) and the corresponding real value associated with k in the golden standard (p(k)), if $|p(k) - q(k)| < |p(k) - r(k)|$, then*

$$\mathbb{M}(p(k),q(k)) < \mathbb{M}(p(k),r(k))$$

Property 5.c [Fair penalization on perfect match] *Given two probability distributions q(k), and p(k) representing the probability of an item k being classified with the positive label by two different models (q(k) and r(k)) and the corresponding real value associated with k in the golden standard (p(k)); given two items k_i and k_j, if p$(k_i) =$ q(k_i), p$(k_j) =$ q(k_j) and p$(k_i) \neq$ p(k_j), then*

$$\mathbb{M}(p(k_i), q(k_i)) = \mathbb{M}(p(k_j), q(k_j))$$

Property 6 [Scale sensitivity] *Given two probability distributions p(k) and q(k) representing the probability of an item k being classified with the positive label (q(k)) and the corresponding real value associated with k in the golden standard, (p(k)), We say that $\mathbb{M}$ is scale sensitive (of order β), if there exists a $\beta > 0$, and a real value $c > 0$, such that for all k*

$$\mathbb{M}(cp(k), cq(k)) <= |c|\beta\mathbb{M}(p(k), q(k))$$

If $\mathbb{M}$ is scale sensitive of order $\beta = 1$ then the divergence $\mathbb{M}(\delta, \delta_{1/2})$ can be no more than half the divergence $\mathbb{M}(\delta_0, \delta_1)$. If $\mathbb{M}$ is sum invariant, then the divergence of δ_0 to δ_0 is equal to the divergence of the same distributions shifted by a constant c, i.e., of δ_c to δ_{1+c}.

The above mentions set of peroperties are desired when evaluating soft metrics in order to ensure the fairness and the consistency of the evaluation process. In particular the *Simmetry property* ensure an objective evaluation, independent by the arrangement of the input data (i.e. regardless of whether we evaluate predictions against ground truth or vice versa). The *Boundedness* property guarantee that the evaluation values remains in a defined range, allowing for comparison among different models and facilitating the identification of outlayers. The *Triangle Inequality* property is essential for a consistent evaluation since it guarantee that composed metrics remain coherent and does not leads to contradictory results. Similarly, the *sum invariant* property ensuer the consistency of the metric when combined or aggregated. The *Transitivity* property guarantees consistency in comparisons across different instances or groups. It ensures the consistency of a model performances when comparing across different tasks, datasets, or experimental conditions. Finally, the *scale sensitivity* property guarantee that the metric correctly capture the magnitude of the differences among models. In other words, it ensure that sligh variations in the model's performance are reflected as a minor change in the metric score, while big changes in performance lead to a significant change in the metric score.

4. Metric properties assessment in the binary case

In this Section, we analyze the extent to which the evaluation metrics under consideration and presented in Section 2, satisfy the properties we deem desirable and presented in Section 3, in the case of binary labels. An analysis is performed (Section 4.1), focusing on the selected properties, providing theoretical background and practical examples when the defined properties are not fulfilled. In Section 4.2 a graphical representation of the metrics behaviour at different target distributions is shown. Furthermore, the figure is used as a visual support to discuss some metrics' properties. Finally, in Section 4.3 we compare metrics behaviour in the real case scenario of the LeWiDi competition.

4.1. Properties assessment and examples

In this section, properties are discussed with respect to selected metrics. Table 1 summarizes the properties satisfied by the metrics.

Property 1 All the selected metrics satisfy the *simmetry* property (P1), i.e. inverting target and prediction does not affect the result, except for Cross Entropy.
Cross Entropy, in fact, is *asymmetric*, given its relation to Kullback-Leibler Divergence. Cross Entropy is related to KL-Divergence as follows:

$$\mathbb{H}(p,q) = D_{KL}(p||q) + \mathbb{E}(p) \tag{7}$$

where $\mathbb{H}$ is the Cross Entropy of distribution *p* and *q*, $D_{KL}(p||q)$ is the KL-Divergence and $\mathbb{E}(p)$ is the Entropy of the distribution *p*. Since $\mathbb{E}(p)$ can be considered as a constant, Cross Entropy follows the same asymmetry of KL-Divergence. The definition of Cross Entropy reported in equation (1) leads to the following inequality:

$$-\sum_i p_i log q_i \neq -\sum_i q_i log p_i \tag{8}$$

Example 1 shows two distributions for which the symmetry property is not fulfilled by Cross Entropy: in the proposed example, $\mathbb{H}(p(k_1), q(k_1)) \neq \mathbb{H}(p(k_2), q(k_2))$, although $p(k_1) = q(k_2)$ and $q(k_1) = p(k_2)$.

Table 1: Properties of Evaluation Metrics (Binary Case)

Metric	Properties								
	P1	P2	P3	P4	P5	P5a	P5b	P5c	P6
Cross Entropy	✗	✓	✗	✗	✗	✓	✗	✗	✗
Manhattan Distance	✓	✓	✓	✓	✓	✓	✓	✓	✓
Euclidean Distance	✓	✓	✓	✓	✓	✓	✓	✓	✓
Jensen-Shannon Divergence	✓	✓	✓	✗	✗	✓	✗	✓	✓

Example 1 *[Symmetry violation]*

k	Target p(k)	Prediction q(k)
k_1	[0.83; 0.17]	[0.5; 0.5]
k_2	[0.5; 0.5]	[0.83; 0.17]

Cross Entropy values:

$\mathbb{H}(p(k_1), q(k_1))$ = 0.6931

$\mathbb{H}(p(k_2), q(k_2))$ = 0.9791

Property 2 In the binary case, all the selected metrics satisfy the *Boundedness* property (P2), i.e. they can only assume finite values. To note that Cross Entropy is left-bounded by definition, also given its relationship with the KL-Divergence. Commonly, it is bounded by introducing a smoothing that affects the extremants. But the scaling technique that is chosen to make the metric *bounded* has a great effect on the interval where the $\mathbb{H}$ values are distributed.

Property 3 The *Triangle inequality* property (P3) is satisfied by all metrics except Cross Entropy. When comparing two binary distributions, the asymmetry and sensitivity to specific distribution values of Cross Entropy can lead to instances where the distance between two distributions is not guaranteed to be less than or equal to the sum of their distances to a third distribution. Therefore, Cross Entropy does not consistently satisfy the *triangle inequality* property (P3).

Example 2 reports an example in which the Triangle Inequality property is not fulfilled by the Cross Entropy. Triangle Inequality property implies that the sum of the Cross Entropies for two *consecutive* predictions should be greater than or equal to the Cross Entropy between the target distribution and the direct prediction. However, in the proposed example, $\mathbb{H}(p(k), q(k)) + \mathbb{H}(q(k), r(k))$ is less than $\mathbb{H}(p(k), r(k))$ and contradicts the Triangle Inequality property for Cross Entropy.

Example 2 *[Triangle Inequality violation]*

Target p(k)	Prediction q(k)	Prediction r(k)
[0.7, 0.3]	[0.95, 0.05]	[1, 0]

Cross Entropy values:

$\mathbb{H}(p(k), r(k))$ = 8.2893

$\mathbb{H}(p(k), q(k)) + \mathbb{H}(q(k), r(k))$ = 2.3162

Property 4 The *Transitivity* property is satisfied by all metrics, except the Cross Entropy and the Jensen-Shannon divergence that do not consistently satisfy it.

Example 3 shows how despite $\mathbb{H}(p(k), q(k)) < \mathbb{H}(p(k), r(k))$, the expected transitivity property ($\mathbb{H}(q(k), r(k)) < \mathbb{H}(p(k), r(k))$) is not satisfied by Cross Entropy. Similarily, for Jensen-Shannon Divergence: despite $\mathbb{JSD}(p(k), q(k)) < \mathbb{JSD}(p(k), r(k))$, the expected transitivity property ($\mathbb{JSD}(q(k), r(k)) < \mathbb{JSD}(p(k), r(k))$) is not satisfied.

Example 3 (P4)*[Transitivity violation]*

Target p(k)	Prediction q(k)	Prediction r(k)
[0.9, 0.1]	[0.7, 0.3]	[1, 0]

Cross Entropy values:

$\mathbb{H}(p(k), q(k))$ = 0.4414

$\mathbb{H}(p(k), r(k))$ = 2.7631

$\mathbb{H}(q(k), r(k))$ = 8.2893

Jansen-Shannon values:

$\mathbb{JSD}(p(k), q(k))$ = 0.1801

$\mathbb{JSD}(p(k), r(k))$ = 0.1897

$\mathbb{JSD}(q(k), r(k))$ = 0.3425

Property 5.a In the binary case, all the selected metrics satisfy the *Minimum penalization at perfect match* property. Indeed for each possible target, the perfect match (the exact prediction of the target) assumes the minimum values possible for the target considered. (See also Figure 1 and relative discussion in Section 4.2)

Property P5.b Cross Entropy tends to penalize predictions that perfectly match the target distribution when the target distribution itself is characterized by a large entropy, resulting in an *unfair penalization*. This is because, as shown in Equation 7, when *p* is 'highly entropic', $\mathbb{E}(p)$ is large. Example 4 shows how Cross Entropy tends to unfairly penalize probability distributions close to the boundaries. Despite the error performed in the prediction q(k) is smaller then the one performed by the prediction r(k), $\mathbb{H}(p(k), q(k))$ is bigger than $\mathbb{H}(p(k), r(k))$ and contraddicts the *Fair penalization* property.

The Jensen-Shannon tends to penalize those cases that are less entropic (disregarding which distribution, target, or prediction, is more entropic than the other). Therefore, the Jensen-Shannon measure does not fulfill the *fair penalization* property (P5b). An example of distributions for which the property is not fulfilled is reported in Example 4. In fact, despite the error performed in the prediction q(k) is smaller then the one performed by the prediction r(k), $\mathbb{JSD}(p(k), q(k))$ is bigger than $\mathbb{JSD}(p(k), r(k))$ and contraddicts the *Fair penalization* property.

Example 4 (P5b)
Unfair penalization

Target p(k)	Prediction q(k)	Prediction r(k)
[0.9, 0.1]	[1, 0]	[0.7, 0.3]

Cross Entropy values:

$\mathbb{H}(p(k), q(k))$ = 2.7631
$\mathbb{H}(p(k), r(k))$ = 0.4414

Jensen-Shannon values:

$\mathbb{JSD}(p(k), q(k))$ = 0.1897
$\mathbb{JSD}(p(k), r(k))$ = 0.1801

Property P5.c Another effect of the entropy in the distribution on the Cross Entropy emerges when comparing the scores associated to different distributions that correctly predict the target. Example 5 reports an example showing that the *Fair penalization on perfect match* property is not satisfied: despite both distributions correctly predict the target, $\mathbb{H}(p(k_1), q(k_1))$ is not equal to $\mathbb{H}(p(k_2), q(k_2))$, due to the corresponding entropy in the distributions.

Example 5 (P5c)
Unfair penalization on perfect match

k	Target p(k)	Prediction q(k)
k_1	[0.5, 0.5]	[0.5; 0.5]
k_2	[0.9; 0.1]	[0.9; 0.1]

Cross Entropy values:

$\mathbb{H}(p(k_1), q(k_1))$ = 0.6932
$\mathbb{H}(p(k_2), q(k_2))$ = 0.3251

Property 6 Considering two binary distributions *p* and *q*, and a positive real value c; let *p′* and *q′* be scaled versions of *p* and *q* by the constant factor c: $p' = c \cdot p$ and $q' = c \cdot q$.

Cross Entropy: In the binary classification scenario, the Cross Entropy distance does not fulfill the scale sensitivity property. Substituting the scaled distributions into the Cross Entropy distance formula (Eq. 1), we obtain:

$$\begin{aligned}\mathbb{H}(p', q') &= -\sum_k c \cdot p(k) \log(c \cdot q(k)) \\ &= -c \cdot \sum_k p(k) \log(c \cdot q(k))\end{aligned} \tag{9}$$

By comparing this with $|c| \cdot \mathbb{H}(p, q)$ we obtain:

$$|c| \cdot \mathbb{H}(p, q) = |c| \cdot -\sum_k p(k) \log(c \cdot q(k)) \tag{10}$$

The two expressions are not directly proportional. Therefore, the Cross Entropy distance does not satisfy the scale sensitivity property.

Manhattan distance: In the binary classification scenario, the Manhattan distance satisfies the scale sensitivity property.

Considering two binary distributions *p* and *q*, the Manhattan distance between them is defined as shown in Eq. 3.

Substituting the scaled distributions into the Manhattan distance formula (Eq. 3), we obtain:

$$\begin{aligned}\mathbb{L}_1(p', q') &= \sum_i |c \cdot p(k) - c \cdot q(k))| \\ &= c \cdot \sum_k |p(k) - q(k)|\end{aligned} \tag{11}$$

By comparing this with $|c| \cdot \mathbb{L}_1(p, q)$ we obtain:

$$|c| \cdot \mathbb{L}_1(p, q) = |c| \cdot \sum_k |p(k) - q(k)|$$

Indicating that the Manhattan distance scales linearly with the constant factor c, fulfilling the scale sensitivity property with a sensitivity order (β) of 1.

Euclidean Distance: In the binary classification scenario, the Euclidean distance fulfills the scale sensitivity property.

Considering two binary distributions $p and q$, the Euclidean distance between them is defined as shown in Eq. 4.

Substituting the scaled distributions into the Manhattan distance formula (Eq. 4), we obtain:

$$\begin{aligned}\mathbb{L}_2(p', q') &= \sqrt{\sum_k (c \cdot p(k) - c \cdot q(k))^2} \\ &= c \cdot \sqrt{\sum_k (p(k) - q(k))^2}\end{aligned} \tag{12}$$

By comparing this with $|c| \cdot \mathbb{L}_2(p, q)$ we obtain:

$$|c| \cdot \mathbb{L}_2(p, q) = |c| \cdot \sqrt{\sum_k (p(k) - q(k))^2}$$

Indicating that the Euclidean distance scales linearly with the constant factor c, fulfilling the scale sensitivity property with a sensitivity order (β) of 1.

Jensen-Shannon: In the binary classification scenario, the Jensen-Shannon distance fulfills the scale sensitivity property.

Considering two binary distributions p and q, the Jensen-Shannon distance between them is defined as shown in Eq. 5.

Substituting the scaled distributions into the Manhattan distance formula (Eq. 5), we obtain:

$$\begin{aligned}\mathbb{JSD}(p', q') &= \frac{1}{2}\left(D_{KL}(p' \parallel m') + D_{KL}(q' \parallel m')\right) \\ &= \frac{1}{2}\left(D_{KL}(c \cdot p \parallel c \cdot m) + \right. \\ &\left. D_{KL}(c \cdot q \parallel c \cdot m)\right) \\ &= \frac{1}{2}\left(c \cdot D_{KL}(p \parallel m) + c \cdot D_{KL}(q \parallel m)\right)\end{aligned} \tag{13}$$

where $m'(k) = \frac{1}{2}(p'(k) + q'(k))$ and $m(k) = \frac{1}{2}(p(k) + q(k))$.

By comparing this with $|c| \cdot \mathbb{JSD}(p, q)$ we obtain:

$$\begin{aligned}|c| \cdot \mathbb{JSD}(p, q) = &\frac{1}{2}\left(|c| \cdot D_{KL}(p \parallel m) + \right. \\ &\left. |c| \cdot D_{KL}(q \parallel m)\right)\end{aligned} \tag{14}$$

Indicating that the Jensen-Shannon distance scales linearly with the constant factor c, fulfilling the scale sensitivity property with a sensitivity order (β) of 1.

4.2. Metrics graphical representation

Figure 1 shows distinct plots for each metric, with the x-axes representing the prediction values and the y-axes representing the corresponding distance values (or score) based on the metric under consideration. These plots provide a detailed visual representation of the metrics behaviors at different target values. Moreover, we can visually explore the properties, and in the following we discuss P5.a, P5.b and P5.c.

In Figure 1 we can observe how all the selected metrics satisfy the *Minimum penalization at perfect match property* (P5.a): for each target's curve plotted, the minimum values of the curve corresponds to the perfect match, i.e. the exact prediction of the tar get.

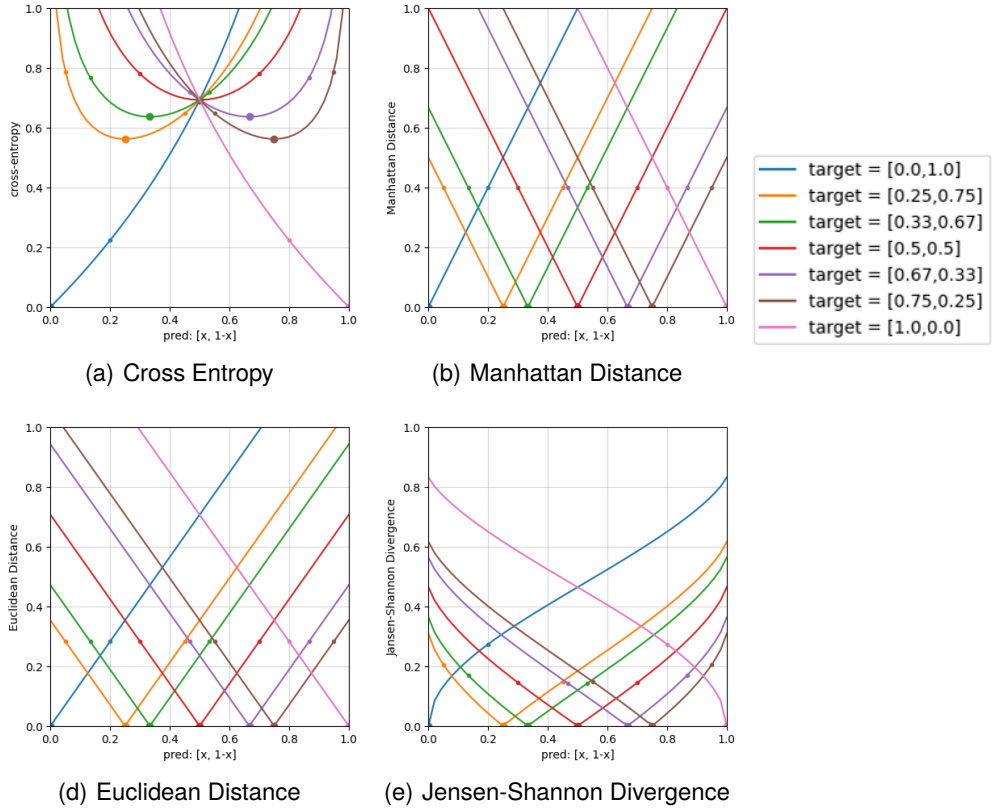
To demonstrate the influence of prediction errors on metric performance (P5.b), distance values when a nominal error of 0.2 appears in the forecast are highlighted with points within the same plots. This intentional perturbation enables an investigation of the metric's robustness in the presence of slight prediction mistakes, evaluating the ability of fair evaluation across a range of targets. Horizontal alignment between two prediction points that are equally distant from the target, indicate that property P5.b is respected (see the cases of Manhattan Distance and Euclidean Distance, Figure 1 b and c). Conversely, deviations from this horizontal alignment implies unfair penalizations (see the cases of Cross Entropy and Jensen Shannon Divergence, Figure 1a and d).

Finally, to contribute to a more detailed understanding of the *Fair penalization on perfect match* property (P5.c), within each plot, dots are used to highlight the resulting score, when the target is correctly predicted. The alignment of all dots along a horizontal axis (such as in the case of Figure 1b,c and d), indicates a fair penalty for perfect matches across targets. Deviations from this horizontal alignment, such in the case of Figure 1a (Cross Entropy) imply diverse penalization levels for perfect matches on diverse targets, revealing disparities in the metric's treatment of different target values.

4.3. Impact on a Leaderboard: The LeWiDi Case Study

In this section, we aim to investigate the application of some of the discussed evaluation metrics to a real case scenario. To this end, we exploit the data from a recent shared task, the Learning With Disagreements task (LeWiDi) (Leonardelli et al., 2023) proposed at the 2023 edition of SemEval [1]. The challenge proposed by the task foresees to

[1] https://semeval.github.io/



(a) Cross Entropy (b) Manhattan Distance

(d) Euclidean Distance (e) Jensen-Shannon Divergence

Figure 1: Metric Visualization: Each plot demonstrates the sensitivity of the analyzed metric to varying target values. X-axes represent prediction values, while Y-axes depict corresponding distance values according to the metric. Dots highlight distance values for accurately predicted targets, while points represent distance values when a nominal error of 0.2 appears in the prediction.

model the disagreements among annotators in four textual datasets that encompass different binary classification tasks (e.g. hate speech, offensive language, sexism detection). Teams competing in the shared task were asked to model annotators agreement/disagreement, represented in the form of soft labels: the probability of each item to be assigned to one class or the other is given by the agreement among annotators on the label. In the official competition to evaluate the performance of participants, Cross Entropy was considered the main evaluation metric. Here, for each of the four datasets that were part the LeWiDi task, rankings were recalculated for the evaluation metrics considered, and statistical difference was assessed from top to bottom using the Wilcoxon signed-rank test. Pairwise comparison among the different evaluation metrics, in terms of percentage of teams for which rank changed and the mean rank's position change, are summarized in Table 2. Results shown report the average value across the four datasets of the LeWiDi challange.

Table 2: Pairwise comparison of evaluation metrics rankings: percentage of teams ranked differently and mean position's change across the LeWiDi datasets

Evaluation metrics compared	% of teams re-ranked	Mean position change $\pm std$
Cross Entropy vs Manhattan distance	79%	2.1$\pm$2
Cross Entropy vs Euclidean distance	73%	2$\pm$2.1
Cross Entropy vs J-S Divergence	75%	2$\pm$2
Manhattan distance vs Euclidean distance	2%	0.1$\pm$0.2
Manhattan distance vs J-S Divergence	21%	0.4$\pm$0.6
J-S Divergence vs Euclidean distance	23%	0.4$\pm$0.6

From Table 2 we can observe how Cross Entropy rankings are substantially different from all the other metrics considered (although the mean position change is relatively small). On the contrary, the other metrics produce more homogeneous results, with Manhattan distance and Euclidean distance exhibiting almost no difference. This confirms that

the metrics' differences in adhering to the properties outlined above, exert a certain influence on the application of the metrics in the real cases.

5. Multiclass Classification

Additional analyses have been performed considering the Multiclass Classification scenario. The most promising metrics, selected through the binary classification analysis (i.e., Manhattan distance and Euclidean distance), have been evaluated with respect to further desirable properties defined in the scope of multiclass classification.

Property 7 *[Non-Invariance with respect to the Most Probable Label]* Given three probability distributions $q(k)$, $r(k)$, and $p(k)$ representing the probability of an item k being classified with the positive label by two different models ($q(k)$ and $r(k)$) and the corresponding real value associated with k in the golden standard ($p(k)$), let $\mathbb{M}(p, q)$ and $\mathbb{M}(p, r)$ denote the distance measure between the two probability distributions and the golden standard if the most probable label in q corresponds to the target distribution p, and the most probable label in r does not correspond to the target distribution p, then $\mathbb{M}(p(k), q(k)) < \mathbb{M}(p(k), (r(k)))$.

The proposed property is not fulfilled by the selected metrics. For instance, Example 6 reports an example in which two different predictions lead to the same value, according to the Manhattan distance. However, $r(k)$ leads to a wrong classification, while $q(k)$ still preserves the ground truth of the target distribution.

Example 6

Target p(k)	Prediction q(k)	Prediction r(k)
[0, 0.1,0.1,0.8]	[0.1, 0.3,0.2,0.4]	[0,0.1,0.5,0.4]

Manhattan Distance values:

$\mathbb{L}_1(p(k), q(k))$ = 0.8

$\mathbb{L}_1(p(k), r(k))$= 0.8

Similarly, Example 7 reports an example in which, despite the most probable label in the second prediction (r(k)) does not correspond to the most probable label in the target prediction (p(k)), it is considered closer, according to the Euclidean distance, with respect to the other prediction (q(k)). However, in the last prediction, the most probable label corresponds to the most probable label in the target prediction.

Example 7

Target p(k)	Prediction q(k)	Prediction r(k)
[0, 0.1,0.4,0.5]	[0.1, 0.2,0.3,0.4]	[0,0.1,0.5,0.4]

Euclidean Distance values:

$\mathbb{L}_2(p(k), q(k))$ = 0.2

$\mathbb{L}_2(p(k), r(k))$= 0.1414

Property 8 *[Positional Error Sensitivity for Multiple Labels]* Given three probability distributions $q(k)$, $r(k)$ and $p(k)$ representing the probability of an item k being classified with the positive label by two different models ($q(k)$ and $r(k)$) and the corresponding real value associated with k in the golden standard ($p(k)$), if $\sum_i |p_i(k) - q_i(k)| \leq \sum_i |p_i(k) - r_i(k)|$, then $\mathbb{M}(p(k), q(k)) \leq \mathbb{M}(p(k), (r(k)))$.

The Manhattan distance confers equivalent significance to a substantial error on a single label and to minor distributed errors across multiple labels relative to the target distribution. In other words, even if a prediction leads to performing the smallest number of errors (implying a more realistic prediction that is close to the target one) it has the same distance of a probability distribution that spreads the wrong prediction across the remaining labels (having a distribution that is characterized by a higher entropy). On the other hand, the Euclidean distance penalizes more a single large error on a given label than small distributed errors on multiple labels.

An example of these behaviors is shown in Example 8. Even if a prediction results in the fewest number of errors (implying a more realistic prediction that is close to the target one), it achieves an equal or lower distance score (according to the Manhattan and the Euclidean distance respectively), than a probability distribution that spreads the incorrect prediction across the remaining labels. This indicates that the largest-scaled probability value will outperform the rest.

Example 8

Target p(k)	Prediction q(k)	Prediction r(k)
[0,0,0,0,0,1]	[0,0,0,0.2,0.8]	[0,0.05,0.05,0.05,0.8]

Manhattan Distance values:

$\mathbb{L}_1(p(k), q(k))$ = 0.4

$\mathbb{L}_1(p(k), r(k))$= 0.4

Euclidean Distance values:

$\mathbb{L}_2(p(k), q(k))$ = 0.2828

$\mathbb{L}_2(p(k), r(k))$= 0.2236

The unfulfillment of this property can lead to some cases in which the Euclidean distance penalizes

less completely misclassified distributions than partial (erroneoulsly) label distributions, as shown in Example 9.

Example 9

Target p(k)	Prediction q(k)	Prediction r(k)
[0,0,0,0,0.3,0.7]	[0.25,0.25,0.25,0.25,0,0]	[0,0,0,0,1,0]

Euclidean Distance values:
$\mathbb{L}_2(p(k), q(k))$ = 0.911
$\mathbb{L}_2(p(k), r(k))$= 0.99

The identification of unique properties for Multiclass Classification problems is crucial due to the intricate nature of multiclass categorization itself. Multiclass settings frequently have hierarchical structures or allow for potential label relationships. The complexity of multiclass issues is further increased in multilabel classification, where multiple labels for instance are allowed. Specific properties for each classification problem might be defined, for instance, to deal with the concept of label similarity, to attribute a lower penalization for failures in predicting similar labels with respect to errors in predicting dissimilar labels. The proposed property offers a preliminary insight into the study of multiclass classification, highlighting the need for a more sophisticated understanding.

6. Related Work

We are grateful to one of the reviewers of this paper for directing us towards (Geng, 2016), which we had never previously encountered and appears to come from an entirely different research community. The objectives of that paper are, however, very different from ours, and closer to those of (Uma et al., 2021b). Geng considers six approaches to what he calls Label Distribution Learning and we would call Learning from Disagreement, and compares their performance on 16 datasets, none of which are of NLP tasks (1 is artificial, 11 are biological datasets, 3 are image understanding datasets, and 1 is movie ratings). To do this, he selects six metrics supporting a comparison between label distributions, chosen among 41 (!) measures proposed in previous literature–this selection is made in order to maximize diversity between the metrics. There is essentially no overlap between the metrics considered in the paper, and no proposal regarding the properties such metrics should satisfy, or analysis of the extent to which they satisfy them. This said, that paper does point out to the existence of an extensive literature on soft evaluation metrics we should investigate in the future.

7. Conclusion and Future Directions

In this paper, we propose a set of properties that soft evaluation metrics should have in order to allow for a fair comparison of computational models, and assess the extent to which plausible candidate metrics satisfy these properties. Our analysis suggests that Manhattan distance and Euclidean distance are the most suitable metrics for a robust and fair soft evaluation for binary classification problems, since they adhere to all the desired properties. Our investigation of the LeWiDi real case scenario gave us some indication as to the impact of the adoption of different metrics in a real-case scenario, showing differences in the rankings definitions and thus implying the importance of selecting the best evaluation metric for ensuring a fair evaluation. Further preliminary analysis in the Multiclass Classification domain demonstrated however the unsuitability of the analyzed metrics to provide a fair comparison of models in this scenario. Future works will concentrate on Multiclass Classification and will include the definition of properties in accordance with the different task specifics (e.g. hierarchical, multilabel, etc). The performed analysis suggests the need for a novel metric that overcomes the limitations that arise in Multiclass Classification evaluation.

Ethical issues

This study analyzes the impact of metrics on a real-case scenario. Data from Learning With Disagreements task (LeWiDi) have been exploited. However, no sensitive information is used nor reported within the paper.

Limitations

The investigation of the application of the explored metrics limits to one real-case scenario (Learning With Disagreements task (LeWiDi) at SemEval 2023). The achieved results highlight a relationship among the entropy of the dataset and the impact of a variation of the evaluation metric on the leaderboard. The four LeWiDi datasets exhibit diverse characteristics such as types, languages, goals (misogyny, hate speech, offensiveness detection), and annotation methods and represent therefore a solid case-study. However, additional analysis on real-case scenarios would provide a deepen understanding of the studied phenomena.

Acknowledgments

The work of Paolo Rosso was in the framework of the FairTransNLP-Stereotypes research project (PID2021-124361OB-C31) funded by MCIN/AEI/10.13039/501100011033 and by ERDF, EU A way of making Europe. The work of Massimo Poesio is supported in part by the AINED Fellowship Grant *Dealing with Meaning Variation in NLP*, NGF.1607.22.002. Elisa Leonardelli's work has been partly supported by the Precrisis EU project (GA 101100539 - ISF-2022-TF1-AG-PROTECT).

The work of Elisabetta Fersini has been partially funded by MUR under the grant REGAINS, *Dipartimenti di Eccellenza 2023-2027* of the Department of Informatics, Systems and Communication at the University of Milano-Bicocca and by the European Union – NextGenerationEU under the National Research Centre For HPC, Big Data and Quantum Computing - Spoke 9 - Digital Society and Smart Cities (PNRR-MUR))

Bibliographical References

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2019. A new measure of polarization in the annotation of hate speech. In *AI*IA - XVIIIth International Conference of the Italian Association for Artificial Intelligence*, Lecture Notes in Computer Science, page 588–603. Springer.

Anthony McEnery and others. 2004. *The EMILLE/CIIL Corpus*. EMILLE (Enabling Minority Language Engineering) Project. distributed via ELRA: ELRA-Id W0037, ISLRN 039-846-040-604-0.

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. We need to consider disagreement in evaluation. In *Proc. of the ACL-IJCNLP Workshop on Benchmarking: Past, Present and Future*.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the ACL*, 10:92–110.

Xin Geng. 2016. Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*, 28(7):1734–1748.

Khalid Choukri and Niklas Paullson. 2004. *The OrienTel Moroccan MCA (Modern Colloquial Arabic) database*. distributed via ELRA: ELRA-Id ELRA-S0183, ISLRN 613-578-868-832-2.

Elisa Leonardelli, Gavin Abercrombie, Dina Almanea, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Verena Rieser, Alexandra Uma, and Massimo Poesio. 2023. SemEval-2023 task 11: Learning with disagreements (LeWiDi). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2304–2318, Toronto, Canada. Association for Computational Linguistics.

Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. In *Proc. of EMNLP*.

Rebecca J. Passonneau, Vikas Bhardwaj, Ansaf Salleb-Aouissi, and Nancy Ide. 2012. Multiplicity and word sense: evaluating and learning from multiply labeled word sense annotations. *Language Resources and Evaluation*, 46(2):219–252.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Barbara Plank, Dirk Hovy, and Anders Sogaard. 2014. Linguistically debatable or just plain wrong? In *Proc. EACL*.

Massimo Poesio and Ron Artstein. 2005. The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proc. of ACL Workshop on Frontiers in Corpus Annotation*, pages 76–83.

Roventini, Adriana and Marinelli, Rita and Bertagna, Francesca. 2016. *ItalWordNet v.2*. ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli", National Research Council, in Pisa, ISLRN 532-206-426-067-2. PID http://hdl.handle.net/20.500.11752/ILC-62. Note: You don't really need both an ISLRN and another PID, but it can't hurt.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Speecon Consortium. 2011. *Catalan Speecon database*. SpeeCon. Speecon Project, distributed via ELRA: ELRA-Id S0327, Speecon resources, 1.0, ISLRN 935-211-147-357-5.

Alexandra Uma, Tommaso Fornaciari, Anca Dumitrache, Tristan Miller, Jon Chamberlain, Barbara Plank, and Massimo Poesio. 2021a. Semeval-2021 task 12: Learning with disagreements. In *Proc. of SEMEVAL*. Association for Computational Linguistics.

Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021b. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.