CARMA 2026

8th Int. Conf. on Advanced Research Methods and Analytics
1-3 July 2026, Valencia, Spain
DOI: https://doi.org/10.4995/CARMA2026.2026.22444

Qualitative Analysis of Misclassifications

Joan Fernández-Navarro[1], Ana Debón[1], Josep Domenech[1]
[1]eSMART Research Center, Universitat Politècnica de València, Spain

Abstract

This paper proposes a qualitative framework for the interpretation of misclassifications in supervised learning under noisy and imperfect label conditions. Traditional evaluation metrics assume that observed labels represent the true underlying categories, which can lead to misleading performance assessments when labels are noisy or misaligned with real-world phenomena. The proposed approach reinterprets misclassification as a heterogeneous outcome and introduces a structured taxonomy to distinguish between model errors, label inconsistencies, data ambiguity, and partially valid predictions. The framework is operationalised through a stratified sampling strategy, iterative qualitative adjudication of error cases, and aggregation of results into adjusted performance measures. The results highlight the importance of combining quantitative and qualitative evaluation in noisy-label environments.

Keywords: *label noise; misclassification analysis; weak supervision; data quality*

1. Introduction

Recent advances in machine learning have enabled the application of classification models to increasingly complex real-world data environments. In many applied settings, particularly those involving administrative records, user-generated content, or web-derived information, model evaluation relies on reference labels that are implicitly treated as accurate representations of the underlying phenomenon. However, this assumption is often unrealistic. Labels may be noisy, incomplete, or conceptually misaligned with the target of interest, reflecting limitations in data collection processes rather than true ground truth.

This issue is commonly framed within the broader problem of label noise, where discrepancies between observed and true labels can arise from measurement error, outdated information, or structural mismatches between classification systems and real-world entities. In such contexts, standard evaluation practices, based on direct comparisons between predicted and observed labels, become problematic. By construction, metrics such as accuracy or F1-score treat all disagreements as equivalent errors, regardless of their underlying cause.

As a result, observed misclassifications often conflate fundamentally different phenomena. A mismatch between prediction and label may reflect ambiguity in the input data, errors in the labeling process, the presence of multiple valid labels, or genuine model failure. Without

Article in Press

This work is licensed under a Creative Commons License CC BY-NC-SA 4.0
Editorial Universitat Politècnica de València

distinguishing between these cases, conventional evaluation frameworks risk producing misleading assessments of model performance and obscuring the true sources of error.

This paper addresses this limitation by proposing a qualitative framework for the interpretation of classification errors under noisy labels. Rather than adopting a binary notion of correctness, the framework conceptualises misclassification as a heterogeneous outcome and introduces a structured taxonomy that distinguishes between different types of discrepancies. These include data ambiguity, non-representative inputs, label errors, alignment with secondary or unobserved categories, partially consistent predictions, and genuine model errors. By explicitly separating these mechanisms, the framework enables a more nuanced and informative evaluation of classification models in applied settings.

The contribution of the paper is threefold. First, it develops a generalisable framework for the systematic interpretation of classification errors in the presence of imperfect labels. Second, it demonstrates how this framework can be operationalised through qualitative validation, providing a practical approach to disentangling different sources of misclassification. Third, it illustrates the empirical relevance of the framework through an application to firm activity classification using web data, showing that a substantial share of observed errors can be reinterpreted once label noise and data limitations are considered.

By shifting the focus from aggregate performance metrics to the interpretation of individual prediction outcomes, this paper contributes to a growing literature that questions the reliability of standard evaluation practices in noisy, real-world environments. The results suggest that, in such settings, the notion of classification error should be reconsidered, and that combining quantitative and qualitative approaches is essential for obtaining a more accurate and meaningful assessment of model performance.

2. Background

The evaluation of classification models typically assumes that observed labels correspond to the true underlying categories. Under this assumption, standard metrics such as accuracy, precision, recall, and F1-score treat any disagreement between predictions and labels as a model error. While this framework is appropriate in controlled settings, it becomes problematic in real-world environments where labels are subject to label noise (Frénay & Verleysen, 2014; Northcutt et al., 2021; Song et al., 2022).

In such contexts, discrepancies between predicted and observed labels may arise from multiple sources. First, labels themselves may be inaccurate, incomplete, or conceptually misaligned with the target of interest, reflecting imperfections in data collection or annotation processes (Frénay & Verleysen, 2014; Ratner et al., 2017). Second, input data may provide only a partial or ambiguous representation of the underlying phenomenon, introducing additional uncertainty

Article in Press

(Zhu & Wu, 2004). Third, the mapping between observed data and labels may be inherently non-unique, particularly in settings characterised by multi-dimensional or overlapping categories (Tsoumakas & Katakis, 2007).

These factors imply that the set of observed misclassifications, defined as cases where $y_i \neq \hat{y}_i$, is inherently heterogeneous. However, standard evaluation frameworks do not distinguish between different sources of disagreement, treating all errors as equivalent. This limitation can lead to biased or incomplete assessments of model performance, especially when label noise is systematic rather than random (Northcutt et al., 2021; Rolnick et al., 2017).

While the literature on noisy labels has primarily focused on improving robustness during training, relatively less attention has been paid to the interpretation of errors at the evaluation stage (Song et al., 2022). Existing approaches to error analysis are often qualitative but lack a formal structure that allows for systematic categorisation and comparison across settings (Ribeiro et al., 2016; Wu et al., 2019).

In response, this paper adopts a different perspective by treating misclassification as an outcome that requires interpretation rather than a binary failure. Building on insights from the literature on noisy labels, weak supervision, and qualitative error analysis, we propose a structured framework that distinguishes between different types of discrepancies according to their underlying source. This framework provides the conceptual basis for a more nuanced evaluation of classification models in the presence of imperfect labels.

3. Qualitative Framework for Error Interpretation

We propose a structured qualitative error analysis framework designed to move beyond binary notions of correctness in supervised classification settings and to explicitly incorporate the possibility that observed misclassifications arise from heterogeneous sources. Rather than treating all discrepancies between predicted and observed labels as equivalent, the framework decomposes the evaluation process into a sequence of operational stages that allow for systematic reinterpretation of errors and the derivation of adjusted performance estimates.

Article in Press

The framework is composed of three main components: (i) construction of a representative error sample, (ii) qualitative adjudication of individual cases, and (iii) aggregation of adjudicated outcomes into revised performance indicators. Each stage is designed to progressively transform raw misclassification outputs into structured information about the nature of prediction errors.

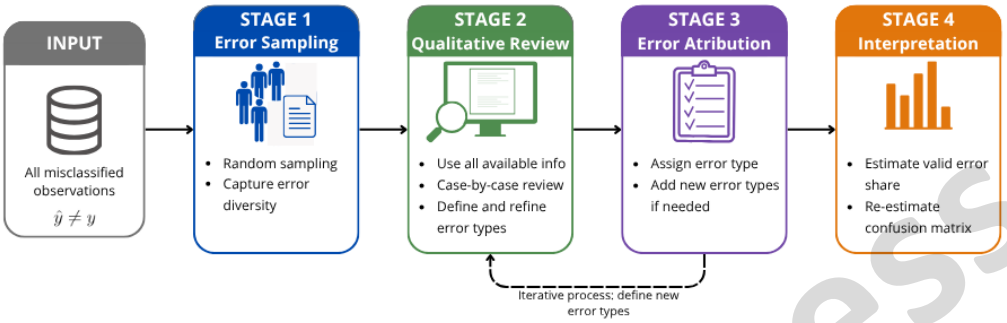


Figure 1. Qualitative error analysis framework.

3.1. Error sampling

The first step consists of defining the population of interest as all instances for which the predicted label differs from the observed label in the evaluation set. From this population, we construct a subset that serves as the basis for qualitative analysis. The key requirement in this stage is that the subset captures the diversity of the observed error patterns rather than constituting a simple convenience sample. For this reason, sampling is performed randomly within each true class, ensuring that all classes are represented in the qualitative evaluation.

This class-level sampling strategy serves two complementary objectives. First, it allows the framework to capture heterogeneous forms of misclassification that may emerge differently across classes, thereby improving the structural coverage of the error space. Second, it prevents highly imbalanced classes from dominating the sample, which would otherwise bias the qualitative analysis towards the most frequent categories and obscure less common but potentially informative error patterns. The objective is therefore not statistical inference in the classical sense, but the construction of a balanced and representative view of the different mechanisms underlying observed misclassifications.

3.2. Qualitative review and iterative error attribution

Once the sample has been defined, each observation is analysed individually using all available auxiliary information. The objective of this stage is to move beyond the binary notion of correctness and to identify the underlying mechanism that explains each apparent

Article in Press

misclassification. In this sense, the focus shifts from determining whether $y_i = \hat{y}_i$ to understanding the structural reason why $\hat{y}_i \neq y_i$.

Before starting the case-by-case analysis, an initial set of error types is defined to guide the adjudication process. This initial taxonomy serves as an analytical scaffold that ensures comparability across cases and provides a structured starting point for interpretation. At a minimum, this taxonomy must include a set of core mechanisms that are expected to be present in most real-world classification settings with noisy labels and heterogeneous inputs. These include model error, ground truth inconsistency, label ambiguity, missing input and partially coherent predictions.

Importantly, this initial set is not treated as a fixed or exhaustive classification scheme. Instead, the framework explicitly allows for the taxonomy to evolve during the evaluation process. As individual cases are examined, it is often the case that certain patterns of discrepancy recur that are not fully captured by the pre-specified categories. When such systematic patterns are identified across multiple observations, new error types can be introduced into the taxonomy.

During this process, each case is assigned to one error categories that best describe the observed discrepancy. The goal is not only to label the error, but to ensure that the assigned category provides a meaningful and interpretable explanation of the misclassification.

The iterative nature of the framework is a key component of its design. As more cases are analysed, the taxonomy becomes progressively refined, both in terms of granularity and interpretability.

3.3. Aggregation and performance re-interpretation

After all cases in the sample have been classified into error types, the results are aggregated to quantify the proportion of misclassifications that can be considered functionally valid under a broader interpretation of correctness.

Let p_{valid} denote the proportion of sampled misclassified instances that fall into categories deemed valid or partially valid. This quantity provides an empirical estimate of the extent to which standard evaluation metrics underestimate model performance due to label noise or structural ambiguity.

This information is then used to construct an adjusted interpretation of performance metrics. Instead of modifying the model directly, we reinterpret the confusion matrix by reallocating a fraction of false positives and false negatives into true positives according to the estimated p_{valid}. This yields an adjusted performance measure that reflects “functional correctness” rather than strict label agreement. The key idea is not to replace standard metrics, but to complement them with a corrected perspective that accounts for uncertainty in the evaluation data

Article in Press

Overall, the framework operationalises error analysis as a structured and reproducible pipeline rather than an ad hoc inspection procedure. It begins by isolating the full set of misclassifications, constructs a representative and stratified subset for in-depth analysis, reinterprets each case through a structured and potentially evolving taxonomy of error types, and finally aggregates these findings to produce adjusted and more informative performance estimates.

Importantly, the scope of the framework extends beyond the evaluation of predictive performance alone. By explicitly disentangling different sources of discrepancy, it also enables the analysis of the data generation and labeling process itself. In particular, it allows for the identification of systematic patterns in the construction of labels, including recurring inconsistencies across specific classes, potential biases introduced during annotation, and cases of persistent divergence between observed labels and external evidence.

In this sense, the framework can also be used as a diagnostic tool for data quality assessment. It makes it possible to detect situations in which labeling decisions may reflect conventions, administrative constraints, or even implicit heuristic rules applied by human annotators, rather than strictly observable underlying phenomena. This includes, for example, systematic misalignment in specific categories, over-reliance on secondary or proxy information, or repeated patterns of ambiguous labeling that suggest structural uncertainty in the annotation process.

By incorporating this additional layer of analysis, the framework therefore not only improves the interpretability of model evaluation but also provides insights into the reliability and internal consistency of the underlying dataset, extending its utility beyond model assessment towards a broader understanding of data quality and labeling behaviour in real-world classification systems.

4. Empirical evidence

To illustrate the proposed framework, we apply it to a real-world classification problem involving the assignment of firms to economic activity categories using unstructured textual information extracted from corporate websites.

4.1. Sample construction

Following the first stage of the framework, we construct a stratified sample of 36 firms selected from the set of observations incorrectly classified under standard evaluation criteria. The sampling strategy ensures balanced coverage across CNAE sections. This design guarantees that the sample is not dominated by frequent classes or isolated error patterns, and that it reflects the full diversity of misclassification cases observed in the model.

Article in Press

4.2. Qualitative adjudication of misclassifications

Each of the 36 cases was individually reviewed using the firm's available web presence. Using the iterative taxonomy defined in the framework, each case was classified into one of the predefined error categories. This process revealed that a substantial proportion of apparent misclassifications did not correspond to clear model failures. Instead, many cases were associated with ambiguous or multi-activity firms, inconsistencies in the registered CNAE codes, or web content that only partially reflected the firm's actual economic activity.

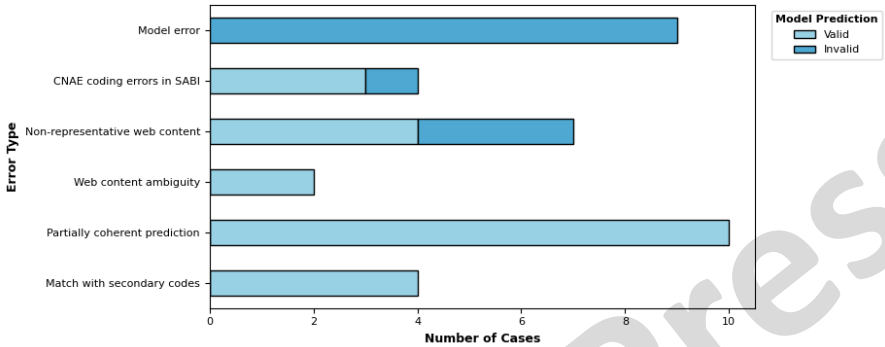


Figure 2. Distribution of valid and invalid predictions by type of error.

Overall, 23 out of the 36 cases (63.9%) were judged to be functionally valid predictions, in the sense that the model output corresponded to a plausible or secondary dimension of the firm's activity. The remaining cases were classified as clear model errors, where no meaningful correspondence could be established between the prediction and either the web evidence or the registered label.

4.3. Reinterpretation of model performance

The results of the qualitative analysis have direct implications for the interpretation of standard evaluation metrics. Under strict label-matching criterion, the model achieves an F1-score of approximately 0.56 on the test set. However, the qualitative evidence suggests that a non-negligible share of these "errors" are not true failures of the model but rather reflect label noise or structural ambiguity in the data.

Under the assumption that the proportion of functionally valid misclassifications observed in the qualitative sample generalises to the full evaluation set, we obtain an adjusted interpretation of performance in which a significant fraction of false positives and false negatives can be reclassified as correct predictions. This yields an indicative corrected F1-score in the range of 0.75–0.80, substantially higher than the original estimate. While this adjustment is not intended as a formal replacement of the standard metric, it provides a more realistic upper bound of model performance under imperfect ground truth conditions.

Article in Press

5. Conclusions

This paper has proposed a structured qualitative framework for the analysis of classification errors in settings where ground truth labels may be noisy, ambiguous, or only partially aligned with the underlying phenomenon. In contrast to standard evaluation approaches, which treat all discrepancies between predicted and observed labels as uniform errors, the proposed methodology explicitly distinguishes between different sources of misclassification and reinterprets evaluation as a structured diagnostic process.

The empirical application shows that this distinction is not merely conceptual. In the firm classification task based on web-derived data, a substantial share of apparent misclassifications corresponds to functionally valid or partially valid predictions when examined in detail. As a result, the effective performance of the model appears higher than suggested by standard metrics, while also revealing systematic patterns of disagreement linked to administrative inconsistencies and heterogeneous firm behaviour.

Overall, the results highlight the importance of complementing quantitative evaluation with structured qualitative analysis in noisy-label environments, where traditional metrics may provide a limited view of model quality.

References

Frénay, B., & Verleysen, M. (2014). Classification in the presence of label noise: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, *25*(5), 845–869.

Northcutt, C., Jiang, L., & Chuang, I. (2021). Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, *70*, 1373–1411.

Ratner, A., Bach, S., Ehrenberg, H., Fries, J., Wu, S., & Ré, C. (2017). Snorkel: Rapid training data creation with weak supervision. *VLDB*, *11*(3), 269–282.

Ribeiro, M., Singh, S., & Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. *KDD*, 1135–1144.

Rolnick, D., Veit, A., Belongie, S., & Shavit, N. (2017). Deep learning is robust to massive label noise. *arXiv Preprint arXiv:1705. 10694*.

Song, H., Kim, M., Park, D., & Others. (2022). Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, *34*(11), 8135–8153.

Tsoumakas, G., & Katakis, I. (2007). Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, *3*(3), 1–13.

Wu, T., Ribeiro, M., Heer, J., & Weld, D. (2019). Errudite: Scalable, reproducible, and testable error analysis. *ACL*, 747–763.

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

Article in Press