Human Rationale Aggregation for Balancing Performance and Explainability in Text Classification

Hideki Tsukamoto[1], Jiyi Li[2,*]

[1]University of Yamanashi, Kofu, Japan
[2]Hokkaido University, Sapporo, Japan

Abstract

This study explores the impact of different rationale aggregation methods on the performance and explainability of text classification models in natural language processing (NLP). Explainability is a key component of trustworthy AI, especially for text classification tasks involving harmful or socially sensitive content. Human rationales, token-level annotations indicating reasoning behind class labels, can enhance both model performance and explainability. However, multiple human rationales often exist for the same instance due to subjective annotation differences, requiring effective aggregation. Using the HateXplain dataset, we evaluate several aggregation methods, including Average, And, Or, and Majority Voting (Mv), and propose a novel method, SampleOneAgg. Unlike the original SampleOne method, which samples one reference translation, SampleOneAgg dynamically samples one rationale supervision label from a set generated by the binary aggregation methods And, Or, and Mv for each training instance at each epoch. Experimental results on HateXplain show that SampleOneAgg maintains competitive classification performance while improving several explanation-quality metrics, suggesting that dynamic human rationale aggregation can help build more interpretable and trustworthy text classification models.

Keywords

Text Classification, Explainable AI, Trustworthy AI, Human Rationales, Rationale Aggregation

1. Introduction

For text classification in natural language processing (NLP), explainability refers to a model's ability to provide human-understandable reasons for its predictions. Trustworthy AI requires models to be not only accurate but also transparent and understandable to humans. This requirement is particularly important in socially sensitive text classification tasks, such as hate speech or offensive language detection, where model decisions may affect users or online communities. In this context, explainability plays an essential role in helping users and developers understand why a model makes a particular prediction. Human rationales, i.e., token-level annotations that indicate the basis for a class label, have been widely used to improve both model performance and explainability [1, 2, 3, 4, 5]. Recent studies have pointed out that it is often inappropriate to assume a single correct rationale for each instance, since annotators may highlight different tokens based on their own perspectives [2]. Therefore, rationale aggregation is not merely a preprocessing step, but a design choice that can influence the transparency, reliability, and human alignment of the resulting model. To address this issue, several datasets, including HateXplain, provide multiple human rationales for each instance [6, 7, 4]. In HateXplain, each instance is annotated with three rationale vectors. In this paper, an annotator refers to a human labeler, a class label refers to the instance-level category, a rationale annotation refers to a token-level vector provided by an annotator, and an aggregation method combines multiple rationale annotations into one rationale supervision label.

When using such data for model training, these multiple rationales must be aggregated into a single rationale vector. Previous studies have adopted different rationale aggregation methods. Mathew et al.

TRUST-AI: The Second European Workshop on Trustworthy AI. Organized as part of the International Joint Conference on Artificial Intelligence - IJCAI/ECAI 2026. August 2026, Bremen, Germany.
*Corresponding author.
garfieldpigljy@gmail.com (J. Li)
0000-0003-4997-3850 (J. Li)

Table 1
Basic Aggregation Methods

Rationales	Aggregation	Aggregated Rationales
[1, 0, 1, 0, 0, 0]	And	[0, 0, 1, 0, 0, 0]
[1, 1, 1, 0, 0, 0]	Or	[1, 1, 1, 0, 0, 1]
[0, 0, 1, 0, 0, 1]	Mv	[1, 0, 1, 0, 0, 0]

[4] used the Average method, while Subramaniam et al. [8] examined logical And and Or operations. In addition, Majority Voting (Mv) [9], which is commonly used for label aggregation in NLP, can also be applied to rationale aggregation.

In this paper, we investigate how different rationale aggregation methods affect both classification performance and explainability. We also propose SampleOneAgg, a novel method that randomly selects one binary-based aggregation method (And, Or, or Mv) and its corresponding aggregated rationale labels during each training epoch. For class labels, we use Mv for aggregation and focus on the effect of rationale aggregation methods. Our contributions are as follows. (1) We investigate how different human rationale aggregation methods affect the trade-off between classification performance and explanation quality. (2) We propose SampleOneAgg, a dynamic aggregation method that samples among multiple binary rationale aggregation methods during training to reflect the diversity of human rationales. (3) We evaluate the methods on HateXplain using both predictive performance metrics and explanation metrics covering plausibility and faithfulness, providing empirical insights for trustworthy text classification.

2. Human Rationale Aggregation for Explainable Text Classification

2.1. Basic Aggregation Methods

We first briefly introduce the basic aggregation methods, including And [8], Or [8], and Mv [9]. Table 1 illustrates these methods with an example.

Subramaniam et al. [8] proposed aggregation methods based on the And and Or operations, showing that the And operation improved model performance, while the Or operation enhanced explainability. However, both methods have limitations. For example, consider three rationale labels for an instance: [1, 0, 0, 0, 0], [1, 1, 1, 0, 0, 0], and [0, 0, 1, 0, 0, 1]. Applying the And operation yields [0, 0, 0, 0, 0, 0], which fails to capture any rationale. Conversely, the Or operation results in [1, 1, 1, 0, 0, 1], which not only includes valid rationales but also tokens selected by only one annotator. Such tokens may be unreliable and could degrade the quality of the aggregated rationale used as a supervision label during training.

However, Mv is still not perfect, and And and Or may also be better in some cases. The Majority Voting (Mv) [9] method selects tokens deemed important by the majority of the three annotators. This approach excludes tokens chosen by only one annotator, potentially uninformative rationales, while capturing more rationales than the And operation. However, Mv is not flawless; in certain cases, the And or Or methods may yield better results depending on the characteristics of the data and the annotation agreement. This observation motivates a dynamic aggregation method that does not rely on a single fixed aggregation rule throughout training.

2.2. SampleOneAgg Method

In machine translation tasks, multi-reference methods are used when multiple target translations are available in the ground truth [10, 11, 12]. SampleOne [10] is a typical multi-reference method. In each training epoch, one translation is randomly selected from the available references and used as the ground truth for that iteration. By leveraging multiple reference translations, this approach enables training that better accounts for the diversity and variability inherent in natural language translation.

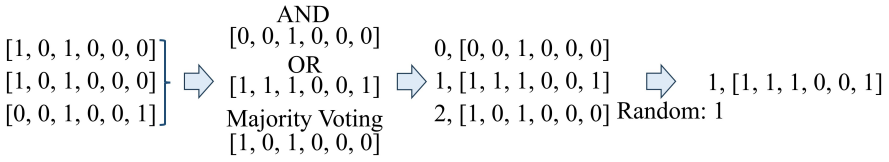

Figure 1: Example of SampleOneAgg.

This consideration of diversity aligns with the motivation for constructing datasets that include multiple rationales in our setting. Inspired by the multi-reference idea, we propose a novel rationale aggregation method. Unlike the original SampleOne approach, which randomly selects a reference translation, our proposed SampleOneAgg method randomly selects a basic rationale aggregation method and its corresponding aggregated labels.

SampleOneAgg samples only from binary aggregation methods. Specifically, And, Or, and Mv each produce a binary rationale vector, whereas Average produces soft values. For example, for a token with annotator labels $[1, 1, 0]$, Average assigns a value of $2/3$, while for $[1, 0, 0]$, it assigns a value of $1/3$. We therefore exclude Average from the candidate set so that all sampled rationale labels have the same binary format. An example of the SampleOneAgg method using And, Or, and Mv is shown in Figure 1.

- First, three aggregated rationales are generated using the And, Or, and Mv methods based on the three original annotator rationales.
- Next, these three aggregated rationales are grouped together to form a new rationale set.
- Finally, for each training instance at each epoch, one rationale is independently sampled from the rationale set generated by And, Or, and Mv. The sampled rationale is used as the rationale supervision label for that instance in the current epoch. Thus, the supervision signal can vary across epochs, while the candidate rationale set remains fixed for each instance.

3. Experiments

3.1. Experimental Settings

We use the HateXplain dataset [4] for our experiments. The dataset was collected from Twitter and Gab, consisting of 9,055 instances from Twitter and 11,093 from Gab, totaling 20,148 instances. Among these, 919 instances exhibit disagreement in class labels across the three annotators (i.e., each annotator assigned a different sentiment label) and are therefore excluded. Consequently, 19,229 instances are used in our experiments. We split the dataset into training, validation, and test sets with an 8:1:1 ratio, resulting in 15,383, 1,922, and 1,924 instances, respectively. The same fixed split is used for all aggregation methods to ensure a fair comparison. For class labels, we apply majority voting across the three annotators and keep the resulting class labels fixed across all experiments. Therefore, the comparison focuses on the effect of different human rationale aggregation methods.

3.2. Evaluation Metrics

Performance: To evaluate model performance, we employ three metrics, Accuracy, Macro-F1 (F1), and AUROC, across the three classes: Hate, Offensive, and Normal.

Explainability: To evaluate model explainability, we adopt the evaluation metrics defined in ERASER [1], a standard benchmark for assessing model-generated rationales in NLP tasks such as text classification. These metrics are broadly categorized into Plausibility and Faithfulness.

- **Plausibility** evaluates how convincing the model-generated rationales are to humans. We use three metrics for this purpose: IoU F1, Token F1, and Area Under the Precision-Recall Curve (AUPRC).

  - IoU F1 measures the similarity between the model-generated rationale and the three human-provided rationales in the dataset using the Intersection over Union (IoU) metric. Rationale pairs with an IoU value greater than 0.5 are counted as True Positives (TP), and the F1 score is subsequently computed based on these matches.
  - Token F1 compares the model-predicted rationale with the three human-provided rationales by identifying tokens marked as important by both. Tokens overlapping between the model and human rationales are counted as True Positives (TP), and the F1 score is then computed accordingly.
  - AUPRC represents the Area Under the Precision-Recall (PR) Curve, which quantifies the trade-off between precision and recall across different threshold settings.
- **Faithfulness** is assessed using two metrics: Comprehensiveness and Sufficiency. Higher comprehensiveness and lower sufficiency values indicate better faithfulness.
  - Comprehensiveness measures how much the generated rationale contributes to the model's prediction. It is computed as the difference in the predicted probability for the correct label between the original input and the input with the rationale removed. A larger value indicates that removing the rationale significantly impacts the prediction, suggesting higher comprehensiveness.
  - Sufficiency assesses whether the rationale alone is adequate to reproduce the model's prediction. It is computed as the difference in the predicted probability for the correct label between the original input and the rationale-only input. A smaller value indicates better sufficiency.

In this experiment, we evaluate both the performance and explainability of the proposed aggregation methods by comparing model results when trained with rationale supervision labels generated using these different aggregation methods. We adopt BERT as the backbone model for our experiments. The text classification model incorporates rationale classification as an auxiliary task to enhance its learning and interpretability. The input text is fed into a text classification model, which simultaneously predicts both class labels and rationale labels. The predicted class and rationale labels are used to compute the cross-entropy losses $Loss_{class}$ and $Loss_{rationale}$, respectively, based on the aggregated class and rationale labels. The total loss $Loss_{total}$ is a weighted sum of $Loss_{class}$ and $Loss_{rationale}$, defined as $Loss_{total} = Loss_{class} + \lambda * Loss_{rationale}$, where λ is the weight hyperparameter.

For the hyperparameters, we followed the settings used in [4] and [8] for BERT. The model was trained using AdamW with a batch size of 16 for 5 epochs, a learning rate of 2e-5, a dropout rate of 0.1, and λ of 100. All results are reported from a single run for each aggregation method. Since SampleOneAgg involves random sampling during training, evaluating the stability of the results across multiple random seeds is left for future work.

3.3. Experimental Results

Table 2 presents the performance and explainability results. First, the Average method achieves the highest classification performance, but it performs worse on rationale plausibility metrics such as IoU F1, Token F1, and AUPRC. This suggests that non-binary averaged rationale labels may help classification learning but provide less clear supervision for generating human-aligned rationales. Therefore, in the following discussion, we mainly focus on comparisons among the binary-based methods. In the following discussion, bold and underlined values indicate the first- and second-best performances among the remaining methods, respectively.

For the performance results, And, Mv, and SampleOneAgg demonstrate relatively comparable performance, while Or slightly underperforms. And and SampleOneAgg achieve the highest AUROC, indicating strong discriminative ability across classes. SampleOneAgg shows a balanced performance overall. Or has the lowest performance across all metrics, likely due to including too many noisy or less-informative tokens in its aggregated rationales. Mv performs comparably to And, supporting the

Table 2
Performance and explainability results. Average is reported as a non-binary reference baseline, while best and second-best marks are assigned among binary aggregation methods.

	Accuracy ↑	F1 ↑	AUROC ↑	IOU F1 ↑	Token F1 ↑	AUPRC ↑	Comp. ↑	Suff. ↓
Average	0.697	0.692	0.854	0.132	0.448	0.664	0.494	0.096
And	0.685	0.679	**0.851**	0.139	0.531	0.820	0.554	0.051
Or	0.676	0.670	0.848	0.141	**0.579**	**0.873**	**0.573**	0.054
Mv	0.684	0.677	0.849	0.141	0.562	0.859	0.558	0.072
SampleOneAgg	**0.690**	**0.685**	**0.851**	**0.144**	0.569	0.872	0.526	**0.040**

idea that majority-based filtering provides a stable middle ground between conservative and permissive rationale selection. Overall, SampleOneAgg achieves competitive performance, suggesting that dynamic sampling is a promising alternative to using a single fixed aggregation method.

For the explainability results, SampleOneAgg achieves the highest IoU F1 and a high Token F1, showing that its generated rationales align closely with human rationales. Its AUPRC is also high and comparable to Or, indicating strong precision-recall performance in rationale detection. Comprehensiveness for Or and Mv is higher than for SampleOneAgg, meaning their rationales contribute slightly more to prediction confidence when removed. However, this difference is modest. SampleOneAgg achieves the lowest Sufficiency, indicating that the model can make confident predictions using only the rationale tokens and showing favorable faithfulness in this comparison. Nevertheless, the differences from the strongest fixed aggregation baselines are modest and should be interpreted cautiously, especially because each method was evaluated in a single run.

For both performance and explainability, And yields precise but overly restrictive rationales, leading to solid performance but limited explainability. Or enhances explainability at the cost of performance due to noisy rationales. Mv provides balanced results across both performance and explainability dimensions. SampleOneAgg, by dynamically sampling among binary aggregation methods, achieves a desirable trade-off, maintaining competitive performance while yielding faithful and plausible explanations. These results suggest that incorporating diversity in rationale aggregation can be beneficial for improving the balance between performance and interpretability.

4. Conclusion

In this paper, we studied human rationale aggregation as an important design choice for explainable and trustworthy text classification. We propose a novel rationale aggregation method, SampleOneAgg, which dynamically samples one rationale supervision label from the set of binary aggregated rationales for each training instance at each epoch. Experimental results show that SampleOneAgg achieves competitive classification performance while improving several explanation-quality metrics, suggesting that dynamically leveraging multiple aggregation methods may support more interpretable models. These findings highlight the importance of considering annotator rationale diversity when using human rationales to train trustworthy and interpretable NLP models.

Limitations

In this study, to simplify the experimental setup, we aggregate class labels using majority voting and focus on the effect of rationale aggregation. If both class labels and rationale labels are treated as multiple annotations, the setting becomes more complex and should be explored in future work. Second, our experiments report single-run results. Because SampleOneAgg randomly samples rationale supervision labels during training, the results may vary across different random seeds. Future work should conduct multiple-seed experiments and report mean and standard deviation to better assess the stability of the proposed method. Third, our experiments are conducted only on HateXplain with

BERT as the backbone model. Future work should evaluate other datasets with human rationales and other backbone models to further validate the generalizability of our findings.

Ethics Statement

This study uses the publicly available HateXplain dataset, which contains hate speech and offensive language. We use the dataset only for research purposes and do not attempt to identify users or infer sensitive attributes. We acknowledge that models trained on such data may inherit biases from the dataset and annotations. Our work focuses on improving the transparency of model decisions rather than deploying the model for real-world content moderation.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP23K28092.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT in order to: Grammar and spelling check. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] J. DeYoung, S. Jain, N. F. Rajani, E. Lehman, C. Xiong, R. Socher, B. C. Wallace, ERASER: A benchmark to evaluate rationalized NLP models, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 4443–4458.

[2] S. Wiegreffe, A. Marasovic, Teach me to explain: A review of datasets for explainable natural language processing, in: J. Vanschoren, S. Yeung (Eds.), Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks, volume 1, 2021.

[3] I. Arous, L. Dolamic, J. Yang, A. Bhardwaj, G. Cuccu, P. Cudré-Mauroux, Marta: Leveraging human rationales for explainable text classification, Proceedings of the AAAI Conference on Artificial Intelligence 35 (2021) 5868–5876.

[4] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, A. Mukherjee, Hatexplain: A benchmark dataset for explainable hate speech detection, Proceedings of the AAAI Conference on Artificial Intelligence 35 (2021) 14867–14875.

[5] L. Resck, M. M. Raimundo, J. Poco, Exploring the trade-off between model performance and explanation plausibility of text classifiers using human rationales, in: Findings of the Association for Computational Linguistics: NAACL 2024, 2024, pp. 4190–4216.

[6] D. Wadden, S. Lin, K. Lo, L. L. Wang, M. van Zuylen, A. Cohan, H. Hajishirzi, Fact or fiction: Verifying scientific claims, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 7534–7550.

[7] M. Kutlu, T. McDonnell, T. Elsayed, M. Lease, Annotator rationales for labeling tasks in crowdsourcing, Journal of Artificial Intelligence Research 69 (2020) 143–189.

[8] A. Subramaniam, A. Mehra, S. Kundu, Exploring hate speech detection with hatexplain and bert, arXiv preprint arXiv:2208.04489 (2022).

[9] R. Snow, B. O'Connor, D. Jurafsky, A. Ng, Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks, in: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, 2008, pp. 254–263.

[10] R. Zheng, M. Ma, L. Huang, Multi-reference training with pseudo-references for neural translation and text generation, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 3188–3197.
[11] J. Effendi, S. Sakti, K. Sudoh, S. Nakamura, Multi-paraphrase augmentation to leverage neural caption translation, in: Proceedings of the 15th International Conference on Spoken Language Translation, 2018, pp. 181–188.
[12] H. Khayrallah, B. Thompson, M. Post, P. Koehn, Simulated multiple reference training improves low-resource machine translation, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 82–89.